NIST Special Publication 1275v2

NIST Conference Papers
Fiscal Year 2018
Volume 2:
Communications Technology Laboratory
Information Technology Laboratory
Material Measurement Laboratory

Compiled and edited by:
Information Services Office

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1275v2

NIST
National Institute of
Standards and Technology
U.S. Department of Commerce

NIST Special Publication 1275v2

NIST Conference Papers
Fiscal Year 2018
Volume 2:
Communications Technology Laboratory
Information Technology Laboratory
Material Measurement Laboratory

Compiled and edited by:
Resources, Access, and Data Team
Information Services Office

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1275v2

January 2022

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

National Institute of Standards and Technology Special Publication 1275v2
Natl. Inst. Stand. Technol. Spec. Publ. 1275v2, 663 pages (January 2022)
CODEN: NSPUE2

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1275v2

Foreword

NIST is committed to the idea that results of federally funded research are a valuable national resource and a strategic asset. To the extent feasible and consistent with law, agency mission, resource constraints, and U.S. national, homeland, and economic security, NIST will promote the deposit of scientific data arising from unclassified research and programs, funded wholly or in part by NIST, except for Standard Reference Data, free of charge in publicly accessible databases. Subject to the same conditions and constraints listed above, NIST also intends to make freely available to the public, in publicly accessible repositories, all peer-reviewed scholarly publications arising from unclassified research and programs funded wholly or in part by NIST.

This Special Publication represents the work of Communications Technology Laboratory, Information Technology Laboratory, and Material Measurement Laboratory researchers at professional conferences, as reported in Fiscal Year 2018.

More information on public access to NIST research is available at https://www.nist.gov/open.

Key words

NIST conference papers; NIST research; public access to NIST research.

Table of Contents

Memory and Motor Processes of Password Entry Error

Franklin P. Tamborello, II (franklin.tamborello.ctr@nrl.navy.mil)
National Reseaerch Council Postdoctoral Research Associate
Washington, DC, USA

Kristen K. Greene (kristen.greene@nist.gov)
National Institute of Standards and Technology[1]
Gaithersburg, MD, USA

Passwords are tightly interwoven with the digital fabric of our current society. Unfortunately, passwords that provide better security generally tend to be more complex, both in length and composition. Complex passwords are problematic both cognitively and motorically, leading to both memory and motor errors during recall and entry. It is important that we better understand and disentangle the two error sources, as password entry errors can have significant negative consequences, such as being locked out of a critical information system. We present a computational cognitive model of password recall and typing, with memory and motor errors each contributing to password entry error. With this synthesis we can study human-computer interaction issues involving the usability of computer access control systems, specifically the password as an authentication mechanism. Ultimately we hope to make science-based recommendations for password policies that promote the use of passwords that are more usable.

INTRODUCTION

Despite widespread recognition that character-based passwords are a deeply problematic method of user authentication (Honan, 2012), they are tightly interwoven with the digital fabric of our current society. The ubiquity of passwords is true both for personal and work place accounts, as is the challenge of complying with a variety of password policies (Shelton, 2014; Choong & Theofanos, 2015). People are forced to remember—or in some other way keep track of—a large and ever-increasing number of passwords as they interact with a variety of systems and accounts each day (Florencio & Herley, 2007; Choong, Theofanos, & Liu, 2014).

In addition to an increasing number of passwords, people must also contend with passwords of increasing length. Computer security specialists suggest increasing the length of passwords; this increases their entropy, or randomness, which makes them more computationally expensive to guess. Furthermore, passwords are increasing in complexity as well as length. For most systems—particularly systems in higher-security enterprise environments—passwords containing only lowercase letters are not permitted. In addition to lowercase letters, the inclusion of uppercase letters, numbers, and special characters is also required, as using all four character categories is often recommended for increasing password security (United States Department of Homeland Security, 2009).

Most password requirements also prohibit the use of words, as dictionary attacks on passwords are so successful, even since the late 1970s (Morris & Thompson, 1979). This means that higher-entropy passwords can differ greatly from the natural language words used in studies on skilled typing and transcription typing (e.g., Coover, 1923; Gentner, 1981; Salthouse, 1984; Salthouse, 1986). While words follow orthographic rules and are predictable given neighboring semantic content, passwords should ideally be as random as possible to help mitigate guessing. While non-word strings of random letters have been included in prior transcription typing research (e.g., Salthouse, 1984), the numbers and special characters suggested for passwords were not.

Although there are longterm research efforts underway to replace passwords (National Strategy for Trusted Identities in Cyberspace, 2011), widespread implementation will take some time. Furthermore, even as newer identity management systems and authentication technologies such as biometrics become more prevalent, legacy systems may remain reliant upon passwords. Therefore, balance between usability and security in password policies remains important.

Unfortunately, due to privacy and security concerns, it can be difficult to collect real-world password data. To collect laboratory data from large numbers of participants across a variety of password requirement combinations would require prohibitively large investments of time and money. Usable security is certainly not the only domain where access to human data can be challenging, and as in other domains, computational cognitive modeling offers a promising alternative to augment existing behavioral research.

Drawing upon theories from cognitive science and experimental psychology can help understand the roles that human cognition and motor movement play in generating, rehearsing, recalling, and typing passwords on various devices. Unifying theories of memory and motor error can help inform recommendations for password policies that better address both the limits and capabilities of human performance. By supplementing behavioral data from prior password studies with predictive models of human performance, we can test theories and hypotheses in ways that neither research method can do alone.

In particular, we are interested in whether existing theories and models can disentangle memory from motor errors for those complex, system-generated passwords suggested or required in higher-security enterprise

[1] Disclaimer: Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that the products mentioned are necessarily the best available for the purpose.

Greene, Kristen; Tamborello, Frank.
"Memory and Motor Processes of Password Entry Error."
Paper presented at 2015 Annual Meeting of the Human Factors and Ergonomics Society, Los Angeles, CA, United States. October 26, 2015 - October 30, 2015.

environments. In such enterprise environments, passwords differ from words in several important ways, which means that traditional memory, transcription typing, and mobile text entry literature and theory may not be completely sufficient to inform and test predictive models of password typing. The usable security literature may address this somewhat, yet many password studies do not report sufficiently detailed data for model validation purposes.

REVIEW

There have certainly been many studies on memory in general (e.g., Miller, 1956; Baddeley & Hitch, 1974; Unsworth & Engle, 2007), and password memorability in particular (Vu, Bhargav-Spantzel, & Proctor, 2003; Forget & Biddle, 2008; Chiasson et al., 2009). There is also a large existing body of literature examining expert typing and transcription typing from the 1920s to the 1980s (e.g., Coover, 1923; Gentner, 1981; Salthouse, 1984; Salthouse, 1986). There has been a comprehensive cognitive model of transcription typing, Bonnie John's TYPIST model (1996), which quantified 19 of the 29 phenomena reviewed by Salthouse (1986), as well as two additional phenomena. However, these studies did not include stimuli similar enough of complex passwords to suit our modeling goals.

Although the typing literature and models do well at examining the cognitive and perceptual-motor facets of typing, there are certain distinctions between passwords and words that may not be fully addressed by existing theory and research. For example, the cost of errors and error recovery can differ significantly between typing for communicative purposes, such as composing emails, and typing for authentication tasks (i.e., password entry). Typos in communication can be embarrassing, but typos in passwords can cause failed authentication attempts, which in turn cause accounts to be locked. Users are sensitive to the time (and frustration) cost of unlocking an account, which may impact their speed-accuracy tradeoff function specifically for password entry in comparison with other text entry tasks. This may be particularly true on mobile devices, where users cannot rely on the now common predictive algorithms for password entry. There is a rich body of mobile text entry literature examining factors such as the effect of devices (Castellucci & MacKenzie, 2011), motion (Nicolau & Jorge, 2012), and age (Nicolau & Jorge, 2012) on how people type words or phrases, but again, such stimuli are not representative of the complex passwords we are interested in modeling.

One important difference between general text entry and password entry is the lack of visual feedback during password entry tasks. On desktop computers, text is masked immediately as it is typed. On mobile devices, the character just typed is generally visible for a moment[2] before being masked. An additional difference between general text entry and complex password entry is the required navigation back and forth between multiple onscreen keyboards that password entry requires of the user. Passwords requiring a number of onscreen keyboard changes, or screen depth changes, can have disproportionately large effects across both entry times and error rates (Greene, Gallagher, Stanton, & Lee, 2014).

Studies using password-like stimuli and masked text can help to address the aforementioned literature gaps and provide much-needed data to inform computational cognitive models of the often onerous password entry task. There have been both desktop (Stanton & Greene, 2014) and mobile studies (Greene, Gallagher, Stanton, & Lee, 2014; Gallagher, 2015) using such complex password-like stimuli. As our current focus is on modeling desktop password entry errors, we focus much of our review on the desktop study and model that motivated our work.

Stanton and Greene (2014) examined the usability of system-generated passwords by having participants memorize a series of ten passwords and type them repeatedly using a desktop computer. Participants were given one password at a time. For each password, there was a set of three task phases: practice, verification, and entry. During the practice phase, participants could practice typing the password as many (or as few) times as they wished. The password was visible, and typed text was also visible during the practice phase. During verification, typed text was still visible, but the password was not. Participants had to enter the memorized password correctly during the verification phase in order to move on to the entry phase. During the entry phase, participants had to type the the memorized password ten times. After the series of three phases (practice, verification, and entry) was completed for each of the ten passwords, there was a surprise recall test. For the surprise recall test, typed text was visible.

The Stanton and Greene (2014) study examined the fundamentals of desktop password typing, contributing baseline data on human performance with stimuli representative of the complex, system-generated passwords found in higher-security enterprise environments. Most relevant for the current work were Stanton and Greene's (2014) error findings: at 45% of the total error corpus, incorrect capitalization errors were by far the most prevalent. Incorrect capitalization, or shifting, errors were almost three times as likely as the next most prevalent error category (missing character errors, or omissions, were 17% of the total error corpus).

The nature of the most common error category (incorrect capitalization, or shifting errors) is interesting for several reasons. The high frequency of incorrect capitalization errors was particularly important given the fact that most modern password policies—and certainly those in higher-security enterprise environments—require at least one uppercase letter. Additionally, most special characters (which are also required by many password policies) require a shift action. Twenty-one of the total 32 possible special characters require shifting; only 11 special characters can be executed without requiring a shift action. Finally, of greatest interest for our modeling efforts is the fact that based purely on the behavioral data reported in Stanton and Greene (2014), it cannot be fully determined whether those errors were memory errors or motor execution errors (or a combination of both).

Greene and Tamborello (2015) began modeling work to disambiguate memory from motor errors using a single password from the Stanton and Greene (2014) stimuli set. They report a cognition-only ACT-R model of password

[2] This is a setting that can be changed; for increased security to help protect against shoulder-surfing, mobile keyboard settings allow the user to turn the momentary visibility feature off for password text fields.

Greene, Kristen; Tamborello, Frank.
"Memory and Motor Processes of Password Entry Error."
Paper presented at 2015 Annual Meeting of the Human Factors and Ergonomics Society, Los Angeles, CA, United States. October 26, 2015 - October 30, 2015.

rehearsal, finding that recall errors alone were insufficient to fully explain the incorrect capitalization errors of interest. They also report an expansion of ACT-R's native typing abilities to support password-specific typing needs, giving ACT-R the ability to type capital letters and symbols, and to err while doing so. Such modifications were necessary to explore the role of motor error during desktop password entry, as the canonical ACT-R architecture is limited to perfect typing performance and would not predict the motor execution errors expected with typing complex passwords. Furthermore, the canonical ACT-R architecture does not support case-sensitivity in typing, nor does its typing vocabulary support all possible symbols; without such capacity, it would be impossible to model typing complex passwords.

ACT-R

We use the ACT-R cognitive architecture (Anderson et al, 2004) to model user password recall and typing. ACT-R is a hybrid symbolic and subsymbolic computational cognitive architecture that takes as inputs knowledge (both procedural and declarative about how to do the task of interest) and a simulated environment in which to run. It posits several modules, each of which perform some aspect of cognition (e.g., long-term declarative memory, vision). Each module has a buffer into which it can place a symbolic representation that is made available to the other modules. ACT-R contains a variety of computational mechanisms and the output of the model is a time stamped series of behaviors including individual attention shifts, speech output, button presses, and the like. It can operate stochastically and so models may be non-deterministic.

NEW CONTRIBUTION

Our model works by incorporating and coordinating two distinct systems underlying prospective memory and motor operations. The former operates on the principle of associative spreading activation (Anderson et al., 2004) while the latter builds upon the motor models embodied in EPIC (Meyer & Kieras, 1997A & B) and ACT-R (Anderson et al, 2004).

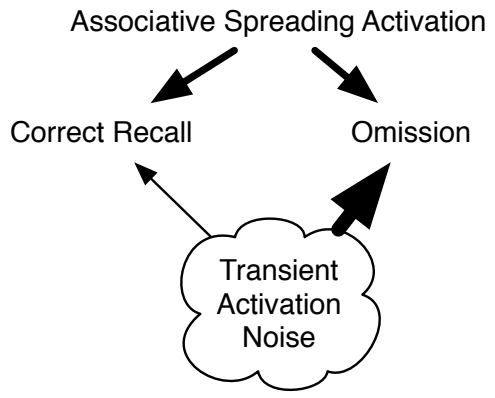


Figure 1. The role of noise in the model's memory processes: Associative spreading activation is the prospective memory process underlying selection of correct actions. When transient activation noise, a fundamental property of human memory, spikes during prospective retrieval it can lead to an omission.

Password Sequence Recall

Sequential tasks require prospective memory to remember what comes next. Our model uses this memory process, selecting the next character using the current character to prime retrieval.

Selecting the next character. Sequence memory is a prospective memory task, using a representation of the current character to associatively prime retrieval of a memory representation of the next character. We use ACT-R's spreading activation mechanism to implement prospective memory. Furthermore, activation propagates from active buffer contents to long-term memory according to what we assume to be learned association from each context to its subsequent action (Botvinick & Plaut, 2004).

Memory Errors

Memory errors arise out of the interaction of noise with the processes of normal task execution (Figure 1).

Omission. We assume that association is somewhat imprecise in that there is not a clean one-to-one mapping of cue to target. Instead, some association "bleeds" over from the target to a handful of subsequent items, with each subsequent item receiving less association than the one coming before it in sequence. The model may omit a character when transient noise is such that it simultaneously suppresses activation of the correct next step and enhances activation of one of these subsequent items.

Investigating the source of password entry errors is a perfect application opportunity for cognitive modeling to shed light on the root cause of error that was intractable to ascertain through prior behavioral data alone. By implementing support for an ACT-R model that can type capital letters, one could then test different models to see whether those incorrect capitalization errors were memory errors or motor execution errors (where a shift key press had been attempted but simply not executed properly, such as by prematurely releasing the shift key). The ability to type capital letters raises interesting theoretical questions. For each letter of the alphabet, do people have two distinct versions in their memory, one lowercase and one uppercase? Or is an uppercase letter encoded as the lowercase plus a required shift action?

Implementation Issues in ACT-R

In order to support modeling of incorrect capitalization typing errors, two limitations in ACT-R first required addressing: missing special characters and lack of case-sensitivity in typing.

Missing Special Characters. Of the non-alphanumeric characters available on typical American English keyboards, ACT-R previously included support only for the period, semicolon, slash, and quote (Bothell, 2014, see "key" on page 320 of the ACT-R Reference Manual). Therefore, in order to enable modeling typing of the remaining special characters, we added support for all remaining ASCII printable characters not previously supported by ACT-R.

Greene, Kristen; Tamborello, Frank.
"Memory and Motor Processes of Password Entry Error."
Paper presented at 2015 Annual Meeting of the Human Factors and Ergonomics Society, Los Angeles, CA, United States. October 26, 2015 - October 30, 2015.

Lack of Case-Sensitivity. Regardless of whether calling ACT'R's "press-key" motor module request (Bothell, 2014, see page 317 of the ACT-R Reference Manual) with a capital or lowercase letter, the output will be the same in ACT-R's current instantiation. This is somewhat problematic for modeling incorrect capitalization errors, which requires that ACT-R be capable of press-and-hold capability for the left and right shift keys, combined with a simultaneous key press of a second key (i.e., chorded typing). Therefore we added to ACT-R a capability to type key chords and output case-sensitive text, as described in the following section.

Stochastic Typing Extension for ACT-R

The standard ACT-R distribution (Anderson, et al, 2004; Anderson 2007) does not predict any typing errors as a matter of motor error (Bothell, 2014). However, real humans, even very skilled typists, are imperfect, and tend to err at rates from 0.5% to 35% (Salthouse, 1986; Panko, 2008; Landauer, 1987). We wished to explain password entry errors, but because some errors are due to memory processes and some are due to motor processes, we had to extend our modeling framework of choice, ACT-R, so that it, too, would be capable of such motor errors. Furthermore, we needed to implement the low-frequency, non-alphanumeric characters that information systems often require their users to incorporate into their passwords as a matter of security policy, e.g. "*" or "?". Source code for the ACT-R stochastic typing extension may be downloaded from https://github.com/usnistgov/CogMod.

Motor Errors in Typing

Our typing extension for ACT-R redefines some of ACT-R's existing code so that any requested typing action can stochastically result in the output of a typed key other than the one intended. It adapts the ellipsoid motor movement error equation of May (2012) and Gallagher and Byrne (2013), producing greater error along the axis of movement than off the axis, the off-axis error being scaled to .75 of the on-axis. However, because here the units are keys rather than pixels as in May's study, and ACT-R assumes most keys are the same width, the width term in May's equation is simplified to 1.

Hold-Key. Because typing non-alphanumeric characters typically involves holding a shift key while striking another key, and standard ACT-R provides no way to hold any such modifier key, it was necessary to invent such a method. Our errorful typing extension provides two motor module request extensions (see "extend-manual-requests" on page 325 of the ACT-R Reference Manual, Bothell, 2014) to enable the holding and releasing of modifier keys such as shift.

The new hold-key motor module request acts like press-key, translating the requested key to be held into a peck movement (Bothell, 2014, pp. 315-316) with the appropriate features. Once the hold-key motor movement is executed, ACT-R will have a state indicating that the appropriate key is being held. This state in turn causes ACT-R to now output a different character for the same press-key requests that follow for the given keys. The model can request the release-key function to release the given modifier key and end the modifier key state.

Nonalphanumeric Characters. With a shift key held, ACT-R can now type non-alphanumeric ASCII characters such as "*" and "?." It can now also type capital letters as well as lower-case letters, a critical feature for case-sensitive passwords lacking in standard ACT-R.

DISCUSSION

As in other domains, computational cognitive modeling can be a useful tool in the usable security research field, where behavioral data from prior password studies can be supplemented with predictive models of human performance. Although the study that motivated our work was focused on passwords for higher-security enterprise environments, our work has implications beyond that restrictive environment. By extending a widely used cognitive architecture to address motor errors in a way it previously did not, we contribute to the growing corpus of typing models (e.g., John, 1988; John, 1996; Das & Stuerzlinger, 2007; Gallagher, 2015; Gallagher & Byrne, 2015; Greene & Tamborello, 2015), all of which act together to test and expand the ACT-R theory.

Memory Errors

The kinds and frequencies of sequence memory errors arise from the fundamental properties of that memory system. Work on this problem from other domains (e.g. Anderson et al, 2004; Botvinick and Plaut, 2004) lend strong support to the memory account we use here, associative spreading activation.

Motor Errors

Motor errors are their own important contributor to password entry error, as the shifting errors in Stanton and Greene's (2014) study so strikingly exemplify. Moreover, as mobile touchscreen computers continue to gain importance it will become necessary to understand the mechanics of motor errors involved with that interface and how they contribute to password entry errors.

REFERENCES

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. Psychological Review, 111(4), 1036-60. doi:10.1037/0033-295X.111.4.1036

Anderson, J. R. (2007). *How can the human mind exist in the physical universe?* New York, NY: Oxford University Press. Retrieved from Google Scholar.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In Bower, G. (ed.) Recent Advances in Learning and Motivation, vol. 8, pp. 47–90. Academic Press, New York.

Bothell, D. (2014). ACT-R 6.0 reference manual. ACT-R Research Group. Retrieved from act-r.psy.cmu.edu

Botvinick, M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. Psychol Rev, 111(2), 395-429. doi:10.1037/0033-295X.111.2.395

Castellucci, S. J., & MacKenzie, I. S. (2011). Gathering text entry metrics on android devices. In CHI 2011 Extended Abstracts on Human Factors in Computing Systems, pp. 1507–1512.

Coover, J. E. (1923). A method of teaching typewriting based upon a psychological analysis of expert typing. National Education Association 61, 561–567

Chiasson, S., Forget, A., Stobert, E., Van Oorschot, P., & Biddle, R. (2009). Multiple password interference in text

Greene, Kristen; Tamborello, Frank.
"Memory and Motor Processes of Password Entry Error."
Paper presented at 2015 Annual Meeting of the Human Factors and Ergonomics Society, Los Angeles, CA, United States. October 26, 2015 - October 30, 2015.

passwords and click-based graphical passwords. In Proceedings of the 16th ACM Conference on Computer and Communications Security, pp. 500–511.
Choong, Y., Theofanos, M., & Liu, H. (2014). United States Federal Employees' Password Management Behaviors - a Department of Commerce Case Study. National Institute of Standards and Technology Interagency Report (NISTIR) 7991.
Choong, Y., & Theofanos, M. F. (2015). What 4,500+ People Can Tell You – Employees' Attitudes toward Organizational Password Policy Do Matter. To Appear in Proceedings of the 3rd International Conference on Human Aspects of Information Security, Privacy and Trust, in the 17th International Conference on Human-Computer Interaction.
Das, A., & Stuerzlinger, W. (2007). A cognitive simulation model for novice text entry on cell phone keypads. Proceedings of the 14th European Conference on Cognitive Ergonomics: invent! explore!, 141-147.
Florencio, D., & Herley, C. (2007). A large-scale study of web password habits. In Proceedings of the 16th International Conference on World Wide Web (WWW), pp. 657-666. ACM, New York.
Forget, A., & Biddle, R. (2008). Memorability of persuasive passwords. In CHI 2008 Extended Abstracts on Human Factors in Computing Systems, pp. 3759–3764.
Gallagher, M. A. (2015). Modeling Password Entry on Mobile Devices: Please Check Your Password and Try Again. Doctoral Dissertation, Rice University, Houston TX.
Gallagher, M. A., & Byrne, M. D. (2015). Modeling Password Entry on a Mobile Device. To appear in Proceedings of the 2015 International Conference on Cognitive Modeling.
Gallagher, M. A., & Byrne, M. D. (2013). The devil is in the distribution: Refining an ACT-R model of a continuous motor task. In *Proceedings of the 12th International Conference on Cognitive Modeling*. Ottawa, Canada.
Gentner, D. (1981). Skilled finger movements in typing. Center for Information Processing, University of California, San Diego. CHIP Report 104
Greene, K. K., Gallagher, M. A., Stanton, B. C., & Lee, P. Y. (2014). I Can't Type That! P@$$w0rd Entry on Mobile Devices. In Human Aspects of Information, Security, Privacy, and Trust. Lecture Notes in Computer Science, Vol. 8533, pp 160-171.
Greene, K. K., & Tamborello, F. P. (2015). Password Entry Errors: Memory or Motor? To appear in Proceedings of the 2015 International conference on Cognitive Modeling.
Honan, M. (2012). Kill the password: Why a string of characters can't protect us anymore. Wired.
John, B.E. (1988). Contributions to Engineering Models of Human-Computer Interaction, Department of Psychology, Carnegie-Mellon University, Ph.D. thesis.
John, B.E. (1996). TYPIST: A theory of performance in skilled typing. Human Computer Interaction, 11, 321-355.
Landauer, T. K. (1987). Relations between cognitive psychology and computer systems design. In J. M. Carroll (Ed.), Interfacing thought: Cognitive aspects of human-computer interaction (pp. 1-25). Cambridge, MA: MIT Press.
May, K. (2012). A model of error in 2D pointing tasks. Undergraduate Honors Thesis, Rice University, Houston, TX.
Meyer, D. M., & Kieras, D. K. (1997). A computational theory of executive control processes and human multiple-task performance: Part 1. Basic mechanisms. Psychological Review, 104, 3-65. Retrieved from Google Scholar.
Meyer, D. M., & Kieras, D. K. (1997). A computational theory of executive control processes and human multiple-task performance: Part 2. Accounts of psychological refractory-period phenomena. Psychological Review, 104, 749-791. Retrieved from Google Scholar.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review 63(2), 81–97. Morris, R., & Thompson, K. (1979). Password Security: A Case History. Communications of the ACM, 22(11): 594-597.
Morris, R., & Thompson, K. (1979). Password Security: A Case History. Communications of the ACM, 22(11): 594-597.
National Strategy for Trusted Identities in Cyberspace. Enhancing Online choice, Efficiency, Security, and Privacy. (2011). Retrieved online from http://www.whitehouse.gov/sites/default/files/rss_viewer/NSTICstrategy_041511.pdf
Nicolau, H., & Jorge, J. (2012). Elderly text-entry performance on touchscreens. In Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility. ACM, Boulder.
Nicolau, H., & Jorge, J. (2012). Touch typing using thumbs: understanding the effect of mobility and hand posture. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2683–2686.
Panko, R. R. (2008.). Basic error rates. [Web page] Retrieved from http://panko.shidler.hawaii.edu/HumanErr/Basic.htm
Salthouse, T. (1984). Effects of age and skill in typing. Journal of Experimental Psychology 113(3), 345–371
Salthouse, T. (1986). Perceptual, cognitive, and motoric aspects of transcription typing. Psychological Bulletin 99(3), 303–319.
Shelton, D. C. (2014). Reasons for Non-Compliance with Mandatory Information Assurance Policies by a Trained Population. Doctoral Dissertation, Capitol Technology University.
Stanton, B. C., & Greene, K. K. (2014). Character Strings, Memory and Passwords: What a Recall Study Can Tell Us. In Human Aspects of Information, Security, Privacy, and Trust. Lecture Notes in Computer Science, Vol. 8533, pp 195-206.
United States Department of Homeland Security. (2009). United States Computer Emergency Readiness Team (US-CERT), Security tip (ST04-002): Choosing and protecting passwords. Retrieved online from http://www.us-cert.gov/cas/tips/ST04-002.html
Unsworth, N., & Engle, R. W. (2007). The foundations of remembering: Essays in honor of Henry L. Roedgier III, pp. 241-258. Psychology Press, New York.
Vu, K., Bhargav-Spantzel, A., & Proctor, R. (2003). Imposing password restrictions for multiple accounts: Impact on generation and recall of passwords. In Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society (HFES), pp. 1331–1335.

Greene, Kristen; Tamborello, Frank.
"Memory and Motor Processes of Password Entry Error."
Paper presented at 2015 Annual Meeting of the Human Factors and Ergonomics Society, Los Angeles, CA, United States. October 26, 2015 - October 30, 2015.

Towards Probabilistic Identification of Zero-day Attack Paths

Xiaoyan Sun[1], Jun Dai[2], Peng Liu[1], Anoop Singhal[3], John Yen[1]

[1] Penn State University, University Park, PA 16802, USA
[2] California State University, Sacramento, CA 95819, USA
[3] National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
xzs5052,pliu,jyen@ist.psu.edu, jun.dai@csus.edu, anoop.singhal@nist.gov

Abstract. Zero-day attacks continue to challenge the enterprise network security defense. A zero-day attack path is formed when a multi-step attack contains one or more zero-day exploits. Detecting zero-day attack paths in time could enable early disclosure of zero-day threats. In this paper, we propose a probabilistic approach to identify zero-day attack paths and implement a prototype system named Pr0bA. A *System Object Instance Dependency Graph* (*SOIDG*) is first built from system calls to capture the intrusion propagation. To further reveal the zero-day attack paths hiding in the SOIDG, our system constructs an SOIDG-based Bayesian network. By leveraging intrusion evidence, the Bayesian network can quantitatively compute the probabilities of object instances being infected. The object instances with high infection probabilities reveal themselves and form the candidate zero-day attack paths. The experiment results show that our system can successfully identify zero-day attack paths and the paths are of manageable size.

1 Introduction

Defending against zero-day attacks is one of the most fundamentally challenging problems yet to be solved. Zero-day attacks are usually enabled by unknown vulnerabilities. The information asymmetry between what the attacker knows and what the defender knows makes zero-day exploits extremely difficult to detect. Signature-based detection assumes that a signature is already extracted from detected exploits. Anomaly detection [1–3] may detect zero-day exploits, but this solution has to cope with high false positive rates.

Recently, one noticeable research progress is based on a key observation that in many cases identifying zero-day attack paths is substantially more feasible than identifying individual zero-day exploits. A *zero-day attack path* is a multi-step attack path which includes one or more zero-day exploits. When not every exploit in a zero-day attack path is zero-day, part of the path can already be detected by commodity signature-based IDS. That is, the defender can leverage one weakness of the attacker: in many cases he is unable to let an attack path be completely composed of zero-day exploits.

Both alert correlation [4,5] and attack graphs [6–9] are limited in identifying zero-day attack paths. They both can identify the non-zero-day segments (i.e., "islands") of a zero-day attack path; however, none of them can automatically bridge these islands into a meaningful path, especially when different segments may belong to totally irrelevant attack paths.

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

To address these limitations, Dai et al. proposed to use (data and control) dependencies between OS-level objects (e.g., files, processes, sockets) to bridge the non-zero-day islands so that the zero-day segments can be revealed [10]. Nevertheless, this approach has a main limitation, namely the explosion in the number and size of zero-day attack path candidates. The forward and backward tracking from intrusion detection points can result in a large number of candidate paths, especially when lots of intrusion detection points are available. In addition, a candidate path can be too big because it preserves every tracking-reachable object. With the large number and size, discerning from the candidates and verifying the real zero-day attack paths becomes unpractical. As a consequence, in many cases this approach may generate a big "haystack" for the defender to find a "needle" in it.

In this paper, we propose a probabilistic zero-day attack path identification approach to address the explosion problem. The goal is to make the "haystack" orders of magnitude smaller. Our approach is to 1) establish a *System Object Instance Dependency Graph* (*SOIDG*) to capture the intrusion propagation, where an instance of an object is a "version" of the object with a specific timestamp; 2) build a Bayesian network (BN) based on the SOIDG to leverage the intrusion evidence collected from various information sources. Intrusion evidence can be the abnormal system and network activities that are noticed by human admins or security sensors such as Intrusion Detection Systems (IDSs). With the evidence, the SOIDG-based BN can quantitatively compute the probabilities of object instances being infected. Connected through dependency relations, the instances with high infection probabilities form a path, which can be viewed as a candidate zero-day attack path. As a result, the SOIDG-based BN can significantly narrow down the set of suspicious objects and make the manual verification of the zero-day attack paths feasible.

This approach is proposed based on the following insights: 1) A BN is able to capture cause-and-effect relations, and thus can be used to model the infection propagation among instances of different system objects: the cause is an already infected instance of one object, while the effect is its infection to an innocent instance of another object. We name this cause-and-effect relation as a type of *infection causality*, which is formed due to the interaction between the two objects in a system call operation. 2) An SOIDG can reflect the infection propagation process by capturing the dependencies among instances of different system objects. 3) Based on above insights, a BN can be constructed on top of the SOIDG because they couple well with each other: the dependencies among instances of different system objects can be directly interpreted into infection causalities in the BN. The BN's graphical nature makes it fit well with SOIDG.

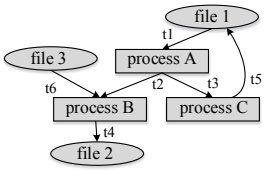
We made the following contributions.

- To the best of our knowledge, this work is the first probabilistic approach towards zero-day attack path identification.
- We proposed constructing Bayesian network at the low system object level by introducing System Object Instance Dependency Graph.
- We have designed and implemented a system prototype named Pr0bA, which can successfully identify the zero-day attack paths.

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

```
t1: process A reads file 1
t2: process A creates process B
t3: process A creates process C
t4: process B writes file 2
t5: process C writes file 1
t6: process B reads file 3
```

(a) simplified system call log in time-order



(b) SODG

Fig. 1: An SODG generated by parsing an example set of simplified system call log. The label on each edge shows the time associated with the corresponding system call.

2 Motivation and Approach Overview

2.1 System Object Dependency Graph

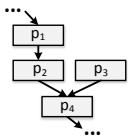
This paper classifies OS-level entities in UNIX-like systems into three types of objects: processes, files and sockets. The operating system performs a set of operations towards these objects via system calls such as read, write, etc. For instance, a process can read from a file as input, and then write to a socket. Such interactions among system objects enable intrusions to propagate from one object to another. Generally an intrusion starts with one or several seed objects that are created directly or indirectly by attackers. The intrusion seeds can be processes such as compromised service programs, or files such as viruses, or corrupted data, etc. As the intrusion seeds interact with other system objects via system call operations, the innocent objects can get infected. We call this process as *infection propagation*. Therefore the intrusion will propagate throughout the system, or even propagate to the network through socket communications.

To capture the intrusion propagation, previous work [10, 16, 17] has explored constructing System Object Dependency Graphs (SODGs) by parsing system call traces. Each system call is interpreted into three parts: a source object, a sink object, and a dependency relation between them. The objects and the dependencies respectively become nodes and directed edges in SODGs. For example, a process reading a file in the system call *read* indicates that the process (sink) depends on the file (source). The dependency is denoted as *file→process*. Similar rules (Table 5 in Appendix) as used in previous work [10, 16, 17] can be adopted to generate dependencies from system calls. Fig. 1b is an example SODG generated by parsing the simplified system call log shown in Fig. 1a.

2.2 Why use Bayesian Network?

The SODG can be used directly to identify the candidate zero-day attack paths through forward and backward tracking from intrusion detection points. However, such tracking will result in an explosion in the number and size of zero-day attack path candidates. The explosion is two-fold. First, in addition to real zero-day attack paths, the number of false positive path candidates is proportional to the number of false alerts. Second, an individual candidate path may contain too many objects for security analysts to comprehend, because it preserves every tracking-reachable object. Therefore, discerning from the candidates and verifying the real paths becomes difficult.

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.



CPT at node p_2

	p_1=T	p_1=F
p_2=T	0.9	0.01
p_2=F	0.1	0.99

Fig. 2: An example Bayesian network.

A Bayesian network can effectively deal with the explosion problem. The key reason is that a BN can quantitatively compute the probabilities of objects being infected through incorporating intrusion evidence from a variety of information sources. By only focusing on the objects with high infection probabilities, the set of suspicious objects can be significantly narrowed down. The candidate zero-day attack paths formed by the high-probability objects through dependency relations can be of manageable size.

The BN is a probabilistic graphical model that represents the cause-and-effect relations. It is formally defined as a Directed Acyclic Graph (DAG) that contains a set of nodes and directed edges, where a node denotes a variable of interest, and an edge denotes the causality relations between two nodes. The strength of such causality relation is indicated using a conditional probability table (CPT). Fig. 2 shows an example BN and the CPT tables associated with p_2. Given p_1 is true, the probability of p_2 being true is 0.9, which can be represented with $P(p_2 = T|p_1 = T) = 0.9$. Similarly, the probability of p_4 can be determined by the states of p_2 and p_3 according to a CPT table at p_4. BN is able to incorporate the collected evidence by updating the posterior probabilities of interested variables. For example, after evidence $p_2 = T$ is observed, it can be incorporated by computing probability $P(p_1 = T|p_2 = T)$.

2.3 Problems of Constructing Bayesian Network based on SODG

SODG has the potential to serve as the base of BN construction. For one thing, BN has the capability of capturing cause-and-effect relations in infection propagation. For another thing, SODG reflects the dependency relations among system objects. Such dependencies imply and can be leveraged to construct the infection causalities in BN. For example, the dependency *process A→file 1* in an SODG can be interpreted into an infection causality relation in BN: *file 1* is likely to be infected if *process A* is already infected. In such a way, an SODG-based BN can be constructed by directly taking the structure topology of SODG.

However, several drawbacks of the SODG prevent it from being the base of BN. First, an SODG without time labels cannot reflect the correct information flow according to the time order of system call operations. This is a problem because the time labels cannot be preserved when constructing BNs based on SODGs. Lack of time information will cause incorrect causality inference in the SODG-based BNs. For example, without the time labels, the dependencies in Fig. 1b indicates infection causality relations existing among *file 3*, *process B* and *file 2*, meaning that if *file 3* is infected, *process B* and *file 2* are likely to be infected by *file 3*. Nevertheless, the time information shows that the system call operation *"process B reads file 3"* happens at time *t6*, which is after the

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

operation *"process B writes file 2"* at time *t4*. This implies that the status of *file 3* has no direct influence on the status of *file 2*.

Second, the SODG contains cycles among nodes. For instance, *file 1*, *process A* and *process C* in Fig. 1b form a cycle. By directly adopting the topology of SODG, the SODG-based BN inevitably inherits cycles from SODG. However, the BN is an *acyclic* probabilistic graphical model that does not allow any cycles.

Third, a node in an SODG can end up with having too many parent nodes, which will render the CPT assignment difficult and even impractical in the SODG-based BN. For example, if *process B* in Fig. 1b continuously reads hundreds of files (which is normal in a practical operating system), it will get hundreds of file nodes as its parents. In the corresponding SODG-based BN, if each file node has two possible states that are *"infected"* and *"uninfected"*, and the total number of parent file nodes are denoted as n, then the CPT table at *process B* has to assign 2^n numbers in order to specify the infection causality of the parent file nodes to *process B*. This is impractical when n is very large.

To address the above problems, we propose a new type of dependency graph, *System Object Instance Dependency Graph*, which is a mutation of SODG.

2.4 System Object Instance Dependency Graph

In SOIDG, each node is not an object, but an instance of the object with a certain timestamp. Different instances are different "versions" of the same object at different time points, and can thus have different infection status.

Definition 1. *System Object Instance Dependency Graph (SOIDG)*
If the system call trace in a time window $T[t_{begin}, t_{end}]$ is denoted as Σ_T and the set of system objects (mainly processes, files or sockets) involved in Σ_T is denoted as O_T, then the SOIDG is a directed graph $G_T(V, E)$, where:

- V is the set of nodes, and initialized to empty set $\varnothing$;
- E is the set of directed edges, and initialized to empty set $\varnothing$;
- If a system call $syscall \in \Sigma_T$ is parsed into two system object instances src_i, $sink_j$, $i, j \geq 1$, and a dependency relation dep_c: $src_i \rightarrow sink_j$ (according to dependency rules in Table 5), where src_i is the i^{th} instance of system object $src \in O_T$, and $sink_j$ is the j^{th} instance of system object $sink \in O_T$, then $V = V \cup \{src_i, sink_j\}$, $E = E \cup \{dep_c\}$. The timestamps for *syscall*, dep_c, src_i, and $sink_j$ are respectively denoted as *t_syscall, t_*dep_c*,t_*src_i*, and t_*$sink_j$. The *t_*dep_c inherits *t_syscall* from *syscall*. The indexes i and j are determined before adding src_i and $sink_j$ into V by:
  - For $\forall$ src_m, $sink_n \in V$, $m, n \geq 1$, if i_{max} and j_{max} are respectively the maximum indexes of instances for object *src* and *sink*, and;
  - If $\exists$ $src_k \in V$, $k \geq 1$, then $i = i_{max}$, and *t_*src_i stays the same; Otherwise, $i = 1$, and *t_*src_i is updated to *t_syscall*;
  - If $\exists$ $sink_z \in V$, $z \geq 1$, then $j = j_{max}$+1; Otherwise, $j = 1$. In both cases *t_*$sink_j$ is updated to *t_syscall*; If $j \geq 2$, then $E = E \cup \{dep_s$: $sink_{j-1} \rightarrow sink_j\}$.
- If $a \rightarrow b \in E$ and $b \rightarrow c \in E$, then c transitively depends on a.

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

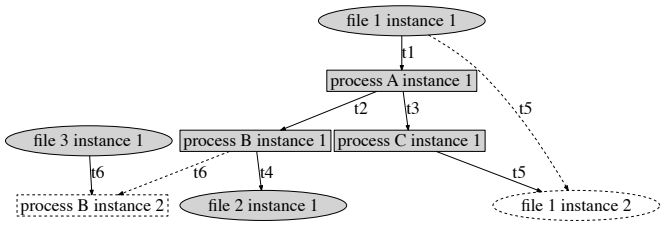


Fig. 3: An SOIDG generated by parsing the same set of simplified system call log as in Fig. 1a. The label on each edge shows the time associated with the corresponding system call operation. The dotted rectangle and ellipse are new instances of already existed objects. The solid edges and the dotted edges respectively denote the contact dependencies and the state transition dependencies.

According to Definition 1, for *src* object, a new instance is created only when no instances of *src* exist in the SOIDG. For *sink* object, however, a new instance is created whenever a *src*→*sink* dependency appears. The underlying insight is that the status of the *src* object will not be altered by the *src*→*sink*, while the status of *sink* will be influenced. Hence a new instance for an object should be created when the object has the possibility of being affected. A dependency dep_c is added between the most recent instance of *src* and the newly created instance of *sink*. We name dep_c as *contact dependency* because it is generated by the contact between two different objects through a system call operation.

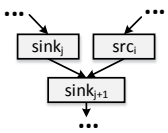
In addition, when a new instance is created for an object, a new dependency relation dep_s is also added between the most recent instance of the object and the new instance. This is necessary and reasonable because the status of the new instance can be influenced by the status of the most recent instance. We name dep_s as *state transition dependency* because it is caused by the state transition between different instances of the same system object.

The SOIDG can well tackle the problems existing in the SODG for constructing BNs. It can be illustrated using Fig. 3, an SOIDG created for the same simplified system call log as in Fig. 1a. First, the SOIDG is able to reflect correct information flows by implying time information through creating object instances. For example, instead of parsing the system call at time *t6* directly into *file 3*→*process B*, Fig. 3 parsed it into *file 3 instance 1*→*process B instance 2*. Comparing to Fig. 1b in which *file 3* has indirect infection causality on *file 2* through *process B*, the SOIDG in Fig. 3 indicates that *file 3* can only infect *instance 2* of *process B* but no previous instances. Hence in this graph *file 3* does not have infection causality on *file 2*.

Second, SOIDGs can break the cycles contained in SODGs. Again, in Fig. 3, the system call at time *t5* is parsed into *process C instance 1*→*file 1 instance 2*, rather than *process C*→*file 1* as in Fig. 1b. Therefore, instead of pointing back to *file 1*, the edge from process C is directed to a new instance of *file 1*. As a result, the cycle formed by *file 1, process A* and *process C* is broken.

Third, the mechanism of creating new *sink* instances for a relation *src*→*sink* prevents the nodes in SOIDGs from getting too many parents. For example,

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.



CPT at node $sink_{j+1}$

	$sink_j$=Infected		$sink_j$=Uninfected	
	src_i=Infected	src_i=Uninfected	src_i=Infected	src_i=Uninfected
$sink_{j+1}$=Infected	1	1	τ	ρ
$sink_{j+1}$=Uninfected	0	0	$1-\tau$	$1-\rho$

Fig. 4: The infection propagation models.

process B instance 2 in Fig. 3 has two parents: *process B instance 1* and *file 3 instance 1*. If *process B* appears again as the *sink* object in later *src→sink* dependencies, new instances of *process B* will be created instead of directly adding *src* as the parent to *process B instance 2*. Therefore, a node in the SOIDG only has 2 parents at most: one is the previous instance for the same object; the other one is an instance for a different object that the node depends on.

3 SOIDG-based Bayesian Networks

To build a BN based on an SOIDG and compute probabilities for interested variables, two steps are required. First, the CPT tables have to be specified for each node via constructing proper infection propagation models. Second, evidence from different information sources has to be incorporated into BN for subsequent probability inference.

3.1 The Infection Propagation Models

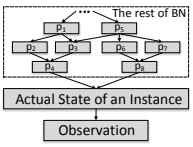
In SOIDG-based BNs, each object instance has two possible states, *"infected"* and *"uninfected"*. The strength of the infection causalities among the instances has to be specified in corresponding CPT tables. Our infection propagation models in this paper deal with two types of infection causalities, *contact infection causalities* and *state transition infection causalities*, which correspond to the contact dependencies and state transition dependencies in SOIDGs.

Contact Infection Causality Model. This model captures the infection propagation between instances of two different objects. Fig. 4 shows a portion of BN constructed when a dependency *src→sink* occurs and the CPT table associated with $sink_{j+1}$. When $sink_j$ is uninfected, the probability of $sink_{j+1}$ being infected depends on the infection status of src_i, a *contact infection rate* τ and an *intrinsic infection rate* ρ, $0 \leq \tau, \rho \leq 1$.

The intrinsic infection rate ρ decides how likely $sink_{j+1}$ gets infected given src_i is uninfected. In this case, since src_i is not the infection source of $sink_{j+1}$, if $sink_{j+1}$ is infected, it should be caused by other factors. So ρ can be determined by the prior probabilities of an object being infected, which is usually a very small constant number.

The contact infection rate τ determines how likely $sink_{j+1}$ gets infected when src_i is infected. The value of τ determines to which extent the infection can be propagated within the range of an SOIDG. In an extreme case where $\tau = 1$, all the object instances will get contaminated as long as they have contact with the infected objects. In another extreme case where $\tau = 0$, the infection will be confined inside the infected object and does not propagate to any other contacting object instances. Our system allows security experts to tune the value of τ based on their knowledge and experience.

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.



CPT at node Observation

	Actual=Infected	Actual=Uninfected
Observation=True	0.9	0.15
Observation=False	0.1	0.85

False negative rate

False positive rate

Fig. 5: Local observation model.

Since a large number of system call traces with ground truths are often unavailable, it is very unlikely to learn the parameters of τ and ρ using statistical techniques. Hence, currently these parameters have to be assigned by security experts. We will evaluate the impact of τ and ρ in Section 6.2.

State Transition Infection Causality Model. This model captures the infection propagation between instances of the same objects. We follow one rule to model this type of causalities: an object will never return to the state of *"uninfected"* from the state of *"infected"*[4]. That is, once an instance of an object gets infected, all future instances of this object will remain the infected state, regardless of the infection status of other contacting object instances. This rule is enforced in the CPT tables as exemplified in Fig. 4. If $sink_j$ is infected, the infection probability of $sink_{j+1}$ keeps to be 1, no matter whether src_i is infected or not. If $sink_j$ is uninfected, the infection probability of $sink_{j+1}$ is decided by the infection status of src_i according to the contact infection causality model.

3.2 Evidence Incorporation

BN is able to incorporate security alerts from a variety of information sources as the evidence of attack occurrence. Numerous ways have been developed to capture intrusion symptoms, which can be caused by attacks exploiting both known vulnerabilities and zero-day vulnerabilities. A tool Wireshark [12] can notice a back telnet connection that is instructed to open; an IDS such as Snort [13] may recognize a malicious packet; a packet analyzer tcpdump [14] can capture suspicious network traffic, etc. In addition, human security admins can also manually check the system or network logs to discover other abnormal activities that cannot be captured by security sensors. As more evidence is fed into BN, the identified zero-day attack paths get closer to real facts.

In this paper, we adopt two ways to incorporate evidence. First, add evidence directly on a node by providing the infection state of the instance. If human security experts have scrutinized an object and proven that an object is infected at a specific time, they can feed the evidence to the SOIDG-based BN by directly changing the infection status of the corresponding instance into *infected*. Second, leverage the local observation model (LOM) [22] to model the uncertainty towards observations. Human security admins or security sensors may notice suspicious activities that imply attack occurrence. Nonetheless, these observations often suffer from false rates. As shown in Fig. 5, an observation node can be added as the direct child node to an object instance. The implicit causality

[4] This rule is formulated based on the assumptions that no intrusion recovery operations are performed and attackers only conduct malicious activities.

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

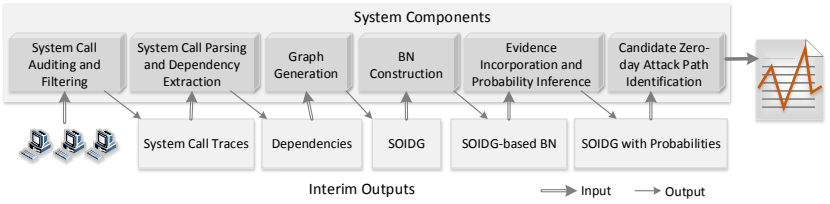


Fig. 6: System design.

relation is that the actual state of the instance can likely affect the observation to be made. If the observation comes from security alerts, the CPT inherently indicates the false rates of the security sensors. For example, $P(Observation = True \mid Actual = Uninfected)$ shows the false positive rate and $P(Observation = False \mid Actual = Infected)$ indicates the false negative rate.

4 System Design

Fig. 6 shows the overall system design, which includes 7 components.

System call auditing and filtering. System call auditing is performed against all running processes and should preserve sufficient OS-aware information. Subsequent system call reconstruction can thus accurately identify the processes and files by their process IDs or file descriptors. The filtering process basically prunes system calls that involve redundant and very likely innocent objects, such as the dynamic linked library files or some dummy objects. We conduct system call auditing at run time towards each host in the enterprise network.

System call parsing and dependency extraction. The collected system call traces are then sent to a central machine for off-line analysis, where the dependency relations between system objects are extracted according to Table 5.

Graph generation. The extracted dependencies are then analyzed line by line for graph generation. The generated graph can be either host-wide or network-wide, depending on the analysis scope. A network-wide SOIDG can be constructed by concatenating individual host-wide SOIDGs through instances of the communicating sockets. Algorithm 1 is the basis algorithm for SOIDG generation, which is designed according to the logic in Definition 1.

BN construction. The BN is constructed by taking the topology of an SOIDG. The instances and dependencies in an SOIDG become nodes and edges in BN. Basically the nodes and the associated CPT tables are specified in a *.net* file, which is one file type that can carry the SOIDG-based BN.

Evidence incorporation and probability inference. Evidence is incorporated by either providing the infection state of the object instance directly, or constructing an local observation model (LOM) for the instance. After probability inference, each node in the SOIDG receives a probability.

Candidate Zero-day Attack Paths Identification. To reveal the candidate zero-day attack paths from the mess of SOIDG, the nodes with high probabilities are to be preserved, while the link between them should not be broken. We implemented Algorithm 2 on the basis of depth-first search (DFS) algorithm [24]

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

to tag each node in the SOIDG as either possessing high probability itself, or having both an ancestor and a descendant with high probabilities. The tagged nodes are the ones that actually propagate the infection through the network, and thus should be preserved in the final graph. Our system allows a probability threshold to be tuned for recognizing high-probability nodes. For example, if the threshold is set at 80%, only instances that have the infection probabilities of 80% or higher will be recognized as the high-probability nodes.

5 Implementation

The whole system includes online system call auditing and off-line data analysis. For system call auditing, we share with Patrol [10] the same component that is implemented with a loadable kernel module. For the off-line data analysis, our prototype is implemented with approximately 2915 lines of gawk code that constructs a *.net* file for the SOIDG-based BN and a *dot*-compatible file for visualizing the candidate zero-day attack paths in Graphviz [25], and 145 lines of Java code for probability inference, leveraging the API provided by the BN tool SamIam [23].

An SOIDG can be too large due to the introduction of instances. Therefore, in addition to system call filtering, we also develop several ways to prune that SOIDGs while not impede reflecting the major infection propagation process.

One helpful way is to ignore the repeated dependencies. It is common that the same dependency may happen between two system objects for a number of times, even through different system call operations. For example, *process A* may write *file 1* for several times. In such cases, each time the *write* operation occurs, a new instance of *file 1* is created and a new dependency is added between the most recent instance of *process A* and the new instance of *file 1*. If the status of *process A* is not affected by any other system objects during this time period, the infection status of *file 1* will not change neither. Hence the new instances of *file 1* and the related new dependencies become redundant information in understanding the infection propagation. Therefore, a repeated $src \rightarrow sink$ dependency can be ignored if the *src* object is not influenced by other objects since the last time that the same $src \rightarrow sink$ dependency appeared.

Another way to simplify an SOIDG is to ignore the root instances whose original objects have never appear as the *sink* object in a $src \rightarrow sink$ dependency during the time period of being analyzed. For instance, *file 3* in Fig. 3 only appears as the *src* object in the dependencies parsed from the system call log in Fig. 1a, so *file 3 instance 1* can be ignored in the simplified SOIDG. Such instances are not influenced by other objects in the specified time window, and thus are not manipulated by attackers, neither. Hence ignoring these root instances does not break any routes of intrusion sequence and will not hinder the understanding of infection propagation. This method is helpful for situations such as a process reading a large number of configuration or header files.

A third way to prune an SOIDG is to ignore some repeated mutual dependencies, in which two objects will keep affecting each other through creating new instances. One situation is that a process can frequently send and receive

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

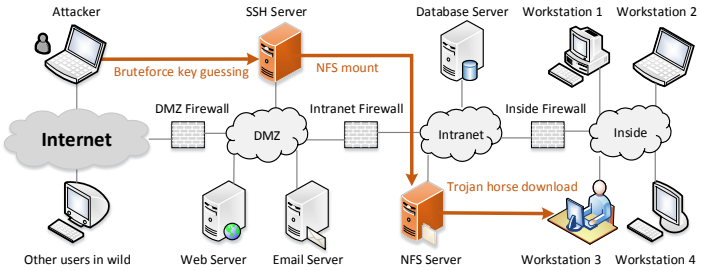


Fig. 7: Attack scenario.

messages from a socket. For example, in one of our experiments, 107 new instances are created respectively for the process (*pid:6706, pcmd:sshd*) and the socket (*ip:192.168.101.5, port: 22*) due to their interaction. Since no other objects are involved during this procedure, the infection status of these two objects will keep the same through all the new instances. Thus a simplified SOIDG can preserve the very first and last dependencies while neglect the middle ones. Another situation is that a process can frequently take input from a file and then write the output to it again after some operations. The middle repeated mutual dependencies could also be ignored in a similar way.

6 Experiments

6.1 Attack Scenario

We built a test-bed network and launched a three-step attack towards it. Fig. 7 illustrates the attack scenario, which is similar to the one in [10]. Step 1, the attacker exploits vulnerability CVE-2008-0166 to gain root privilege on SSH Server through a brute-force key guessing attack. Step 2, since the export table on NFS Server is not set up appropriately, the attacker can upload a malicious executable file to a public directory on NFS. The malicious file contains a Trojan-horse that can exploit CVE-2009-2692. The public directory is shared among all the hosts in the test-bed network. Step 3, once the malicious file is mounted and installed on the Workstation 3, the attacker is able to execute arbitrary code on Workstation 3. To capture the intrusion evidence for subsequent BN probability inference, we deployed security sensors in the test-bed, such as firewalls, Snort, Tripwire, Wireshark, Ntop [26] and Nessus. For sensors that need configuration, we tailored their rules or policy files to match our hosts.

Since zero-day exploits are not readily available, we emulate zero-day vulnerabilities with known vulnerabilities. For example, we treat CVE-2009-2692 as a zero-day vulnerability by assuming the current time is Dec 31, 2008. In addition, the configuration error on NFS is also viewed as a special type of unknown vulnerability because it is ruled out by vulnerability scanners like Nessus. The strategy of emulation also brings another benefit. The information for these "*known* zero-day" vulnerabilities can be available to verify the correctness of our experiment results.

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

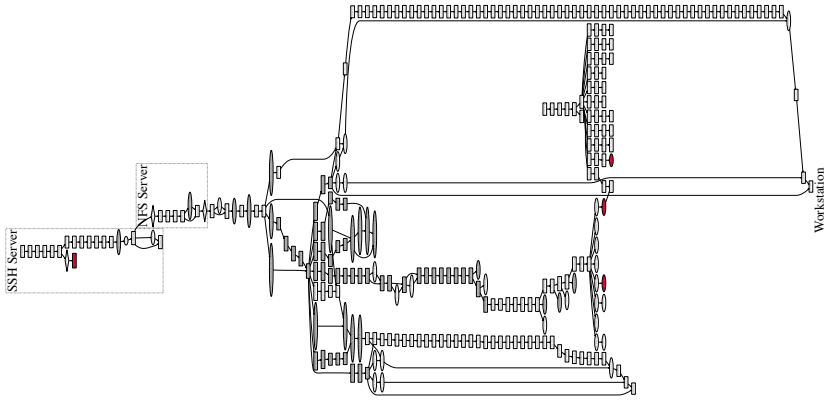


Fig. 8: The zero-day attack path in the form of SOIDG.

6.2 Experiment Results

While conducting the three-step attack, we simultaneously log the system calls on each host and collect the security alerts. After analyzing a total number of 143120 system calls generated by three hosts, an SOIDG-based BN with 1853 nodes and 2249 edges is constructed. The evidence as in Table 1 is collected and fed into BN.

Correctness. Given the evidence, Fig. 8 illustrates the identified candidate zero-day attack paths in the form of an SOIDG, with the contact infection rate τ as 0.9, the intrinsic infection rate ρ as 0.001, and the probability threshold of recognizing high-probability nodes as 80%. The processes, files, and sockets are denoted with rectangles, ellipses, and diamonds respectively. We mark the evidence with red color and the nodes that are verified to be malicious with grey color. Therefore, Fig. 8 shows that our approach can successfully reveal the actual zero-day attack path. It is worth noting that although no evidence is provided on NFS Server, but the identified attack path can still demonstrate how NFS Server contributes to the overall intrusion propagation: the file *workstation_attack.tar.gz* is uploaded from SSH Server to the */exports* directory on NFS Server, and then downloaded to */mnt* on Workstation 3. More importantly, the identified path can expose key objects that are related to the exploits of zero-day vulnerabilities. For example, the identified system objects on NFS Server can alert system admins for possible configuration errors because SSH Server should not have the privilege of writing to the */exports* directory. As another example, the object *PAGE0: memory(0-4096)* on Workstation is also exposed as highly suspicious on the identified attack path. Page-zero is actually what triggers the null pointer dereference and enables attackers gain privilege on Workstation 3. Therefore, exposing the page-zero object can help system admins to further diagnose how the intrusion happens and propagates.

An additional merit of our approach is that the SOIDG-based BN can clearly show the state transitions of an object using instances. By matching the in-

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

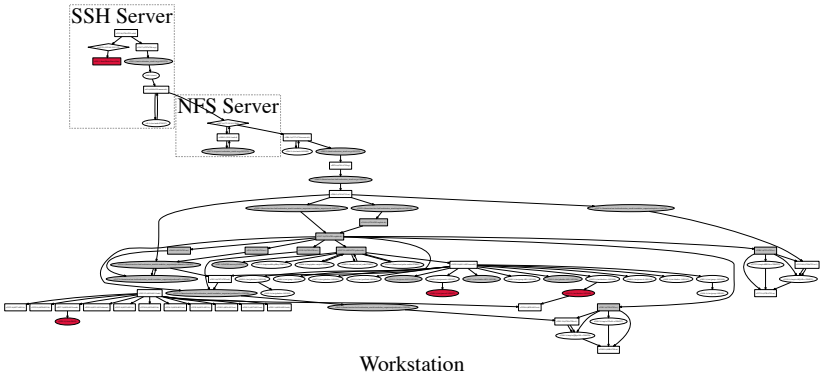


Fig. 9: The zero-day attack path in the form of SODG.

stances and dependencies back to the system call traces, it can even find out the exact system call that causes the state-changing of the object. For example, the node x2086.4*:(6763:6719:tar)* in Fig. 8 represents the fourth instance of process (*pid:6763, pcmd:tar*). Previous instances of the process are considered as innocent because of their low infection probabilities. The process becomes highly suspicious only after a dependency occurs between node x2082.2*:(/home/user/test-bed/workstation_attack.tar.gz:1384576)* and node x2086.4. Matching the dependency back to the system call traces reveals that the state change of the process is caused by "*syscall:read, start:827189, end:827230, pid:6763, ppid:6719, pcmd:tar, ftype:REG, pathname:/home/user/test-bed/workstation_attack.tar.gz, inode:1384576*", a system call indicating that the process reads a suspicious file.

Table 1: The Collected Evidence

ID	Host	Evidence
E1	SSH Server	Snort messages "potential SSH brute force attack"
E2	Workstation	Tripwire reports "/virus is added"
E3	Workstation	Tripwire reports "/etc/passwd is modified"
E4	Workstation	Tripwire reports "/etc/shadow is modified"

Size of Candidate Zero-day Attack Paths. If all the instances belonging to the same object are merged into one node, we will generate a zero-day attack path in the form of SODG as shown in Fig. 9. This path contains only objects and can be used for verification when details regarding instances are not needed. The main candidate path identified by Patrol contains 175 objects, while the path by our system is composed of only 77 objects, and thus can be verified with ease. Considering that the total number of objects involved in original SOIDG is only 913, the 56% reduction of path size is substantial. Our further investigation shows that when the time period of being analyzed is longer, our system can generate candidate paths much smaller than Patrol without hurting the correctness of the paths.

Influence of Evidence. We choose a number of nodes in Fig. 8 as the representative interested instances. Table 2 shows how the infection probabilities

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

of these instances change after each piece of evidence is fed into BN. We assume the evidence is observed in the order of attack sequence. The results show that when no evidence is available, the infection probabilities for all nodes are very low. When *E1* is added, only a few instances on SSH Server receive probabilities higher than 60%. After *E2* is observed, the infection probabilities for instances on Workstation 3 increase, but still not much. As *E3* and *E4* arrive, 5 of the 9 representative instances on all three hosts become highly suspicious. Therefore, the evidence makes the instances on the actual attack paths emerge gradually from the "sea" of instances in the SOIDG. However, it is also possible that the arrival of some evidence may decrease the probabilities of certain instances, so that these instances will get removed from the final path. In a word, as more evidence is collected, the revealed zero-day attack paths become closer to the actual fact.

Table 2: The Influence of Evidence

Evidence	***SSH Server***			***NFS Server***		***Workstation***			
	x4.1	x10.1	x253.3	x1007.1	x1017.1	x2006.2	x2083.1	x2108.1	x2311.32
No Evi.	0.56%	0.51%	0.57%	0.51%	0.54%	0.54%	0.51%	0.51%	1.21%
E1	63.76%	57.38%	79.13%	57.38%	46.54%	41.92%	37.75%	24.89%	26.93%
E2	63.76%	57.38%	79.13%	57.38%	46.94%	42.58%	38.34%	27.04%	30.09%
E3	86.82%	78.14%	80.76%	84.50%	75.63%	81.26%	79.56%	75.56%	81.55%
E4	86.84%	78.16%	80.77%	84.53%	75.65%	81.3%	79.59%	75.60%	81.66%

Influence of False Alerts. We assume that *E4* is a false alarm generated by Tripwire and evaluate its influence to the BN output. Table 3 shows that when only one piece of evidence exists, the observation of *E4* will at least greatly influence the probabilities of some instances on Workstation 3. However, when other evidence is fed into BN, the influence of *E4* decreases. For instance, given just *E1*, the infection probability of $x2006.2$ is 97.78% when *E4* is true, but should be 29.96% if *E4* is a false alert. Nonetheless, if all other evidence is already input into BN, the infection probability of $x2006.2$ only changes from 81.13% to 81.3% if *E4* becomes a false alert. Therefore, the impact of false alerts can be reduced substantially if sufficient evidence is collected.

Table 3: The Influence of False Alerts

Evidence		x4.1	x10.1	x253.3	x1007.1	x1017.1	x2006.2	x2083.1	x2108.1	x2311.32
Only E1	E4=True	98.46%	88.62%	81.59%	98.20%	88.30%	97.78%	97.67%	90.23%	94.44%
	E4=False	56.33%	50.70%	78.60%	48.65%	37.60%	29.96%	24.92%	10.89%	12.48%
All Evidence	E4=True	86.84%	78.16%	80.77%	84.53%	75.65%	81.3%	79.59%	75.60%	81.66%
	E4=False	86.74%	78.06%	80.76%	84.41%	75.54%	81.13%	79.42%	75.39%	81.38%

Sensitivity Analysis and Influence of τ and ρ. We also performed sensitivity analysis and evaluated the impact of the contact infection rate τ and the intrinsic infection rate ρ by tuning these numbers. ρ is usually set at a very low value, so our experiment results are not very sensitive to the value of ρ. Since τ decides how likely $sink_j$ get infected given src_i is infected in a $src_i \rightarrow sink_j$ dependency, the value of τ will definitely influence the probabilities produced by BN. If a node is marked as infected, other nodes that are directly or indirectly connected to this node should expect higher infection probabilities when

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

τ is bigger. Our experiments show that adjusting τ within a small range (e.g. changing from 0.9 to 0.8) does not influence the output probabilities much, but a major adjustment of τ (e.g. changing it from 0.9 to 0.5) can largely affect the probabilities. However, we still argue that although τ influences the produced infection probabilities, it will not greatly affect the identification of zero-day attack paths. Our rationale is that the probability threshold of recognizing high-probability nodes for zero-day attack paths can be adjusted according to the value of τ. For example, when τ is a small number such as *50%*, even nodes that have low infection probabilities of around *40%* to *60%* should be considered as highly suspicious because it is hard for an instance to get infected with such a low contact infection rate.

Complexity. One concern of adopting SOIDG is that it can become too large due to introduction of instances. However, the techniques of pruning the SOIDG can significantly reduce the number of instances. Table 4 summarizes the total number of instances in SOIDGs for each host before and after the pruning. It shows that the number of instances can be reduced to an acceptable value.

The experiment results also show that the off-line data analysis is very efficient. Considering that our system shares the system call logging component with Patrol, we will not repeat the evaluation of its run-time performance overhead. We only evaluate time cost for the off-line data analysis, which includes the time for SOIDG-based BN generation, probability inference and zero-day attack path identification. The time cost for probability inference depends on the algorithm employed in SamIam. The time complexity can be $O(|V|^2)$ for both SOIDG-based BN generation and zero-day attack path identification, because the DFS algorithm is applied towards every node in the SOIDG. For our experiments, Table 4 already shows the time required for constructing the SOIDG-based BN for each host, so the total time of BN construction comes to around 27 seconds. For a BN with approximately 1854 nodes, assuming that the evidence is already fed into BN and the algorithm used is *recursive conditioning*, the average time cost is 1.57 seconds for BN compilation and probability inference, and 59 seconds for zero-day attack path identification. Combining all the time required together, the average data analysis speed is 280 KB/s, which is reasonable comparing to the system call generation speed of around 1.03 KB/s [10]. The average memory used for compiling a BN with approximately 1854 nodes is 4.32 Mb.

Table 4: The Impact of Pruning the SOIDG

	SSH Server		***NFS Server***		***Workstation***	
	before	after	before	after	before	after
number of syscalls in raw data trace	82133		14944		46043	
size of raw data trace (MB)	13.8		2.3		7.9	
number of extracted object dependencies	10310		11535		17516	
number of objects	349		20		544	
number of instances(nodes) in SOIDG	10447	745	11544	39	17849	1069
number of dependencies(edges) in SOIDG	20186	968	19863	37	34549	1244
number of contact dependencies	9888	372	8329	8	17033	508
number of state transition dependencies	10298	596	11534	29	17516	736
average time for graph generation(s)	14	11	6	5	13	11
.net file size(KB)	2000	123	2200	8	3600	180

7 Related Work

The work that is most related to us is Patrol. Our work differs from Patrol in several aspects. First, Patrol relies on "shadow indicators" to distinguish zero-day attack paths from other candidate paths. However, investigating and crafting shadow indicators requires human analysts' or even the whole community's efforts. Instead, our approach solely relies on the collected intrusion evidence to generate a zero-day attack path. Second, Patrol identifies the candidate zero-day attack paths by tracking from a trigger point. If the trigger points are provided by security sensors with high false rates, the identified paths can also suffer from certain false rates. In constrast, our system does not perform any tracking, but only relies on the computed probabilities. By taking various evidence in, the SOIDG-based BN can cope with false rates to a large extent. Third, Patrol only conducts qualitative analysis and treats every object on the identified paths as having the same malicious status. Scrutinizing every object on the path to verify its status is a daunting job, especially when the identified path is very big. Compared to Patrol, the SOIDG-based BN quantifies the infection status of system objects with probabilities. By only focusing on system objects with relatively high probabilities, we can significantly reduce the set of suspicious objects, and make the subsequent verification of zero-day attack paths practical.

Other related work includes system call dependency tracking and zero-day attack identification. System call dependency tracking is first proposed in [16] to help the understanding of intrusion sequence. It is then applied for alert correlation in [4, 5]. Instead of directly correlating these alerts, our system takes the alerts as evidence and quantitatively compute the infection probabilities of system objects. [27] conducts an empirical study to reveal the zero-day attacks by identifying the executable files that are linked to exploits of known vulnerabilities. A zero-day attack is identified if a malicious executable is found before the corresponding vulnerability is disclosed. Attack graphs have been employed to measure the security risks caused by zero-day attacks [19–21]. Nevertheless, the metric simply counts the number of required unknown vulnerabilities for compromising an asset, rather than detects the actually occurred zero-day exploits. Our system takes an approach that is quite different from the above work.

8 Limitation and Conclusion

The current system still has some limitations. For example, when some attack activities evade the system calls (although difficult, but possible), or the attack time span is much longer than the analyzed time period, the constructed SOIDG may not reflect the complete zero-day attack paths. In such cases, our system can only reveal partial of the paths.

This paper proposes to use Bayesian networks to identify the zero-day attack paths. For this purpose, a System Object Instance Dependency Graph is built to serve as the basis of Bayesian networks. By incorporating the intrusion evidence and computing the probabilities of objects being infected, the implemented system Pr0bA can successfully reveal the zero-day attack paths at run-time.

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

Disclaimer

References

1. V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. ACM Computing Surveys (CSUR), 2009.
2. C. Kruegel, D. Mutz, F. Valeur, and G. Vigna. On the detection of anomalous system call arguments. ESORICS, 2003.
3. S. Bhatkar, A. Chaturvedi, and R. Sekar. Dataflow anomaly detection. IEEE S&P, 2006.
4. S. T. King, Z. M. Mao, D. G. Lucchetti, P. M. Chen. Enriching intrusion alerts through multi-host causality. NDSS, 2005.
5. Y. Zhai, P. Ning, J. Xu. Integrating IDS alert correlation and OS-Level dependency tracking. IEEE Intelligence and Security Informatics, 2006.
6. S. Jajodia, S. Noel, and B. O'Berry. Topological analysis of network attack vulnerability. Managing Cyber Threats, 2005.
7. P. Ammann, D. Wijesekera, and S. Kaushik. Scalable, graph-based network vulnerability analysis. ACM CCS, 2002.
8. X. Ou, W. F. Boyer, and M. A. McQueen. A scalable approach to attack graph generation. ACM CCS, 2006.
9. X. Ou, S. Govindavajhala, and A. W. Appel. MulVAL: A Logic-based Network Security Analyzer. USENIX security, 2005.
10. J. Dai, X. Sun, and P. Liu. Patrol: Revealing zero-day attack paths through network-wide system object dependencies. ESORICS, 2013.
11. Symantec Report. http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v19_21291018.en-us.pdf
12. Wireshark. https://www.wireshark.org/.
13. Snort. https://www.snort.org/.
14. Tcpdump. http://www.tcpdump.org/.
15. Tripwire. http://www.tripwire.com/.
16. S. T. King, and P. M. Chen. Backtracking intrusions. ACM SIGOPS, 2003.
17. X. Xiong, X. Jia, and P. Liu. Shelf: Preserving business continuity and availability in an intrusion recovery system. ACSAC, 2009.
18. Nessus. http://www.tenable.com/products/nessus-vulnerability-scanner.
19. L. Wang, S. Jajodia, A. Singhal, and S. Noel. k-zero day safety: Measuring the security risk of networks against unknown attacks. ESORICS, 2010.
20. M. Albanese, S. Jajodia, A. Singhal, and L. Wang. An Efficient Approach to Assessing the Risk of Zero-Day Vulnerabilities. SECRYPT, 2013.
21. L. Wang, S. Jajodia, A. Singhal, P. Cheng, and S. Noel. k-Zero day safety: A network security metric for measuring the risk of unknown vulnerabilities. IEEE Transactions on Dependable and Secure Computing, 2014.
22. P. Xie, J. H. Li, X. Ou, P. Liu, and R. Levy. Using Bayesian networks for cyber security analysis. DSN, 2010.

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

23. SamIam. http://reasoning.cs.ucla.edu/samiam/.
24. R. Tarjan. Depth-first search and linear graph algorithms. SIAM journal on computing 1, 1972.
25. GraphViz. http://www.graphviz.org/.
26. Ntop. http://www.ntop.org/.
27. L. Bilge, and T. Dumitras. Before we knew it: an empirical study of zero-day attacks in the real world. ACM CCS, 2012.

Appendix

Table 5: System Call Dependency Rules

Dependency	Events	System calls
process→file	a process creates or writes a file	write, pwrite64, rename, mkdir, fchmod, chmod, fchownat, etc.
file→process	a process reads or executes a file	stat64, read, pread64, execve, etc.
process→process	a process creates or kill a process	vfork, fork, kill, etc.
process→socket	a process writes a socket	write, pwrite64, send, sendmsg, etc.
socket→process	a process reads a socket	read, pread64,recv, recvmsg, etc.
socket→socket	socket communication	sendmsg, recvmsg, etc.

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

Algorithm 1 Algorithm of SOIDG Generation

Require: set D of system object dependencies
Ensure: the SOIDG graph $G(V, E)$

for each *dep:* $src \rightarrow sink \in D$ **do**
  look up the most recent instance src_k of src, $sink_z$ of $sink$ in V
  if $sink_z \notin V$ **then**
    create new instances $sink_1$
    $V \leftarrow V \cup \{sink_1\}$
    if $src_k \notin V$ **then**
      create new instances src_1
      $V \leftarrow V \cup \{src_1\}$
      $E \leftarrow E \cup \{src_1 \rightarrow sink_1\}$
    else
      $E \leftarrow E \cup \{src_k \rightarrow sink_1\}$
    end if
  end if
  if $sink_z \in V$ **then**
    create new instance $sink_{z+1}$
    $V \leftarrow V \cup \{sink_{z+1}\}$
    $E \leftarrow E \cup \{sink_z \rightarrow sink_{z+1}\}$
    if $src_k \notin V$ **then**
      create new instances src_1
      $V \leftarrow V \cup \{src_1\}$
      $E \leftarrow E \cup \{src_1 \rightarrow sink_{z+1}\}$
    else
      $E \leftarrow E \cup \{src_k \rightarrow sink_{z+1}\}$
    end if
  end if
end for

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

Algorithm 2 Algorithm of Candidate Zero-day Attack Paths Identification

Require: the SOIDG graph $G(V, E)$, a vertex $v \in V$
Ensure: the candidate zero-day attack path $G_z(V_z, E_z)$

function $DFS(G, v, direction)$
  set v as *visited*
  if $direction = ancestor$ **then**
    set $next_v$ as parent of v that $next_v \rightarrow v \in E$
    set *flag* as *has_high_probability_ancestor*
  else if $direction = descendant$ **then**
    set $next_v$ as child of v that $v \rightarrow next_v \in E$
    set *flag* as *has_high_probability_descendant*
  end if
  for all $next_v$ of v **do**
    if $next_v$ is not labeled as *visited* **then**
      if the probability for $next_v$ $prob[next_v] \geq threshold$ **or** $next_v$ is marked as *flag* **then**
        set *find_high_probability* as $True$
      else
        $DFS(G, next_v, direction)$
      end if
    end if
    if *find_high_probability* is $True$ **then**
      mark v as *flag*
    end if
  end for
end function
for all $v \in E$ **do**
  $DFS(G, v, ancestor)$
  $DFS(G, v, descendant)$
end for
for all $v \in V$ **do**
  if $prob[v] \geq threshold$ **or** (v is marked as *has_high_probability_ancestor* **and** v is marked as *has_high_probability_descendant*) **then**
    $V_z \leftarrow V_z \cup v$
  end if
end for
for all $e : v \rightarrow w \in E$ **do**
  if $v \in V_z$ **and** $w \in V_z$ **then**
    $E_z \leftarrow E_z \cup e$
  end if
end for

Jun, Dai; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yen, John.
"Towards Probabilistic Identification of Zero-day Attack Paths."
Paper presented at 2016 IEEE Conference on Communications and Network Security (CNS), Philadelphia, PA, United States. October 17, 2016 - October 19, 2016.

Formal Definition of Edge Computing: An Emphasis on Mobile Cloud and IoT Composition

Charif Mahmoudi[†‡], Fabrice Mourlin[†] and Abdella Battou[‡]
[†]Advanced Network Technologies Division, National Institute of Standards and Technology
[‡]Algorithmic, Complexity and Logic Laboratory, University of Paris-Est Créteil
charif.mahmoudi@nist.gov, fabrice.mourlin@u-pec.fr, abdella.battoug@nist.gov

Abstract—Within the Edge computing umbrella, mobile cloud computing is an emerging area where two trends come together to compose its major pillars. On one hand, the virtualization affecting the data centers hypervisors. On the other hand, device's mobility, especially Smart Phones, which proved to be the most effective and convenient tools in human life. This emerging area is then changing the game in terms of mobility of workspaces and the interaction with the connected devices and sensors. This paper provides a formal specification of the Mobile cloud component using the π-calculus. The proposed model defines the mobile cloud component, the virtual device representation, and interaction that leads to application offloading and device composition. This paper describe our contribution that enables the composition of virtual devices from physical devices, sensors, and actuators available on the network. Moreover, we present a model of application offloading and virtual devices networking on mobile clouds. Our architectural model is inspired from the Cloudlet based system. In addition to the formal specifications and architecture this paper presents a case studies showing the structural congruence between a locally executed application and an offloaded version of that same application.

Keywords— formal definition; migration; mobile; mobile cloud computing; offloading; virtualization; virtual device representation, fog, internet of things

I. INTRODUCTION

Mobile devices are increasingly having an essential usage in human life as the most effective and convenient communication tools. The unbounded time and place usage introduced by those devices allows mobile users to accumulate a rich experience of various services and applications. The execution of those services is not limited to the mobile device itself, more and more applications use nowadays remote servers via wireless networks to interact with services. Architectures based on the n-tiers computing have become a powerful trend in the development of IT technology as well as in the commerce and industry fields on mobile computing [1]. Such a systems can accept any (finite) number of layers (or tiers). Where each tier like presentation, application processing, and data management functions is physically separated from the others.

However, mobile devices have considerable hardware limitations. Mobile computing faces many challenges in attempting to provide the various applications living on a single device with limited resources such as battery, storage, and bandwidth. Communication challenges like mobility and security arise too. Those challenges motivate the delegation of the resources-consuming application modules to remote servers using the cloud service platforms. Google offers one of the major solutions called AppEngine [2]. Such a solution is allowing developers without previous understanding or knowledge of cloud technology infrastructure to deploy services and use the cloud. These platforms execute the deployed services and expose them as a remote service. That enables delegation of massive computation pieces of the mobile software to the cloud infrastructure.

As one component of the Edge computing, current mobile cloud architectures are based on cloud computing abstractions (IaaS, PaaS and SaaS) [3] and adapt this concepts for a deployement at the edge of the network. This architecture addresses the virtualization and distribution of the deployed services. However, the mobility aspect is not designed for the nomadic usage of mobile devices. The lack of specific formalism to address mobile virtualisation contribute to the heterogeneity of the actual solutions. Indeed, the virtualisation of devices and services is following the server architectures that are not suitable for the mobile platforms. This is due to the heterogeneity of the hardware architectures and the available resources. Another limitation is the lack of specific representation of the mobile devices on the cloud. The deployed services artefacts are a classical web service. There is no specific representation that makes abstraction for the application offloading and the location management. Moreover, using a generic representation makes the remote services implementations dependant of both the cloud platform and the devices capability. In term of development, this constraint implies that the software component developed as a remote cannot be reused in the client side. In addition, interfaces that exposes the same services may be deffirent from an implementation to an other.

Our contribution aims to define an additional abstraction level on the cloud to specify a structure that represents mobile devices. It enables a common interface to communicate with differents devices like mobile devices, sensors and actuators. Communications addressed to the devices are translated to the specific protocols by this representation. And the responses are stored on a cache which is the virtual state of the device. This representation act also as a "mobile-friendly" platform within the cloud. Indeed, the representation is built on emulation capabilities that offer a compliant environment with the physical device on which the representation is associated.

We distinguish three kinds of representations depending on their association (or not) with the physical devices. The first type of representations are those associated with simple sensors or actuators. They are the simplest forms for the representation where there act as a cached proxy with a common interface. The

Battou, Abdella; Mahmoudi, Charif; Mourlin, Fabrice.
"Formal Definition of Edge Computing: An Emphasis on Mobile Cloud and IoT Composition."
Paper presented at The Third IEEE International Conference on Fog and Mobile Edge Computing, Barcelona, Spain. April 23, 2018 - April 26, 2018.

second type is the representations associated with the mobile devices. This representation offers offloading capabilities, and keeps a cached state of the differents sensors and actuators available on the mobile device. We consider this second type of representations as a composition of resources associated with a mobile device and it is not the exact image of the device. It can be used as an extension to the resources available locally on the device. The third type of representations have no direct association with a physical device. It is a composition of multiple representations. We define this representation as a composition of resources distributed over the network which transforms the mobile device into a sort of "super device" by eliminating the physical limitation. The composite representation adds smartness to the devices by enabling the composition of multiple devices in a smooth way.

This paper describes, in Section II, the various challenges faced during virtualization and how they are addressed in work related work. We describe the existing cloud techniques that are useful for mobile cloud computing and presents the formalism efforts that are related to the differents virtualization aspects. We end this section with an introduction to the π-calculus which is the formalism used for our definition. In Section III, we expose our definition of the Virtual Device Representation(VDR) using a formal language and how we address the orchestration and the networking for these VDRs. In section IV, we will expose our proposed architecture for Mobile Cloud Computing (MCC) and the approach that we use in order to have a high performance mobile cloud network. The last Section describes a case study for an MCC platform highlighting the structural congruence between the system when a mobile application is running directly on the device and when this same application is offloaded to a mobile cloud.

II. Related work

A. Mobile Device Challenges

The role of mobile devices has expanded into the modern workplace. Workplaces are not limited to the office and the warehouse anymore. They have expanded to include airport terminals, loading docks and delivery trucks, physician waiting rooms, and even playing fields and family gatherings. Mobile devices have erased workplace boundaries, and as a result, employees can connect with their corporate networks almost anytime, anywhere.

With the emergence of Fog, the network connections between edge devices and the cloud are reconsidered as part of the computational processes being done close to the edge devices called edge computing. Mobile Cloud is one of the implementations of the Edge to incorporate the network needed to get processed data to its destination. The mobile technology, which is easing access to business data and applications is also providing various means of communications. These features continue to be embraced by users. For IT point of view, mobile technology and the unprecedented pace of change in the mobile arena will generate new IT management challenges. Indeed, as mobile innovation continues, machine-to-machine (M2M) connectivity (or Internet of Things) will further accelerate mobile opportunity [4] and transform how people, enterprises, and governments interact with the many aspects of modern life.

Several trends -- and the way companies react to them -- will create challenges for IT, as organizations attempt to exercise some control over devices that are not necessarily designed to be secure and manageable. With careful planning and an understanding of best practices and Mobile Device Management (MDM) [5] options. IT can go a long way toward meeting those challenges. With a well-implemented MDM strategy, enterprises can enforce corporate security policies without stifling user productivity.

B. Cloud and Virtualization

The virtualization is used for abstracting the Operating System (OS) and applications from the physical hardware to build a more cost-efficient, agile and simplified server environment. There are two types of virtualization and many major uses of virtualization.

1) Virtualization types

Two kinds of virtualization are used to simulate the machine hardware and allow the execution of a guest OS. First is emulation where VM emulates (or simulates) complete hardware if the unmodified guest OS for a different PC cannot be run. There are some hypervisors specialized on "emulation" like Bochs, VirtualPC for Mac and Qemu [6]. Second is full/Native where VM simulates "enough" hardware to allow an unmodified guest OS to be run in isolation. This virtualization type requires that the same hardware CPU if used by the VM and the hypervisor. This type is supported also by many hypervisors like, VMWare Workstation [7] and Microsoft Hyper-V [8].

2) Virtualization usage

By using virtualization, multiple VM instances containing operating systems can run on a single physical server or a single VM can use hardware from multiple physical servers, each with access to the underlying server's computing resources. The virtualization is used to addresses the resources waste caused by the fact that the host servers operate at less than 15 percent of capacity, leading to server sprawl and complexity. According to VMware statistics [9], virtualization can deliver 80 percent greater utilization of resources on the server and 10:1 or better server consolidation ratio.

The objective of this kind of virtualization -- considered as a subset of server virtualization-- is to provide an abstraction of the networking resources into a logical model that have the same behavior as the physical resources. The virtual networking resources are divided in two categories: first is the physical resources virtualization like vRouter (Router) and vSwitch (Switch), the second is the resources appliances like FWaS (Firewall) and LBaaS (Load balancer). This network virtualization approach is called Network Functions Virtualization (NFV) [12]. It aims to consolidate and deliver the networking components needed to support a fully virtualized infrastructure and shared by multiple tenants in a secure and isolated manner.

Existing efforts aims formalizing the cloud services interactions [10] and orchestration [11]. However, those efforts do not address the virtualization aspect of such cloud systems.

Battou, Abdella; Mahmoudi, Charif; Mourlin, Fabrice.
"Formal Definition of Edge Computing: An Emphasis on Mobile Cloud and IoT Composition."
Paper presented at The Third IEEE International Conference on Fog and Mobile Edge Computing, Barcelona, Spain. April 23, 2018 - April 26, 2018.

C. Virtualization Formalizm

In parallel with the pragmatic work on the networking, there is many existing efforts on the definition of formalism dedicated to networking. U. Montanari and M. Sammartino have worked on a proper extension [13] of the π-calculus. The resulting process calculi provide both an interleaving and a concurrent networking oriented semantics. A. Singh et al have also worked on an extension called ω-calculus [14] that formally modeling and reasoning about mobile ad hoc wireless networks. These works focus on the reasoning and the verification of the networking protocols and does not address the virtualization aspect. This lack of networking virtualization formalism motivates our high-level definition of network virtualization in the next section.

III. Virtual Device Representation

In our approach, the VDR aims to address the mobile cloud computing virtualization paradigm. We have identified three types of VDRs, and each type has a specific role within the mobile cloud.

1. Sensor VDR (SVDR): it represents a physical sensor or actuator within to the mobile cloud.
2. Device VDR (DVDR): it represents a physical mobile device within to the mobile cloud.
3. Composite VDR (CVDR): it represents a composition of SVDR, DVDR, and mobile cloud resources.

In this section, we present the different aspects turning around the VDR by giving our definition of the VDR, a formal definition using the Higher-Order π-Calculus (HOπC) [15], stressing the orchestration mechanism for the VDRs, and the networking aspect. Our choice for the HOπC is motivated by the need of expressing the mobility of the VDRs in the mobile cloud, also the mobility of mobile applications between the physical devices and the VDRs. In our definition, we do not use the network related extensions of the π-calculus for two reasons: first, those extensions do not address the higher-order paradigm, next, there are designed to express networking protocols not the virtualization-oriented communication.

A. Definition

VDR is defined as composition of resources (CPU, RAM, and Storage), devices, and sensors. It is a software composite component that provides emulation of the behavior of the physical hardware that it represents.

A VDR is a small VM instance used in cloud computing, typically hosting a mobile OS and exposing management services that emulate a display screen and/or a keyboard. As same as the physical handheld computing device that it represents, it can run various types of mobile applications (known as apps) and it have a network connection.

A VDR can be associated to a physical device nor sensor in this case, a 1:1 association is control their interactions (ex: SVDR and DVDR). We call this category "Emulated VDR" A VDR can be free of any hardware association, in this case, it is a composed VDR (CVDR) that aggregates it components hardware associations and have then a 0:n association to the hardware devices. This category of VDR is called "Native VDR".

B. Formal Specification

The VDR operates according to an event driven architecture. Every interaction is initiated by a message sent from a driver (further to hardware sensing activity) nor a service call. We define an event vector representing all interfaces of a VDR. This event vector, illustrated in (1) contains channels that are used to exchange messages

$$\overrightarrow{ev} \stackrel{\text{def}}{=} [camera_m, micro_n, nfc_o, keyboard_p, \dots] \quad (1)$$

The event vector $\overrightarrow{ev}$ is used only for the interactions between VDRs, the interactions between DVDR on one hand SVDR and the physical device on the other hand, are using a service based channel called ws that represents a web service based exchange.

$$VDR(\overrightarrow{ws}) \stackrel{\text{def}}{=} (\nu\ \overrightarrow{ev}) \begin{pmatrix} (\lambda\ \overrightarrow{ev}\ ws_i)SVDR \\ + \\ (\lambda\ \overrightarrow{ev}\ ws_j)DVDR \\ + \\ (\lambda\ \overrightarrow{ev})CVDR \\ + \\ \emptyset \end{pmatrix} \quad (2)$$

We define the generalization called VDR as a nondeterministic choice between the three types of VDRs as illustrated in (2)

The term $VDR(\overrightarrow{ws})$ have a vector $\overrightarrow{ws}$ of web services channels as parameter, these channels are shared with the mobile cloud system and are transmitted to the specific VDRs to allow the communication with the physical devices. The term VDR creates a new $\overrightarrow{ev}$ vector containing the channels that are used to interface the specific VDRs. We benefit in the VDRs definition of the use abstractions where $(\lambda\ \overrightarrow{ev}\ ws_i)SVDR$ is a natural way to write $SVDR(\overrightarrow{ev}, ws)$, and so on for the two other VDR types. The specific VDR is activated iff the corresponding element in the $\overrightarrow{ws}$ vector is a valid channel and not an empty process $\emptyset$.

The SVDR is activated behind the physical sensor connection event. Once connected, the physical sensor sends the identification data to the SVDR through the ws channel. This data is persisted inside the SVDR using the term $DevId$ defined in (5) that give back the identification data if requested through the right event channel.

$$SVDR(\overrightarrow{ev}, ws) \stackrel{\text{def}}{=} ws(id).\tau.\begin{pmatrix} DevId(ev_i, id) \\ | VirtualSensor(\overrightarrow{ev}, ws) \end{pmatrix} \quad (3)$$

As illustrated in (3), the term $SVDR$ uses the term $VirtualSensor$ defined in (4) to dispatch the data perceived by the physical sensor using the event channel. At this level, we consider the mapping between the sensor and the matching channel as an invisible action represented by τ. Two possible behaviors can be adapted by the term $VirtualSensor$ as illustrated in (4): if a Stop command (15) is received, the process will end, else, the dispatching action is executed. The parallel composition of the term $SVDR$ allows the administrator to

Battou, Abdella; Mahmoudi, Charif; Mourlin, Fabrice.
"Formal Definition of Edge Computing: An Emphasis on Mobile Cloud and IoT Composition."
Paper presented at The Third IEEE International Conference on Fog and Mobile Edge Computing, Barcelona, Spain. April 23, 2018 - April 26, 2018.

retrieve the VDR identifier using the environment channel ev_i and don't impact the execution of the virtual sensor.

$$VirtualSensor(\overrightarrow{ev}, ws) \stackrel{\text{def}}{=} ws(sens).\begin{pmatrix}[sens = Stop]\, Stop \\ + \\ \tau.\overline{ev_i}\langle sens\rangle.VirtualSensor(\overrightarrow{ev}, ws)\end{pmatrix} \quad (4)$$

$$DevId(req, id) \stackrel{\text{def}}{=} req(cb).\overline{cb}\langle id\rangle.DevId(req, id) \quad (5)$$

Messages sent through the ws channel are initiated by the mobile device or the sensor. However, these messages are forwarded to the target VDR by the networking infrastructure defined in (19) and (22).

The DVDR specification respects the same fundamentals as the SVDR. As illustrated in (6), it uses the term $DevId$ to persist and give back the device identifier and use term called $VirtualDevice$ to manage the virtual device behavior. However, the DVDR can run applications instead of SVDR that only proxy the sensor events.

$$DVDR(\overrightarrow{ev}, ws) \stackrel{\text{def}}{=} ws(id).\tau.\begin{pmatrix}DevId(ev_i, id) \\ |VirtualDevice(\overrightarrow{ev}, ws)\end{pmatrix} \quad (6)$$

We need to dissociate between sensing events sent from the device embedded sensors and the application offloading requests. To do that, we define a type called App (7) that encapsulate the offloaded application.

$$App\big(BackEndProc(ws)\big) \stackrel{\text{def}}{=} BackEndProc(ws) \quad (7)$$

We define in (8) the term $VirtualDevice$ that execute the offloaded application if need, else, it proxies the sensing data.

$$VirtualDevice(\overrightarrow{ev}, ws) \stackrel{\text{def}}{=} ws(msg).\left(\begin{array}{c}[msg = Stop]\, Stop \\ + \\ \begin{pmatrix}case\ msg\ of \\ : App\big(P(x)\big) \Rightarrow P(x) \\ : msg \Rightarrow \tau.\overline{ev_i}\langle msg\rangle\end{pmatrix} \\ .VirtualDevice(\overrightarrow{ev}, ws)\end{array}\right) \quad (8)$$

We used to this definition the syntactic sugar introduced by R. Milner in [16] by using the "case of" instruction to distinguish between the offloading action represented by $App\big(P(x)\big)$ and sensing actions. Where we run the higher-order parameter $P(x)$ within the DVDR on the offloading action, elsewhere, we proxy the message to the corresponding event channel as we do for SVDR. The service channel used for the communication between the physical device and the offloaded application is set as parameter x before the offloading action, this channel is different from the service channel that connects the DVDR and the physical device.

The CVDR in (9) have no direct association with a physical device, its interactions pass through a SVDR nor a DVDR. The term $CVDR$ is defined as an aggregation of SVDR and DVDR that are sharing the same events vector.

$$CVDR(\overrightarrow{ev}) \stackrel{\text{def}}{=} (\nu\ id)$$

$$DevId(ev_i, id)|CompositeDevice(\overrightarrow{ev}) \quad (9)$$

A identifier is created the term $CVDR$ and returned trough the right event channel ev_i.using the term $DevId$.

$$CompositeDevice(\overrightarrow{ev}) \stackrel{\text{def}}{=} ev_i(\vec{e}).CompositeDevice(\overrightarrow{ev}\hat{}\vec{e}) \quad (10)$$

The term $CompositeDevice$ defined in (10) is used to aggregate the events channels $\overrightarrow{ev}$ (the event channel associated with the actual $CompositeDevice$) and $\vec{e}$ (the event channel associated with the VDR to add to this composition) using the concatenation operator ^.

C. Orchestration

In an MCC context, orchestration is the automation of the management and coordination tasks of the services and components. In addition to the interconnection processes running across heterogeneous systems, the localization of services is an important issue. Processes and VDRs must cross multiple organizations, systems and firewalls.

The mobile cloud orchestration aims to automate the configuration, coordination and management of VDRs and VDRs interactions in such an environment. The process involves automating workflows required for the composition of VDRs and the offloading of mobile Apps. Involved tasks include managing virtualization and emulation in server runtimes, directing the communication flow of Apps among VDRs and dealing with exceptions to typical workflows.

In our approach, the orchestrator is composed by three main components as illustrated in (11), we define these three components as common orchestration tasks: 1) Configuration where the cloud orchestrator manages the storage, compute, and networking. In this paper, we do not focus on the resources allocation algorithm (compute and storage), this aspect will be stressed in a future publication. A high-level specification of the networking mechanism is presented in the next sub section. 2) Provisioning where the cloud orchestrator manages the VDRs by providing the run, suspend, and terminate operations. 3) Security where the cloud orchestrator manages the monitoring, and reporting. We describe the details of this aspect also on a separate paper where we describe our implementation and detailed algorithms.

$$Orchestrator(\overrightarrow{api}) \stackrel{\text{def}}{=} Configuration(\overrightarrow{api})|\text{Provisioning}(\overrightarrow{api}) \quad (11)$$
$$|(\nu\ data)Monitoring\big(\overrightarrow{api}, \overrightarrow{data}\big)$$

In the term $Configuration$ in (12), we illustrate the use of the configuration api that is used for the allocation of resources, the deallocation (free) of resources, and to suspend the execution. The api is a vector in two-dimensional space. The contravariant indicates the target module (ex: api^c where ^{c} stand for $\boldsymbol{C}onfiguration$). The covariant indicates the service called within the module (ex: api_a where $_a$ stand for $\boldsymbol{a}llocate$). The system administrator will use the vector $\overrightarrow{api}$

$$Configuration(\overrightarrow{api}) \stackrel{\text{def}}{=}$$

$$\begin{pmatrix} api_a^c(allocate).\tau.(\nu\ res)\overline{allocate}\langle res\rangle \\ |api_f^c(free).\tau \\ |api_s^c(suspend).\tau \end{pmatrix} .Configuration(\overrightarrow{api}) \quad (12)$$

The term Provisioning in (13) uses also an api to ask the configuration module for resource allocation. Once allocated, it delegates the creation of the VDR to the term Run defined in (14). We use the abstraction of the resources information returned by the term $Configuration$ to communicate this information to the term Run that is preconfigured with the two parameters before its reception through the channel api_r^p.

$$\text{Provisioning}(\overrightarrow{api}) \stackrel{\text{def}}{\equiv} \begin{pmatrix} \begin{pmatrix} api_r^p(Run).(\nu\ allocate)\overline{api_a^c}\langle allocate\rangle \\ |allocate(res).(\lambda\ res)Run(\ ) \end{pmatrix} \\ |api_s^p(suspend).\overline{api_s^c}\langle suspend\rangle \\ \left|\begin{pmatrix} api_t^p(terminate).terminate(ws) \\ .\overline{ws}\langle Stop\rangle.\overline{api_f^c}\langle terminate\rangle \end{pmatrix}\right. \end{pmatrix} .\text{Provisioning}(\overrightarrow{api}) \quad (13)$$

The suspension is delegated to the term $Configuration$ where it is represented as an invisible action τ. The provisioning sends the term $Stop$ to the VDR to terminate its execution. We use for that the ws channel sent through the channel $terminate$.

The term Run defined in (14) composes a vector depending on the type of the VDR that the initiator wants to create. After the creation of the VDR, it creates and sends a new identifier using the ws channel to start the new created VDR.

$$Run(ws, \overrightarrow{type}) \stackrel{\text{def}}{\equiv} \tau.\begin{pmatrix} [type_v = type_s]\ VDR(ws\hat{}\emptyset\hat{}\emptyset) \\ |[type_v = type_d]\ VDR(\emptyset\hat{}ws\hat{}\emptyset) \\ |[type_v = type_c]\ VDR(\emptyset\hat{}\emptyset\hat{}\emptyset) \end{pmatrix}\Big| (\nu\ id)\overline{ws}\langle id\rangle \quad (14)$$

$$Stop(\ ) \stackrel{\text{def}}{\equiv} \emptyset \quad (15)$$

To keep our definitions as clear as possible, we didn't integrate the communications between the VDRs and the monitoring module defined in $Monitoring$. We can easily imagine that after each communication on the events vector $\overrightarrow{ev}$ channels, an information must be sent to the monitoring module using the api_{put}^m channel. This information is stored in data vector $\overrightarrow{data}$ on the recursive call in (16) to the term $Monitoring$.

$$Monitoring(\overrightarrow{api}, \overrightarrow{data}) \stackrel{\text{def}}{\equiv} \begin{pmatrix} api_{put}^m(datum).\tau.(\nu\ id)api_{ret}^m(id) \\ |api_{get}^m(id).api_{res}^m(data_{id}) \end{pmatrix} .Monitoring(\overrightarrow{api}, \overrightarrow{data}\hat{}datum) \quad (16)$$

D. Networking

On our mobile cloud approach, multiple tenants can use the same physical infrastructure. The network virtualization simplifies the multi-tenancy. The shared infrastructure allows independence of the VDRs regarding the physical host on which it's located. The VDR should be movable between the hosts based on the need. We commit our networking definition to allow VDRs across 2 different Layer 3 (L3) networks look like they are in the same Layer 2 (L2) domain.

The proposed virtual networking model allows the provisioning module (13) to manage the virtual network component like a VDR and hide the complexity from the user. The model allows also to bypass the scale perspective 4096 VLAN limit as proposed on VXLAN by the Internet Engineering Task Force (IETF) RFC 7348 [17]. Our model definition is composed from two terms: $vSwitch$ defined in (19) and $vRouter$ defined in (22).

For our network modelling, we define the structure of the packet transiting on the networking infrastructure. The vector $\overrightarrow{ethernet}$ in (17) represents the L2 frame where the names $ethernet_{dst}$ and $ethernet_{src}$ are the channels corresponding to the ws_x used by the VDRs in (2). $ethernet_{ip}$ contains the information needed by the $vRouter$ and the message as $ip_{payload}$. The names that composes the vectors in (17) and (18) are abbreviations of header fields of the packets as described in the IETF RFC 791.

$$\overrightarrow{ethernet} \stackrel{\text{def}}{\equiv} [dst, src, tag, type, \overrightarrow{ip}, check] \quad (17)$$

$$\overrightarrow{ip} \stackrel{\text{def}}{\equiv} \begin{bmatrix} version, ihl, tos, len, id, flag, frag, ttl \\ , proto, check, src, dst, opt, payload \end{bmatrix} \quad (18)$$

Given that our objective is not to stress the networking protocols but to point out the communications between the virtual components, we abstract all network behavior that is not directly related to the virtualization as non-observable operations τ.

$$\begin{aligned} &vSwitch\ (cntl, \overrightarrow{adr}) \stackrel{\text{def}}{\equiv} \\ &Control\ (vSwitch\ (cntl, \overrightarrow{adr}), cntl, \overrightarrow{adr}) \\ &\left|adr_i(\overrightarrow{ethernet}).\tau.\overline{ethernet_{dst}}\ \langle ethernet_{ip_{payload}}\rangle\right. \\ &\qquad .vSwitch\ (cntl, \overrightarrow{adr}) \end{aligned} \quad (19)$$

$$\begin{aligned} &Control\ (Target, cntl, \overrightarrow{adr}) \stackrel{\text{def}}{\equiv} \\ &cntl_{connect}(link).Target \\ &\left|cntl_{disconnect}(link).(\nu\ \vec{p})\right. \\ &\quad \left(Disconnect(Target, \overrightarrow{adr}, (\nu\ \vec{p}), link, O\ )\right) \end{aligned} \quad (20)$$

$$\begin{aligned} &Disconnect(Target, \overrightarrow{old}, \overrightarrow{adr}, port, i\ ) \stackrel{\text{def}}{\equiv} \\ &\quad [i = \|\overrightarrow{old}\|](\lambda\ \overrightarrow{adr})Target \\ &\quad \left|[old_i = port]\right. \\ &\qquad Disconnect(c, \overrightarrow{old}, \overrightarrow{adr}, port, i+1\ ) \\ &\quad \left|Disconnect(c, \overrightarrow{old}, \overrightarrow{new}\hat{}port, port, i+1\ )\right. \end{aligned} \quad (21)$$

The term $vSwitch$ defined in (19) represents the virtualization of the L2 switch. It is modelled as a congruency

Battou, Abdella; Mahmoudi, Charif; Mourlin, Fabrice.
"Formal Definition of Edge Computing: An Emphasis on Mobile Cloud and IoT Composition."
Paper presented at The Third IEEE International Conference on Fog and Mobile Edge Computing, Barcelona, Spain. April 23, 2018 - April 26, 2018.

between a control (20) that manage the VDRs connections and a L2 network bridge.

The term $Control$ has three parameters, the first one is higher-order called $Target$ that is used to pass the terms $vSwitch$ and $vRouter$. The second is called $cntl$ and it is used as a channel to control connections of the VDRs. The third one the vector containing connected VDRs channels. The $Target$ parameter is passed also to the term $Disconnect$ defined in (21), we use the abstraction λ to override the addresses vector $\overrightarrow{adr}$ that is still a free name in $Target$ when a device is disconnected.

$$vRouter\ (ipAdr, cntl, \overrightarrow{adr}) \stackrel{def}{=}$$

$$Control\left(\begin{array}{c} vRouter\ (ipAdr, cntl, \overrightarrow{adr}) \\ , cntl, \overrightarrow{adr} \end{array}\right)$$

$$\left|\begin{array}{c} adr_i(\overrightarrow{ethernet}).\tau.\left(\begin{array}{c} [ipAdr = ethernet_{ip_{dest}}] \\ \overline{ethernet_{dst}}\langle\overrightarrow{ethernet}\rangle \\ + \\ \overline{ethernet_{ip_{dest}}}\langle\overrightarrow{ethernet}\rangle \end{array}\right) \\ .vRouter\ (cntl, \overrightarrow{adr}\hat{}link) \end{array}\right. \quad (22)$$

The term $vRouter$ defined in (22) represents the virtualization of the L3 routing. It is modelled as a congruency between a control (20) that manage the virtual switches nor VDRs connections and a L3 network bridge. In this model, we don't illustrate some features like IP forwarding to keep our definition clear.

The management of the networking infrastructure in exposed as a part of the provisioning API. To do so, we illustrate in (23) an extension of the term Provisioning defined initially in (13).

$$\text{Provisioning}(\overrightarrow{api}) \stackrel{def}{=}$$

$$\left|\begin{array}{c} ... \\ ... \\ ... \\ api^p_{vsCreate}(ret).(\nu\ cntl) \\ \left(\begin{array}{c} \tau.(\nu\ \overrightarrow{adr})vSwitch\ (cntl, \overrightarrow{adr}) \\ |\ \overline{ret}\langle cntl\rangle \end{array}\right) \\ |api^p_{vsConnect}(cntl, adr).\tau.cntl_{connect}(adr) \\ \left(\begin{array}{c} |api^p_{vsDisconnect}(cntl, adr).\tau.cntl_{disconnect}(adr) \\ ... \end{array}\right) \\ .Provisioning(\overrightarrow{api}) \end{array}\right. \quad (23)$$

In (12), we describe the Switch related provisioning API, the channel $api^p_{vsCreate}$ is used to create the virtual switch and return the control channel $cntl$ to the initiator of the request. The Router provisioning API is like the Switch one, the two differences is that the api^p_{vs*} channels are defined as api^p_{vr*} and the $api^p_{vrCreate}$ is used to create a virtual router, to keep our definition clear, we omit this part of the definition.

The previous terms are formally defined in the objective to model a new architecture of cloudlet. The definitions are useful not only for this current work but also for all software researcher in cloud computing domain.

IV. Architecture

The definition presented in the previous section is made

$$Control\ (Target, cntl, \overrightarrow{adr}) \stackrel{def}{=}$$
$$cntl_{connect}(link).Target$$
$$\left|\begin{array}{c} cntl_{disconnect}(link).(\nu\ \vec{p}) \\ \left(Disconnect(Target, \overrightarrow{adr}, (\nu\ \vec{p}), link, O)\right) \end{array}\right.$$
$$Disconnect(Target, \overrightarrow{old}, \overrightarrow{adr}, port, i) \stackrel{def}{=}$$
$$[i = \|\overrightarrow{old}\|](\lambda\ \overrightarrow{adr})Target$$
$$\left|\begin{array}{c} [old_i = port] \\ Disconnect(c, \overrightarrow{old}, \overrightarrow{adr}, port, i+1) \\ Disconnect(c, \overrightarrow{old}, \overrightarrow{new}\hat{}port, port, i+1) \end{array}\right.$$

based on the state-of-art regarding the MCC research [18] [19] that converge into the Cloudlet-based MCC. The cloudlets are defined as trusted and resource-rich network computers that offer bridging capabilities to the Internet and is available for use by nearby mobile devices through a direct and well-connection. In this section, we describe our Cloudlet-based architecture by illustrating some of the technical aspects that was abstracted in the formal definition. We also link the technical implementation with their correspondent formal model. Moreover, we introduce our contribution to the migration pattern and stress the projection of the ACID (Atomicity, Consistency, Isolation, and Durability) properties from the formal model to the implementation model.

A. Cloudlet-based MCC

In our approach, we have identified the need of a set of rules and regulations, as a protocol, which determine how data and processes are transmitted between the different components of the MCC. The Fig. 1 illustrate our vision of the MCC that is composed by three layers: the first is the Device Layer (DL) composed by physical sensor and mobile devices. The second is the Cloudlet Layer (CL) that is composed from the network of Cloudlets, each Cloudlet may contain the VDRs, Virtual Service Representation (VSR), and local services. The third layer is the Internet Layer (IL) composed by the central Cloud that contains Cloud services and needed registries in addition to Internet services like the media sensors.

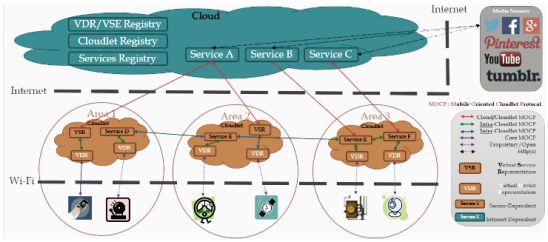


Fig. 1. Global architecture

In the CL, we define a networking infrastructure based on the NFV. As illustrated in Fig. 2, the device is connected to the VDR through a vRouter defined in (22) and a vSwitch defined in (19). The networking infrastructure is managed using the cloud orchestrator API, in our implementation model, we use OpenStack [20] that contains a powerful networking module called Neutron. Is module is based on Open vSwitch [21]. This implementation and provide a ReST [22] API for the creation

Battou, Abdella; Mahmoudi, Charif; Mourlin, Fabrice.
"Formal Definition of Edge Computing: An Emphasis on Mobile Cloud and IoT Composition."
Paper presented at The Third IEEE International Conference on Fog and Mobile Edge Computing, Barcelona, Spain. April 23, 2018 - April 26, 2018.

and the managing of the provided virtual networking infrastructure.

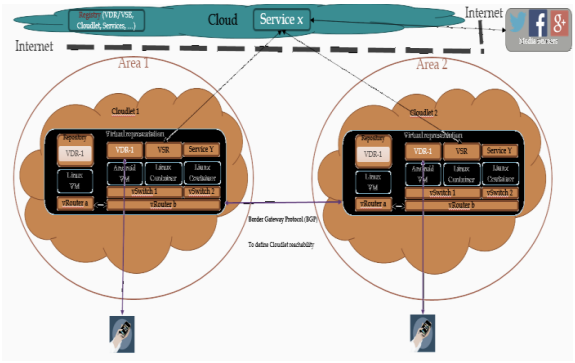


Fig. 2. Cloudlet structure and networking

B. *Mobile application offloading*

As a part of our contribution with the Mobile Oriented Cloudlet Protocol (MOCP), the formal definition of this paper focus on the communications especially used by the Core MOCP for the migration of the Apps from the physical device to the VDRs. Our implementation model, as illustrated in Fig. 3, extends the formal definition in (6) by adding technical details to the abstract definition. The two components of the VDR are the *Device Descriptor* that is modelled by the $DevId$ in (5) and the *Virtual device* is modelled in (8). The *Backend app* is modelled as the higher-order parameter $BackEndProc$ in (7). The *OSGi* [23] *container* operations are considered as non-observable operations.

Our offloading approach differs from the actual overlays oriented [24] approaches. We consider the Backend application as an ACID service that can migrate from one host to another one. Our definition of the DVDR in (8) allows the ACID properties by isolating the *Backend app* in an atomic process, which runs that makes durable impact on the target VDR. These properties are extended to the implementation model by using the OSGi framework that isolates the class-loading inside the JVM and guarantees a strict lifecycle of the *Backend app* bundle. This lifecycle management guarantees the consistency of the service execution. The Apache Felix [25] OSGi implementation in used in our architecture due to an Android porting effort that Apache has been supporting since the version 1.3. This mechanism works with stateless Backend services that provides a response after for the Frontend Cloudlet Android Application Package (CAPK) request, and then requires no further attention. Regarding the stateful Backend service where subsequent Frontend CAPK requests depend on the results of the first request, they are more difficulties to manage because a single action typically involves more than one request. We thus need another isolation level in top of the OSGi.

To address the issue of the state management, we use a *chroot* of ArchLinux that provides an additional layer of abstraction using the Docker package available with this distribution. We are working on the integration of Docker on Android to bypass the need of a *chroot* and to allow a native isolation support on Android.

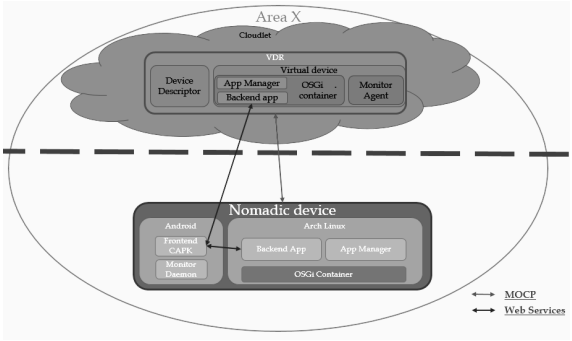


Fig. 3. DVDR implementation model

V. Case Studies

Our case of study aims to show the structural congruence between a Backend app offloaded in a VDR and the same backend app running in the device. Our objective is to illustrate that a Backend App (7) that runs in a VDR are identical up to structure parallel composition to the Backend App which runs in a mobile device. This result is obtained after the reduction of both systems to an identical system.

A. *Mobile device*

We first define the terms $FrontEnd$ which represents the Frontend CAPK and $BackEnd$ which represents the Backend app used in our study. Those terms are composing the mobile devices defined in (26) and (27).

The term $FrontEnd$, defined in (24), is a model of a "web view" which sends messages to the Backend using the channel ws, once the response received from the Backend, the Frontend execute another iteration as a recursion. This term has also the $touch$ channel as parameter to communicate with the user defined in (29).

$$\begin{aligned} &FrontEnd(touch, ws) \stackrel{\text{def}}{=} (\nu\ cb) \\ &\qquad touch(event).\tau.\overline{ws}\langle event, cb\rangle \\ &\qquad |cb(res).FrontEnd(ws) \end{aligned} \tag{24}$$

The term $BackEnd$, defined in (25), react to the message sent by the Frontend. If the abstraction $intra$ binds to the same channel as the parameter ws, the Backend app is executed locally to the mobile device. Else, the Backend send a message containing a copy of itself to the corresponding VDR and terminate the local execution. The execution continues into the VDR after the offloading.

$$\begin{aligned} &BackEnd(ws) \stackrel{\text{def}}{=} (\lambda\ intra) \\ &ws(event, cb).\tau \\ &\qquad \cdot\begin{pmatrix} [intra = ws].\overline{intra}\langle\ \ \rangle \\ + \overline{ws}\langle App((\lambda\ ws)BackEnd(ws))\rangle.\emptyset \end{pmatrix} \\ &\qquad\qquad |intra(\ \ ).\tau.(\nu\ res)\overline{cb}\langle res\rangle \end{aligned} \tag{25}$$

We define two parallel composition as models for the mobile devices. The first mobile device is defined in (26) as the parallel execution of a Frontend and a locally executed Backend. The

second mobile device is defined in (27) as the parallel execution of a Frontend and a Backend which is configured to be offloaded to the VDR.

$$Devicelocal(ws, touch) \stackrel{\text{def}}{=} FrontEnd(touch, ws) \mid (\lambda\, ws) BackEnd(ws) \quad (26)$$

$$DeviceRemote(ws, touch) \stackrel{\text{def}}{=} (\nu\, local)(FrontEnd(touch, local) \mid (\lambda\, local) BackEnd(ws)) \quad (27)$$

To keep the clarity of our specification, we omit the details of the definition of the term $Admin$, we define just the signature in (28). It is important to note that this term send all needed messages using the vector $\overrightarrow{api}$. It starts the networking infrastructure and the VDRs.

$$Admin(ws, \overrightarrow{api}) \stackrel{\text{def}}{=} \cdots \quad (28)$$

The term $user$ defined in (29) represents a device user executing a single action by sending an event to the Frontend through the channel $touch$ that represents the device's touch screen. We have defined a simple action for the user to have a system which can be reduced manually by a human is a reasonable time slot.

$$user(touch) \stackrel{\text{def}}{=} (\nu\, event)\overline{touch}\langle event\rangle \quad (29)$$

B. Systems

To verify the structural congruence, we define two systems as parallel composition of the mobile user, mobile device, administrator, and the orchestrator. The term $SystemMig$ defined in (30) represents the system that will give raise to a Backend offloading after some reductions.

$$SystemMig \stackrel{\text{def}}{=} (\nu\, ws)\left((\nu\, touch)\left(\begin{matrix} user(touch) \\ \mid DeviceRemote(ws, touch)\end{matrix}\right) \mid \left(\begin{matrix} (\nu\, \overrightarrow{api})\left(\begin{matrix} Admin(ws, \overrightarrow{api}) \\ \mid Orchestrator(\overrightarrow{api})\end{matrix}\right)\end{matrix}\right)\right) \quad (30)$$

The term $SystemLocal$ defined in (31) represents the system that initiate a Backend after some reductions.

$$SystemLocal \stackrel{\text{def}}{=} (\nu\, ws)\left((\nu\, touch)\left(\begin{matrix} user(touch) \\ \mid Devicelocal(ws, touch)\end{matrix}\right) \mid \left(\begin{matrix} (\nu\, \overrightarrow{api})\left(\begin{matrix} Admin(ws, \overrightarrow{api}) \\ \mid Orchestrator(\overrightarrow{api})\end{matrix}\right)\end{matrix}\right)\right) \quad (31)$$

C. Structural congruence

We have performed some computations steps to fully to reach a stable system starting from $SystemMig$. We call this stable state reached after those reductions $SystemMig'$ where $SystemMig \xrightarrow{-\overline{touch}\langle event\rangle,\ldots} SystemMig'$.

We have applied the operation to the $SystemLocal$. However, the reduction of this system is simpler by dint of no offloading related reductions. Also, we obtain $SystemLocal'$ where $SystemLocal \xrightarrow{-\overline{touch}\langle event\rangle,\ldots} SystemLocal'$.

Only some bound names and non-observables actions composes the difference between the two reduced systems. We have thus find that $SystemMig' \equiv SystemLocal'$.

The structural congruence is commutative and associative. We can then write:

$$\begin{aligned} &\text{given that} && SystemMig \equiv SystemMig' \\ &\text{and} && SystemLocal \equiv SystemLocal' \\ &\text{and} && SystemMig' \equiv SystemLocal' \\ &\text{then} && SystemMig \equiv SystemLocal \end{aligned} \quad (32)$$

VI. Conclision and future works

In this paper, we present our formal definition of the MCC. This specification focus on the communications interactions on the MCC. Moreover, architectural aspects dedicated to the realization of a MCC solution are described. The case studies proof the structural congruence between offloading and local execution of a mobile application and shows the transparency of the offloading in our MCC system. On our future work, we will focus on two aspects of the MCC. First one is a formal definition of a metric to define a unit to measure the applications migration. The second aspect is the definition of the data collection and algorithm to calculate the application offloading cost.

Disclaimer

Any mention of commercial products or organizations is for informational purposes only; it is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the products identified are necessarily the best available for the purpose. The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

References

[1] B. Varghese and R. Buyya, "Next generation cloud computing: New trends and research directions.," Future Generation Computer Systems, vol. 79, pp. 849-861, 2014.

[2] D. Sanderson, Programming google app engine: build and run scalable web apps on google's infrastructure., O'Reilly Media, Inc., 2009.

[3] A. Hosseinian-Far, M. Ramachandran and C. L. Slack, "Emerging Trends in Cloud Computing, Big Data, Fog Computing, IoT and Smart Living," in Technology for Smart Futures, vol. 53, Springer, 2018, pp. 29-40.

[4] W. Geng, S. Talwar, K. Johnsson, N. Himayat and K. D. Johnson, "M2M: From mobile to embedded internet.," IEEE Communications Magazine, vol. 49, no. 4 , pp. 36-43, 2011.

[5] L. Liu, R. Moulic and D. Shea, "Cloud service portal for mobile device management," IEEE 7th International Conference on e-Business Engineering (ICEBE), pp. 474-478, 2010.

Battou, Abdella; Mahmoudi, Charif; Mourlin, Fabrice.
"Formal Definition of Edge Computing: An Emphasis on Mobile Cloud and IoT Composition."
Paper presented at The Third IEEE International Conference on Fog and Mobile Edge Computing, Barcelona, Spain. April 23, 2018 - April 26, 2018.

[6] D. Bartholomew, "Qemu a multihost multitarget emulator," Linux Journal, no. 145, p. 3, 2006.

[7] E. Bugnion, S. Devine, M. Rosenblum, J. Sugerman and E. Y. Wang, "Bringing virtualization to the x86 architecture with the original vmware workstation," ACM Transactions on Computer Systems (TOCS), vol. 30, no. 4, p. 12, 2012.

[8] A. Velte and T. Velte, Microsoft virtualization with Hyper-V, McGraw-Hill, Inc., 2009.

[9] B. Walters, "VMware virtual platform," Linux journal, vol. 63, p. 6, 1999.

[10] N. M. K. Chowdhuryr and R. Boutaba, "A survey of network virtualization," Computer Networks, vol. 54, no. 5, pp. 862-876, 2010.

[11] J. Jiulei, L. Jiajin, H. Feng, W. Yan and S. Jie, "Formalizing Cloud Service Interactions," Journal of Convergence Information Technology, vol. 7, no. 13, 2012.

[12] C. Mahmoudi, Orchestration d'agents mobiles en communauté, Universite Paris-Est Creteil, 2014.

[13] M. Ugo and S. Matteo, Network conscious pi-calculus, Pisa: Universita di Pisa, 2012.

[14] A. Singh, C. Ramakrishnan and S. A. Smolka, "A process calculus for mobile ad hoc networks," Science of Computer Programming, vol. 75, no. 6, pp. 440-469, 2010.

[15] R. Milner, P. Joachim and W. David, "A calculus of mobile processes," Information and computation , vol. 100, no. 1, pp. 1-40, 1992.

[16] R. Milner, The polyadic π-calculus: a tutorial, Berlin Heidelberg: Springer, 1993.

[17] T. Sridhar, L. Kreeger, D. Dutt, C. Wright, M. Bursell, M. Mahalingam, P. Agarwal and K. Duda, Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks, IETF, 2014.

[18] H. T. Dinh, C. Lee, D. Niyato and P. Wang, "A survey of mobile cloud computing: architecture, applications, and approaches.," Wireless communications and mobile computing, vol. 13, no. 18, pp. 1587-1611, 2013.

[19] E. Cuervo, A. Balasubramanian, D.-k. Cho, A. Wolman, S. Saroiu, R. Chandra and P. Bahl, "MAUI: making smartphones last longer with code offload," Proceedings of the 8th international conference on Mobile systems, applications, and services, pp. 49-62, 2010.

[20] A. Corradi, M. Fanelli and L. Foschini, "VM consolidation: A real case based on OpenStack Cloud," Future Generation Computer Systems, vol. 32, pp. 118-127, 2014.

[21] B. Pfaff, J. Pettit, K. Amidon, M. Casado, T. Koponen and S. Shenker, Extending Networking into the Virtualization Layer., Hotnets, 2009.

[22] R. Fielding, "Representational state transfer," Architectural Styles and the Design of Netowork-based Software Architecture, pp. 76-85, 2000.

[23] Alliance, OSGi, Osgi service platform, release 3, IOS Press, Inc., 2003.

[24] B.-G. Chun, S. Ihm, P. Maniatis, M. Naik and A. Patti, "Clonecloud: elastic execution between mobile device and cloud," Proceedings of the sixth conference on Computer systems, pp. 301-314, 2011.

[25] Felix, Apache, "Apache Felix-welcome," Apache Software Fundation, 2 03 2018. [Online]. Available: http://felix.apache.org. [Accessed 2 03 2018].

Battou, Abdella; Mahmoudi, Charif; Mourlin, Fabrice.
"Formal Definition of Edge Computing: An Emphasis on Mobile Cloud and IoT Composition."
Paper presented at The Third IEEE International Conference on Fog and Mobile Edge Computing, Barcelona, Spain. April 23, 2018 - April 26, 2018.

Testing Spectrum Sensing Networks by UAV

Daniel G. Kuester, Ryan T. Jacobs, Yao Ma, and Jason B. Coder
U.S. Department of Commerce
National Institute of Standards and Technology
Communications Technology Lab, RF Technology Division
Boulder, CO, USA
daniel.kuester@nist.gov, ryan.jacobs@nist.gov, yao.ma@nist.gov, jason.coder@nist.gov

***Abstract*—We consider here the use of hovering unmanned aerial vehicles (UAVs) to test spectrum sensing capabilities of Citizen's Broadband Radio Service (CBRS) systems at 3.5 GHz. Aircraft transmit synthesized LTE or pulsed radar modulation to excite spectrum sensing nodes on the ground. This kind of test could be useful either to validate a system during development or to check a deployed system in the field. Technical challenges that will need to be addressed include rotor blade effects on signal fidelity, waveform parameter selection, and understanding of positioning estimation and control.**

I. Introduction

The Federal Communication Commission (FCC) is developing rules for a proposed new CBRS in the 3.55 GHz – 3.7 GHz band [1]. The rules establish three tiers of service. The highest priority tier is allocated to incumbent users, notably including the U.S. Navy, who will continue to operate shipborne radar systems under the existing band allocation. The new second tier CBRS is allocated to priority access licenses (PALs) intended for commercial broadband service, and assigned through a competitive bidding process. The new lowest priority tier CBRS is called General Authorized Access (GAA), where users may access channels opportunistically.

Devices transmitting in CBRS tiers must vacate the band during use by users in higher-priority tiers. Incumbent users, as the highest tier, are to have the strongest protections against interference. This is to be achieved by means of a spectrum access system (SAS). The SAS tries to monitor the spectrum in order to detect incumbent or PAL users in the band in time and space. On detection, the SAS directs PAL and/or GAA devices to vacate the band.

Widespread deployment of a CBRS network will require buy-in from both incumbents and the potential CBRS industry. Strong testing for the effectiveness of SAS spectrum sensing is one piece of this problem. This type of test could validate system performance with less operator time and effort (and therefore expense). It could also help to understand the repercussions of CBRS network design and policy on a larger scale than is practical in human-performed tests that are difficult to repeat.

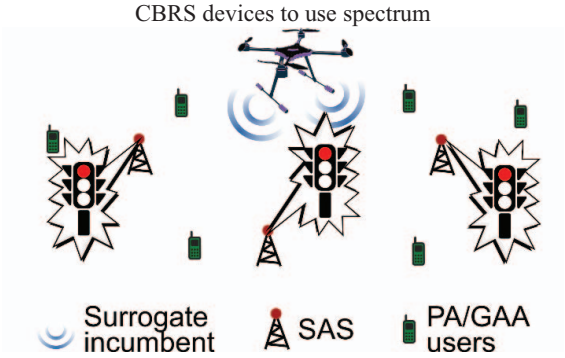
We consider here unmanned aerial vehicles (UAVs) outfit with calibrated transmitters as tools in this measurement problem. The aircraft is to transmit surrogate "radar-like" signals to mimic first–tier incumbent users or broadband waveforms like LTE to mimic second–tier PAL service.

II. UAV Transmission Platform

A. Overview

Initial UAV work is a demonstration system composed of a consumer quad-rotor UAV mounted with a 3.5 GHz calibrated transmitter payload. The transmitter is preloaded with recorded or synthesized modulation waveforms, and played back at preprogrammed intervals.

(a) SAS detects no upper-tier use, and permits lower-tier CBRS devices to use spectrum



(b) SAS detects incumbent or PAL users, and directs lower-tier CBRS devices to vacate spectrum

Fig. 1. Expected device behavior in the 3.5 GHz CBRS spectrum sensing test scenarios.

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao.
"Testing Spectrum Sensing Networks by UAV."
Paper presented at 2016 United States National Committee of URSI National Radio Science Meeting (USNC-URSI NRSM), Boulder, CO, United States. January 6, 2016 - January 9, 2016.

TABLE I. TRANSMITTER SPECIFICATIONS

Parameter	*Nominal*	*Measured*
RF peak conducted power out	30 dBm	TBD
Waveform sample depth	73 MS	TBD
Nyquist Modulation bandwidth	14 MHz	TBD

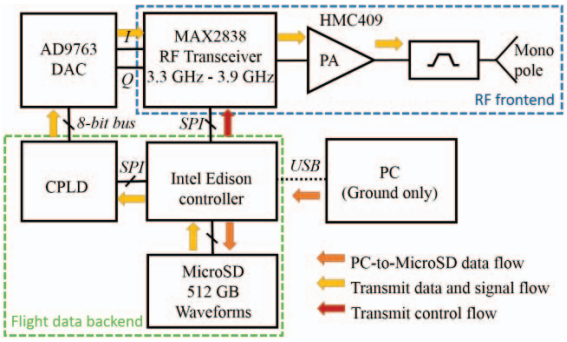


Fig. 2. Block diagram for a prototype transmitter to be mounted on the aircraft.

B. *Test Signal Synthesis*

The UAV is illustrated in test operation in Figure 1. Figure 1a shows the expected ground network response when the UAV transmitter is silent. Turning on the transmitter leads to the expected behavior shown in Figure 1b. The specific response of each tier depends on whether the SAS receives incumbent or PAL tier transmission.

When the UAV synthesizes LTE modulation, the SAS should identify the spectrum use, and direct PAL or GAA users to behave appropriately. If the UAV synthesizes radar waveforms, the SAS should direct both PAL and GAA users to vacate the band. Failure to do so would invite improvements to the CBRS network implementation.

Incumbent Navy radar systems and modulation parameters are not generally available to the public, so a surrogate imitation waveform needs to be identified and synthesized. Previous work by the NTIA [2] identified parameters of several types of marine radar, and may serve as a suitable basis.

C. *Payload Design*

The RF transmitter payload architecture under development is illustrated in Figure 2. The transmitter is intended to transmit a carrier with "playback" of a custom baseband waveform that is preloaded on UAV flash memory. The MicroSD card stores up to 73 Msamples of quadrature (*I* and *Q*) baseband. The transceiver supports up to 14 MHz of Nyquist bandwidth. This combination allows custom modulation waveform storage and playback for up to 5 seconds. Further specifications are listed in Table I. RF bandwidth and power are most limited by the tight size, weight and power (SWAP) constraints of the platform.

One characteristic to investigate will be the effect of rotor blade modulation on the integrity of the transmit signal. Previous work [3] has identified modulation in full-sized helicopters. Common small UAVs have plastic rotor blades, so we believe this effect is likely to be small.

D. *Positioning Uncertainty*

We need to characterize the positioning uncertainty of the UAV in order to ensure test repeatability. In this context, it is important to understand the uncertainty of both aircraft control and logged position. The aircraft comes integrated with GPS for control and position logging, corresponding with position uncertainty on the order of 1 m to 2 m. We may be able to improve this accuracy by other means, which could include differential GPS or optical tracking.

III. CONCLUSION

Robust testing of spectrum sensing systems by UAV has potential to improve the amount and quality of test data available compared to human time and effort on foot or by car. Upcoming efforts at NIST may help to evaluate the practicality of this type of RF test at the current state of the art in unmanned aircraft.

REFERENCES

[1] Federal Communications Commission, "Report and Order and Second Further Notice of Proposed Rulemaking," GN Docket No. 12-354, April 2015.

[2] FH Sanders, JE Carroll, GA Sanders, RL Sole, "Effects of Radar Interference on LTE Base Station Receiver Performance," NTIA Report 14-499, May 2014.

[3] AC Polycarpou, CA Balanis, A Stefanov, "Helicopter Rotor-Blade Modulation of Antenna Radiation Characteristics," *IEEE Trans. Ant. Prop.*, Vol. 49, No. 5, May, 2001.

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao.
"Testing Spectrum Sensing Networks by UAV."
Paper presented at 2016 United States National Committee of URSI National Radio Science Meeting (USNC-URSI NRSM), Boulder, CO, United States. January 6, 2016 - January 9, 2016.

MIMO-OFDM Transmissions Invoking Space-Time/Frequency Linear Dispersion Codes Subject to Doppler and Delay Spreads

Jiayi Zhang, Hamid Gharavi, Bin Hu
National Institute of Standards and Technology
Gaithersburg, MD 20899-8920, USA
Email: {jiayi.zhang,hamid.gharavi,bin.hu}@nist.gov

***Abstract*—Linear dispersion codes (LDC) can support arbitrary configurations of transmit and receive antennas in multi-input multi-output (MIMO) systems. In this paper, we investigate two transmit diversity applications of LDC for orthogonal frequency division multiplexing (OFDM) systems in order to achieve space-time/frequency (ST/SF) diversity gains when transmitting over time-/frequency-selective fading channels. LDC-aided ST/SF-OFDM is flexible in configuring various numbers of transmit antennas and time-slots or frequency-tones. Our results show that the ST-OFDM scheme is sensitive to exploiting diversity gains, subject to the impact of varying channel Doppler spreads; while the performance of SF-OFDM is mainly subject to delay spread. Particularly, when the transmitter employs more than two antennas, the LDC-aided ST/SF-OFDM outperforms the orthogonal block codes (e.g. Tarokh's codes) aided ST/SF-OFDM, when communicating over higher Doppler/delay spread.**

I. Introduction

Multi-input multi-output (MIMO) [1] is a most attractive multi-antenna technique that has been adopted by many emerging wireless communication standards, such as IEEE 802.11n and 3GPP LTE, owing to the achievable antenna array, multiplexing, and diversity gain. In order to improve link reliability, the diversity gain enabled at transmission can be exploited by space-time coding, while the diversity gain achieved at the receiver may benefit from maximum ratio combining (MRC) [2]–[5]. These gains are obtained without increasing the transmission power by employing multiple transmit and/or receive antennas. Particularly, Hassibi's linear dispersive codes (LDC) [4], [6]–[8] allow arbitrary configurations in space-time coding for high-rate MIMO transmissions.

Meanwhile, broadband communication plays an increasingly important role in meeting the growing demand for high-speed multimedia transmissions in our daily lives. However, when the bandwidth of a signal exceeds the coherent bandwidth of the wireless channel, the small-scale fading imposed on the signal becomes frequency-selective rather than frequency-flat. Such a fading time-dispersion incurs inter-symbol interference (ISI) to the air-interface and therefore degrades the link performance [9]. Orthogonal frequency division multiplexing (OFDM) [10] is one of the transceiver techniques designed to combat ISI. When the number of subcarriers in OFDM is sufficiently larger than the number of taps, the frequency-selective channel can be decomposed into mutually independent frequency-flat fading channels on each subcarrier with the aid of OFDM transmission. Moreover, a center length of cyclic prefix (CP) or zero-padding (ZP) should be inserted between any two adjacent OFDM blocks to mitigate inter-block interference (IBI) incurred by multi-path fading [11]. As a result, each received signal can be recovered at the low-complexity single-tap equalizer without inter-carrier interference (ICI), thanks to the orthogonality between adjacent subcarriers with flat fading [12].

For transmit diversity aided MIMO systems, a simple time-reversed space-time block coding scheme [13] was proposed in the context of a broadband MIMO channel to combat ISI. Such a large time-reversal frame requires slow channel varying, which is not suitable for mobile wireless communications [9]. In [14], [15], the space-time block coding (STBC) including the LDC schemes were investigated when communicating over uncorrelated and correlated frequency-flat fading channels. With the aid of a multi-antenna employed at the transmitter, many of transmit diversity schemes have been invented to combat the frequency-selective fading incurred by the high-speed data rate and also achieve diversity gain at the same time [16]–[19]. Specifically, the space-time block coding (STBC) schemes that were used in frequency-flat fading channels may be applied to each subcarrier to achieve space-time diversity and combat channel time-dispersion [20]. In parallel, rather than exploiting the achievable diversity by crossing spatial antennas and time-slots, the alternative approach may benefit from OFDM's multi-carrier feature relying on so-called space-frequency block coding (SFBC) schemes by exploiting the diversity crossing transmit antennas and subcarriers [21]. A further combined version of the above two schemes is known as space-time-frequency block coding, which can exploit diversity in all three domains [22]–[25]. Furthermore, the authors in [25], [26] propose various LDC-aided OFDMs to achieve space-frequency diversity for the constant fading channel within a single OFDM block; while in [27] the LDC is designed to obtain diversity from the time and frequency domain rather than the spatial domain. However, to the best of our knowledge, there has not been a comprehensive investigation assessing LDC in space-time (ST) and space-frequency (SF) diversigy gain impacts varying channel coherent times and bandwidths.

Gharavi, Hamid; Hu, Bin; Zhang, Jiayi.
"MIMO-OFDM Transmissions Invoking Space-Time/Frequency Linear Dispersion Codes Subject to Doppler and Delay Spreads."
Paper presented at IEEE-wcn.org/, Doha, Qatar. April 3, 2016 - April 6, 2016.

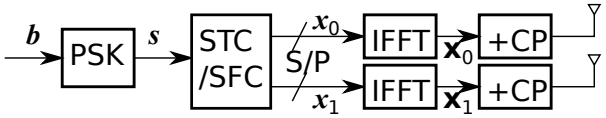


Fig. 1. Transmitter block diagram for transmit diversity aided OFDM

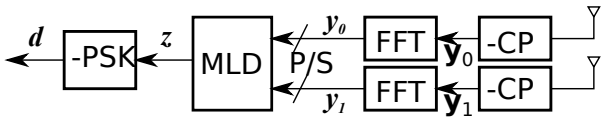


Fig. 2. Receiver block diagram for transmit diversity aided OFDM

In this paper, we investigate transmit diversity for the OFDM system by invoking LDC to take into account the trade-off between ST and SF diversities. Our contributions are highlighted as follows:

- We unify the analysing structure of transmit diversity aided block codes in order to compare LDC with the corresponding Alamouti's code and Tarokh's code [2], [3] in diverse MIMO configurations.
- The LDC, Alamouti's and Tarokh's codes are applied to OFDM in both the ST and SF approaches.
- Our results show that when the channel is constant within the coherent time/bandwidth, the ST-OFDM or SF-OFDM is capable of achieving full diversity gain in space-time or space-frequency domains, respectively.
- We quantify the performance impact with varying Doppler spreads. Results show that the ST-OFDM scheme is sensitive to exploit diversity gains subject to the effect upon varying channel Doppler spreads.
- In parallel, we examine the performance impacts owing to varying numbers of paths in terms of delay spread. As a result, the performance of SF-OFDM is mainly subject to delay spreads.
- Compared to fixed orthogonal block codes, the LDC-aided ST/SF-OFDM is flexible to configure various numbers of transmit antennas and time-slots or frequency-tones.
- When a transmitter employs more than two antennas, the performance of LDC-aided OFDM schemes is less impacted by channel Doppler/delay spreads, as compared with orthogonal block codes.

The rest of this paper is structured as follows. We firstly elaborate on the transceiver system model of MIMO-OFDM in Section II. The ST- and SF-oriented OFDM schemes that achieve transmit diversity will be studied in Section III in both the Alamouti's and a LDC cases. We will present the simulation results in Section IV, followed by closing remarks in Section V.

II. System Model

A. Transmitted Signal

The multi-antenna aided OFDM transmitter is shown in Fig. 1. Specifically, the $N_{\rm b}$-length binary source data bit stream $\boldsymbol{b} = [\boldsymbol{b}_0^T, \boldsymbol{b}_1^T, \cdots, \boldsymbol{b}_{N_{\rm s}-1}^T]^T$ is fed into the $\mathcal{M}$-ary Gray labeled phase-shift keying (PSK) mapper transmitting $\mathcal{Q}$ bits per symbol, where we have $N_{\rm s} = N_{\rm b}/\mathcal{Q}$ and $\mathcal{M} = 2^{\mathcal{Q}}$. Moreover, the $N_{\rm s}$-length modulated symbol sequence $\boldsymbol{s} = [s_0, s_1, \cdots, s_{N_{\rm s}-1}]^T$ is inputed into the transmit diversity module $\mathcal{C}$, which maps the symbols into $N_{\rm Tx}$ antennas and $N_{\rm T}$ time-slots or $N_{\rm C}$ subcarriers in terms of ST/SF coding, which will be further detailed in Section III. The $n_{\rm Tx}$-th output ST/SF module containing the U-symbol block $\mathbf{x}_{n_{\rm Tx}} = [\mathsf{x}_{n_{\rm Tx},0}, \mathsf{x}_{n_{\rm Tx},1}, \cdots, \mathsf{x}_{n_{\rm Tx},(U-1)}]^T$ to be transmitted via the antenna $n_{\rm Tx}$ is then converted from serial-to-parallel (S/P) corresponding to U orthogonal subcarriers in the F-domain. These U-symbols in $\mathbf{x}_{n_{\rm Tx}}$ are transformed by U-point IDFT operation matrix $\boldsymbol{\mathcal{F}}_U^H$ [28] into T-domain at the t-th time-slot for $t = 0, 1, \cdots, T-1$, expressed by

$$\begin{aligned}\boldsymbol{x}_{n_{\rm Tx}}[t] &= \boldsymbol{\mathcal{F}}_N^H \mathbf{x}_{n_{\rm Tx}}[t] \\ &= \left[x_{n_{\rm Tx},0}[t], x_{n_{\rm Tx},1}[t], \cdots, x_{n_{\rm Tx},(U-1)}[t]\right]^T, \end{aligned} \quad (1)$$

The CP is inserted at the beginning of $\boldsymbol{x}_{n_{\rm Tx}}[t]$ by copying the last $L_{\rm CP}$ elements of the $\boldsymbol{x}_{n_{\rm Tx}}[t]$, which results in the $(U+L_{\rm CP})$-element transmitted OFDM symbol block $\tilde{\boldsymbol{x}}_{n_{\rm Tx}}[t]$ at the t-th time-slot via the $n_{\rm Tx}$-th antenna.

B. Signal Representation at the BS Receiver

The multi-antenna aided OFDM receiver is shown in Fig. 2. By satisfying the channel order $L < L_{\rm CP}$, after removing the CP at the receiver, the equivalent U-element T-domain signal block received at the t-th time-slot may be expressed as:

$$\boldsymbol{y}_{n_{\rm Rx}}[t] = \sum_{n_{\rm Tx}=0}^{N_{\rm Tx}-1} \boldsymbol{H}_{n_{\rm Rx},n_{\rm Tx}}[t] \boldsymbol{x}_{n_{\rm Tx}}[t] + \boldsymbol{n}_{n_{\rm Rx}}[t], \quad (2)$$

where $\boldsymbol{H}_{n_{\rm Rx},n_{\rm Tx}}$ denotes the $U \times U$-element T-domain circulant matrix [28] holding the channel impulse response (CIR) between transmit and receive antennas $n_{\rm Tx}$ and $n_{\rm Rx}$. In Eq. (2), $\boldsymbol{n}_{n_{\rm Rx}}$ is the noise imposed at the $n_{\rm Rx}$-th receiver antenna, each element of which has a power of $\mathcal{N}_u = \sigma_{\rm N}^2$.

Hence, after the U-point DFT transforming the signal $\boldsymbol{y}_{n_{\rm Rx}}$ into the F-domain, we have the equivalent symbol block given by

$$\mathbf{y}_{n_{\rm Rx}}[t] = \boldsymbol{\mathcal{F}}_U \boldsymbol{y}_{n_{\rm Rx}}[t] = \sum_{n_{\rm Tx}=0}^{N_{\rm Tx}-1} \mathbf{H}_{n_{\rm Rx},n_{\rm Tx}}[t] \mathbf{x}_{n_{\rm Tx}}[t] + \mathbf{n}_{n_{\rm Rx}}[t]. \quad (3)$$

Since we have $\boldsymbol{H}_{n_{\rm Rx},n_{\rm Tx}} = \boldsymbol{\mathcal{F}}_U^H \mathbf{H}_{n_{\rm Rx},n_{\rm Tx}} \boldsymbol{\mathcal{F}}_U$ according to [28], the $\mathbf{H}_{n_{\rm Rx},n_{\rm Tx}}$ in Eq. (3) is a diagonal matrix with entries $\mathsf{h}_{u,u}$ $(u = 0, 1, \cdots, U-1)$ representing the corresponding F-domain channel transfer function on U subcarriers, leading to a low-complexity one-tap channel equalization method. In Eq. (3), we have $\mathbf{n}_{n_{\rm Rx}} = \boldsymbol{\mathcal{F}}_U \boldsymbol{n}_{n_{\rm Rx}}$ with $\mathcal{N}_u = \sigma_{\rm N}^2$.

Furthermore, we reshape Eq. (3) into an $N_{\rm Rx}$-length multi-antenna received symbol vector for the u-th subcarrier at time-slot t expressed by

$$\check{\mathbf{y}}_u[t] = \check{\mathbf{H}}_u[t] \check{\mathbf{x}}_u[t] + \check{\mathbf{n}}_u[t], \quad (4)$$

Gharavi, Hamid; Hu, Bin; Zhang, Jiayi.
"MIMO-OFDM Transmissions Invoking Space-Time/Frequency Linear Dispersion Codes Subject to Doppler and Delay Spreads."
Paper presented at IEEE-wcn.org/, Doha, Qatar. April 3, 2016 - April 6, 2016.

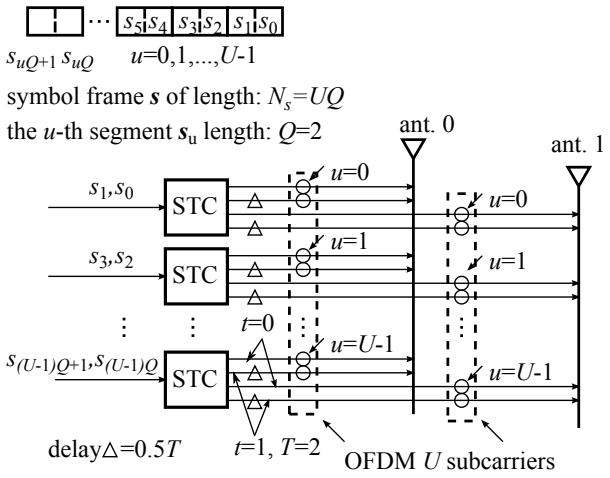


Fig. 3. Schematic diagram of space-time coded OFDM

where $\check{\mathbf{H}}_u[t]$ is a $(N_{\text{Rx}} \times N_{\text{Tx}})$-size MIMO-channel matrix at time-slot t, in which the $(n_{\text{Rx}}, n_{\text{Tx}})$-th entry denotes the F-domain coefficients of $\mathbf{H}_{n_{\text{Rx}},n_{\text{Tx}}}$ on the u-th subcarrier; $\check{\mathbf{x}}_u[t] = \left[\mathsf{x}_{0,u}[t], \mathsf{x}_{1,u}[t], \cdots, \mathsf{x}_{(N_{\text{Tx}}-1),u}[t]\right]^T$ is N_{Tx}-antenna transmitted symbol vector in the F-domain before IDFT at time-slot t. Additionally, $\check{\mathbf{n}}_u[t] = \left[\mathsf{n}_{0,u}[t], \mathsf{n}_{1,u}[t], \cdots, \mathsf{n}_{(N_{\text{Tx}}-1),u}[t]\right]^T$ is N_{Rx}-antenna noise component added at receiver.

III. Transmit Diversity Aided MIMO-OFDM Schemes

In this section, we elaborate two transmit diversity aided OFDM schemes, namely ST coded OFDM and SF coded OFDM, respectively.

A. Space-Time Coded OFDM

In order to achieve the space- and time-diversity, the OFDM may be ST-encoded in a subcarrier-by-subcarrier basis as shown in Fig. 3. Specifically, a N_s-length symbol frame $\boldsymbol{s}$ is divided into U segments, and the u-th segment for $u = 0, 1, \cdots, U-1$ contains Q symbols in $\boldsymbol{s}_u$ for the input of ST encoder. The encoder employs specific ST algorithms procuding the ouput frame for the n_{Tx}-th antenna having $T > 1$ consecutive OFDM blocks $\mathbf{x}_{n_{\text{Tx}}}[t]$ over time-slots $t = 0, 1, \cdots, T-1$ in F-domain. For instance, the Alamouti's g2 ST [2] encoded OFDM blocks for $N_{\text{Tx}} = 2$, $Q = 2$ and $T = 2$ may be expressed as

$$\begin{cases} \mathbf{x}_{n_{\text{Tx}}=0}[t=0] = [s_0, s_2, \cdots, s_{2U-2}]^T, \\ \mathbf{x}_{n_{\text{Tx}}=1}[t=0] = [s_1, s_3, \cdots, s_{2U-1}]^T, \\ \mathbf{x}_{n_{\text{Tx}}=0}[t=1] = -\mathbf{x}_1^*[0], \\ \mathbf{x}_{n_{\text{Tx}}=1}[t=1] = \mathbf{x}_0^*[0], \end{cases} \tag{5}$$

where the u-th element in symbol vector $\mathbf{x}_{n_{\text{Tx}}}[t]$ is conveyed onto u-th subcarriers at time-slot t and emitted via antenna n_{Tx}. Alternatively, when the Tarokh's g4 ST [3] code is invoked in OFDM for $N_{\text{Tx}} = 4$, $Q = 4$ and $T = 8$, we have:

$$\begin{cases} \mathbf{x}_0[0] = [s_0, s_4, \cdots, s_{4U-4}]^T, \\ \mathbf{x}_1[0] = [s_1, s_5, \cdots, s_{4U-3}]^T, \\ \mathbf{x}_2[0] = [s_2, s_6, \cdots, s_{4U-2}]^T, \\ \mathbf{x}_3[0] = [s_3, s_7, \cdots, s_{4U-1}]^T, \end{cases} \begin{cases} \mathbf{x}_0[1] = -\mathbf{x}_1[0], \\ \mathbf{x}_1[1] = -\mathbf{x}_0[0], \\ \mathbf{x}_2[1] = -\mathbf{x}_3[0], \\ \mathbf{x}_3[1] = \mathbf{x}_2[0], \end{cases}$$
$$\begin{cases} \mathbf{x}_0[2] = -\mathbf{x}_2[0], \\ \mathbf{x}_1[2] = \mathbf{x}_3[0], \\ \mathbf{x}_2[2] = \mathbf{x}_0[0], \\ \mathbf{x}_3[2] = -\mathbf{x}_1[0], \end{cases} \begin{cases} \mathbf{x}_0[3] = -\mathbf{x}_3[0], \\ \mathbf{x}_1[3] = -\mathbf{x}_2[0], \\ \mathbf{x}_2[3] = \mathbf{x}_1[0], \\ \mathbf{x}_3[3] = \mathbf{x}_0[0], \end{cases} \begin{cases} \mathbf{x}_0[4] = \mathbf{x}_0^*[0], \\ \mathbf{x}_1[4] = \mathbf{x}_1^*[0], \\ \mathbf{x}_2[4] = \mathbf{x}_2^*[0], \\ \mathbf{x}_3[4] = \mathbf{x}_3^*[0], \end{cases}$$
$$\begin{cases} \mathbf{x}_0[5] = \mathbf{x}_0^*[1], \\ \mathbf{x}_1[5] = \mathbf{x}_1^*[1], \\ \mathbf{x}_2[5] = \mathbf{x}_2^*[1], \\ \mathbf{x}_3[5] = \mathbf{x}_3^*[1], \end{cases} \begin{cases} \mathbf{x}_0[6] = \mathbf{x}_0^*[2], \\ \mathbf{x}_1[6] = \mathbf{x}_1^*[2], \\ \mathbf{x}_2[6] = \mathbf{x}_2^*[2], \\ \mathbf{x}_3[6] = \mathbf{x}_3^*[2], \end{cases} \begin{cases} \mathbf{x}_0[7] = \mathbf{x}_0^*[3], \\ \mathbf{x}_1[7] = \mathbf{x}_1^*[3], \\ \mathbf{x}_2[7] = \mathbf{x}_2^*[3], \\ \mathbf{x}_3[7] = \mathbf{x}_3^*[3]. \end{cases} \tag{6}$$

Furthermore, the LDC encoded OFDM block for the n_{Tx}-th antenna on the u-th subcarrier at time-slot $t = 0, 1, \cdots, T-1$ before IFFT operation is given by

$$\mathbf{x}_{n_{\text{Tx}}}[t] = [[\boldsymbol{\mathcal{B}}_{n_{\text{Tx}}}\boldsymbol{s}_0]_t, [\boldsymbol{\mathcal{B}}_{n_{\text{Tx}}}\boldsymbol{s}_1]_t, \cdots, [\boldsymbol{\mathcal{B}}_{n_{\text{Tx}}}\boldsymbol{s}_{U-1}]_t]^T, \tag{7}$$

where $\boldsymbol{\mathcal{B}}_{n_{\text{Tx}}}$ is the linear dispersion matrix defined in [6] of n_{Tx}-th antenna and $\boldsymbol{s}_u = [s_{uQ}, s_{uQ+1}, \cdots, s_{(u+1)Q-1}]^T$ is the u-th input symbol segment having a legnth of Q for $u = 0, 1, \cdots, U-1$. Then, by using Eq. (1), the ST-OFDM symbols may be transmitted.

The detection of ST-OFDM receiver is also operated by subcarrier-by-subcarrier basis. After the signal transformed into F-domain by Eq. (4), we consider on the symbol blocks over all T time-slots, having an equivalent F-domain signal expression as

$$\underline{\mathbf{y}}_u = \underline{\mathbf{H}}_u \underline{\mathbf{x}}_u + \underline{\mathbf{n}}_u, \tag{8}$$

where each component vector $\underline{\mathbf{y}}_u$, $\underline{\mathbf{x}}_u$ and $\underline{\mathbf{n}}_u$ may be expressed by $\underline{\mathbf{a}}_u = \left[\check{\mathbf{a}}_u^T[0], \check{\mathbf{a}}_u^T[1], \cdots, \check{\mathbf{a}}_u^T[T-1]\right]^T$; while the channel component matrix is given by $\underline{\mathbf{H}}_u = \left[\check{\mathbf{H}}_u^T[0], \check{\mathbf{H}}_u^T[1], \cdots, \check{\mathbf{H}}_u^T[T-1]\right]^T$.

B. Space-Frequency Coded OFDM

As shown in Fig. 4, another method to exploit the diversity in both space and frequency is to invoke SF coding in OFDM system [21]. Specifically, each consecutive Q elements of frame $\boldsymbol{s}$ are SF-encoded having N_{Tx} output blocks, each of which is converyed into $M \leq U$ subcarriers within a single time-slot, i.e. $t = 0, T = 1$. Hence, each OFDM block requires $N = U/M$-set consecutive Q-symbol inputs in order to crossing U subcarriers. For example, the Alamouti's g2 style SF-encoded OFDM blocks for $N_{\text{Tx}} = 2$, $Q = 2$ and $M = 2$ may be expressed as

$$\begin{cases} \mathbf{x}_{n_{\text{Tx}}=0}[t=0] = [s_0, -s_1^*, s_2, -s_3^*, \cdots, s_{U-2}, -s_{U-1}^*]^T, \\ \mathbf{x}_{n_{\text{Tx}}=1}[t=0] = [s_1, s_0^*, s_3, s_2^*, \cdots, s_{U-1}, s_{U-2}^*]^T, \end{cases} \tag{9}$$

where the u-th element in symbol vector $\mathbf{x}_{n_{\text{Tx}}}[t]$ is conveyed onto u-th subcarriers at time-slot t and emitted via antenna

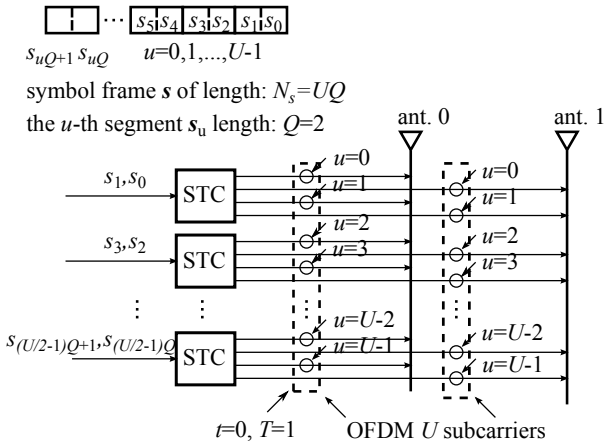


Fig. 4. Schematic diagram of space-frequency coded OFDM

n_{Tx}. When the Tarokh's g4 based SF code with $N_{\mathrm{Tx}} = 4$, $Q = 4$ and $M = 8$ is employed in OFDM, we have:

$$\begin{cases} \mathbf{x}_0[0] = [s_0, -s_1, -s_2, -s_3, s_0^*, -s_1^*, -s_2^*, -s_3^*, \cdots, \\ \quad s_{U-4}^*, -s_{U-3}^*, -s_{U-2}^*, -s_{U-1}^*]^T, \\ \mathbf{x}_1[0] = [s_1, -s_0, -s_3, -s_2, s_1^*, -s_0^*, -s_3^*, -s_2^*, \cdots, \\ \quad s_{U-3}^*, s_{U-4}^*, s_{U-1}^*, -s_{U-2}^*]^T, \\ \mathbf{x}_2[0] = [s_2, -s_3, -s_0, -s_1, s_2^*, -s_3^*, -s_0^*, -s_1^*, \cdots, \\ \quad s_{U-2}^*, -s_{U-1}^*, s_{U-4}^*, s_{U-3}^*]^T, \\ \mathbf{x}_3[0] = [s_3, -s_2, -s_1, -s_0, s_3^*, -s_2^*, -s_1^*, -s_0^*, \cdots, \\ \quad s_{U-1}^*, s_{U-2}^*, -s_{U-3}^*, s_{U-4}^*]^T, \end{cases} \tag{10}$$

Moreover, the LDC encoded OFDM block for the n_{Tx}-th antenna on the u-th subcarrier at time-slot $t = 0$ (before IFFT operation) is given by

$$\mathbf{x}_{n_{\mathrm{Tx}}}[t=0] = \left[[\boldsymbol{\mathcal{B}}_{n_{\mathrm{Tx}}} \boldsymbol{s}_0]^T, [\boldsymbol{\mathcal{B}}_{n_{\mathrm{Tx}}} \boldsymbol{s}_1]^T, \cdots, [\boldsymbol{\mathcal{B}}_{n_{\mathrm{Tx}}} \boldsymbol{s}_{N-1}]^T\right]^T, \tag{11}$$

where $\boldsymbol{s}_u = [s_{uQ}, s_{uQ+1}, \cdots, s_{(u+1)Q-1}]^T$ is the u-th input symbol segment having a legnth of Q for $u = 0, 1, \cdots, U-1$. Consequently, by using Eq. (1), the SF-OFDM symbols can be formed.

At the receiver side, the detection of SF-OFDM operates in subcarrier group-by-group basis. Unlike the Section III-A, after the signal transformed into F-domain by Eq. (4), the symbol blocks for the n-th subcarrier group which cross over subcarriers from nM to $(n+1)M-1$ for $n = 0, 1, \cdots, N-1$ at time-slot $t = 0$ can be expressed by

$$\underline{\mathbf{y}}_n[0] = \underline{\mathbf{H}}_n[0] \underline{\mathbf{x}}_n[0] + \underline{\mathbf{n}}_n[0], \tag{12}$$

where each component vector $\underline{\mathbf{y}}_n[0]$, $\underline{\mathbf{x}}_n[0]$ and $\underline{\mathbf{n}}_n[0]$ can be expressed by $\underline{\mathbf{a}}_n[0] = \left[\check{\mathbf{a}}_{nM}^T[0], \check{\mathbf{a}}_{nM+1}^T[0], \cdots, \check{\mathbf{a}}_{(n+1)M-1}^T[0]\right]^T$; while the channel component matrix is given by $\underline{\check{\mathbf{H}}}_n[0] = \left[\check{\mathbf{H}}_{nM}^T[0], \check{\mathbf{H}}_{nM+1}^T[0], \cdots, \check{\mathbf{H}}_{(n+1)M-1}^T[0]\right]^T$.

TABLE I
SIMULATION PARAMETERS

Channel model	time/frequency-selective time-correlated Rayleigh fading
Bits per symbol	$\mathcal{Q} = 1$
Normalized Doppler frequency	$f_{\mathrm{ND}} = 0.01, \cdots, 0.1$
No. of CIR paths	$L = 1, 2, 4, 8, 16$
No. of subcarriers	$U = 128$
No. of transmitter antennas	$N_{\mathrm{Tx}} = 2, 4$
No. of receiver antennas	$N_{\mathrm{Rx}} = 1$
No. of time-slots per code	$T = 2, 4, 8$
No. of frequency-tone per code	$M = 2, 4, 8$
No. of symbols per code	$Q = 2, 4$

C. Maximum-Likelihood Detection

Based on Eqs. (8) and (12), the equivalent F-domain system model for detection can be represented by [1]

$$\overline{\mathbf{y}} = \overline{\mathbf{H}} \boldsymbol{\Xi} \boldsymbol{s}_n + \overline{\mathbf{n}}_n, \tag{13}$$

where[1] $\overline{\mathbf{y}} = \mathrm{vec}(\underline{\mathbf{y}})$, $\overline{\mathbf{H}} = \boldsymbol{I} \otimes \underline{\mathbf{H}}$ is an equivalent channel matrix with a size of $(N_{\mathrm{Rx}}T \times N_{\mathrm{Tx}}T)$-element for ST-OFDM and $(N_{\mathrm{Rx}}M \times N_{\mathrm{Tx}}M)$-element for SF-OFDM. Most importantly, $\boldsymbol{\Xi}$ is referred to as the dispersion character matrix (DCM) [1], defined by $\boldsymbol{\Xi} = [\mathrm{vec}(\boldsymbol{\mathcal{B}}_0)], \mathrm{vec}(\boldsymbol{\mathcal{B}}_1)], \cdots, \mathrm{vec}(\boldsymbol{\mathcal{B}}_{Q-1})]$. $\boldsymbol{s}_n = [s_0, s_1, \cdots, s_{Q-1}]^T$ is the n-th segment of transmit signal frame $\boldsymbol{s}$ in Eq. (1). Additionally, $\overline{\mathbf{n}}_n = \mathrm{vec}(\underline{\mathbf{n}}_n)$.

Therefore, we obtain the estimated symbol vector $\hat{\boldsymbol{s}}_n$ by maximum likelihood (ML) detection expressed as [1]:

$$\hat{\boldsymbol{s}} = \arg\{\min(\|\overline{\mathbf{y}} - \overline{\mathbf{H}} \boldsymbol{\Xi} \mathbf{a}\|^2)\}, \tag{14}$$

where $\mathbf{a}$ denotes all possible combinations of the Q transmitted symbols in $\boldsymbol{s}_n$.

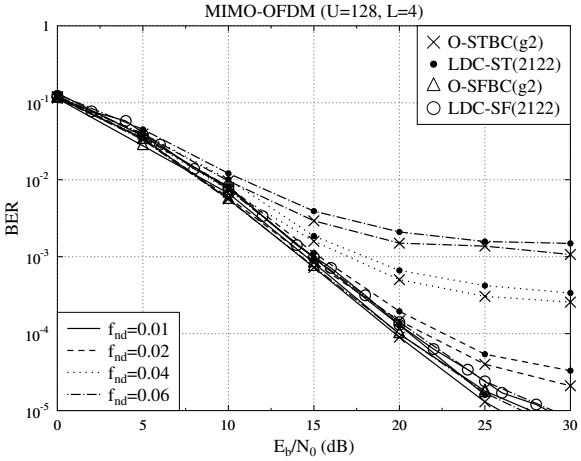
IV. Simulation Results and Discussions

In this section, we evaluate the performance achieved by the varying sets of simulation parameters, which are summarized in Table I.

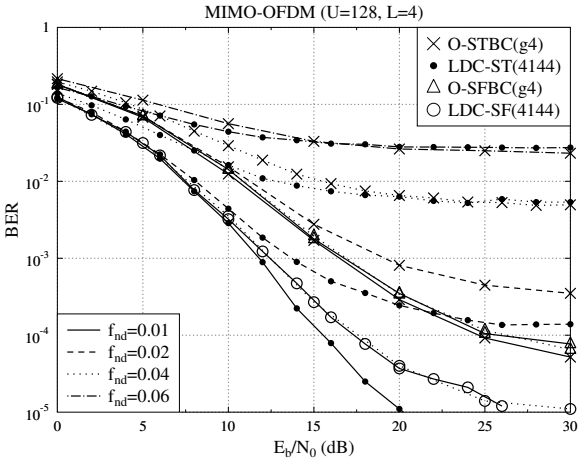
The BER performance results of ST and SF-oriented OFDM invoking LDC $(2122)^2$ or LDC (4144) upon varying the number of normalized Doppler frequency f_{ND} are shown in Fig. 5(a) and Fig. 5(b), respectively. Both LDC (2122) aided ST-OFDM having $N_{\mathrm{Tx}} = T = Q = 2$ and LDC (4144) aided ST-OFDM associated with $N_{\mathrm{Tx}} = 4$, $T = 4$, $Q = 4$ achieve the best performance for $f_{\mathrm{ND}} = 0.01$ when $L = 4$. At the same time the performance degrades when f_{ND} increases up to 0.06. This means the channel becomes even more time-selective with the duration of whole ST-OFDM symbol period for $T = 2$ and 4. By contrast, the performance of orthogonal STBC (g4) aided OFDM for $N_{\mathrm{Tx}} = 4$, $T = 8$, $Q = 4$ is decreased when $L = 4$, due to doubling the time slots for transmissing in comparison to the LDC (4144), upon varying the f_{ND} from 0.01 to 0.06.

[1]We define the vec$(\cdot)$ operation as the vertical stacking of the columns of an arbitrary matrix. $\boldsymbol{I}$ is a identity matrix with size of $(T \times T)$ or $(M \times M)$. $otimes$ is a Kronecker product operator.

[2]We denote that LDC $(N_{\mathrm{Tx}}, N_{\mathrm{Rx}}, T, Q)$ and $(N_{\mathrm{Tx}}, N_{\mathrm{Rx}}, M, Q)$ for ST or SF-encoded OFDM, respectively.

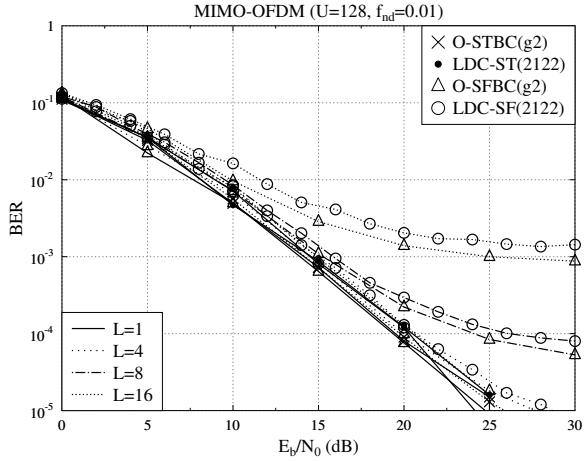


(a) $N_{\mathrm{t}} = 2$

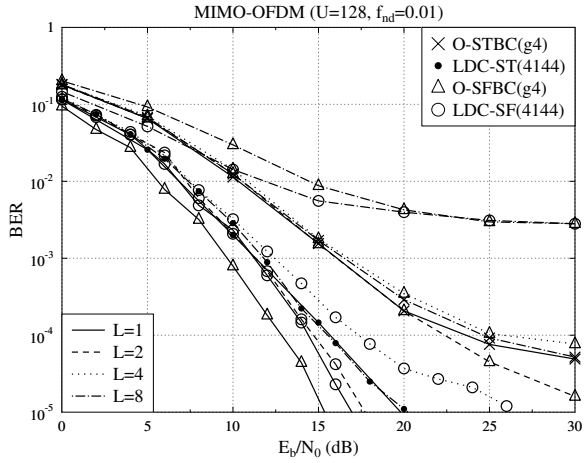


(b) $N_{\mathrm{t}} = 4$

Fig. 5. BER performance of MIMO-OFDM experiencing time-selective fading in terms of varying Doppler spread.



(a) $N_{\mathrm{t}} = 2$



(b) $N_{\mathrm{t}} = 4$

Fig. 6. BER performance of MIMO-OFDM experiencing frequency-selective fading in terms of varying delay spread (multipath).

Meanwhile, Fig. 5 also demonstrate that both the LDC and orthogonal code aided SF-OFDM are capable of achieving a constant performance in low delay spreads with the number of CIR taps $L = 4$ without the impact of increasing f_{ND}.

Furthermore, the performance of ST- and SF-OFDM invoking LDC (2122,4144) upon varying L is shown in Fig. 6. As seen in this figure, the performance of SF-OFDM is not impacted by the varying delay spreads for a given $f_{\mathrm{ND}} = 0.01$. However, SF-OFDM benefits frequency-diversity for neighboring M subcarriers having correlated fading coefficients associated with low frequency-selectivity with $L = 1, 2, 4$ for $N_{\mathrm{Tx}} = 2$ and with $L = 1, 2$ for $N_{\mathrm{Tx}} = 4$. Note that, the performance degrades when increasing L. Particularly, LDC (4144) aided SF-OFDM having $M = 4$ outperforms orthogonal SFBC (g4) aided OFDM with $M = 8$ when communicating over the frequency-selective fading channel of $L = 2, 4$.

V. Conclusions

In this contribution, we investigated ST- and SF-diversity oriented OFDM systems invoking LDC, in order to study the advantages and disadvantages of transmit diversity based MIMO transmission over time-/frequency-selective fading channels. Our results demonstrate that when the channel is constant within the coherent time/bandwidth, ST- or SF-OFDM is capable of achieving full diversity gain in ST or SF domains. The ST-OFDM scheme is sensitive to exploiting diversity gains subject to the impact of varying channel Doppler spreads; while the performance of SF-OFDM is mainly subject to delay spread. Moreover, compared with the orthogonal STBC/SFBC (g4), the LDC-aided ST/SF-OFDM is flexible in configuring various numbers of transmit antenna and time-slots or frequency-tones. When the transmitter employs more than two antennas, the performance of LDC-aided ST/SF-OFDM

Gharavi, Hamid; Hu, Bin; Zhang, Jiayi.
"MIMO-OFDM Transmissions Invoking Space-Time/Frequency Linear Dispersion Codes Subject to Doppler and Delay Spreads."
Paper presented at IEEE-wcn.org/, Doha, Qatar. April 3, 2016 - April 6, 2016.

schemes is less impacted by channel Doppler/delay spreads, as compared with orthogonal block codes.

REFERENCES

[1] L. Hanzo, O. Alamri, M. El-Hajjar, and N. Wu, *Near-capacity Multi-functional MIMO Systems: Sphere-packing, Iterative Detection and Cooperation*. Wiley, 2009.

[2] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1451–1458, Oct. 1998.

[3] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1456–1467, Jul. 1999.

[4] B. Hassibi and B. M. Hochwald, "High-rate codes that are linear in space and time," *IEEE Trans. Inf. Theory*, vol. 48, no. 7, pp. 1804–1824, Jul. 2002.

[5] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*. Cambridge University Press, 2003.

[6] J. Heath, R. W. and A. J. Paulraj, "Linear dispersion codes for MIMO systems based on frame theory," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2429–2441, Oct. 2002.

[7] N. Wu and H. Gharavi, "Asynchronous cooperative MIMO systems using a linear dispersion structure," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 779–787, Feb. 2010.

[8] N. Wu, S. Sugiura, and L. Hanzo, "Coherent versus noncoherent," *IEEE Veh. Technol. Mag.*, vol. 6, no. 4, pp. 38–48, Dec 2011.

[9] C. Xu and H. Gharavi, "A low-complexity solution to decode diversity-oriented block codes in MIMO systems with inter-symbol interference," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3574–3587, Oct. 2012.

[10] L. Hanzo, M. Münster, B.-J. Choi, and T. Keller, *OFDM and MC-CDMA for Broadband Multi-User Communications, WLANs and Broadcasting*. Wiley, 2003.

[11] B. Muquet, Z. Wang, G. B. Giannakis, M. de Courville, and P. Duhamel, "Cyclic prefixing or zero padding for wireless multicarrier transmissions?" *IEEE Trans. Commun.*, vol. 50, no. 12, pp. 2136–2148, Dec. 2002.

[12] J. Zhang, L.-L. Yang, L. Hanzo, and H. Gharavi, "Advances in cooperative single-carrier FDMA communications: Beyond LTE-advanced," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 730–756, 2nd Quarter 2015.

[13] P. Stoica and E. Lindskog, "Space-time block coding for channels with intersymbol interference," in *Proc. ACSSC 2001*, vol. 1, Nov. 2001, pp. 252–256.

[14] A. Kuhestani and P. Azmi, "Design of efficient full-rate linear dispersion space-time block codes over correlated fading channels," *IET Commun.*, vol. 7, no. 12, pp. 1243–1253, Aug 2013.

[15] A. Kuhestani, H. Pilaram, and A. Mohammadi, "Simply decoded efficient full-rate space-time block codes over correlated rician fading channels," *IET Commun.*, vol. 8, no. 10, pp. 1684–1695, July 2014.

[16] G. Bauch, "Space-time block codes versus space-frequency block codes," in *IEEE VTC2003-Spring*, vol. 1, Apr. 2003, pp. 567–571.

[17] W. Zhang, X.-G. Xia, and K. Ben Letaief, "Space-time/frequency coding for MIMO-OFDM in next generation broadband wireless systems," *IEEE Wireless Commun. Mag.*, vol. 14, no. 3, pp. 32–43, Jun. 2007.

[18] L. Hanzo, Y. Akhtman, L. Wang, and M. Jiang, *MIMO-OFDM for LTE, WIFI and WIMAX: Coherent Versus Non-Coherent and Cooperative Turbo-Transceivers*. Wiley (IEEE Press), Oct. 2010.

[19] H. Gharavi and B. Hu, "Cooperative diversity routing and transmission for wireless sensor networks," *IET Wireless Sens. Syst.*, vol. 3, no. 4, pp. 277–288, Dec. 2013.

[20] G. Stuber, J. Barry, S. McLaughlin, Y. Li, M.-A. Ingram, and T. Pratt, "Broadband MIMO-OFDM wireless communications," *Proc. IEEE*, vol. 92, no. 2, pp. 271–294, Feb. 2004.

[21] H. Bolcskei, M. Borgmann, and A. Paulraj, "Impact of the propagation environment on the performance of space-frequency coded MIMO-OFDM," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 3, pp. 427–439, Apr 2003.

[22] A. Molisch, M. Win, and J. Winters, "Space-time-frequency (STF) coding for MIMO-OFDM systems," *IEEE Commun. Lett.*, vol. 6, no. 9, pp. 370–372, Sept 2002.

[23] Z. Liu, Y. Xin, and G. Giannakis, "Space-time-frequency coded OFDM over frequency-selective fading channels," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2465–2476, Oct 2002.

[24] W. Su, Z. Safar, and K. Liu, "Towards maximum achievable diversity in space, time, and frequency: performance analysis and code design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1847–1857, Jul. 2005.

[25] J. Wu and S. Blostein, "High-rate codes over space, time, and frequency," in *Proc. IEEE GLOBECOM 2005*, vol. 6, Dec. 2005, p. 6.

[26] G. V. Rangaraj, D. Jalihal, and K. Giridhar, "Exploiting multipath diversity using space-frequency linear dispersion codes in MIMO-OFDM systems," in *Proc. ICC 2005*, vol. 4, May 2005, pp. 2650–2654.

[27] J. Wu and S. Blostein, "High-rate diversity across time and frequency using linear dispersion," *IEEE Trans. Commun.*, vol. 56, no. 9, pp. 1469–1477, Sep. 2008.

[28] L.-L. Yang, *Multicarrier Communications*. Wiley, 2009.

Microwave Radiometry of Blackbody Radiation

Dazhen Gu[1, 2] and David K. Walker[1]
[1]National Institute of Standards and Technology, Boulder, CO USA
[2]Department of Electrical, Computer, and Energy Engineering, University of Colorado, Boulder, CO
dazhen.gu@nist.gov

***Abstract*—We outline the theoretical formulation of radiometry of the free-space radiation emitted by a blackbody target. Simulation shows a much smaller drop of radiation intensity of a Lambertian source than that of an incoherent source in the near-field region, indicative of a powerful influence produced by the coherence property of the blackbody source. Further, the coupling of the radiation to a radiometer is formulated by the plane-wave scattering theory of the radiation field.**

***Index Terms*—Calibration of remote sensing, coherence propagation, electromagnetic radiation, microwave blackbody, plane-wave expansion, thermal noise.**

I. Introduction

Microwave radiometry has become more and more crucial in remote-sensing instruments due to its significant role in weather forecasting and climate studies. Nearly all the environmental parameters observed by such systems are originally represented by temperature, which is directly extracted from total-power radiometer measurements. The measurement accuracy relies on the radiometric calibration, often including the observation of two known thermal noise sources. One typical configuration consists of a man-made blackbody target in conjunction with the naturally accessible cosmic background. The radiation from both sources are independently collected through the front-end antenna of the radiometer to complete a calibration.

On one hand, the radiation arising from the cosmic background is fairly well known to us. On the other hand, the artificial blackbody usually possesses imperfect properties and measurements of its radiation almost always take place in the near-field (NF) region. As a consequence, precise knowledge of the blackbody NF radiation has a preponderant impact on calibration accuracy. In this paper, we furnish a theoretical model of radiometric measurements of blackbody radiation at close range. The radiation emanating from the blackbody is modeled by coherence-propagation theory and the power coupled to the antenna is based on the plane-wave scattering-matrix theory.

II. Plane-Wave Expansion of Primary Thermal Source

In general, blackbody sources can be compartmentalized into two categories, namely a primary source and a secondary source. Throughout this report, we concentrate on the primary source, which often embodies a thermal object with ideal emission characteristics. Handling the secondary source is actually more straightforward and can be simply inferred from the approaches in what follows.

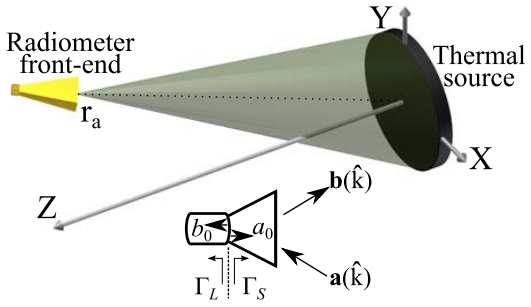


Fig. 1. Illustration of a radiometer collecting radiation from a blackbody and some notations for the plane-wave scattering matrix representing an antenna.

We consider that a radiometric receiver detects the radiation from a planar source located at the plane $z = 0$, as shown in Fig. 1. Although portrayed in a circular shape, the source can have any arbitrary form. We first attempt to obtain the "cumulative spectra" $\mathbf{U}$ of prescribed currents $\mathbf{j}$ in vacuum [1]. It is well known that the vector potential $\mathbf{A}$ relates to $\mathbf{j}$ by

$$\mathbf{A}(\mathbf{r}) = \mu_0 \int_{\mathscr{A}} \mathbf{j}(\mathbf{r}') \frac{\exp(ik|\mathbf{r}-\mathbf{r}'|)}{4\pi|\mathbf{r}-\mathbf{r}'|} d^2r', \tag{1}$$

where μ_0 is the permeability of free space, $\mathbf{r}$ and $\mathbf{r}'$ symbolize the field and the source point, respectively. The integration is carried over the area $\mathscr{A}$ occupied by the blackbody. With the utility of the following identity

$$\frac{\exp(ik|\mathbf{r}-\mathbf{r}'|)}{ik|\mathbf{r}-\mathbf{r}'|} = \frac{1}{2\pi} \int \exp\left(i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')\right) \frac{d^2K}{k\gamma}, \tag{2}$$

the vector potential $\mathbf{A}$ can be represented in terms of the cumulative spectra as

$$\mathbf{A}(\mathbf{r}) = \frac{i}{2\pi\omega} \int \mathbf{U}(\hat{\mathbf{k}}) \exp(i\mathbf{k}\cdot\mathbf{r}) \frac{d^2K}{k\gamma}, \tag{3}$$

where $\mathbf{U}$ is given by

$$\mathbf{U}(\hat{\mathbf{k}}) = \frac{\mu_0 k\omega}{4\pi} \int_{\mathscr{A}} \mathbf{j}(\mathbf{r}') \exp(-i\mathbf{k}\cdot\mathbf{r}') d^2r'. \tag{4}$$

Here, the wave vector $\mathbf{k}$ is composed of $(\alpha\hat{\mathbf{x}} + \beta\hat{\mathbf{y}} + \gamma\hat{\mathbf{z}})$ with its transverse component $\mathbf{K} = \alpha\hat{\mathbf{x}} + \beta\hat{\mathbf{y}}$ and its unit vector $\hat{\mathbf{k}} = \mathbf{k}/k$. After acquiring the expression of $\mathbf{U}(\hat{\mathbf{k}})$, we next obtain the plane-wave spectrum of the blackbody radiation from the relation between the electric field $\mathbf{E}(\mathbf{r})$ and the vector potential $\mathbf{A}(\mathbf{r})$

$$\mathbf{E}(\mathbf{r}) = i\omega \left(\mathbf{A}(\mathbf{r}) + \frac{1}{k^2} \nabla\nabla\cdot\mathbf{A}(\mathbf{r}) \right). \tag{5}$$

Gu, Dazhen.
"Microwave Radiometry of Blackbody Radiation."
Paper presented at Conference on Precision Electromagnetic Measurements, Ottawa, ON, Canada. July 10, 2016 - July 15, 2016.

In view of (3) and (5), the plane-wave spectrum $\mathbf{a}(\hat{\mathbf{k}})$ can be conveniently expressed as

$$\mathbf{a}(\hat{\mathbf{k}}) = (\hat{\mathbf{k}}\hat{\mathbf{k}} - \mathbb{I}) \cdot \mathbf{U}(\hat{\mathbf{k}}), \tag{6}$$

where $\mathbb{I}$ is the unit dyad.

III. Coherence Property of Blackbody Radiation Source

In contrast to the customary postulation that blackbody sources are of complete spatial incoherence, the correlation distance is on the order of the radiation wavelength for any blackbody source with its far-field (FF) radiation following the Lambertian cosine law. In light of its coherence property, the correlation tensor of the blackbody source at any pair of points ($\mathbf{r}'_1$ and $\mathbf{r}'_2$) can be factorized into two terms under the quasi-homogeneous condition [2]:

$$\mathbb{W}_{\mathbf{j}}(\mathbf{r}'_1, \mathbf{r}'_2) = L_{\mathbf{j}}(\tfrac{\mathbf{r}'_1+\mathbf{r}'_2}{2}) C_{\mathbf{j}}(\mathbf{r}'_2 - \mathbf{r}'_1)\mathbb{I}, \tag{7}$$

where $L_{\mathbf{j}}$ is the intensity distribution of $\mathbf{j}$ in the source plane and $C_{\mathbf{j}}$ is the source correlation function.

In spite of a small value, the non-negligible correlation length renders a remarkable influence on the radiation field especially in the NF. In Fig. 2, we show the radiation intensity normalized to its FF value as a function of the separation distance for blackbody targets of various sizes. All the normalized quantities approach 1 asymptotically and the target with a smaller footprint appears relatively brighter in the NF. Furthermore, the inset plot clearly indicates a pronounced difference between a completely incoherent source and a partially coherent source (the blackbody).

IV. Antenna Coupling of Radiation

Without losing generality, an antenna can be characterized by the receiving function $\mathbf{s}(\hat{\mathbf{k}})$. Referring to Fig. 1, the emergent wave amplitude b_0 at the antenna feed is

$$b_0 = \Gamma_S a_0 + \int \mathbf{s}'(\hat{\mathbf{k}}) \cdot \mathbf{a}(\hat{\mathbf{k}}) \frac{d^2K}{k\gamma}. \tag{8}$$

This is equivalent to a source with a reflection coefficient Γ_S excited by a generated wave b_G (the second term on the RHS of (8)). Here, the receiving function is primed to distinguish its translated position and rotated orientation $(\mathbf{r_a}; \mathbb{R}\cdot\hat{\mathbf{x}}, \mathbb{R}\cdot\hat{\mathbf{y}}, \mathbb{R}\cdot\hat{\mathbf{z}})$ from the nominal $(\mathbf{O}; \hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}})$. $\mathbf{s}'(\hat{\mathbf{k}})$ relates to $\mathbf{s}(\hat{\mathbf{k}})$ through

$$\mathbf{s}'(\hat{\mathbf{k}}) = \mathbb{R} \cdot \mathbf{s}(\mathbb{R}^{-1} \cdot \hat{\mathbf{k}}) \exp(i\mathbf{k} \cdot \mathbf{r}_a). \tag{9}$$

$\mathbb{R}$ is the transformation matrix pertaining to the Euler angle.

From the radiometric standpoint, the power received by the radiometer can be found as

$$P_{rec} = \frac{1}{2k^2 Z_0} \frac{(1 - |\Gamma_L|^2)|b_G|^2}{|1 - \Gamma_L \Gamma_S|^2}, \tag{10}$$

where $Z_0 \approx 377\ \Omega$ is the wave impedance in vacuum, Γ_L is the reflection coefficient looking into the radiometer from the antenna feed. With (4), (6), and (8) at our disposal, the essential part of the received power can now be quantified by the ensemble average:

$$\langle b_G^* b_G \rangle = \pi^2 k^2 \mu_0^2 \omega^2 \int \frac{d^2K_1}{k\gamma_1^*} \int \frac{d^2K_2}{k\gamma_2} \tag{11}$$
$$\mathbf{s}'^*(\hat{\mathbf{k}}_1) \cdot (\hat{\mathbf{k}}_1^* \hat{\mathbf{k}}_1^* - \mathbb{I}) \cdot \widetilde{\mathbb{W}}_{\mathbf{j}}(-\hat{\mathbf{k}}_1, \hat{\mathbf{k}}_2) \cdot (\hat{\mathbf{k}}_2 \hat{\mathbf{k}}_2 - \mathbb{I}) \cdot \mathbf{s}'(\hat{\mathbf{k}}_2),$$

where $\widetilde{\mathbb{W}}_{\mathbf{j}}$ is the Fourier transform of $\mathbb{W}_{\mathbf{j}}$ in (7). So long as the antenna receiving function (radiation pattern) is available to us from simulation or measurements, the received power of the radiometer can be predicted by numerical integration from the closed form of (11). Although the computational cost may seem unduly high, a variety of numerical techniques can be applied to make the calculation feasible in some specialized yet commonly encountered circumstances.

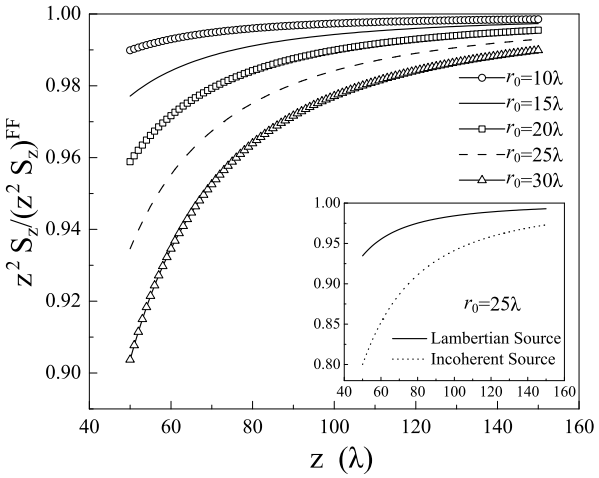


Fig. 2. Normalized radiant intensity $(z^2 \cdot S_z)$ as a function of the separation distance between the blackbody and the on-axis observation point ($\mathbf{r}_a = z\hat{\mathbf{z}}$) for a target radius of 10λ, 15λ, 20λ, 25λ, and 30λ. The inset shows the comparison between the incoherent source and the *Lambertian* source (partially coherent) with a circular profile $r_0 = 25\lambda$.

V. Conclusion

We have established the formulation of blackbody-radiation radiometry in a rigorous way by using coherence-propagation theory and plane-wave scattering theory. The power received by a radiometer is closely tied to the coherence function of the source current in the blackbody through the plane-wave spectrum of the radiation field. The correlation distance at a sub-wavelength scale produces a profound effect on the radiation emerging from the blackbody source. Simulation results including the NF coupling through antennas will be reported at the conference.

References

[1] D. M. Kerns, "Plane-wave scattering-matrix theory of anennas and antenna-antenna interactions," *NBS Monograph* 162, June 1981.

[2] D. Gu and D. K. Walker, "Application of coherence theory to modeling of blackbody radiation at close range," *IEEE Trans. Microw. Theory Techn.*, vol. 63, no. 5, pp. 1475-1488, May 2015.

Gu, Dazhen.
"Microwave Radiometry of Blackbody Radiation."
Paper presented at Conference on Precision Electromagnetic Measurements, Ottawa, ON, Canada. July 10, 2016 - July 15, 2016.

Development of a NIST WR10 Radiometer

Jack Surek[1], Chunyue Cheng[2], Dazhen Gu[1] and David Walker[1]

[1]National Institute of Standards and Technology, Communications Technology Laboratory
325 Broadway Street, Boulder, CO 80305, USA
jack.surek@nist.gov
[2]Beijing Institute of Radio Metrology and Measurement
NO.52 Yongding Road
Haidian District, Beijing 100854, China

***Abstract* — This paper describes the development of a waveguide radiometer to measure noise from millimeter wave electronic components from 75 GHz to 110 GHz. The radiometer will estimate the noise temperature of a device under test (DUT) based on comparison with room temperature and 77K noise standards. This is a standard physical approach in other NIST microwave radiometers. The radiometer is particularly amenable to performing noise temperature as well as noise parameter measurements for amplifier and transistor characterization. As wireless communications progresses towards millimeter wave systems, noise characterization of related components and subsystems becomes essential. We report our progress in radiometer design, construction and verification for millimeter wave noise metrology at NIST.**

***Index Terms* — Noise metrology, radiometry, millimeter waves, WR10 band, waveguide radiometer**

I. INTRODUCTION

The thermal noise metrology project at NIST in Boulder has a long history of applying radiometry to noise temperature measurement. These endeavors have ranged from lab-based metrology stations for the characterization of electronic noise emitted from active and passive components to blackbody reference targets for satellite-based microwave remote sensing. The present WR10 radiometer development reflects a renewed emphasis in electronic component and circuit noise characterization. This is in part due to a recent mandate to support advanced wireless communications.

A radiometer estimates object temperature by comparing the radiation it receives from this object with one or more standards. For our lab-based WR10 radiometer, the object is a one-port electronic DUT. As in other NIST microwave radiometers [1], DUT noise is compared to noise from a room temperature and a liquid-nitrogen-boiling-point standard using the same calibrated receiver. Each noise source is switched into the receiver input as shown in Fig.1.

Usually the DUT is a diode that is designed for enhanced noise output. Once the noise spectrum from this diode is characterized with the radiometer, two-port electronic components such as waveguide amplifiers and filters can be inserted between this DUT and the radiometer receiver. In this configuration, the diode's noise can provide a unique source of broadband simultaneous frequency excitation.

When measuring noise power, the assumption of linearity is maintained by nibbling a broad noise spectrum in short enough spectral segments to accurately estimate temperature across any peaks. With short enough spectral segments the ratio of the power to the bandwidth of these segments defines their local temperature [2]. Broadband noise from each source is downconverted in spectral segments that range from 10 MHz to 1 GHz wide. Noise power measurement per segment is slow enough to assume that equilibrium has been reached.

Peaks in the noise spectrum of an electronic device represent non-thermal variations in temperature. A focus on these peaks may provide useful information about the interplay between materials, structure and circuit topology. These effects are often attributed to circuit nonlinearity and intermodulation distortion. Injected noise may provide a much more thorough way to explore these energy dependencies. We proceed with the WR10 radiometer development by assuming that its characterization as a linear network, described above, remains a valid basis for observing non-thermal variations.

Kang et al recently reported a WR10 radiometer relying on the same physical approach of estimating DUT temperature based on room temperature and 77K noise standards [3]. Our design uses the same tuned RF receiver architecture and its characterization relies on the same theory outlined over the last 50 years. We do allow for two paths to contrast balanced versus sub-harmonic mixing, but the basic differences in sensitivity and accuracy between our radiometer and other modern designs will likely be small. The specific difference for our work comes from applying this hardware to two-port component testing and from a focus on noise spectral features for what they reveal about component performance. As a specific example, a well-characterized noise diode would provide a signal source to the input of a broadband signal amplifier DUT. The tuned receiver parses emissions from this DUT into short IF spectral segments. In this way DUT sensitivity to the simultaneous excitation of one to many simultaneous instantaneous frequencies (wavelets) in the WR10 band would be revealed in comparison with DUT noise at the same bias without noise diode excitation. The average noise level between these would be normalized, accentuating peaks and valleys. Radiometer characterization for standard operation remains fundamental to making such measurements and we plan to fully report on this at the conference. Beyond this, we hope to report on one such two-port DUT experiment performed at narrow IF bandwidth.

Gu, Dazhen; Surek, Jack; Walker, Dave.
"Development of a NIST WR10 Radiometer."
Paper presented at Conference on Precision Electromagnetic Measurements, Ottawa, ON, Canada. July 10, 2016 - July 15, 2016.

II. WR10 RADIOMETER

A. Design

The WR10 radiometer relies on swapping in temperature standards, T_{RT} and T_{CT}, as well as the DUT, T_{DUT}, with a 4-port rotary waveguide switch in order to piece out representative noise spectra about a given frequency in the WR10 band by mixing and IF filtering through a shared signal path. Our radiometer allows for downconversion through a balanced mixer or a 1/3 sub-harmonic mixer with a second 4-port rotary waveguide switch in order to compare these. Receiver operating temperature of 23.0° C ± 0.5° C will be maintained with water cooling. However, during this initial checkout phase, components operate from 22.0° C to 24° C.

B. Components and Characterization

Noise temperature standards in Fig.1 are: a WR10 matched load T_{RT}, a NIST custom cryogenic standard T_{CT} that operates at the boiling point of liquid nitrogen, 77 K ± 1 K [4] and a device under test T_{DUT} , often a noise diode.

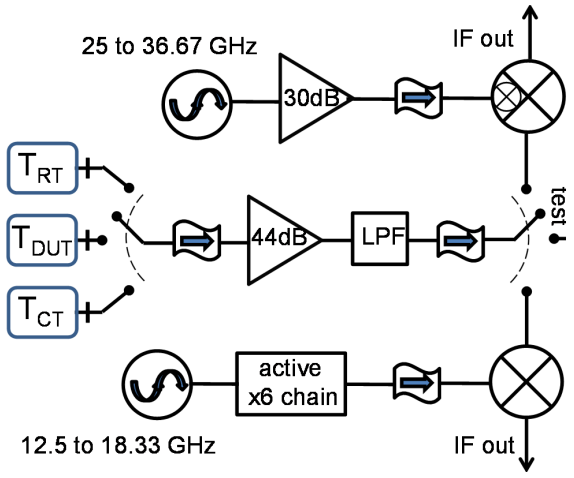


Fig. 1 Architecture of WR10 switching radiometer.

We have two diode noise sources to verify the performance of the WR10 radiometer over its operational life. The vendor specifies these to have an excess noise ratio (ENR) of 12 dB with ± 3 dB flatness across the band. As part of initial testing, the input was switched between these two noise diodes and the matched termination, directly measuring power for these connections at the point "test" in Fig. 1 with a WR10 power sensor and no added filtering. The direct ratios of measured power between each diode and the matched termination provide ENR estimates of 8.0 dB and 8.5 dB, respectively. Two amplifiers, each with an average noise figure of 5.5 dB (per the vendor's specification) are concatenated to make up the 44 dB gain stage in Fig.1. Considering this noise figure and overall amplification, preliminary ENR measurements are reasonable. These results confirm operation of each diode and the amplifier chain. Next steps will be to characterize the gain and noise of the amplifiers in the 44 dB chain in order to de-embed accurate ENR values for these noise diodes, and to measure with short spectral segments using each mixer path, in order to see ENR variations.

Balanced mixer conversion loss was measured with LO consistently offset by 100 MHz from the RF. RF and LO power were monitored with couplers in these paths with these couplings accounted for. Measured IF power is shown offset by +15 dBm in Fig.3. Note that the LO signal is between -2 dBm and +2 dBm while the RF signal is between -10 dBm and -13 dBm in Fig.3. The vendor specifies an LO drive power of between 2 dBm and 4 dBm and an RF signal level of around -10 dBm to limit conversion loss to no more than 12 dB. Fig. 2 shows conversion loss of 12 dB or less up to ~108 GHz even at the low LO and RF powers of our initial tests.

III. CONCLUSIONS

We will report on further progress towards the full WR10 radiometer system characterization and implementation at the conference.

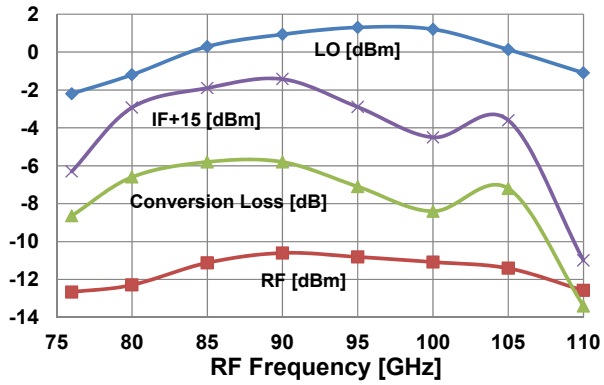


Fig. 2. Balanced mixer conversion loss for the above input powers and with sinusoidal LO less than sinusoidal RF by 100 MHz.

REFERENCES

[1] J. Randa and L. A. Terrell, "Noise Temperature Measurement System for the WR28 Band," NIST Technical Note 1395, August 1997.

[2] J. Randa, "Amplifier and Transistor Noise-Parameter Measurements," *Wiley Encyclopedia of Electrical and Electronics Engineering,* pp. 1–38, online 15 Sept. 2014.

[3] T. Kang, J. Kim, N. Kang and J. Kang, "A Thermal Noise Measurement System for Noise Temperature Standards in W-Band," IEEE Trans. Instrum. Meas., vol. 64, no. 6, pp.1741-1747.

[4] W. C. Daywitt, "Design and Error Analysis for the WR10 Thermal Noise Standard," NBS Technical note 1071, Dec. 1983.

Gu, Dazhen; Surek, Jack; Walker, Dave.
"Development of a NIST WR10 Radiometer."
Paper presented at Conference on Precision Electromagnetic Measurements, Ottawa, ON, Canada. July 10, 2016 - July 15, 2016.

Demultiplexing Spectrum-Sharing Field Sources with Distributed Field Probes

Daniel G. Kuester, *Member, IEEE*, Ryan T. Jacobs, Yao Ma, *Senior Member, IEEE*, and Jason B. Coder
U.S. Department of Commerce
National Institute of Standards and Technology
Communications Technology Laboratory, RF Technology Division
Boulder, Colorado 80305
Email: daniel.kuester@nist.gov, yao.ma@nist.gov, ryan.jacobs@nist.gov, jason.coder@nist.gov

***Abstract*—A complete characterization of multiple-device wireless interactions must include data relatable to the electromagnetic field radiated by each device under test (DUT). If these field sources are separable in time or frequency, they can be demultiplexed with a single probe antenna and time gating or bandpass filtering. Spectrum-sharing coexistence testing, however, may deal with simultaneous co-channel radiation. Communication channels may realize orthogonality in signal modulation or coding, for example, instead of time or frequency. These signals need an alternative to time or frequency as a basis to discriminate between signals in tests.**

We explore here distributed multi-probe detection as a means to address this problem. Simultaneous coherent detection of quadrature baseband at multiple probes provides degrees of freedom necessary to decompose modulated signals with different origins in space. The approximately deterministic simple propagation behavior in an anechoic chamber allows us to estimate channel delay, phase shift, and attenuation parameters between each combination of probes and DUTs. These parameters are sufficient to extrapolate a transfer matrix across frequency that we invert to compute a weighting matrix that deembeds the received superposition of DUT waveforms. We demonstrate experiments that demultiplex three DUTs: a 802.11n Wi-Fi link pair and a source of LTE traffic, all in overlapping channels near 2.4 GHz. The demultiplexed channels show clearly the channel occupancy of each DUT without time-gating or frequency filtering.

I. Introduction

Industry and government are developing technology, standards, and regulation policy to share spectrum allocations at the desirable frequencies below 6 GHz. The results of this work include the proposed Citizen's Broadband Radio Service (CBRS) near 3.5 GHz [1], IEEE 802.22 in TV whitespace bands [2], [3], and the well-known industrial, scientific, and medical (ISM) bands around 900 MHz, 2.4 GHz and 5.8 GHz. Increasing access to this spectrum presents an opportunity to increase wireless data network capacity if existing incumbent users can be protected from interference.

At its most basic, a spectrum sharing scheme needs to enable some type of separation of communication channels separable between different systems. Common mechanisms in ISM bands include frequency hopping and spread spectrum. Television whitespace use includes adaptive access by cognitive radio and a centralized database to minimize interference with regional television broadcasting. The proposed CBRS creates an elaborate three-tier sharing scheme built around a centralized spectrum access system (SAS) that coordinates use by non-government tiers [4]. Many approaches to adaptive spectrum access that use spectrum sensing may need to coexist, including combinations of physical orthogonality (time, frequency, space, or polarization) and/or signal orthogonality (like modulation or coding).

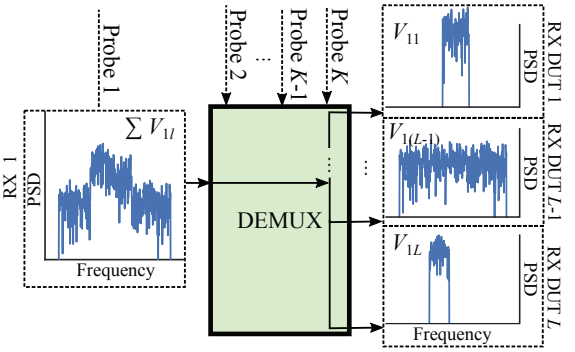


Fig. 1. Spatial demultiplexing ("DEMUX") decomposes a waveform received from multiple transmitter DUTs. The reference input waveform is defined at the reference port connected to probe 1. Coherent baseband receivers at probes 2 through K provide the additional necessary input degrees of freedom.

Spectrum-sensing and adaptive approaches like those above require tight coordination — ensuring proper radiation from each spectrum-sharing device becomes vitally important. Improper transmissions could be the result of product implementation errors, misconfiguration, or abuse, and may degrade communication network availability, data throughput, and latency. The modulation radiated by each transmitter is therefore exactly the fundamental quantity that needs to be captured to test coexistence. This means demultiplexing the signals superimposed over the air into separate channels, even when every radiator transmits different signals that are simulataneously in the same frequency band.

Current standards for over-the-air testing in the electromagnetic compatibility (EMC) and communication communities often approach measurements at the high and low extremes

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao.
"Demultiplexing Spectrum-Sharing Field Sources with Distributed Field Probes."
Paper presented at 2016 IEEE International Symposium on Electromagnetic Compatibility and Signal Integrity, Ottawa, ON, Canada. July 25, 2016 - July 29, 2016.

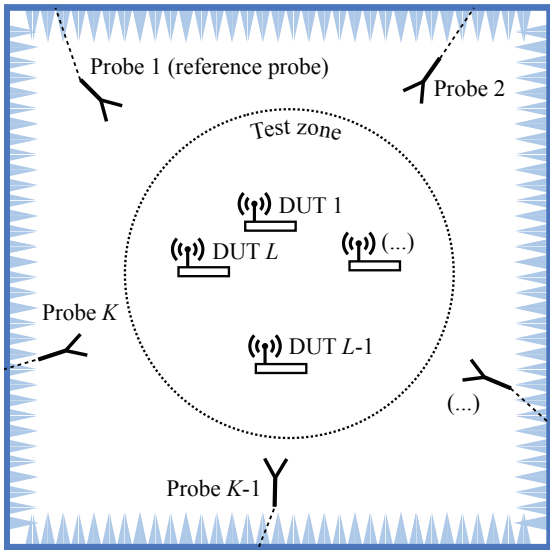


Fig. 2. Example distribution of probe antennas and DUTs in the test zone. They are presumed motionless during test. For demultiplexing, $K > L$.

of the networking stack. A high-level network performance measurand, or Key Performance Indicator (KPI), is what is observable by the user, often throughput or latency [5]. Research and standardization efforts may lead to shared-spectrum KPI tests that point directly to problems that noticeable to end-users and network operators (perhaps the network is "slow" or "won't connect"), but not diagnostic data to troubleshoot underlying causes. Low-level physical signal measurands related to timing, modulation, or power are the usual way to diagnose problems at the device level. How do we apply these methods to over-the-air spectrum sharing tests if they all summed together when they arrive at our signal analyzer?

We propose in this paper a technique to separate baseband waveforms detected from multiple co-channel DUTs, illustrated in Figs. 1 and 2. We establish a two-parameter matrix model to extrapolate multichannel anechoic interactions across frequency (Section II). We develop a method to estimate these parameters from measured data, and invert the probe matrix over frequency to determine probe weights that demultiplex the DUTs (Section III). Last, we demonstrate application to coexistence work by demultiplexing coexisting LTE and Wi-Fi signals (Section IV).

II. Multichannel System Model

This section defines the physical model and notation that underlie the rest of the paper. We reduce the interactions between each probe-DUT pair across a band of interest to a propagation delay and a complex response coefficient.

The environment is an anechoic chamber like the one illustrated in Fig. 2. There are K probe antennas, each connected to a separate channel of a coherent measurement receiver. The receiver acquires complex in-phase and quadrature (IQ) baseband waveforms on all of K channels simultaneously. The test zone in the center of the chamber has $L < K$ DUT transmitters that, during tests, may potentially all transmit simultaneously at the same frequencies.

A. CW Signaling Case - Arbitrary Scattering Environment

First consider an unknown quiet scattering environment. For this subsection only, assume the lth DUT radiates continuous-wave (CW) with magnitude and phase represented voltage phasor V_l^{DUT}. The wave propagates to each of the K receive ports, scaled by transmit-to-receive (DUT-to-probe) response coefficient h_{kl}. The transmit-receive transfer function encapsulates all linear single-frequency attenuation and phase shift: propagation, scattering, impedance mismatch, antenna transduction, receiver frequency response, etc. The component of the voltage phasor received at the k^{th} probe from lth DUT is represented by the phasor V_{kl}. The wave undergoes linear and time-invariant propagation with added receive channel noise, N_k:

$$V_{kl} = h_{kl} V_l^{\text{DUT}} + N_k. \tag{1}$$

We need an equation in terms of probe (not DUT) voltages for over the air (OTA) tests. To this end, define the receive transfer function as relative to a reference probe defined at $k = 1$: $V_{1l} = h_{1l} V_l^{\text{DUT}}$. Solve for V_l^{DUT} and substitute into (1) to find

$$V_{kl} = \frac{h_{kl}}{h_{1l}} V_{1l} + N_k = H_{kl} V_{1l} + N_k. \tag{2}$$

The response coefficient $H_{kl} = h_{kl}/h_{1l}$ for $h_{1l} \neq 0$ relates received phasors, independent of the magnitude or phase of the transmitter. It is no longer strictly a transfer function, because it only relates received signals.

Now let all transmit devices excite CW at the same frequency. Each receive channel is related to the reference receive channel by weighted superposition of each reference antenna transmitter component l:

$$V_k = \sum_{l=1}^{L} H_{kl} V_{1l} + N_k. \tag{3}$$

This is equivalent to matrix multiplication,

$$\mathbf{V} = \mathbf{H}\mathbf{V}_1 + \mathbf{N}. \tag{4}$$

Here $\mathbf{V}$ is the $1 \times K$ row vector of probe receive phasors V_k; $\mathbf{H}$ is the $K \times L$ probe response matrix with elements H_{kl}; $\mathbf{V}_1$ is the $1 \times L$ row vector comprising the demultiplexed voltages at the reference probe, V_{1l}; and $\mathbf{N}$ is the $1 \times K$ row vector of independent and identically distributed (i.i.d.) receive channel noise samples N_k.

The phasor could be normalized as a node voltage in simulation or microwave parameter system like pseudowaves for measurement [6]. The choice of pseudowaves leads to response functions that are a subset of $(K + L)$-port scattering parameters.

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao.
"Demultiplexing Spectrum-Sharing Field Sources with Distributed Field Probes."
Paper presented at 2016 IEEE International Symposium on Electromagnetic Compatibility and Signal Integrity, Ottawa, ON, Canada. July 25, 2016 - July 29, 2016.

B. Modulated Case - Free Field Environment

Each DUT is a communication device, not a swept-frequency CW test instrument; we need to model the channel response for modulated signals. The model will stay tractable in our test scope by exploiting the free-field propagation approximated by the anechoic chamber.

The ideal transmit-receive response in the the $(k,l)^{\text{th}}$ pair in free space is [7]

$$h_{kl}(\omega) = \frac{1}{r_{kl}} \frac{K_k}{\text{AF}_{kl}^{\text{DUT}} \text{AF}_{kl}} e^{jk_0 r_{kl}}. \tag{5}$$

The variables are $\text{AF}_{kl}^{\text{DUT}}$, the complex-valued [8] antenna factor of the DUT antenna l toward the receive antenna k; AF_{kl}, the complex-valued antenna factor of the receive antenna k toward the transmit antenna l; r_{kl}, the separation distance between the $(k,l)^{\text{th}}$ probe-DUT pair; K_k, a calibration factor encapsulating physical constants and corrections for hardware losses and receiver equalization in the receive path of the k^{th} probe; and $k_0 = 2\pi/\lambda_0$, the free-space wavenumber.

The receive impulse response between the k^{th} probe and the 1^{st} probe, with the TEM phase relation $\omega\tau_{kl} = k_0 r_{kl}$, is

$$\begin{aligned} H_{kl}(\omega) &= \frac{h_{kl}(\omega)}{h_{1l}(\omega)} = \frac{\tau_{1l}}{\tau_{kl}} \frac{K_k}{K_1} \frac{\text{AF}_{1l}^{\text{DUT}}}{\text{AF}_{kl}^{\text{DUT}}} \frac{\text{AF}_{1l}}{\text{AF}_{kl}} e^{j\omega(\tau_{kl}-\tau_{1l})} \\ &\approx \overline{H}_{kl} e^{j\omega\Delta\tau_{kl}}. \end{aligned} \tag{6}$$

The complex transfer coefficient $\overline{H}_{kl}$ encapsulates all of the τ, K, and AF terms, which we assume to be frequency-invariant. Each $(k,l)^{\text{th}}$ probe-DUT response is therefore assumed to be a delay $\Delta\tau_{kl}$ in addition to the complex response coefficient $\overline{H}_{kl}$. These two parameters are also the basis of delay-and-sum beamforming [9], though we use it in this paper to excite a mode in the center of the test zone, not form a beam.

The response $H_{kl}(\omega)$ still fits nicely into a matrix equation. The dependence on frequency means (4) now has to be evaluated at each frequency:

$$\mathbf{V}(\omega) = \mathbf{H}(\omega)\mathbf{V}_1(\omega) + \mathbf{N}(\omega). \tag{7}$$

Moving forward, practical use of this expression will depend on determining the delays and response coefficients that characterize $H(\omega)$.

If a calibrated reference probe is available, we can similarly demultiplex the incident co-polarized electric field over frequency, $E_{1l}(\omega)$. The complex antenna factor needs to be known the direction toward each DUT, AF_{1l}, with any necessary impedance and level adjustments K_1:

$$E_{1l}(\omega) = K_1 \text{AF}_{1l}(\omega) V_{1l}(\omega). \tag{8}$$

C. Error Sources

The two-parameter delay and weight model in (6) and (7) requires that a $\Delta\tau_{kl}$ exists that leaves $\overline{H}_{kl}$ invariant across the detection bandwidth. This requires ideal reflectionless transverse electromagnetic (TEM) propagation like the ideal free field.

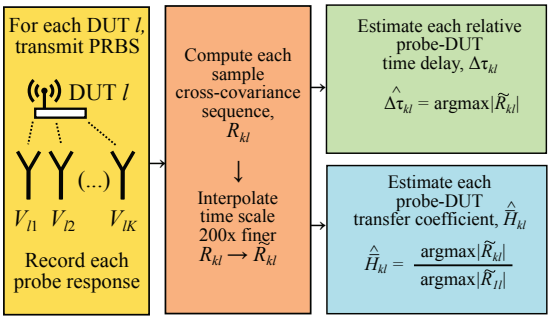


Fig. 3. Summary of the transfer matrix estimation process.

An anechoic chamber approximates this behavior, but other effects arise as error sources, including 1) Reflection interactions in the test zone, 2) Non-TEM near-field interactions for small r_{kl}/k_0, or 3) frequency variation in the ratios $\text{AF}_{kl}^{\text{DUT}}/\text{AF}_{1l}^{\text{DUT}}$, $\text{AF}_{kl}/\text{AF}_{1l}$, and K_k/K_1. The principal impact of these errors on our demultiplexing process is distorted crosstalk between output channels.

III. Probe Weighting

This section details detection and least-squares weighting estimation that we use to demonstrate implementing demultiplexing in this paper. The process is summarized in Fig. 3.

A. Acquiring Alignment Waveforms

All receive channels acquire and store a detection trace composed of M samples of complex IQ baseband voltage at sampling period T_s. Samples are time-synchronized and phase-locked. Sample acquisition occurs at $t[m] = mT_s$ for each $V_{kl}[m]$.

The procedure for collecting channel response calibration data is as follows. At each DUT for $l = \{1, 2, ..., L\}$,

1) Disable all DUTs.
2) Transmit a pseudo-random bit sequence from the l^{th} DUT
3) Simultaneously acquire the receive waveform on all receive channels, $V_{1l}[m], V_{2l}[m], \ldots, V_{Kl}[m]$.

The process results in one M-sample trace for every transmit-receive pair. The $V_{kl}[m]$ are the complete set of calibration data.

B. Probe Response Alignment

We need to estimate $\hat{\overline{H}}_{kl}$ and $\hat{\Delta\tau}_{kl}$ from the alignment data, which we will use to extrapolate the transfer matrix $\mathbf{H}$. The process proposed here was developed intuitively, and is therefore almost certainly suboptimal. Significant improvement may be achieved in further efforts in the future.

The sample cross-correlation sequence $R_{kl}[m]$ between the k^{th} and 1^{st} probe channels excited by the l^{th} DUT iswith respect to the reference antenna is

$$R_{kl}[n] = \sum_{m=1}^{M} (V_{1l}[m]^*) V_{kl}[m+n] \tag{9}$$

$$= \mathcal{F}^{-1}\left\{\mathcal{F}\left(V_{k1}[m]\right)^* \mathcal{F}\left(V_{kl}[m]\right)\right\}. \tag{10}$$

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao.
"Demultiplexing Spectrum-Sharing Field Sources with Distributed Field Probes."
Paper presented at 2016 IEEE International Symposium on Electromagnetic Compatibility and Signal Integrity, Ottawa, ON, Canada. July 25, 2016 - July 29, 2016.

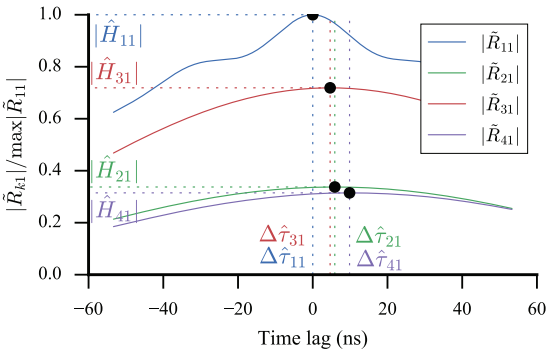


Fig. 4. Example probe channel parameter estimation values taken for DUT 1 (the Wi-Fi AP) from the interpolated cross-correlation sequence $\tilde{R}_{k1}$.

where $\mathcal{F}\{\cdot\}$ denotes discrete Fourier transform and indices n correspond with time lag $\tau[n]$.

The peak of $|R_{kl}[n]|$ is located near the delay we wish to estimate, $\Delta\tau_{kl}$. Digitization results in a quantization error bounded by $\epsilon_\tau = \pm T_s$, resulting in a frequency-domain phase progression error $\Delta\phi$. The worst case is $|\Delta\phi| = 180°$ at the acquisition band edges. To mitigate this we upsample $R_{kl}[n]$ to produce an oversampled $\tilde{R}_{kl}[n]$ on a new time grid $\tilde{\tau}[n] = T_i/T_s\tau[n]$. If $R_{kl}[n]$ has even symmetry about its peak (as under the idealized conditions of Section II), upsampling reduces the maximum phase error to $|\Delta\phi| = (T_i/T_s) \times 180°$. We fix the oversampling factor to $T_i/T_s = 1/1000$ in this work to keep $|\Delta\phi|$ well below 1° across the band.

Each delay pair is estimated by the correlation peak,

$$\Delta\hat{\tau}_{kl} = \underset{\tilde{\tau}}{\operatorname{argmax}} \left|\tilde{R}_{kl}[n]\right|. \tag{11}$$

This is the discrete naïve cross-correlation method of [10] (sampling period T_i). Further robustness to noise may be realized in the future by implementing the complete cross-correlation method from the same source.

We estimate the weight parameter at the corresponding magnitude peaks:

$$\hat{\overline{H}}_{kl} = \frac{\tilde{R}_{kl}\left[\underset{n}{\operatorname{argmax}}\left|\tilde{R}_{kl}[n]\right|\right]}{\tilde{R}_{1l}\left[\underset{n}{\operatorname{argmax}}\left|\tilde{R}_{1l}[n]\right|\right]}. \tag{12}$$

Figure 4 demonstrates the process with data from Section IV.

C. Demultiplexing to Separate DUTs

Now let all DUTs transmit to begin testing. Each measurement receiver channel acquires M_{test} samples. Each transfer function estimator behaves as

$$\hat{H}_{kl}(\omega_n) = \hat{\overline{H}}_{kl} \exp(-j\omega_n \hat{\Delta\tau}_{kl}). \tag{13}$$

The sampling rate can be different from that used to estimate the delay and weighting coefficients in (11)-(12).

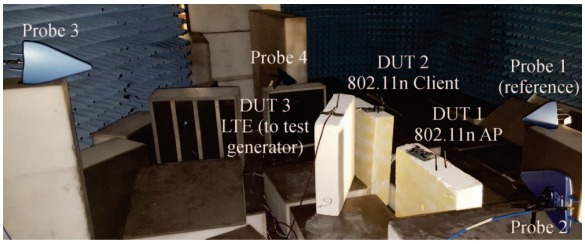


Fig. 5. Our demonstration test used $K = 4$ probes (connected to a 4-channel vector signal transceiver) to demultiplex $L = 3$ DUTs: a pair of Wi-Fi devices and a signal generator sending PRBS LTE.

TABLE I
DUT PARAMETERS

DUTs 1 and 2 (Wi-Fi)		
Protocol standard	802.11n	
Channel center frequency	2.442	GHz
Channel bandwidth	20	MHz
Transmit power setting	20 dBm	
Antenna type	3 cm monopole	
Transport	UDP	
Maximum transport unit	1500	bits
Data payload	Unknown	
DUT 3 (LTE Downlink Generator)		
Protocol standard	LTE FDD	
Center frequency	2.453	GHz
Bandwidth	20	MHz
Modulation type	64QAM	
Transmit power setting	19	dBm
Antenna type	3 cm monopole	
Data payload	PRBS	

The probe weighting matrix from (7) is related to the probe response matrix by inverse:

$$\begin{aligned}\hat{\mathbf{V}}_1(\omega_n) &= \hat{\mathbf{H}}^+(\omega_n)\mathbf{V}(\omega_n) \\ &= \hat{\mathbf{W}}(\omega_n)\mathbf{V}(\omega_n),\end{aligned} \tag{14}$$

where $(\cdot)^+(\omega_n)$ is matrix pseudo-inverse [11] computed at each ω_n. Since each H_{kl} is normalized to h_{1l}, each $\hat{H}_{1l} = 1$, and there is no information in the first row of $\overline{\mathbf{H}}$. Therefore, $\overline{H}_{1l} = 1$ has rank $K - 1$, and $K \geq L + 1$ receive channels are necessary to demultiplex L transmitters.

The time domain baseband can be recovered by IFFT for each row of $\hat{\mathbf{V}}_1(\omega_n)$.

IV. Application to a Coexistence Test

We demonstrate here the results of an initial experiment with this technique applied to LTE and Wi-Fi coexistence.

A. Test Setup

Figure 5 shows the semi-anechoic chamber configured for a Wi-Fi and long-term evolution (LTE) coexistence test.

Like Fig. 2, the DUTs are near the center of the chamber, and the probes are located around the perimeter. This topology provides line-of-sight between each pair of DUTs and maintains the standard practice of the test zone at the center of the chamber for the DUTs [12]. Separation between DUTs

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao.
"Demultiplexing Spectrum-Sharing Field Sources with Distributed Field Probes."
Paper presented at 2016 IEEE International Symposium on Electromagnetic Compatibility and Signal Integrity, Ottawa, ON, Canada. July 25, 2016 - July 29, 2016.

TABLE II
DETECTION SYSTEM PARAMETERS

Probes 1-4		
Antenna type	LPDA	
Manufacturer-specified antenna factor	$\lvert AF_{kl}\rvert = 46$	m^{-1}
2:1 VSWR bandwidth	$600 < f < 6000$	MHz
Acqusition		
Number of channels	4	
Center frequency	$f_c = 2.45$	GHz
Sampling rate	$1/T_s = 60$	MHz
Acquisition count	$M = 10^6$	
Sample size	32	bits
Cross-correlation interpolation factor	$T_s/T_i = 1000$	

is approximately 1 m, and each probe is approximately 1.5 m from the nearest DUT. The short $D = 3$ cm dipoles make these separations much greater than $2D^2/\lambda$.'

Detailed DUT and probe system parameters are in Tables I-II. The probe antennas are all the same make and model of log-periodic dipole antenna (LPDA). Their manufacturer specifies 5 dBi of gain at boresight and linear polarization. Each is oriented to include all DUTs within $\pm 45°$ of boresight at approximately vertical polarization (co-polarized with the DUTs).

Each Wi-Fi DUT is commercial development board hardware with the same make and model. One is configured as an 802.11n access point (AP), and the other is an 802.11n client. A software interface running on a laptop outside the chamber controls the devices during test. We use a bandwidth test mode to radiate by generating uplink (client-to-AP) or downlink (AP-to-client) traffic. These are actual 802.11n devices that use clear channel assessment to sense spectrum; at this close range we expect they will reliably wait to transmit until LTE vacates the channel.

A commercial radio frequency (RF) signal generator instrument excites a carrier modulated with LTE. The modulation data load is PRBS traffic on 10 resource blocks. We pulse the LTE signal at 10 ms period at 50% duty cycle to leave vacant time on the channel for the Wi-Fi system. The signal generator synthesizes the LTE without feedback from the channel, so only the Wi-Fi devices have the sensing information for opportunistic channel use in this test.

B. Calibration Self-Validation

We implemented the probe response alignment described in Section III, and applied the demultiplexing procedure to the original alignment data (baseband traces for each of the 3 DUTs radiating alone). This serves as a validation step, checking that power is dominated by the diagonal terms.

The results are in Fig. 6. Ideally, each l^{th} DUT waveform would appear only at the l^{th} demultiplexer output, and only noise at the other outputs. The crosstalk level averages 21 dB below the desired signal in its column. This figure of merit is the isolation, and needs to be high enough to help a measurement instrument or post-processing tool decode the information transmit by each DUT. The isolation contributes to the dynamic range of demultiplexed detection.

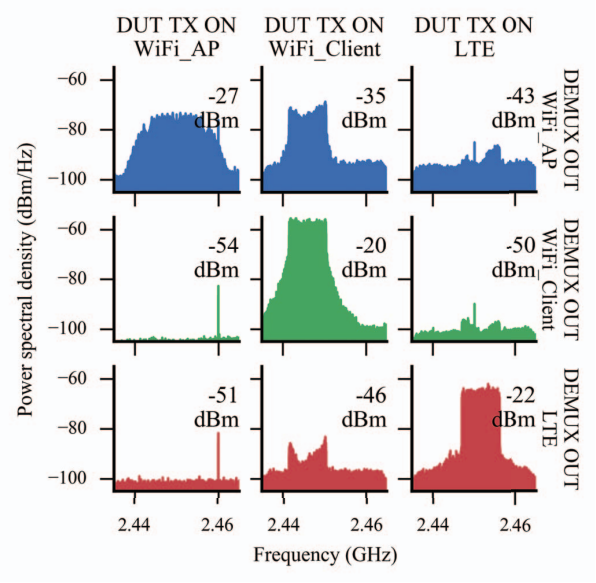


Fig. 6. Demultiplexed and residual cross-talk spectra, shown with total channel power. The average cross-talk level is 21 dB below the desired "through" channel (the diagonal plots in the grid).

TABLE III
CHANNEL UTILIZATION RATES DECOMPOSED BY DUT

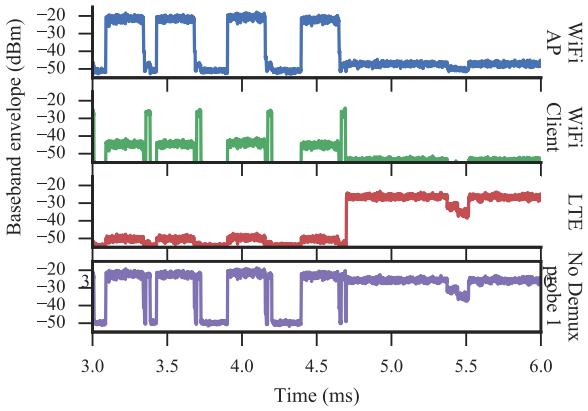
	Wi-Fi AP	Wi-Fi Client	LTE	Empty
802.11n downlink test	31%	3%	50%	16%
802.11n uplink test	4%	30%	50%	16%

C. Device-by-Device Channel Occupancy Tests

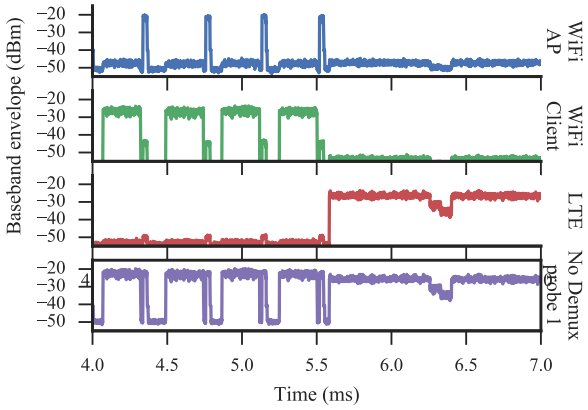
One use of the demultiplexed output in this application is to study the channel utilization by each DUT jointly over time (during simultaneous co-channel operation). We ran two scenarios to demonstrate this idea, shown in the power envelopes of Fig. 7: a Wi-Fi uplink test with LTE interference, and a Wi-Fi downlink test with LTE interference. For clarity, power levels are averaged in 100 µs bins. Each set of plots shows the three DUT output channels, plus the "no demux" raw data received at the reference antenna for comparison. We emphasize that no spectral filtering or time gating has been applied here to separate the channels.

The principal remaining uncertainty for estimating channel occupancy here is whether a waveform feature is produced by cross-talk the indicated DUT. Figure 6 suggests that the strongest residual cross-talk will be from the Wi-Fi AP channel to the Wi-Fi Client channel. This holds true in both scenarios of Fig. 7. Therefore, assuming that each DUT power envelope at this time-scale is fully "on" or "off," we can define an occupancy power threshold above the cross-talk level for each channel. Moving one-by-one by output channel, we can decompose total channel occupancy by device, even during co-channel transmission. These data are shown in Table III (averaged over one 10 ms on-off period that we applied to the

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao.
"Demultiplexing Spectrum-Sharing Field Sources with Distributed Field Probes."
Paper presented at 2016 IEEE International Symposium on Electromagnetic Compatibility and Signal Integrity, Ottawa, ON, Canada. July 25, 2016 - July 29, 2016.



(a) 802.11n downlink test with LTE downlink interference



(b) 802.11n uplink test with LTE downlink interference

Fig. 7. The demultiplexed signal envelopes of multiple spectrum-sharing DUTs in two spectrum-sharing coexistence test scenarios. Spatial demultiplexing divided the raw probe data (violet) into separate channels (red, green blue), the waveform components detected from each DUT.

LTE). In each case, the Wi-Fi sender channel occupancy is about 30%, in contrast with the receiving device at 3% to 4%. The LTE generator occupies the channel at a fixed 50% rate as configured on the instrument.

Channel occupancy for multiple co-channel DUTs demonstrated here could be difficult to determine (or define) from test data without demultiplexing. Are the dips in the raw data near 5.5 ms an instance of destructive interference from a collision event, or a feature of a single DUT? Answering this question would need *a priori* information about the constituent signals. However, the demultiplexed outputs show straightforwardly that the dip is a feature of the LTE waveform.

V. Conclusion

We demonstrated spatial demultiplexing of co-channel radiators for testing spectrum sharing and wireless coexistence. The implementation is a simple multiple-input multiple-output (MIMO) receiver. More study will be needed to characterize the processing impacts on the fidelity of the demultiplexed receive channels. Modulation and protocol analysis software may then be able to decode the waveforms for protocol-level insights into a spectrum-sharing scenario.

The data here was post-processed in the frequency domain. However, similar performance might be realized in real-time digital signal processing by developing of a suitable fractional delay filter. The demultiplexing technique could then be integrated aboard a multi-channel detection instrument by including "calibration" or "alignment" feature determine probe weightings. The most challenging practical aspect here may be controlling the DUTs to obtain a channel alignment signal; consumer devices are typically designed to transmit only as part of bi-directional communication with a larger network.

We were careful here to discuss the process and results in terms of detection. Developing this technique into a proper measurement tool will require significant effort in uncertainty analysis for each demultiplexer output channel. This will require characterizing and propagating uncertainty arising from sources such as reflections in the test zone, detector nonlinearity, and antenna responses that do not fit the time delay and coefficient model.

Acknowledgments

The authors are grateful for the valuable ideas and feedback provided by Paul Hale and Peter Jeavons (also with NIST).

References

[1] T. Wheeler, M. Clyburn, J. Rosenworcel, A. Pai, and M. O'Rielly, "Further Notice of Proposed Rulemaking," U.S. Federal Communications Commission, Washington, D.C., Tech. Rep. 15-47, 2015.

[2] "IEEE 802 LAN/MAN Standards Committee 802.22 WG on WRANs (Wireless Regional Area Networks)," Institute of Electrical and Electronics Engineers, New York, NY, Tech. Rep. 802.22.1, 2011.

[3] "Third Memorandum Order and Opinion," Federal Communications Commission, Washington, D.C., Tech. Rep. 12-36, 2012.

[4] M. M. Sohul, M. Yao, T. Yang, and J. H. Reed, "Spectrum Access System for the Citizen Broadband Radio Service," *IEEE Commun. Mag.*, no. 7, pp. 18–25, 2015.

[5] "Key Performance Indicators (KPI) for the Evolved Packet Core (EPC) v. 11," European Telecommunication Standards Institute, Valbonne, FR, Tech. Rep. TS 132 455, 2012.

[6] D. Williams, "Traveling waves and power waves: Building a solid foundation for microwave circuit theory," *IEEE Microw. Mag.*, vol. 14, no. 11, pp. 38–45, 2013.

[7] A. A. Smith, R. F. German, and J. B. Pate, "Calculation of Site Attenuation From Antenna Factors." *IEEE Trans. Electromagn. Compat.*, vol. EMC-24, no. 3, pp. 301–316, 1982.

[8] S. Ishigami, H. Iida, and T. Iwasaki, "Measurements of complex antenna factor by the near-field 3-antenna method," *IEEE Trans. Electromagn. Compat.*, vol. 38, no. 3, pp. 424–432, 1996.

[9] D. Johnson, *Array Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 1993.

[10] K. Coakley and P. Hale, "Alignment of noisy signals," *IEEE Trans. Instrum. Meas.*, vol. 50, no. 1, pp. 141–149, 2001.

[11] R. Penrose, "A Generalized Inverse for Matrices," *Math. Proc. Cambridge Philos. Soc.*, vol. 51, no. 7, pp. 406–413, 1955.

[12] A. Simmons and W. Emerson, "An Anechoic Chamber Making Use of a New Broadband Absorbing Material," in *IRE Int. Conv. Rec.*, vol. 1, 1958, pp. 34–41.

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao.
"Demultiplexing Spectrum-Sharing Field Sources with Distributed Field Probes."
Paper presented at 2016 IEEE International Symposium on Electromagnetic Compatibility and Signal Integrity, Ottawa, ON, Canada. July 25, 2016 - July 29, 2016.

Developing Models for Type-N Coaxial VNA Calibration Kits within the NIST Microwave Uncertainty Framework*

Jeffrey A. Jargon, Dylan F. Williams, and Paul D. Hale

National Institute of Standards and Technology, 325 Broadway, M/S 672.03, Boulder, CO 80305 USA
Email: jeffrey.jargon@nist.gov, Tel: +1.303.497.4961

***Abstract* — We developed models for Type-N coaxial vector network analyzer (VNA) calibration kits within the NIST Microwave Uncertainty Framework. First, we created physical models of commercially-available standards that support multiline thru-reflect-line (TRL) and open-short-load-thru (OSLT) calibrations, and included error mechanisms in each of the standards' constituent parameters that were utilized in the NIST Microwave Uncertainty Framework to propagate uncertainties. Next, we created a measurement-based model of a commercial electronic calibration unit (ECU) by characterizing the scattering parameters of its internal states with a multiline TRL calibration. Finally, we calibrated a network analyzer using the three calibration methods, and compared measurements, including uncertainties, made on a number of verification devices. We show that the three calibrations agreed to within their respective uncertainties.**

***Index Terms* — calibration, coaxial, electronic calibration unit, physical models, uncertainty, vector network analyzer, verification.**

I. Introduction

The multiline, thru-reflect-line (TRL) calibration [1] is perhaps the most fundamental and accurate vector network analyzer (VNA) calibration for coaxial circuits. Multiline TRL calibrations measure the propagation constant of the line standards so that the characteristic impedance can be transformed to a selected reference impedance, and offer high bandwidth and accuracy through the use of multiple transmission-line standards. However, a set of coaxial lines, some relatively long, is required to obtain a wide-band measurement. Coaxial airlines also require considerable care to ensure good connections without damaging the standards. Furthermore, a set of lines can be costly, and measurements are time-consuming.

Other types of VNA calibrations make use of compact, lumped-element standards, the most common being open-short-load-thru (OSLT) and line-reflect-match (LRM) methods [2]. They provide calibration procedures that are easier to perform, oftentimes at the cost of lower accuracy.

Over the years, electronic calibration units (ECUs) have become a viable alternative to the aforementioned methods. First proposed in 1993 [3], these units provide the advantage of requiring only one connection and are capable of rapidly switching among a large variety of reflection coefficients and low-loss transmission coefficients. Recently, newer commercial units have been shown to be stable enough to be used in place of mechanical verification artifacts [4, 5].

In this paper, we utilize the NIST Microwave Uncertainty Framework [6-10] to develop physical models of commercially available Type-N multiline TRL and OSLT coaxial calibrations kits, and then use the multiline TRL calibration to create a traceable measurement-based model of an ECU. The NIST Microwave Uncertainty Framework utilizes parallel sensitivity and Monte-Carlo analyses, and enables us to capture and propagate the significant *S*-parameter measurement uncertainties and statistical correlations between them [11]. By identifying and modeling the physical error mechanisms in the calibration standards, we can determine the statistical correlations between both the scattering parameters at a single frequency and uncertainties at different frequencies. These uncertainties can then be propagated to measurements of the devices under test. In the following sections, we describe our methodology in further detail, and compare measurements and uncertainties made on a number of verification devices.

II. Model Development

We began by modeling the multiline TRL calibration standards (an offset short, and five airlines of varying lengths) for purposes of determining uncertainties. Table I lists the line lengths and associated uncertainties for the multiline TRL standards, and Table II lists the other sources of uncertainty for the standards. Our values and distributions of the uncertainties come from a variety of sources, including manufacturers' specifications and an IEEE standard [12].

The NIST Microwave Uncertainty Framework was employed to construct models for the calibration standards. The airline and offset-short standards were modeled with closed-form expressions for coaxial lines of finite metal conductivity [13]. The framework was also used for automatically propagating the uncertainties to the calibrated verification devices in conjunction with the calibration engine, StatistiCAL™ [14, 15], which utilizes a "mix-and-match" philosophy to VNA calibrations.

Next, the OSLT standards were modeled with the values and uncertainties listed in Tables II and III. We modeled the load standard as a simple 50 Ohm resistor after observing that the magnitudes of the measured reflection coefficients for both the male and female connectors were less than -30 dB at most frequencies. The offset lengths of the open and short standards

Hale, Paul; Jargon, Jeffrey; Williams, Dylan.
"Developing Models for Type N Coaxial VNA Calibration Kits within the NIST Microwave Uncertainty Framework."
Paper presented at 87th ARFTG, San Francisco, CA, United States. May 27, 2016 - May 27, 2016.

were estimated from the respective phase delays measured with the multiline TRL calibration.

Finally, we created a measurement-based model of our ECU by characterizing the *S*-parameters of its internal states with a multiline TRL calibration. The added complication here was that our airlines had male connectors on both ports, so a male-to-male adapter was required to measure the insertable ECU. We then de-embedded the calibrated adapter to properly characterize the ECU.

Table I. Lengths and uncertainties of the Type-N TRL standards.

Line Designation	**Length (mm) ± Uncertainty (Distribution)**
Airline 1	36.966 ± 0.005 (Rectangular)
Airline 2	43.472 ± 0.005 (Rectangular)
Airline 3	49.959 ± 0.005 (Rectangular)
Airline 4	56.442 ± 0.005 (Rectangular)
Airline 5	99.904 ± 0.005 (Rectangular)
Offset Short	18.955 ± 0.005 (Rectangular)

Table II. Physical error mechanisms of the Type-N standards.

Mechanism (units)	**Value ± Uncertainty (Distribution)**
Inner Cond. Diameter (mm)	3.04 ± 0.0026 (Rectangular)
Outer Cond. Diameter (mm)	7.0 ± 0.0051 (Rectangular)
Pin Diameter (mm)	1.651 ± 0.0127 (Rectangular)
Pin Depth (mm)	0.051 ± 0.051 (Rectangular)
Metal Conductivity (S/m)	$7.9\times10^6 \pm 4\times10^6$ (Rectangular)
Relative Dielectric Constant	1.000535 ± 0
Dielectric Loss Tangent	0 ± 0

Table III. Physical error mechanisms of the Type-N OSLT standards.

Mechanism (units)	**Value ± Uncertainty (Distribution)**
Male Open Offset Length (mm)	6.504 ± 0.005 (Rectangular)
Female Open Offset Length (mm)	1.944 ± 0.005 (Rectangular)
Open Conductance (1/Ω)	0 ± 0
Open Capacitance (pF)	0 ± 0
Male Short Offset Length (mm)	5.321 ± 0.005 (Rectangular)
Female Short Offset Length (mm)	0.000 ± 0.005 (Rectangular)
Short Resistance (Ω)	0 ± 0
Short Inductance (nH)	0 ± 0
Load Resistance (Ω)	50.0 ± 0.1 (Rectangular)
Load Inductance (nH)	0.0 ± 0.1 (Rectangular)

III. Measurement Comparison

Once the multiline TRL, OSLT, and electronic calibration standards were defined, we used the three sets of standards to calibrate the measurements of verification devices for comparison purposes. Once again, for the TRL calibration, the calibrated adapter was required for measuring the insertable devices. Figures 1-5 show calibrated S-parameters and corresponding 95 % confidence bounds calculated from the sensitivity analysis performed in the NIST Uncertainty Framework for a 20 dB attenuator, a 50 dB attenuator, an airline, and a Beatty line. Dashed curves in the figures correspond to confidence bounds.

In each of the figures, the three calibrated measurements agreed to within their respective uncertainties at most frequencies, although the OSLT-calibrated measurements were visibly noisier. The uncertainties of the OSLT-calibrated measurements were also larger in general. For instance, the mean confidence intervals for $|S_{21}|$ of the 20-dB attenuator were approximately ± 0.07 dB with multiline TRL, ± 0.15 dB with OSLT, and ± 0.07 dB with the ECU.

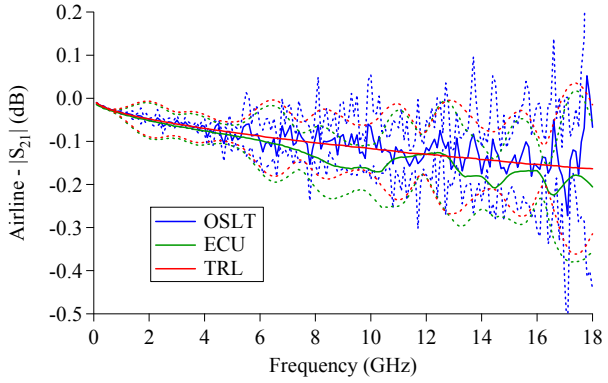


Fig. 1. Comparing measurements and 95% confidence intervals of the airline's transmission coefficients.

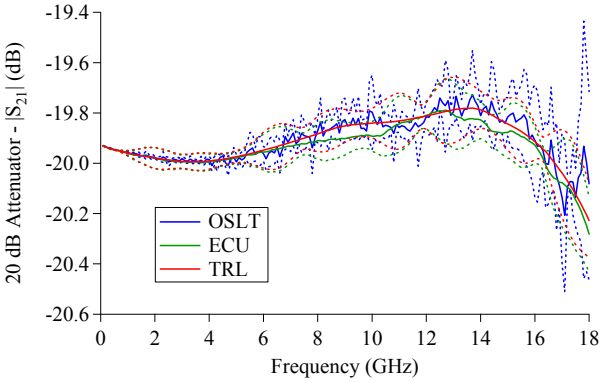


Fig. 2. Comparing measurements and 95% confidence intervals of the 20 dB attenuator's transmission coefficients.

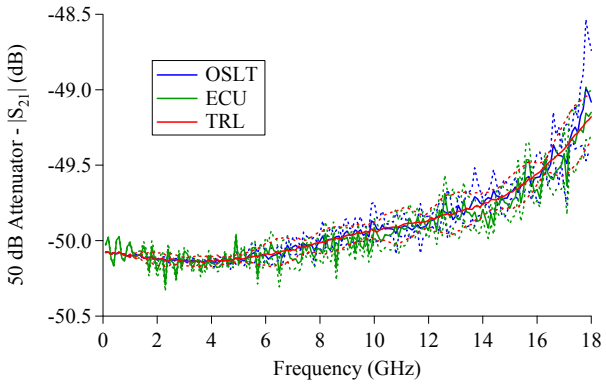


Fig. 3. Comparing measurements and 95% confidence intervals of the 50 dB attenuator's transmission coefficients.

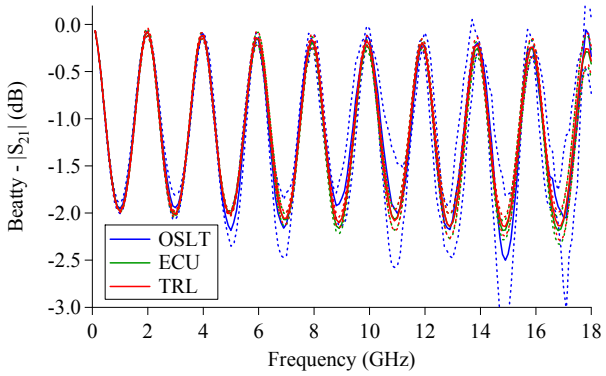


Fig. 4. Comparing measurements and 95% confidence intervals of the Beatty line's transmission coefficients.

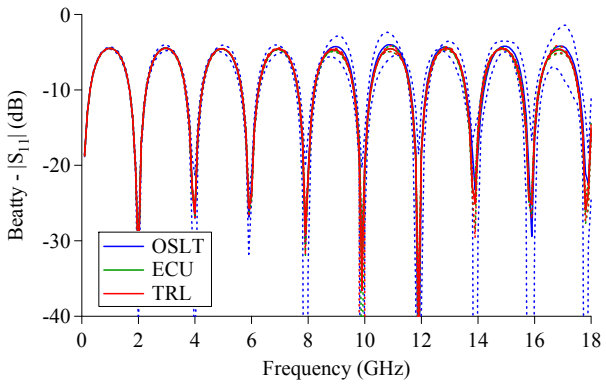


Fig. 5. Comparing measurements and 95% confidence intervals of the Beatty line's reflection coefficients.

IV. Conclusions

We have developed physical and measurement-based models of Type N coaxial calibration kits for vector network analyzers that support multiline TRL, OSLT, and electronic calibrations within the NIST Microwave Uncertainty Framework. The verification devices measured with the three calibration approaches agree to within their respective uncertainties. Although other sources of uncertainty may be included in a final uncertainty analysis, we believe these minor additions will not significantly affect the overall uncertainties.

The principle advantage of characterizing an ECU and providing uncertainties is that the unit can be used as a working set of standards that requires only a single connection to calibrate the VNA, saving both time and wear-and-tear on the VNA test ports as well as the TRL and OSLT standards.

Acknowledgement

*This work was supported by the U.S. Department of Commerce, and is not subject to U.S. copyright. The authors thank Ronald Ginley for the use of his multiline TRL calibration kit, and David Walker for the use of his verification kit.

References

[1] R. B. Marks, "A multiline method of network analyzer calibration," *IEEE Trans. Microwave Theory Tech.*, vol. 39, no. 7, pp. 1205–1215, July 1991.

[2] D. K. Rytting, "Network analyzer error models and calibration methods," *52nd ARFTG Conference*, Short Course on Computer-Aided RF and Microwave Testing and Design, Dec. 1998.

[3] V. Adamian, "A novel procedure for network analyzer calibration and verification," *41st ARFTG Conference*, Spring 1993.

[4] D. F. Williams, A. Lewandowski, D. LeGolvan and R. Ginley, "Electronic vector-network-analyzer verification," *IEEE Microwave Magazine*, pp. 118-123, Oct. 2009.

[5] D. F. Williams, A. Lewandowski, D. LeGolvan, R. Ginley, C. M. Wang and J. Splett, "Use of electronic calibration units for vector-network-analyzer verification," *74th ARFTG Conference*, Dec. 2009.

[6] D. F. Williams, NIST Microwave Uncertainty Framework, Beta Version, http://www.nist.gov/ctl/rf-technology/related-software.cfm, 2015.

[7] J. A. Jargon, D. F. Williams, T. M. Wallis, D. X. LeGolvan, and P. D. Hale, "Establishing traceability of an electronic calibration unit using the NIST Microwave Uncertainty Framework," *79th ARFTG Microwave Measurement Conference*, Montreal, CANADA, Jun. 2012.

[8] J. A. Jargon, U. Arz, and D. F. Williams, "Characterizing WR-8 waveguide-to-CPW probes using two methods implemented within the NIST Uncertainty Framework," *80th ARFTG Microwave Measurement Conference*, San Diego, CA, Nov. 2012.

[9] J. A. Jargon, D. F. Williams, P. D. Hale, and M. D. Janezic, "Characterizing a noninsertable directional device using the NIST Uncertainty Framework," *83rd ARFTG Microwave Measurement Conference*, Tampa Bay, FL, Jun. 2014.

[10] J. A. Jargon, C. H. Cho, D. F. Williams, and P. D. Hale, "Physical models for 2.4 mm and 3.5 mm coaxial VNA calibration kits developed within the NIST Microwave Uncertainty Framework," *85th ARFTG Microwave Measurement Conference*, Phoenix, AZ, May 2015.

[11] A. Lewandowski, D. F. Williams, P. D. Hale, C. M. Wang, and A. Dienstfrey, "Covariance-matrix-based vector-network-analyzer uncertainty analysis for time-and frequency-domain measurements," *IEEE Trans. Microwave Theory Tech*., vol. 58, no. 7, pp. 1877-1886, July 2010.

[12] IEEE Standard 287-2007 - Standard for precision coaxial connectors (DC to 110 GHz).

[13] A. Lewandowski, "Multi-frequency approach to vector-network analyzer scattering-parameter measurements," Ph.D. Thesis, Warsaw University of Technology, 2010.

[14] D. F. Williams, StatistiCAL VNA Calibration Software Package, http://www.nist.gov/ctl/rf-technology/related-software.cfm, 2015.

[15] D.F. Williams, C.M. Wang, and U. Arz, "An optimal vector-network-analyzer calibration algorithm," *IEEE Trans. Microwave Theory and Tech*., vol. 51, no. 12, pp. 2391-2401, Dec. 2003.

A Clustering-Based Device-to-Device Communication to Support Diverse Applications

Amirshahram Hematian*, Wei Yu*, Chao Lu*, David Griffith§, Nada Golmie§
*Towson Univ., USA
ahemat1@students.towson.edu,{wyu,clu}@towson.edu
§National Institute of Standards and Technology, USA
{david.griffith,nada.golmie}@nist.gov

ABSTRACT

In this paper, we address the issue of how to leverage Wi-Fi Direct (as an outband solution) to enable the Device-to-Device (D2D) communication that can offload massive data traffic from the LTE (Long Term Evolution)-based cellular network and support other applications. Particularly, we develop a clustering-based scheme that automatically finds the best candidates to remain connected to the LTE network while the rest of the devices can be disconnected directly from the LTE-based cellular network. By doing so, we can reduce the signal interference, increase the average throughput and spectral efficiency of the network, and also reduce unnecessary data traffic that can be transmitted locally by D2D communications instead of going through the LTE-based cellular network. Devices in established clusters can indirectly communicate with the LTE network via the cluster head, which can be dynamically selected and remains connected to the LTE network directly. Using the real-world cellular data collected from a public database related to the deployment of LTE networks, we show the effectiveness of our proposed scheme in traffic offloading in the cellular network. We also discuss how to use our developed techniques to support Internet-of-Things (IoT) applications such as smart grid communications.

CCS Concepts

•Networks → Wireless access points, base stations and infrastructure; Network performance evaluation; Mesh networks; Mobile networks; Network mobility; Peer-to-peer networks; Wireless access networks; Network architectures; Network algorithms;

Keywords

Device-to-Device Communication; Clustering; Traffic Offloading; Applications

RACS '16 , October 11 - 14, 2016, Odense, Denmark

ACM 978-1-4503-4455-5/16/10…$15.00
DOI: http://dx.doi.org/10.1145/2987386.2987391

1. INTRODUCTION

Generally speaking, D2D communication refers to a technology that empowers devices (User Equipments (UE), smart meters, sensors, etc.) to send and receive data directly without going through the core wireless network infrastructure via base stations (or access points) [3]. The D2D communication can be either inband or outband. Inband refers to the case where the D2D communication utilizes the same spectrum that devices use to communicate to the base station, while the outband D2D communication refers to the case where the spectrum used for D2D communication does not coincide with the one used by base station communication. Wi-Fi Direct [1] is one known outband D2D communication technology, which operates at Industrial, Scientific and Medical (ISM) radio bands. Notice that the question of how to effectively share spectrum resources and overcome interferences from devices (UE, smart meters, sensors, base stations, etc.) remains a challenging issue in inband D2D communication.

In this paper, we focus on the investigation of the outband D2D communication technique and demonstrate its feasibility by offloading traffic in cellular networks and supporting other applications. Our paramount contributions are listed as following.

First, we outline our developed clustering-based scheme to enable the D2D communication for devices that are close to each other. In this way, massive traffic can be transmitted via D2D communication in local areas and traffic transmitted via the core cellular network can be reduced. The main idea behind the clustering scheme is to create small Wi-Fi networks for communications between devices in local areas while these devices remain connected indirectly to the cellular network via the head of the clusters. Within each cluster, the cluster head directly connects to the LTE network and is dynamically selected based on various factors (quality of reception, bandwidth, etc.).

Second, having implemented Wi-Fi Direct as an outband solution for enabling D2D communication within a simulated LTE network, we demonstrate the effectiveness of our scheme via a case study: offloading traffic in the cellular network as an example. We leverage the Vienna LTE-A system level simulator [12][1] and have collected real-world deploy-

[1]Certain commercial equipment, instruments, or materials are identified in this chapter in order to specify the experimental procedure adequately. Such identification is not

Golmie, Nada; Griffith, David; Hematian, Amirshahram; Lu, Chao; Yu, Wei.
"A Clustering-Based Device-to-Device Communication to Support Diverse Applications."
Paper presented at International Conference on Research in Adaptive and Convergent Systems (RACS '16), Odense, Denmark. October 11, 2016 - October 14, 2016.

ment information of base stations via OpenCellID website [5]) to demonstrate the effectiveness of our proposed scheme. The experimental data shows that our developed scheme can be used to significantly improve the network performance by offloading traffic in the cellular network. Our proposed scheme is generic, and can be applied to support other types of wireless networks and other applications. We discuss how to use our developed scheme to extend the performance improvement to smart grid communications.

The remainder of the paper is organized as follows: We introduce the background and related work in Section 2. In Section 3, we present our schedule in detail. In Section 4, we show the experimental results to validate the effectiveness of our proposed scheme in offloading traffic in the cellular network. We conclude the paper in Section 5.

2. BACKGROUND AND RELATED WORK

In this section, we give the background and related work of D2D communication and Wi-Fi Direct.

2.1 D2D Communication

D2D communication in LTE networks refers to directly routing data traffic among mobile User Equipment (UEs) when UEs are close to each other. By doing so, network performance measured by energy efficiency, throughput, delay, as well as spectrum efficiency can be improved. Such a communication technique has been considered as a viable solution to deploy the cellular network infrastructure in rural areas, support public safety applications when the network infrastructure breaks down during a disaster, and support the monitoring and control of numerous applications (smart grid, etc.). For example, in a public safety application, when a disaster strikes, the mobile devices of emergent response personnel can directly communicate with each other via D2D communication so that the data traffic on the wireless network raised by growing traffic demands can be offloaded, or in the event that wireless infrastructure is not even available.

There have been a number of research efforts on developing D2D communication techniques to improve the spectral efficiency of wireless networks [9, 4]. Sharing a widely used spectrum in the cellular network (also called inband D2D communication) can be problematic because of the interference between the communication spectrum used for both D2D communication and cellular communication. Notice that the inband D2D communication in cellular networks requires additional efforts and changes to the components of cellular networks. Also, how to efficiently manage the shared spectrum allocated for D2D communication remains an open issue. In contrast, because the cellular network uses a different spectrum from the D2D communication, interference will not occur.

With respect to outband-based D2D communication techniques, unlicensed spectrum is commonly used for supporting the communication of D2D links. In these techniques, while no interference issue exists between D2D communication and cellular communication, mobile devices are normally required to have an extra wireless communication interface to support the different wireless communication implementations (Wi-Fi Direct [6, 2], ZigBee [13], Bluetooth [10], etc.) through an unlicensed spectrum.

intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

2.2 Wi-Fi Direct

Wi-Fi Direct is a Wi-Fi standard, which tends to enable wireless devices to easily connect with each other without the support of wireless access points [1]. Wi-Fi Direct can negotiate the link with a Wi-Fi management system, which assigns each device a wireless Access Point (AP) (known as Software Access Point (Soft AP)). By using Soft AP, a Wi-Fi Direct-enabled device becomes multi-role, hosting small networks and clients of other Wi-Fi networks, and supports multi-hop communication. By using multi-hop technology in Wi-Fi Direct-enabled networks, the coverage of a small local network can be easily extended by adding another device to the network. The throughput of Wi-Fi Direct-enabled networks can be enhanced due to shorter communication hops required to send and receive data. In addition, the battery life of devices will be extended due to low power for data transmission between nearby devices even though the destination of the data is far away.

In our proposed clustering-based scheme in Section 3, we assume that all UEs support Wi-Fi Direct and every cluster is a Wi-Fi Direct network with a mesh-based topology to support multi-hopping data transmission. Also, after clustering and selecting the head of the cluster, every cluster is connected to the cellular network via the head of the cluster.

3. CLUSTERING-BASED D2D SCHEME

In this section, we first give an overview of our clustering-based D2D scheme and then present the detailed design and workflow of our proposed scheme.

3.1 Overview

To reduce the traffic overload in the network and improve the bandwidth efficiency, D2D communication is an effective solution to offload traffic from the cellular network and utilize the network resources more efficiently. Recall that in this paper, we consider the use of Wi-Fi Direct as an outbound solution to provide D2D communication, which requires fewer changes in the LTE-based cellular network. By using Wi-Fi Direct and transmitting data locally among nearby mobile users, we can offload data traffic from the cellular network and prevent interference to cellular users as the Wi-Fi Direct and cellular network operate in different frequency bands. By doing so, the available bandwidth for each mobile user can be increased by offloading the local data traffic from the global cellular network. The local data can be transmitted in a multi-hop manner in Wi-Fi Direct networks, where each UE requires a low transmission power to send and receive data, via a multi-hop fashion even when the source and destination of the data are significantly far away in the same Wi-Fi Direct network.

The D2D communication in a mesh-based Wi-Fi Direct network (denoted as cluster) not only provides improved coverage because of multi-hop data forwarding, but also enhances the throughput of the network due to shorter hops and extend battery life because many users and meters are located nearby each other. Nonetheless, if any UE requires data to be transmitted globally, the head of the cluster in the Wi-Fi Direct network can provide the communication to the cellular network indirectly based on the Wi-Fi direct network that it is already connected to.

Golmie, Nada; Griffith, David; Hematian, Amirshahram; Lu, Chao; Yu, Wei.
"A Clustering-Based Device-to-Device Communication to Support Diverse Applications."
Paper presented at International Conference on Research in Adaptive and Convergent Systems (RACS '16), Odense, Denmark. October 11, 2016 - October 14, 2016.

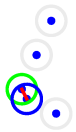


Figure 1: Discovering Neighbours (Discoverer UE in Blue, Neighbours in Range in Green, Out-of-Range Neighbours in Grey)

To enable the D2D communication in the cellular network, we propose a clustering-based scheme, which creates clusters for small Wi-Fi Direct networks. All the UEs within each cluster can directly and indirectly communicate with each other and transfer data without necessarily going through the core cellular network. In each cluster, the node that remains directly connected to the cellular network is the head of the cluster, which is dynamically selected based on the quality of reception, bandwidth, and other factors. In the following, we explain how the clustering-based scheme gathers the statistical data, creates the clusters, merges the clusters, and then selects the heads for clusters.

3.2 Cluster-Based D2D Communication

The main idea behind the proposed clustering method is to enable the local D2D communication so that each UE can communicate with other UEs based on established small mesh-based Wi-Fi networks dedicated for D2D communication among UEs while these UEs remain connected indirectly to the LTE network via the head of the cluster. In each cluster, the node that remains directly connected to the LTE network is defined as the head of the cluster, which is dynamically selected based on the quality of reception, bandwidth and other metrics.

Our developed clustering scheme consists of the following steps:

Step 1. Discovering devices and finding the nearest neighbours: The first step gives UEs that have the Wi-Fi Direct capability a way to discover each other, as well as services that they support. For example, a UE with Wi-Fi Direct capability can see all compatible devices in a given area and then narrow down the list of devices that enable the Wi-Fi Direct D2D communication. Mobile users can decide whether to join the clustering service or not by turning the service on and off on their UE. In the simulation tool that we used [12], as there was no Wi-Fi Direct feature provided, we have implemented a module to define the Wi-Fi range for each UE and perform the discovery. We assume that all discovered UEs are willing to join the D2D-based communication based clusters created. To simulate the Wi-Fi discovery process and obtain a list of available neighbours for each UE, we create a discovery list that can store up to 256 neighbours for each UE. As shown in Figure 1, every neighbour is listed in each UE record only when it is located within the Wi-Fi range of the current UE. The discovery list is generated based on the distance between the current UE and their neighbours, considering the maximum Wi-Fi range given prior to the simulation.

Step 2. Creating clusters for D2D communication: Once the UE that has not joined a cluster has generated a list of discovered nearby devices using the Wi-Fi Direct discovery service, the clustering process begins. Every UE starts from the beginning of the discovery list and creates a new cluster of its own if the UE and its neighbours in the discovery list have not already been assigned to one cluster. Otherwise, every UE will join all the clusters it finds in the discovery list that its neighbours have already connected to. For example, if UE_1 has three UEs (UE_2, UE_3 and UE_4) in its discovery list and none of them has yet joined any cluster, the UE_1 will create a new cluster and allow its three neighbours to join. On the other hand, if any of the three UEs (say UE_3) has already connected to a cluster, UE_1 will not create a new cluster, but instead will join all of the same clusters as its neighbors. By doing so, the UE can join multiple clusters at a same time. Then, later in the merging process, all clusters that are sharing UEs will be merged and become one cluster.

Step 3. Merging clusters: When a UE is joined to more than one cluster, all those clusters could be merged into one cluster. By broadcasting the cluster information, it is possible to let the head of each cluster know that a new UE has joined multiple clusters. Then, the head of clusters can proceed another round of the head election process and one of them becomes the head of the newly formed merged cluster. Figure 2 illustrates an example of merged clusters in regions. As we can see from this example, the Wi-Fi coverage of every UE is portrayed in green circles. Wherever these circles intersect each other, a cluster is created and starts to grow until there are no more circles close enough to expand the cluster any more. In one case, several small clusters could be created inside another cluster and those clusters remain unmerged as there is no UE in between to connect these clusters.

The clustering-based D2D communication in the cellular network needs to be periodically updated due to the fact that mobile users could change their locations dynamically. As a result, the selection process of the new head of cluster needs to be conducted periodically. Since UEs are mobile, they may lose connection from one cluster and join another cluster. Consequently, the clustering and role changing operations must be performed continuously for each cluster as well. On one hand, two or more clusters may be merged because of those clusters are close enough. On the other hand, one cluster could be broken into two or more clusters because UEs may go far enough as a group (or alone) to lose connection from the original cluster and then create their own clusters. In addition, even in some cases there may be a few small clusters (inner clusters) created inside a larger cluster (outer cluster) as the inner clusters cannot reach any of UEs associated with the outer cluster. In Figure 3, we show a black arrow pointing to an inner cluster that cannot merge with the outer cluster unless some of UEs from either inner or outer clusters move toward the UEs from another cluster and make a D2D connection available. Once this bridge connection is established, the inner cluster will merge into the outer cluster.

Step 4. Indirect LTE connection: To reduce the data traffic load of the cellular network, the local data traffic associated with UEs can be transmitted via D2D communications within the clusters. Since we do not want to completely remove the UEs from the LTE network, we select one of the UEs in each cluster that is denoted as the head of the cluster, which remains connected to the LTE network. In each D2D cluster, only the head of the cluster remains connected to both the cluster and cellular network directly. Every other node within the cluster can indirectly connect to the cellular

Golmie, Nada; Griffith, David; Hematian, Amirshahram; Lu, Chao; Yu, Wei.
"A Clustering-Based Device-to-Device Communication to Support Diverse Applications."
Paper presented at International Conference on Research in Adaptive and Convergent Systems (RACS '16), Odense, Denmark. October 11, 2016 - October 14, 2016.

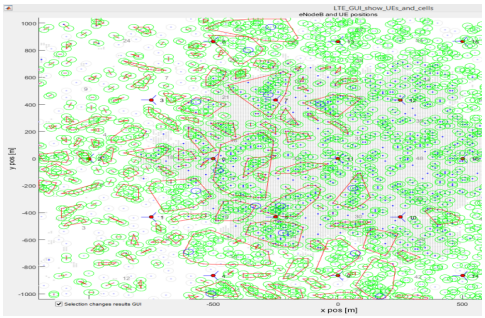


Figure 2: Merged Clusters

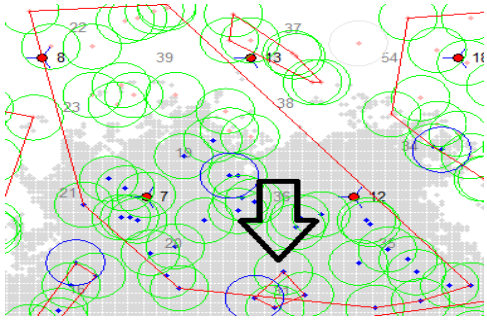


Figure 3: Out of Range Clusters Not Merged

network through the head of the cluster, which functions as a relay node. Notice that the communication between the cluster member and the cluster head may be a direct link between them or through multiple one-hop links.

Step 5. Changing cluster head: At the beginning of the cluster head selection process, each UE considers itself the head of the cluster. Nonetheless, there might be another UE inside the same cluster that has better connecting characteristics. In this case, the head of the cluster will be changed to the latter UE. The re-election process of cluster heads is a periodic process that gathers the statistical information about all UEs connected to a single cluster. At the end, the best candidate will be selected to be the next head of the cluster. To make the selection, the statistical information about signal quality, coverage, and other factors can be obtained via the network. The re-election process of cluster heads can be controlled by the service provider via the cellular network, or be autonomously completed inside the cluster by the head of the cluster. The maximum average throughput and maximum average spectral efficiency are two major factors that the re-election process is based on.

Our proposed scheme produces additional clusters and later merges them due to the fact that each UE only sees the other UEs under its own Wi-Fi coverage at the beginning. By joining other clusters, the heads of the clusters will find out that they are now connected to each other and can merge down and release one of the heads (known as "chaining phenomenon", in particular with the single-linkage clustering [7]). Notice that the complexity of our proposed clustering scheme is $O(n^3)$, where n is the number of UEs.

Due to the fact that the proposed scheme is implemented as part of an LTE network simulator, it sees the UEs from the network perspective. The complexity of this method of implementation is much higher in comparison with running the proposed scheme on each UE individually, where each UE only sees the nearby UEs, making the complexity much lower and computation more efficient. In other words, instead of doing the clustering on the LTE network side, we can let the UEs carry out the clustering themselves in parallel and then let the LTE network know the structure of clusters. This information can be later provided for the LTE network by the heads of clusters for relaying data between LTE and Wi-Fi direct networks.

From the network operation perspective, there are two types of D2D communications: controlled and autonomous. The former is under the supervision of the cellular network, and the latter is totally independent. Although we use the controlled type in our implementation, in order to reduce the complexity and share the computation power required between the UEs, our proposed clustering scheme can be implemented with the autonomous type of D2D clustering as well, which is independent from the cellular network and uses the UEs computation power to reduce the computation power needed, and let the UEs carry out the clustering themselves independently.

3.3 Supporting Other Applications

Our proposed scheme is generic and can support diverse applications. Particularly, IoT has attracted significant attention and can be considered to a networking infrastructure which can connect massive amounts of physical objects belonging to numerous critical infrastructure systems and others. For example, in the smart grid, the geographically distributed meters, sensors, actuators and controllers are tightly integrated through communication networks and computational cores, enabling the secured and efficient operations of the power grid [8, 14, 15]. We can leverage our developed clustering-based scheme to establish mesh-based Wi-Fi Direct networks to connect smart meters (sensors) in the smart grid. Then, the head of a cluster will provide indirect communication links between meters (sensors) in the cluster and the operation center, either through the cellular network or by connecting to another nearby cluster.

We have collected real-world data for both eNodeBs and UEs from OpenCellID (similar to the Case Study in Section 4) and smart meters (as UEs) from Google. For the eNodeBs, we need to define a ROI to obtain the first 1000 tower locations and for the smart meters, we need to generate complete postal addresses to obtain the locations of the smart meters one by one. To use real-world data for smart meters and bring them to the simulation, we used Google geo-coding APIs and generated HTML "get" requests in Matlab simulation code to gather the locations of the smart meters. Since the exact locations of smart meters that are actually installed are not available to the public, we use the locations of the addresses that we gather from Google geo-coding APIs and assume they already installed a smart meter. In the implementation, those coordinates of addresses are converted to a two dimensional map using built-in functions in Matlab that receive latitude and lon-

Golmie, Nada; Griffith, David; Hematian, Amirshahram; Lu, Chao; Yu, Wei.
"A Clustering-Based Device-to-Device Communication to Support Diverse Applications."
Paper presented at International Conference on Research in Adaptive and Convergent Systems (RACS '16), Odense, Denmark. October 11, 2016 - October 14, 2016.

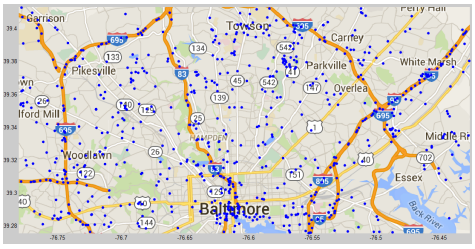


Figure 4: Site Locations of Baltimore City Area

gitude and translate them into a 2D position to show on a 2D map (also gathered from Google). Due to limited space, the detail evaluation of supporting other applications can be found the extended version of the paper.

4. CASE STUDY: TRAFFIC OFFLOADING IN CELLULAR NETWORKS

To show the effectiveness of our proposed scheme in the application case of offloading traffic for cellular users, we leverage the Vienna LTE-A system level simulator [12] and collected the real-world deployment information of base stations via the OpenCellID website [5] to carry out our performance evaluation.

Evaluation Setting: In our simulation, we generate LTE cells in a honeycomb structure and UEs are randomly deployed throughout the cells. To make our study to reflect the real-world practice, we have implemented a tool to obtain the deployment information of base stations from cellular network providers (AT&T, Verizon, T-Mobile, etc.) and have established realistic cellular networks to carry out the performance evaluation.

To measure the effectiveness of Wi-Fi Direct-based D2D communication on the realistic LTE-based cellular network, we consider the performance metrics, including average UE throughput, average spectral efficiency, and UE wideband Signal to Interference and Noise Ratio (SINR). Generally speaking, the throughput can be computed in symbols per second. The downlink spectral efficiency of the communication system can be measured in Bits Per Channel Use (bpcu) [11]. SINR, as the measurement of signal quality, can be used to quantify the relationship between RF conditions and throughput in the experimented network. By comparing metrics in the case without enabling D2D communication to the case with D2D communication, we can observe the improvement of performance of our proposed scheme based on these metrics.

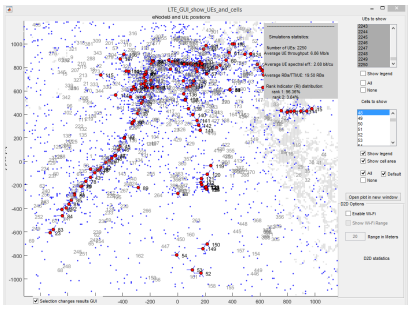


Figure 5: 150 Site Locations

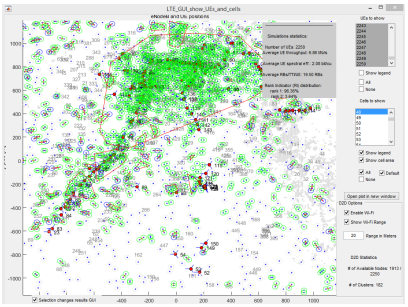


Figure 6: Clusters of 150 Sites

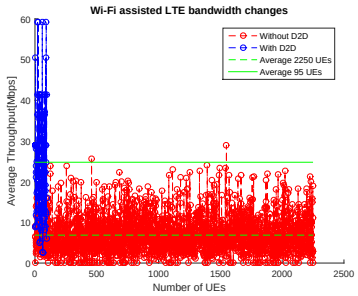


Figure 7: Average Throughput of UEs

Real-World Data Collection: To use the real-world data for the location of each eNodeB, we have retrieved the site locations from an open worldwide database called OpenCellID [5]. This website provided APIs for making HTTP (Hypertext Transfer Protocol) requests in different formats, e.g., XML (Extensible Markup Language) and Comma Separated Values (CSV). Based on the Region Of Interest (ROI), we can define the intended location and retrieve 1000 site locations in each HTTP request. As an example, we have retrieved the locations of eNodeB in Baltimore city, the state of Maryland. Figure 4 shows 1000 eNodeBs.

In our simulation, we choose the first 150 site locations from Baltimore city area and these site locations in the list downloaded from the online database point to highway 695, Pikesville, and Parkville as shown in Figure 5. In our simulation, we use these locations of eNodeB and randomly generate 2250 UEs to run the simulations. After applying our developed clustering-based D2D communication scheme, we create clusters for randomly deployed UEs. Figure 6 shows an example of merged clusters for UEs in the simulated area.

Evaluation Results: We run the simulation for both the LTE only communication and Wi-Fi Direct assisted LTE using our proposed method to create clusters. Figure 7 shows a magnificent increase in the average throughput of UEs when the D2D communication is used. As we can see from Figure 7, the number of clusters (blue curve) was 95 (182 clusters merged down to 95) and 1913 UEs out of 2250 were able to join clusters, while the rest of the UEs were out of range. Each cluster has one UE (head of cluster), which connects to the LTE network directly. By doing so, the average throughput of UEs can be significantly improved in compared with the case where D2D communication is not used (red curve).

We also evaluate the network performance comparing with other metrics when either D2D communication is enabled or

Golmie, Nada; Griffith, David; Hematian, Amirshahram; Lu, Chao; Yu, Wei.
"A Clustering-Based Device-to-Device Communication to Support Diverse Applications."
Paper presented at International Conference on Research in Adaptive and Convergent Systems (RACS '16), Odense, Denmark. October 11, 2016 - October 14, 2016.

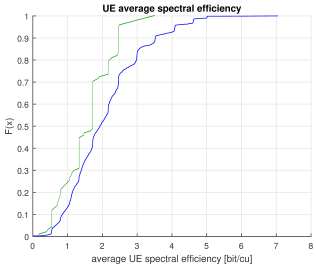


Figure 8: UE Average Spectral Efficiency (with D2D As Green)

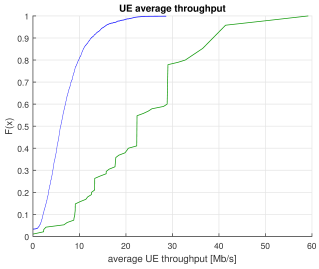


Figure 9: UE Average Throughput (with D2D As Green)

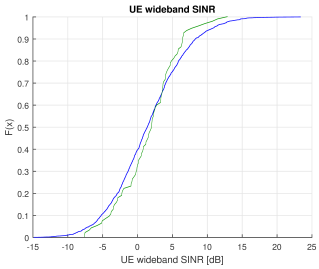


Figure 10: UE Wideband SINR (with D2D As Green)

disabled in the LTE network. Figures 8, 9 and 10 show the CDF of the UE average spectral efficiency, average throughput, and UE wideband SINR, respectively. As we can see, when D2D communication is enabled, the average UE throughput is continuously improved (almost doubled) and this is exactly what we expect, the average UE spectral efficiency is decreased because higher throughput demands higher frequencies that can accommodate lower numbers of bits, and SINR has lower range but still many UEs are in good range of coverage from the base stations.

5. CONCLUSION

In this paper, we investigated Wi-Fi Direct as an outband solution of D2D communication in LTE-based cellular networks. To enable the communication among devices that are nearby each other, provide the ability of offloading traffic transmitted via LTE-based cellular networks, and support communications for IoT applications such as the smart grid, we developed a clustering scheme which could create small Wi-Fi network clusters for nearby devices to remain connected to the LTE-based cellular network. Using real-world data collected from a public database in LTE networks and smart meter information, we demonstrated the effectiveness of our proposed scheme with respect to a comprehensive set of metrics.

6. REFERENCES

[1] *WiFi-Direct*. http://www.wi-fi.org/who-we-are.

[2] A. Asadi and V. Mancuso. Wifi direct and lte d2d in action. In *Proceedings of 2013 IFIP Wireless Days (WD)*, November 2013.

[3] A. Asadi, Q. Wang, and V. Mancuso. A survey on device-to-device communication in cellular networks. *IEEE Communication Survey and Tutorials*, 16(4):1801–1819, April 2014.

[4] A. Asadi, Q. Wang, and V. Mancuso. A survey on device-to-device communication in cellular networks. *IEEE Communications Surveys Tutorials*, 16(4):1801–1819, Fourthquarter 2014.

[5] O. W. by ENAiKOON. *Available at: http://opencellid.org/*.

[6] D. Camps-Mur, A. Garcia-Saavedra, and P. Serrano. Device-to-device communications with wi-fi direct: overview and experimentation. *IEEE Wireless Communications*, 20(3):96–104, 2013.

[7] B. Everitt. *Cluster analysis*.

[8] X. Fang, S. Misra, G. Xue, and D. Yang. Smart grid: The new and improved power grid: A survey. *IEEE Communications Surveys Tutorials*, 14(4):944–980, 2012.

[9] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis. Base-station assisted device-to-device communications for high-throughput wireless video networks. *IEEE Transactions on Wireless Communications*, 13(7), April 2014.

[10] C. Liu, C. Zhang, H. Yao, D. Zeng, Q. Liang, and C. Hu. A gps information sharing system based on bluetooth technology. In *Proceedings of 2014 International Conference on IT Convergence and Security (ICITCS)*, October 2014.

[11] P. K. Rekhi. Throughput calculation for lte tdd and fdd system. *Available at: http://www.slideshare.net/veermalik121/throughput-calculation-for-lte-tdd-and-fdd-system*.

[12] M. Taranetz, T. Blazek, T. Kropfreiter, M. Muller, S. Schwarz, and M. Rupp. Runtime precoding: Enabling multipoint transmission in lte-advanced system-level simulations. *IEEE Access*, 3:725–736, June 2015.

[13] Y. Yan, P. Yang, X.-Y. Li, Y. Zhang, J. Lu, L. You, J. Wang, J. Han, and Y. Xiong. Wizbee: Wise zigbee coexistence via interference cancellation with single antenna. *IEEE Transactions on Mobile Computing*, 14(12):2590–2603, December 2015.

[14] W. Yu, D. An, D. Griffith, Q. Yang, and G. Xu. Towards statistical modeling and machine learning based energy usage forecasting in smart grid. *ACM International Journal of Applied Computing Review (ACR)*, 15(1):6–16, April 2015.

[15] W. Yu, D. Griffith, L. Ge, S. Bhattarai, and N. Golmie. An integrated detection system against false data injection attacks in the smart grid. *International Journal of Security and Communication Networks (SCN)*, 8(2):91–109, 2015.

Golmie, Nada; Griffith, David; Hematian, Amirshahram; Lu, Chao; Yu, Wei.
"A Clustering-Based Device-to-Device Communication to Support Diverse Applications."
Paper presented at International Conference on Research in Adaptive and Convergent Systems (RACS '16), Odense, Denmark. October 11, 2016 - October 14, 2016.

vPROM: vSwitch Enhanced Programmable Measurement in SDN

An Wang*, Yang Guo†, Songqing Chen*, Fang Hao‡, T.V. Lakshman‡, Doug Montgomery†, Kotikalapudi Sriram†
* George Mason University ,† NIST ,‡ Bell Labs, Nokia

Abstract

While being critical to the network management, the current state of the art in network measurement is inadequate, providing surprisingly little visibility into detailed network behaviors and often requiring high level of manual intervention to operate. Such a practice becomes increasingly ineffective as the networks grow both in size and complexity. In this paper, we propose vPROM, a vSwitch enhanced SDN programmable measurement framework that automates the measurement process, minimizes the measurement resource usage, and addresses several significant technical challenges faced by early works. vPROM leverages the SDN programmability and extends the Pyretic run-time system and OpenFlow network interface to achieve the measurement automation. The required measurement resources are minimized by only acquiring the necessary statistics, made possible with instrumented Open vSwitches [1] with user defined monitoring capability. By decoupling monitoring from routing, vPROM reduces the interference between the measurement applications and other applications, and eliminates the frequent involvement of the controller. A vPROM prototype is implemented with DDoS and port-scan detection applications. The performance of vPROM is evaluated and the comparison results with other existing programmable measurement approaches are also presented.

I. Introduction

SDN is an emerging networking paradigm that enables the programming of the underlying network. Network measurement and monitoring is an important network application that can take advantage of the SDN's programmability. The SDN programmable measurement automates the measurement process, minimizes the resource usage by acquiring only the necessary statistics, and is able to utilize SDN switches as the measurement points across the networks. The SDN programmable measurement measures network traffic by actively installing rules for the flows of interest in the SDN routers' forwarding tables. The flow stats, such as packet and byte counts of the flows of interest, are collected through the flow entry counts. The measurement is controlled by the traffic measurement application programmed using network programming languages, and can be dynamically adjusted based on measurement needs. The initial endeavor on the SDN based programmable measurement has shown promises. In [1], network measurement policies are provided that allow users to query the network and conduct the measurement function such as sub-flow monitoring. NetAssay [2] pursues the so-called intentional network monitoring to capture the minimal set of traffic that satisfies the operator's monitoring goal.

While promising, the current SDN programmable measurement faces significant technical challenges: *(1) The interference between monitoring and other applications, e.g., forwarding, is nontrivial.* Each application has its own goal and a set of policies to enforce. Flow rules installed/removed by one application often interfere with overlapping rules installed/removed by other applications [3]. Hence any changes made by any application may require the run-time system to recompile to solve the conflicts. The newly generated forwarding entries then need to be installed into the switches' forwarding tables - resulting in significant overhead on the run-time system, the controller, as well as the SDN switches. In fact, such frequent recompilation negatively affects the system scalability as shown in [4]; *(2) The programmable measurement may require the continuous involvement of the controller.* For instance, define the *subflows* to be the fine-grained flows that belong to a mega-flow. Subflow monitoring requires the switch to send the first packet of every subflow to the central controller since the specific subflows are not known in advance. Such constant controller involvement is undesirable. *(3) Monitoring packet and byte counts by association with flow entries in the forwarding table is neither flexible nor sufficient for supporting various monitoring applications.* One reason is that the header fields that are of interest for packet forwarding may not always overlap with those that are of interest for monitoring. The chances of no overlap are likely to increase further as the number of header fields continues to grow beyond 40 or so [5]; *(4) The amount of Ternary Content-Addressable Memory (TCAM) at hardware switches is limited.* TCAM, widely used for fast packet forwarding, is expensive and power hungry, which limits its amount inside a physical switch. The available TCAM may not be sufficient for the measurement purpose;

In this paper, we propose to build vPROM, a vSwitch enhanced SDN programmable measurement framework that

[1]Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Chen, Songqing; Guo, Yang; Hao, Fang; Lakshman, T.V.; Montgomery, Douglas; Sriram, Kotikalapudi; Wang, An.
"vPROM: vSwitch Enhanced Programmable Measurement in SDN."
Paper presented at IEEE 25th International Conference on Network Protocols (ICNP), 2017, Toronto, ON, Canada. October 10, 2017 - October 13, 2017.

addresses the aforementioned issues. vPROM runs on the instrumented Open vSwitches [6], [7] that decouples monitoring from forwarding and can support user-defined monitoring capability. Furthermore, we extend Pyretic to Pyretic+ run-time system to parse vPROM applications into flow rule sets for both forwarding and monitoring, and extend OpenFlow to OpenFlow+ in order to allow applications to set up monitoring rules. A client is also built to facilitate the communication between the controller and the run-time system.

A salient feature of vRPOM is the extensive use of Open vSwitches (OVS) as measurement vantage points. OVS often runs on a general purpose computer and acts as the edge router for the virtual machines (VMs) hosted on the same machine. Compared to a physical core router, an OVS routes at a slower speed, encounters a smaller number of flows, and has access to much more memory and CPU resources. In addition, because the flows monitored at an OVS are either originated or terminated at the VMs, some management functionality, e.g., intrusion/anomaly detection, can be migrated from the central administration point to the edge. The instrumented Open vSwitch, called UMON [7], supports the explicit measurement function that decouples the measurement function from the packet forwarding function. The decoupling is achieved via the introduction of the *monitoring flow table*, which separates the monitoring rules from the forwarding rules. Users can thus freely install monitoring rules without worrying about the possible interference with the forwarding rules.

vPROM extends the Pyretic run-time system to Pyretic+ so that a vPROM measurement application programmed in Pyretic can seamlessly utilize the UMON capabilities. The run-time system is modified to automatically identify the measurement capability of a SDN switch, and use the monitoring flow table if the switch is instrumented. A Ryu SDN controller is used in vPROM. A Ryu client is built so that the Pyretic+ run-time system can communicate with the Ryu controller to configure SDN switches and retrieve states from SDN switches. To demonstrate the capability of vPROM, we implemented a prototype and several vPROM applications. The performance of vPROM is evaluated and the comparison results with other existing programmable measurement approaches are also presented.

The paper is organized as follows. Related work is summarized in Section II. The vPROM architecture are described in Section III. A vPROM application is presented in Section IV. Evaluation results are presented in Section V. Concluding remarks are in Section VI.

II. Related work

Network programmability has been studied extensively and several network programming languages, e.g., [8], [1], [9], [10], among others, have been developed. The programming languages offer high level abstractions that make the programming of complex network functions/applications [11], [12], [13] possible. Sophisticated SDN applications can be programmed and run simultaneously without worrying about the intricate interactions among them. The study in [4], however, shows that it may take minutes to compile policies of different applications and generate millions of forwarding rules that need to be installed in the data plane for a realistic large Internet exchange point. In vPROM, measurement points, the vSwitches, are instrumented with the explicit measurement function. The network measurement function is thus decoupled from other functions, eliminating the interactions.

Network measurement has been programmed as SDN applications or query policies [1], [12], [2]. These network measurement applications run on top of the run-time system and the controller and often require repeated involvement of both elements, e.g., when conducting sub-flow monitoring. vPROM addresses this issue by decoupling the monitoring from the forwarding. We further extend the OpenFlow API to allow the measurement applications to directly control measuring switches through the controller. Trumpet [14] takes a different approach and designs its own distributed packet monitors and centralized event monitoring system. Trumpet packet monitors collect stats associated with pre-defined 5-tuple flows. vPROM favors customized monitoring that can dynamically change monitoring resolutions demanded by users/applications. In addition, vPROM leverages the existing open source software and latest research advancement on network programming languages.

III. vPROM design

Fig. 1 depicts the architecture of vPROM. As shown in the figure, vPROM consists of five major components: (1) UMON vSwitches, the instrumented Open vSwitches that provide user-defined monitoring capability and some local application functions being pushed from the central controller to the edge; (2) OpenFlow+, the augmented OpenFlow API that allows the applications to set the monitoring rules at UMON and to control the application threads running at the vSwitches; (3) the Ryu client, which serves as the "interpreter" between the run-time system and the Ryu controller; (4) Pyretic+, the extended Pyretic run-time system that can parse a vPROM app into the flow rule sets for traditional SDN switches and the monitoring rule sets for UMON vSwitches; and (5) vPROM applications programmed using extended Pyretic language. vPROM applications obtain measurement stats from both traditional SDN switches and UMON switches. Below we present the Pyretic+ and OpenFlow+, after an overview of UMON, the instrumented OVS.

A. Background on UMON

The UMON design [7] strives to achieve three goals: (1) decoupling monitoring from forwarding; (2) supporting sub-flow monitoring and monitoring based on non-routing fields; and (3) supporting application threads. To achieve these goals, the major challenge lies in how to implement the decoupling in the existing vSwitch architecture. The packet forwarding pipeline is defined in the Openflow specification [15] and implemented in the Open vSwitch's user space (see top part of Fig. 2). In UMON, a new table, *monitoring flow table*, is designed and implemented to separate monitoring rules from

Chen, Songqing; Guo, Yang; Hao, Fang; Lakshman, T.V.; Montgomery, Douglas; Sriram, Kotikalapudi; Wang, An.
"vPROM: vSwitch Enhanced Programmable Measurement in SDN."
Paper presented at IEEE 25th International Conference on Network Protocols (ICNP), 2017, Toronto, ON, Canada. October 10, 2017 - October 13, 2017.

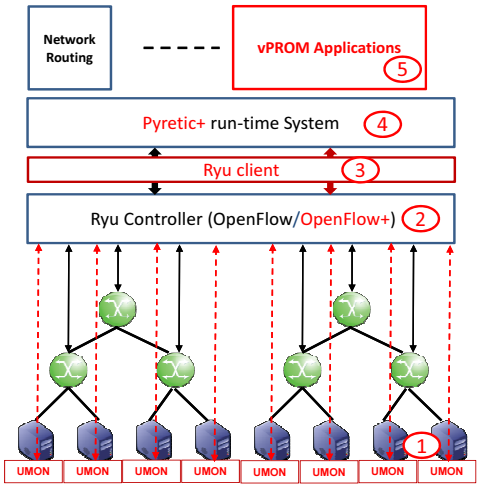


Fig. 1. vPROM framework architecture and key elements.

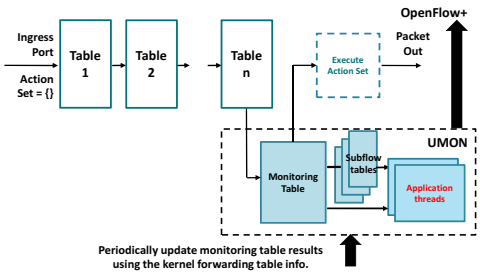


Fig. 2. Packet forwarding pipeline in the UMON, an instrumented Open vSwitch.

forwarding rules, as shown in Fig. 2. Users can thus freely install monitoring rules without worrying about the possible interferences with forwarding rules. The subflow monitoring is also supported by a newly defined *subflow monitoring action*, which acts as a local controller. The subflows subjected to the monitoring are inserted into the subflow table where the monitoring results are gathered and stored. The measurement results can be actively collected by vPROM applications through the OpenFlow+ API. Application threads, such as the port-scan detection threads and DDoS attack detection threads, run in UMON using the locally collected stats. These application threads, running at UMONs distributed across the network, scale up the central vPROM application, and reduce the measurement traffic from the switches to the central controller.

The architecture of the Open vSwitch is more complex than the pipeline as depicted in Fig. 2. It includes a kernel module which caches the flow rules to speed up the packet forwarding. In Section IV-A, we instrumented UMON to support quick large flow detection using coincidence counting scheme [16]. We address the challenge of dividing the tasks between kernel module and user-space modules. Finally, in order to support subflow monitoring and monitoring on non-routing fields, it is infeasible to employ a dedicated flow table in the OpenFlow pipeline to replace the monitoring table.

B. OpenFlow+ Protocol

OpenFlow+ extends the OpenFlow protocol to enable the SDN controller to manage the UMON monitoring table, collect the measurement stats, and start/stop the application threads at UMON vSwitches. The OpenFlow protocol contains three types of messages: *Controller-to-Switch messages*, *Asynchronous messages*, and *Symmetric messages*. The controller-to-switch messages are initiated by the controller and may or may not require a response from the switch. Asynchronous messages are sent by switches to the controller without solicitation. Switches send asynchronous messages to the controllers to signal a packet arrival, change of switch state, or an error. Symmetric messages, such as Hello and Echo, are sent without solicitation in either direction. We next describe the additional messages added in OpenFlow+ and their implementation.

• **Monitoring Table Management**. Each OpenFlow message begins with the OpenFlow header, which includes a `type` field indicating the type of a message. We introduce a new type *OFPT_MONITOR_MOD* to indicate that the message is related to the monitoring table. Table I lists six new commands. Among them, five commands, *OFPMMC_ADD*, *OFPMMC_MODIFY*, *OFPMMC_DELETE*, *OFPMMC_MODIFY_STRICT*, and *OFPMMC_DELETE_STRICT*, are similar to the forwarding flow table modification commands. The last one, *OFPMMC_DELETE_SUBFLOWS*, enables the controller to delete the subflow tables to save the storage space.

Besides the new commands, we add two types of new monitor actions: *OFPAT_MONITOR* for monitoring non-routing fields and subflow monitoring, and actions to control application threads. The *OFPAT_MONITOR* action structure is as follows:

```
struct ofp_action_monitor {
    ovs_be16 type;
    ovs_be32 monitor_flag;
    uint8_t subflow_flag;
    struct ofp_match_header subflow;
    ...
};
```

The field *monitor_flag* allows users to define the monitoring of non-routing fields. For instance, *monitor_flag* values of *OFPMT_SYN*, *OFPMT_SYNACK*, *OFPMT_FIN*, etc., instruct to collect packet/byte counts of TCP SYN, SYN/ACK, and FIN. The two parameters, *subflow_flag* and *subflow_mask*, are for subflow monitoring purpose. The first parameter is a boolean value indicating if subflow monitoring is turned on. If it is on, *struct ofp_match_header subflow* contains the wildcard mask for subflow monitoring. The action for application thread control is described later.

• **Stats collection**. The stats request from the controller to the switch is a new multipart message defined as *OFPMP_MONITOR_STATS*. This stats request allows the controller to collect the stats of the entire monitoring table, or

Chen, Songqing; Guo, Yang; Hao, Fang; Lakshman, T.V.; Montgomery, Douglas; Sriram, Kotikalapudi; Wang, An.
"vPROM: vSwitch Enhanced Programmable Measurement in SDN."
Paper presented at IEEE 25th International Conference on Network Protocols (ICNP), 2017, Toronto, ON, Canada. October 10, 2017 - October 13, 2017.

TABLE I
OPENFLOW+ COMMANDS MANAGING MONITOR TABLE

Command	Functionality
OFPMMC_ADD	add new monitor rules
OFPMMC_MODIFY	modify all matching monitoring rules
OFPMMC_MODIFY_STRICT	modify monitoring rules strictly matching wildcards
OFPMMC_DELETE	delete all matching monitor rules
OFPMMC_DELETE_STRICT	delete monitoring rules strictly matching wildcards
OFPMMC_DELETE_SUBFLOWS	delete subflow tables collected by matching monitor rules

the stats of a specific monitoring rule. The subflow tables associated with the monitoring rules can also be reported when available. We use the following data structure for *OFPMP_MONITOR_STATS*:

```
struct ofp_monitor_stats_request {
   uint8_t type;
   uint8_t with_subflows;
   uint8_t threshold_type;
   ovs_be32 threshold_value;
   /* Followed by an ofp_monitor_match
     structure for exact match rule request. */
   ...
};
```

We define two new *types*: *OFPMR_ALL* and *OFPMR_EXACT*. *OFPMR_ALL* requests the stats of the entire monitoring table, while *OFPMR_EXACT* requests the stats of a specific rule or rules matching the *ofp_monitor_match* field. The field *with_subflows* indicates if the subflow tables should be reported. If *with_subflows* is on, we also control the granularity at which the subflow tables are reported. The field *threshold_value* allows to set up a threshold and only the subflow entries whose byte count or packet count surpasses the threshold will be reported to the controller. The field *threshold_type* defines whether byte count (*OFPMRT_BYTE*) or the packet count (*OFPMRT_PKT*) is chosen in the threshold comparison.

After receiving the stats request, the switch generates a reply message including information concerning the matching monitor rules, the related statistics, and the subflows, if any.

```
struct ofp_monitor_stats {
   uint8_t stat_count;
   ovs_be16 n_subflows;
   /* Monitor rule match */
   /* uint64_t monitor_stats[] */
   /* struct ofp_monitor_subflow flows[] */
   ...
};
```

In the reply message as shown above, the counter *n_subflows* represents the number of subflows from the matching monitor rules. If this value is 0, then there is no subflow reported. Otherwise, all the subflow info will be gathered together within a dynamic array constructed as follows:

```
struct ofp_monitor_subflow {
   ovs_be16 tcp_flags;
   ovs_be64 packet_count;
   ovs_be64 byte_count;
   /* subflow match, struct ofp_match_header */
   ...
};
```

• **Application thread management**. For each application thread, we introduce an action to control this thread. Application threads are implemented as UMON threads that use the measurement stats for various purposes. For instance, we implement the vertical port-scan detection thread, the horizontal port-scan detection thread, and quick large flow detection thread (see Section IV-A). Using the port-scan thread as an example, we introduce the action *OFPAT_PRTSCAN_DETECTION* for its control. The action structure is defined as follows:

```
struct ofp_action_prtscan_detection {
   ovs_be16 type;
   uint8_t detector_switch;
   uint8_t detection_type;
   ovs_be64 interval;
   ovs_be16 vthresh;
   ovs_be16 hthresh;
   struct ofp_match_header submatch;
   ...
};
```

The parameter *detector_switch* is the knob to enable or disable the local detection thread. The detection is achieved by periodic analysis of the subflow stats. The parameter `interval` defines the period at which the port-scan detector runs to analyze the subflow stats. Moreover, we enable two types of scanning behavior detection, i.e., vertical scan and horizontal scan. The parameter `detection_type` dictates which scan is running. For the purpose of detection, this action accepts threshold for each type. Parameters `vthresh` and `hthresh` are thresholds used by vertical and horizontal detection, respectively. During local port-scan detection, whenever suspicious activities are detected by the application thread, we use the Asynchronous message for the application thread to send alert messages to the controller. The data structure for the alert message is as follows:

```
struct ofp_prtscan_alert{
   uint8_t detection_type;
   ovs_be16 n_ports;
   ovs_be16 n_attackers;
   ovs_be16 n_victims;
   ovs_be32 n_subflows;
   /* Monitor rule match */
   /* uint16_t victim_ports[] */
   /* list of attacker ip addresses*/
};
```

Chen, Songqing; Guo, Yang; Hao, Fang; Lakshman, T.V.; Montgomery, Douglas; Sriram, Kotikalapudi; Wang, An.
"vPROM: vSwitch Enhanced Programmable Measurement in SDN."
Paper presented at IEEE 25th International Conference on Network Protocols (ICNP), 2017, Toronto, ON, Canada. October 10, 2017 - October 13, 2017.

The message includes the flow rules where attacks are detected.

All the new commands introduced in OpenFlow+ are compatible with the early versions of OpenFlow. Extra data structures are necessary on both the controller and the switch to support the implementation of OpenFlow+.

C. Ryu Client

The controller client serves as an interface for the run-time system to communicate with the SDN controller. Its main function is to translate the Pyretic messages (of the run-time system) to the OpenFlow messages (used by the SDN controller), and vice versa. We choose to use the Ryu controller in the vPROM framework over the POX controller used by the original Pyretic run-time system. Ryu is a long-term supported project. The Ryu controller continuously upgrades itself to support newer versions of OpenFlow releases, which will allow vPROM to support newer version OpenFlow in the future with minor change to the controller client. In addition, the implementation of OpenFlow+ in Ryu is quite manageable.

In the Ryu client, an OpenFlow+ interface conducts the message translation. Furthermore, the Ryu client allows the Ryu controller to inform the run-time system if a SDN switch is instrumented, i.e., if a switch is a UMON switch, and if so, what edge management threads it supports. Such information will be stored in the run-time system and be used in meeting vPROM app requirement.

The Ryu client also provides the stats collection service for run-time system. The stats collected in UMON switches can be pulled by the Ryu controller. A stats collection module in the Ryu client periodically instructs the Ryu controller to pull the stats. The collected stats are then forwarded to the run-time system and the vPROM applications.

D. Pyretic+ and its run-time system

Pyretic is a Python style network programming language that offers high-level abstractions for users to write compact programs to define what the network switches should do with incoming packets. Pyretic has a corresponding run-time system that takes multiple Pyretic programs as input, compiles them together and generates flow rule sets to be installed at the underlying SDN switches. These flow rule sets satisfy the collective Pyretic programs' requirements. We call the extended Pyretic Pyretic+. Below we first describe how the Pyretic+ language supports UMON switches semantically. We then describe how the Pyretic+ run-time system generates the forwarding rules and monitoring rules separately.

1) Pyretic+ language: Pyretic defines polices and operators [17]. The basic polices includes `match`, `drop`, `identity`, `forward`, `flood`, `if_`, etc., and the operators include `+` (parallel composition), `>>` (sequential composition), etc. Pyretic further defines three query polices:

- `packets(limit=n,group_by=[f1,f2,...])`, which callbacks on every packet received for up to `n` packets identical on fields `f1,f2,...`;
- `count_packets(interval=t,group_by=[f1,f2,...])`, which counts every packet received. Callback every `t` seconds to provide count for each group;
- `count_bytes(interval=t,group_by=[f1,f2,...])`, which counts every byte received. Callback every `t` seconds to provide count for each group.

For instances, in the following example, all TCP traffic incoming from `inport=1` are sub-flow monitored based on their `'srcip'` and `'dstip'`. The traffic is then forwarded to `outport=2`.

```
Q = count_packets(interval=t, group_by=['srcip
    ','dstip'])
match(inport=1) >> if_(match(protocol=6), Q,
    identity) >> fwd(2)
```

To support UMON TCP flagged packets monitoring, Pyretic+ adds the `'tcpflag'` option in the query policies' `group_by` parameter. Using `'tcpflag'` option alone, namely `group_by=['tcpflag']` indicates that the action *OFPAT_MONITOR* as defined in OpenFlow+ is active. In contrast, if the `'tcpflag'` option is used along with other options such as `srcip` and `dstip`, the subflow monitoring will be executed.

New policies are introduced to control individual edge management threads. For instance, the new policy `prtscan_detection` can activate/deactivate local port-scan detector. The parameter options are defined the same as in the action *OFPAT_PRTSCAN_DETECTION* in OpenFlow+. The callback function can also be defined and registered to react to the received alert messages.

2) Pyretic+ run-time system: The Pyretic run-time system compiles the programs and generates an abstract syntax tree (AST) that represents the policies and their inter-relationship as defined by the operators. For example, the abstract syntax tree (AST) as shown in Fig. 3 is derived from the `count_packets` example in Section III-D1. In this figure, all the operator nodes are marked in green and the polices are in yellow. The tree is built by parsing the application programs. The run-time system then generate the flow rule sets for individual SDN switches based on this AST.

In Pyretic+, the run-time system needs to generate both the forwarding rules and monitoring rules for a UMON switch. This is achieved by deriving separate forwarding AST and monitoring AST using the general AST as in Pyretic run-time system. The forwarding rules and monitoring rules are created thereafter.

• **Deriving monitoring AST**. The Algorithm MON-AST-GEN describes how to generate monitoring AST and flow rules. The algorithm starts with finding the query policy nodes and UMON specific policy nodes as defined by the set $\mathbb{C}$. For each identified such node, e.g., policy `Q` in Fig. 3, the while-loop between line 9 and 15 collects all the operator nodes from the identified node up to the top-left node. The nodes posterior to the identified node are ignored since they have no effect on the monitoring policy. The nodes are further processed to remove the nodes operated in parallel with the identified

Chen, Songqing; Guo, Yang; Hao, Fang; Lakshman, T.V.; Montgomery, Douglas; Sriram, Kotikalapudi; Wang, An.
"vPROM: vSwitch Enhanced Programmable Measurement in SDN."
Paper presented at IEEE 25th International Conference on Network Protocols (ICNP), 2017, Toronto, ON, Canada. October 10, 2017 - October 13, 2017.

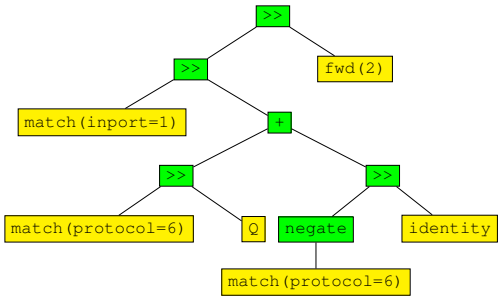
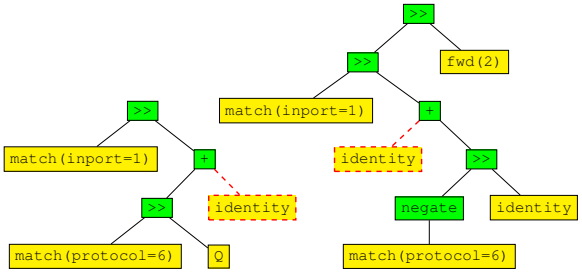


Fig. 3. Abstract syntax tree (AST) of a measurement application example



(a) Monitoring AST (b) Forwarding AST

Fig. 4. Derived forwarding and monitoring ASTs

nodes, as shown between line 14 and 18 in the algorithm. As a result, the sub-trees of any of the operators `intersection`, `sequential` and `difference` are preserved to build monitoring policy. The generated monitoring AST is shown in Fig. 4(a).

```
1:  function MON-AST-GEN(rt_ast)
2:      init policies = LIST()
3:      ℂ=SET([count_packets, count_bytes,
    counts, packets, prtscan_detection])
4:      ℚ=SET([intersection, sequential,
    difference])
5:      𝕃 ← all leave nodes of rt_ast
6:      for l ∈ 𝕃 do
7:          if ISINSTANCE(TYPEOF(l)) ∈ ℂ then
8:              set r_nds = LIST()
9:              while p_node ≠ the top-left node do
10:                 r_nds.append(p_node)
11:                 p_node ← p_node.GETPARENT()
12:             end while
13:             set nd_lst = SET()
14:             for op ∈ r_nds do
15:                 if ISINSTANCE(TYPEOF(op)) ∈ ℚ then
16:                     nd_lst.add(relevant nodes from sub-
    tree of op)
17:                 end if
18:             end for
19:             policies.append(BUILDPOLICY(nd_lst))
20:         end if
21:     end for
22:     return policies
23: end function
```

The function BUILDPOLICY further compiles the monitoring AST into policy, i.e., flow rules, similar to a stack machine compiler. A stack machine uses a last-in, first-out(LIFO) stack to hold the temporary values. Most of its instructions assume the operands are popped from the stack and the operation results are pushed back to the stack. In MON-AST-GEN, BUILDPOLICY creates an empty stack and continuously reads nodes from *nd_lst*. If it reads an operand, the node will be pushed into the stack. Otherwise, operands will be popped from the stack based on the operation types.

The algorithm is run once for each SDN switch since the policies may be different for different switches. For example, monitoring policy `match(protocol=6)>>if_(match(switch=1,srcip='10.0.0.1'),Q,Q)` will generate different rules on switch 1 and other switches.

• **Deriving forwarding AST**. The gist of deriving forwarding AST is to remove the nodes relevant to the monitoring functions. Notice that the forwarding AST is not complementary to the monitoring AST entirely as shown in Figure 4. The node `match(protocol=6)` is unrelated to the forwarding policy. However, node `match(inport=1)` is shared by both ASTs. As a result, algorithm FORWARD-AST-GEN cannot simply remove all the nodes in monitoring AST. Function FORWARD-AST-GEN starts from the query policy nodes and UMON specific policy nodes in the AST. For each such node, the algorithm iterates upward until it hits the first `parallel` operator node. This process will remove all the nodes that are exclusive to the monitoring AST as identified between line 9 and 15 in the algorithm. Finally, function BUILDPOLICY is called to build forwarding rules/policies based on the nodes in the forwarding AST.

IV. vPROM-GUARD: A vPROM USE CASE

To demonstrate vPROM's effectiveness, we build vPROM-GUARD, a vPROM application that detects DDoS and port-scan attacks automatically. Distributed Denial of Service (DDoS) attacks and port-scan attacks are significant threats to the Internet. The challenge in DDoS and port-scan defense is the ability to detect patterns of abusive behaviors amongst a vast sea of benign individual network exchanges. Security monitoring systems often utilize the signature-based and/or the behavior-based approach to detect DDoS attacks. Fine grained packet-level or microflow-level measurement at line rate is often required. Such fine grained real-time measurement is extremely demanding on the hardware and requires sophisticated technologies, resulting in expensive network security middle-boxes.

In contrast, vPROM is a distributed measurement framework that can be programmed and reconfigured in real time to respond to ever changing attack vectors. The key idea of vPROM-GUARD is to *employ efficient attack detectors and monitor the attack cues at a coarse measurement granularity*

Chen, Songqing; Guo, Yang; Hao, Fang; Lakshman, T.V.; Montgomery, Douglas; Sriram, Kotikalapudi; Wang, An.
"vPROM: vSwitch Enhanced Programmable Measurement in SDN."
Paper presented at IEEE 25th International Conference on Network Protocols (ICNP), 2017, Toronto, ON, Canada. October 10, 2017 - October 13, 2017.

```
function FORWARD-AST-GEN(rt_ast)
    init policies = LIST()
    ℂ=SET([count_packets, count_bytes,
counts, packets, prtscan_detection])
    𝕃 ← all leave nodes of rt_ast
    for l ∈ 𝕃 do
        if ISINSTANCE(TYPEOF(l)) ∈ ℂ then
            set nd_lst = SET()
            set r_nds = LIST()
            while p_node ≠ the top-left node do
                if ISINSTANCE(TYPEOF(p_node)) ==
parallel then
                    break
                end if
                r_nds.append(p_node)
                p_node = p_node.GETPARENT()
            end while
            prune subtree of p_node
            nd_lst ← all the relevant nodes
            policies.append(BUILDPOLICY(nd_lst))
        end if
    end for
    return policies
end function
```

when the network is not under attack, and switch to the fine-grained network monitoring and attack detection/validation when suspicious activities are detected. The benefits of such an approach are multifold: (1) the distributed edge measurement and coarse grained measurement level reduce the overall measurement burden on the network; (2) when under attack, only the alerted hosts need to conduct fine granularity measurement and local detection; (3) local detection at edge mitigates the burden of the central detector; and (4) false alarms are more tolerable because the detection is controlled by a program and a false alarm merely triggers the extra fine-grained measurement at vSwitches rather than frequent human interventions. If proven to be effective, vPROM-GUARD has the potential to replace the middle-box solution in a data center with a pure low cost software solution. Next, we present the detection methods used in vRPOM-GUARD.

A. Coincidence counting based large flow detection

Quick detection of large flows at the incipient of DDoS attack is vital for DDoS detection. The authors in [16] developed the Coincidence Base Traffic Estimator (CATE) that can estimate flow rates quickly with provable bounds on estimation error. CATE maintains a predecessor table and a coincidence count table, as shown in Fig. 5. The predecessor table includes the most recently received k packet headers. A flow is defined as $f = r\&m$ with r being the packet header and m the flow mask. Upon the arrival of a new packet, its corresponding flow id f is compared with every flow id in the predecessor table. The number of coincidences for the flow f, l_f, is the number of times the flow f occurs in the predecessor table. If $l_f > 0$ and flow f is not in the Coincidence count table, f is added into the coincidence count table with count of l_f. If f is already in the coincidence count table, then the count for f is incremented by l_f. Let

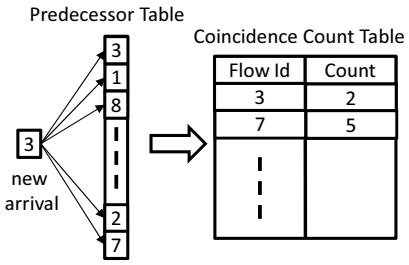


Fig. 5. CATE scheme.

$M(N, f)$ be the number of coincidences for flow f after N arrivals with k comparisons for each arrival. The estimated proportion of traffic from flow f, $\hat{p}_f$, is $\hat{p}_f = \sqrt{\frac{M(N,f)}{Nk}}$.

While the CATE scheme is reasonably simple, instrumenting UMON to support CATE is not trivial. The coincidence counting is conducted for every new arrival. If CATE is implemented in the kernel module of OVS, it can slow down the data path forwarding speed. We adopt a strategy that offloads the CATE from the critical data forwarding path. Specifically, we implement the CATE scheme as a user-level thread in the OVS. A copy of the incoming packet header is made in the kernel module, and a batch of packet headers is periodically delivered to the user-level CATE. User-level CATE executes the coincidence counting upon receiving the newly arrived packet headers. The strategy minimizes the overhead imposed on the UMON data path.

B. Change-point monitoring for attack cues

The authors in [18] developed the *change-point monitoring* for TCP based attack detection. The technique is based on the observation that TCP {SYN, SYN/ACK} and {SYN, FIN} are *request-response pairs* that should be balanced in a normal network environment, and they deviate from the balanced state when under attacks. The Cumulative Sum Method [19], [20] is employed to detect the deviation. Specifically, let q_i and p_i be the number of requests and responses, respectively, in the i-th measurement epoch. The difference, δ_i, is $\delta_i \triangleq q_i - p_i$. Define $\tilde{\delta}_i$ to be the normalized difference,

$$\tilde{\delta}_i = \delta_i / P_i, \tag{1}$$

with $P_i = \alpha P_{i-1} + (1-\alpha)p_i$ and α being a positive constant less than one. To detect the deviation of $\tilde{\delta}_i$ from its mean, which should be close to zero, the *Cumulative Sum* method is used. Define S_i to be the cumulative sum:

$$S_i = (S_{i-1} + \tilde{\delta}_i - t)^+, \tag{2}$$

where t is a constant threshold and $(\cdot)^+$ takes the positive value or zero. The value of t is chosen such that $\tilde{\delta}_i > t$ indicates a

potential attack. When the cumulative sum S_i becomes greater than the threshold T, $S_i > T$, a potential TCP based attack is detected. The value of t and T are design parameters that affect the *attack detectability*, *false alarm interval* and *detection delay*.

In order for the change-point monitoring to detect an attack, $\tilde{\delta}_i$ needs to be greater than t, $\tilde{\delta}_i > t$ when the attack is on. Otherwise $\tilde{\delta}_i - t$ is negative in Eqn (2), and does not contribute to the cumulative sum S_i. Therefore the average number of on-going TCP connections, i.e., the value of P_i, needs to be *comparable* to that of δ_i (see Eqn (1)). Otherwise the attack becomes either not detectable, or the variation in the normal TCP connections is greater than the value of t, which leads to a large number of false alarms (see Eqn (2)). For example, if the goal is to detect if one machine inside a large organization is under attack, conducting the change-point monitoring at the gateway for the entire organization likely does not work since the number of on-going TCP sessions is much larger than the attacking sessions. vPROM-GUARD addresses the issue by conducting the monitoring at individual machines hosting a small number of VMs.

C. Attack detection in vPROM-GUARD

The attack detection in vPROM-GUARD is accomplished in two phases: big flow and coarse-grained indicator/cue monitoring and fine-grained attack detection/validation. In the first phase, vPROM-GUARD periodically detects big flows (via CATE), and collects packet counts of TCP SYN, SYN-ACK, FIN, and RST and runs Cumulative Sum (CUSUM) algorithm. Assume that there are J hosts in total. The change-point monitoring at host j is:

$$\delta_i^j = q_i^j - p_i^j, \quad (3)$$

$$\tilde{\delta}_i^j = \delta_i^j / P_i^j, \quad (4)$$

$$S_i^j = (S_{i-1}^j + \tilde{\delta}_i^j - t)^+, \quad (5)$$

for $j = 1, 2, \ldots, J$. Comparing to the centralized change-point monitoring [18], the distributed change-point monitoring at individual hosts shortens the detection delay and localizes the attacks. For instance, if only host j is under SYN flood attack, then $\delta_i^j = \delta_i$ but $P_i^j \leq P_i$. Thus $\tilde{\delta}_i^j \geq \tilde{\delta}_i$ and $S_i^j \geq S_i$, leading to early detection. Furthermore, the distributed change-point monitoring localizes the detection. Only the hosts whose cumulative sum S_i^j is greater than threshold T need to be further examined.

vPROM-GUARD starts the detection process by turning on CATE monitoring, and installing monitoring rules for TCP SYN, SYN/ACK, FIN, and RST packet counts into UMON at each hosting machine. Then the big flow info and the packets counts are collected periodically by vPROM-GUARD using OpenFlow+ stats collection commands. The change-point monitoring is conducted for individual hosts using the collected stats. If a big flow is detected and the change-point monitoring detects the deviation, a TCP SYN flood attack is likely to be detected. We further install rules to collect TCP flag packet counts associate with this big flow to validate the type of attack. If a big flow is detected but the TCP flag packet counts do not deviate from the balance, it still indicates a potential DDoS attack. The vPROM-GUARD controller can implement whatever policy the users prefer to further classify this flow. Finally, if no big blow is detected but the change-point monitoring issues potential attack alerts to/from a host, the vPROM-GUARD starts the subflow monitoring and port-scan detection threads on that host. The vPROM-GUARD, running at a central location, also periodically collects the subflow stats from the hosts under alert, and runs DDoS detection and port-scan detection across the subflow stats collected from these suspected hosts. This allows the detection of attacks that may spread across multiple hosts.

V. Evaluation

We instrument the Open vSwitch (version 2.3.2) and run it on a four-core, 3.2GHz CPU machine with 10GB memory. The machine is equipped with an Intel NIC with two 10GHz ports. The Ryu controller (version 3.25), Ryu client, Pyretic+, and vPROM apps run on another machine of the same configuration. Both machines use the Ubuntu 14.04.3 LTS kernel. We use a third machine as both the packet generator and the packet sink so as to avoid clock synchronization problem. The packet generator and sink are connected to the vSwitch via two 10GHz ports. The packet generator uses Tcpreplay [21] to replay a data center traffic trace collected by Benson et al. [22]. The trace lasts for a period of about 65 minutes.

A. Comparison of subflow monitoring: vPROM vs Pyretic

Subflow monitoring is an important monitoring capability for applications such as heavy-hitter flow detection and port scanning attack detection. Subflow monitoring is supported in Pyretic by so-called *query policies* [17], which can be conjoined to any of the other policies, e.g., routing policies. It is straightforward to program for the subflow monitoring in Pyretic:

```
m=['srcmac','dstmac','srcport','dstport']
Q=count_packets(interval=t,group_by=m)
match(srcip=A, dstip=B) >> Q
```

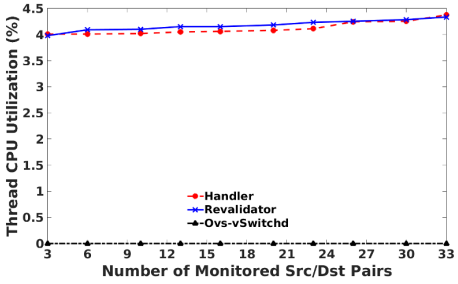
The first line defines the subflow mask that is based on source MAC address, destination MAC address, source port Id, and destination port Id. The function *count_packets()* returns the packet counts every t seconds for each subflow. The megaflow is defined using *match()*. *match(srcip=A, dstip=B)* captures all packets from A to B and hands them to subflow monitoring policy *Q*.

Below we compare the performance of vPROM and Pyretic in supporting subflow monitoring. In the Pyretic experiment, we employ a simple routing that forwards all packets from the input port *in_port* to the output port *out_port* that connects to the sink. The Pyretic routing policy is:

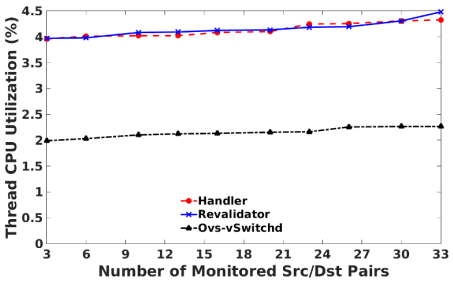
```
match(inport=in_port) >> fwd(out_port)
```

This routing policy runs in parallel with the subflow monitoring policy. The same routing and subflow monitoring are conducted using UMON in vPROM. In addition, since the

Chen, Songqing; Guo, Yang; Hao, Fang; Lakshman, T.V.; Montgomery, Douglas; Sriram, Kotikalapudi; Wang, An.
"vPROM: vSwitch Enhanced Programmable Measurement in SDN."
Paper presented at IEEE 25th International Conference on Network Protocols (ICNP), 2017, Toronto, ON, Canada. October 10, 2017 - October 13, 2017.



(a) vPROM



(b) Pyretic

Fig. 6. Overhead in Data Plane

subflow monitoring is a built-in capability of UMON, vRPOM can directly insert the monitoring rules for each monitored source-destination pair into the monitoring table with the subflow mask of srcmac, dstmac, srcport, and dstport on.

We first evaluate the subflow monitoring overhead imposed on the switches, an UMON vSwtich in vPROM and an non instrumented Open vSwitch in Pyretic. Overhead is measured using the CPU utilization of three types of threads: *handler*, *revalidator*, and *ovs-vswitchd* [6], [7]. *ovs-vswitchd* is a user space daemon that handles the communication with the SDN controller, among other things. Figure 6 depicts the CPU utilization of different threads for vPROM and Pyretic. We vary the number of monitored source-destination pairs, from three to thirty-three, to change the monitoring workload. In all cases, the CPU utilization of all threads increases with the number of monitored pairs. The CPU utilization of threads *handler* and *revalidator* is similar for vPROM and Pyretic, but differs for *ovs-vswitchd* thread. For vPROM, no CPU resources are consumed by thread *ovs-vswitchd* since all packet processing decisions are made locally and there is no need to communicate with the controller. In contrast, the subflow monitoring in Pyretic requires the visibility of every matching subflows. Thread *ovs-vswitchd* needs to forward the matching packets to the controller, resulting in CPU consumption.

Next we measure the control plane overhead. Figure 7 depicts the number of Packet_In messages received at the controller (curves with the left Y-axis) and CPU utilization of the controller (bars with the right Y-axis) against the number of monitored src/dst pairs. Since UMON has no interactions with the controller, the CPU utilization and Packet_In count

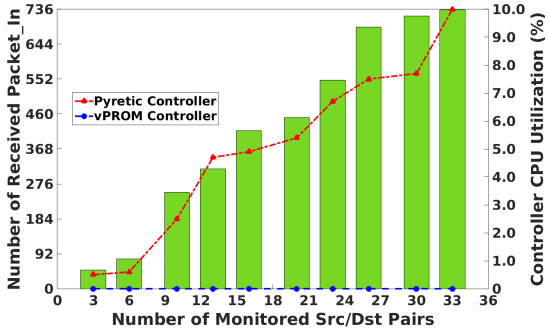


Fig. 7. Overhead in Control Plane

remain at zero throughout the experiment. For Pyretic, the number of Packet_In messages increases when more pairs are monitored. The CPU utilization of the controller also increases proportionally to the number of received Packet_In messages.

To further compare the scalability of both solutions, we increase the number of monitored pairs to stress both the vSwitch and the controller. The test results are shown in Figure 8. In this figure, we present two sets of results. First,

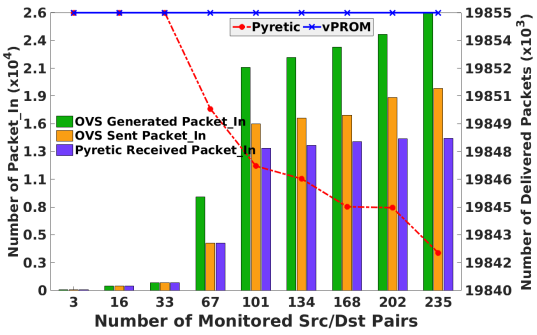


Fig. 8. Stress Test Results

we plot the number of delivered data packets by vPROM and Pyretic (curves with Y-axis on the right). The trace used for the evaluation contains 19,855,388 packets in total. The curves of delivered packets show that vPROM is able to deliver all packets as more and more pairs are monitored, while Pyretic starts to suffer from the packet losses when the number of monitored pairs is greater than 67 pairs.

We investigate the cause of the data path packet losses in Pyrectic. For that we examine the Pyretic's control plane overhead. We plot the number of Packet_In messages generated by Open vSwitch, sent by Open vSwitch, and received by the Pyretic (see bar charts with Y-axis on the left in Figure 8). We observe that Pyretic starts to lose Packet_In messages at both vSwitch and Pyretic (which runs on top of the SDN controller) when the number of monitored pairs surpasses 67 pairs. The difference between the number of OVS generated Packet_In messages and the number of Packet_In messages being sent out indicates the packet loss inside the vSwitch, which is due to the overflow of the Packet_In queue inside the vSwitch.

Chen, Songqing; Guo, Yang; Hao, Fang; Lakshman, T.V.; Montgomery, Douglas; Sriram, Kotikalapudi; Wang, An.
"vPROM: vSwitch Enhanced Programmable Measurement in SDN."
Paper presented at IEEE 25th International Conference on Network Protocols (ICNP), 2017, Toronto, ON, Canada. October 10, 2017 - October 13, 2017.

Meanwhile, the difference between the number of sent-out Packet_In and the number of Packet_In received by the Pyretic application indicates the packet loss at the controller. The controller employs the event queue for dispatching various events, such as Packet_In event, to the applications. The losses are due to the overflow of event queues maintained by Pyretic [1]. The results show that the frequent communications between the vSwitch and the controller greatly degrade the performance of both the vSwitch and the controller. Due to the Packet_In message loss, the Pyretic monitor application can not offer accurate subflow monitoring results. vPROM addresses the problem by instrumenting the vSwitch and localizing the subflow monitoring task.

B. Performance of DPDK UMON

As a software based switch, the performance of vSwitch is important. UMON imposes extra overhead on vSwitch for the monitoring activities. The initial work [7] showed that the overhead from UMON is tolerable. With the recent availability of DPDK [23], further speedup of the packet processing is made possible. DPDK is a set of libraries and drivers that map the memory from the NIC directly to user space, which improves the packet processing performance by up to ten times. DPDK has been adopted by vSwitch to improve its performance. In this section, we examine the performance of UMON in a DPDK accelerated vSwitch.

We use MoonGen [24] for playing back the trace since it can utilize the hardware features of the Intel NICs to enable the sub-microsecond packet time-stamping. The packet generation logic is controlled by *Lua* [25] scripts for flexibility. The trace is split into two groups: background traffic and monitored traffic. Different traffic groups are sent out through different queues, and the monitored traffic is time-stamped at outgoing port and incoming port to measure latency. Thus the measured latency includes the delay on the wire (from the packet generator/sink to the vSwitch, and from vSwitch back to the packet generator/sink) and the delay incurred at the UMON vSwitch. It is an upper bound of the UMON vSwitch delay. We play back the trace at the line rate so as to fill up the 10G pipe.

The results are shown in Figure 9. We vary the number of monitoring pairs and the *revalidator* thread poll intervals to examine their impact. The poll interval determines the frequency the measurement results are reported. There are ten groups of data, each with a different poll interval. The delay increases as the number of monitored pairs increases and as the poll interval decreases. The more monitored pairs, the more data needs to be fetched from the data path. The smaller poll interval causes the *revalidator* thread to be called more frequently. Though the increasing delay suggests degradation of the switch performance, the delay is mostly small. Monitoring 336 source-destination pairs only incurs less than 0.04 millisecond delay.

Next, we investigate the CPU utilization of the *handler* thread and the *revalidator* thread. The results are shown in Fig. 10 and Fig. 11. For each thread, the CPU utilization increases proportionally to the number of monitored pairs and decreases with the increasing poll interval. In addition, comparing results in this figure with those in the Figure 6 when 33 pairs of src/dst are monitored, the CPU utilization for the DPDK accelerated UMON is about 3-4.5 times higher. Such an increase is explained in [26] because the default DPDK configuration uses only one core to run all the standard OVS threads (e.g., main process, handler, revalidator) to avoid wasted resources.

C. vPROM-GUARD attack detection

We use two data traces containing verified attacks to evaluate the effectiveness of vPROM-GUARD. For the SYN Flood attack, we use *Endpoint Traffic* collected from three different deployment points in NUST SEECS labs [27]. In this data-set, eight ports on two hosts are under known SYN Flood attacks. The attacking rate varies from 10 pkts/second to 1000 pkts/second and the average background traffic rate varies between 200 to 650 pkts/second. There are in total 2325 hosts in this data trace. For the port scanning attack, we use the trace collected by Mawilab [28]. In this data-set, one horizontal scanning attack and three vertical scanning attacks are known. We use Emulab in our lab to conduct the experiments. vRPOM-GUARD runs on the vPROM framework at one machine, and two UMON switches and a CATE capable switch controlled by vPROM-GUARD are running on another machine. The CATE capable switch emulates the gateway switch; While the two UMON machines emulate the vSwitches at two host machines in a data center, each hosting about 20 IPs with some of IPs being under attack. We set the polling intervals for all the detection to be one second, and t to be 0.4 and T to be 1. For the CATE scheme, we set the threshold for the large flow detection at 0.05, i.e., a flow is deemed to be large if it is more than 5% of overall traffic rate. The detected big flows are reported to vPROM-GUARD every one second.

vPROM-GUARD manages to detect all attacks in the data traces. Fig. 12(a) shows the eight SYN flood attack detection time. Attacks target the two hosts, with IP address of $87.51.34.132$ (top of Fig. 12(a)) and $69.63.178.11$ (bottom of Fig. 12(a)) with different port ids, as shown in the Y-axis. The horizontal bar indicates the starting and finishing time of the attack, with the vertical line indicating the moment at which the CATE issues a big-flow warning. Once a big flow is detected, a monitoring rule collecting TCP flag packet counts is installed to validate if the large flow is a SYN flood attack. The dot indicates the moment at which the vPROM-GUARD actually validates the attack as SYN flood attack. The lightweight design and implementation of CATE enables vPROM-GUARD to detect such attacks quite efficiently. The average detection time is about 3 seconds, including the attack validation time.

For the port-scan attacks in our trace, they do not generate big enough traffic flows to be detected by CATE. As a result, change-point monitoring and subflow collections are required for the detection. Fig. 12(b) shows the detection time of the vertical and horizontal port-scan attacks. The detection

Chen, Songqing; Guo, Yang; Hao, Fang; Lakshman, T.V.; Montgomery, Douglas; Sriram, Kotikalapudi; Wang, An.
"vPROM: vSwitch Enhanced Programmable Measurement in SDN."
Paper presented at IEEE 25th International Conference on Network Protocols (ICNP), 2017, Toronto, ON, Canada. October 10, 2017 - October 13, 2017.

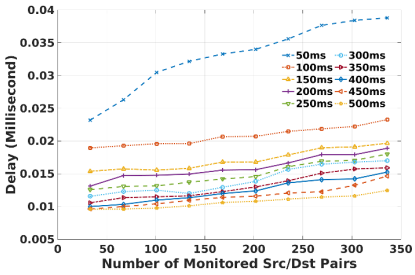


Fig. 9. Packet delay of DPDK UMON

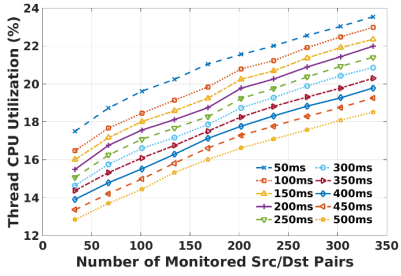


Fig. 10. Revalidator CPU utilization of DPDK UMON

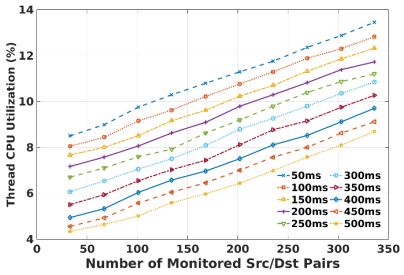
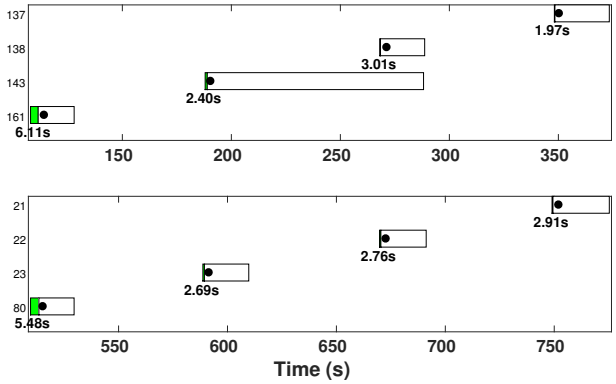
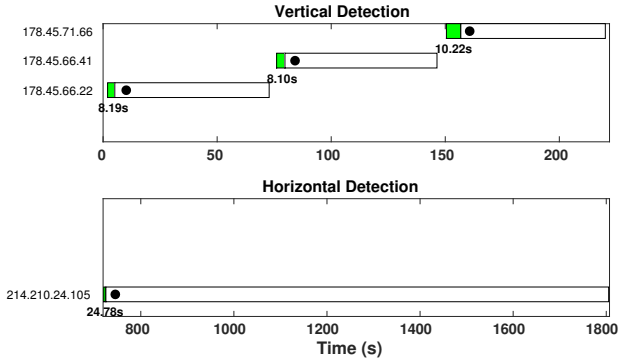


Fig. 11. Handler CPU utilization of DPDK UMON



(a) SYN Flood Attack



(b) Port scanning Attack

Fig. 12. Attack detection in vPROM-GUARD

time for vertical port-scan attack is about 10 seconds. The horizontal port scan attack detection takes about 25 seconds. The horizontal port-scan spreads the attack traffic among multiple IPs, hence a smaller attacking rate for one IP and takes longer time to detect.

VI. Conclusions and Future Work

In this paper, we present the design and implementation of vPROM, a vSwitch enhanced programmable measurement framework that allows users to program the network measurement and control applications. vPROM uses instrumented Open vSwitches (UMON) as the measurement points, and augments the OpenFlow API to OpenFlow+ so that the UMONes can be directly controlled by the applications via the SDN controller. In vPROM, we also extend the Pyretic programming language and run-time system to Pyretic+ and build a controller client in order to support and automate the programmable measurement. To demonstrate its usefulness, we also build the vPROM-GUARD, a DDoS and port-scan attack detection application that demonstrates the major features of vPROM. Performance evaluations and comparisons with other approaches show the advantages of vPROM. Moving forward, we are building more vPROM applications and investigating how to use both UMONes and physical SDN switches as the monitoring points simultaneously. In addition, we are studying to employ the behavior based anomaly detection as the coarse granularity monitoring cues.

VII. Acknowledgement

We appreciate constructive comments from anonymous referees. This work is partially supported by a NIST grant 70NANB16H166, an ARO grant W911NF-15-1-0262, and a NSF grant CNS-1524462. Yang Guo wishes to thank Nien-Fan Zhang (NIST) for helpful discussions on Cumulative Sum Chart Method, and thank Kevin Mills (NIST) and Vladimir Marbukh (NIST) for reviewing the paper and offering helpful comments.

References

[1] J. Reich, C. Monsanto, N. Foster, J. Rexford, and D. Walker, "Modular SDN Programming with Pyretic," *USENIX ;login*, 2013.

[2] S. Donovan and N. Feamster, "Intentional network monitoring: Finding the needle without capturing the haystack," in *Proceedings of the 13th ACM Workshop on Hot Topics in Networks*, 2014.

[3] N. Foster, R. Harrison, M. J. Freedman, C. Monsanto, J. Rexford, A. Story, and D. Walker, "Frenetic: A network programming language," in *ICFP*, 2011.

[4] A. Gupta, R. MacDavid, R. Birkner, M. Canini, N. Feamster, J. Rexford, and L. Vanbever, "An industrial-scale software defined internet exchange point," in *NSDI*, 2016.

[5] P. Bosshart, D. Daly, G. Gibb, M. Izzard, N. McKeown, J. Rexford, C. Schlesinger, D. Talayco, A. Vahdat, G. Varghese, and D. Walker, "P4: Programming protocol-independent packet processors," *SIGCOMM Comput. Commun. Rev.*, 2014.

[6] "Open vSwitch," http://openvswitch.org/.

[7] A. Wang, Y. Guo, F. Hao, T. Lakshman, and S. Chen, "Umon: Flexible and fine grained traffic monitoring in open vswitch," in *Proceedings of the 10th ACM International on Conference on Emerging Networking Experiments and Technologies*, 2015.

Chen, Songqing; Guo, Yang; Hao, Fang; Lakshman, T.V.; Montgomery, Douglas; Sriram, Kotikalapudi; Wang, An.
"vPROM: vSwitch Enhanced Programmable Measurement in SDN."
Paper presented at IEEE 25th International Conference on Network Protocols (ICNP), 2017, Toronto, ON, Canada. October 10, 2017 - October 13, 2017.

[8] N. Foster, M. Freedman, A. Guha, R. Harrison, N. P. Katta, C. Monsanto, J. Reich, M. Reitblatt, J. Rexford, C. Schlesinger, A. Story, and D. Walker, "Languages for software-defined networks," *IEEE Communications Magazine*, 2013.
[9] C. J. Anderson, N. Foster, A. Guha, J.-B. Jeannin, D. Kozen, C. Schlesinger, and D. Walker, "Netkat: Semantic foundations for networks," *SIGPLAN Not.*
[10] M. T. Arashloo, Y. Koral, M. Greenberg, J. Rexford, and D. Walker, "Snap: Stateful network-wide abstractions for packet processing," in *SIGCOMM*, 2016.
[11] H. Kim, J. Reich, A. Gupta, M. Shahbaz, N. Feamster, and R. Clark, "Kinetic: Verifiable dynamic network control," in *12th USENIX Symposium on Networked Systems Design and Implementation (NSDI 15)*, 2015.
[12] S. Narayana, J. Rexford, and D. Walker, "Compiling path queries in software-defined networks," in *Proceedings of the Third Workshop on Hot Topics in Software Defined Networking*, 2014.
[13] A. Gupta, L. Vanbever, M. Shahbaz, S. P. Donovan, B. Schlinker, N. Feamster, J. Rexford, S. Shenker, R. Clark, and E. Katz-Bassett, "SDX: A Software Defined Internet Exchange," in *Proceedings of the 2014 ACM Conference on SIGCOMM*, 2014.
[14] M. Moshref, M. Yu, R. Govindan, and A. Vahdat, "Trumpet: Timely and precise triggers in data centers," in *SIGCOMM*, 2016.
[15] "OpenFlow Switch Specification," https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-spec-v1.4.0.pdf.
[16] F. Hao, M. Kodialam, T. Lakshman, and H. Zhang, "Fast, memory-efficient traffic estimation by coincidence counting," in *INFOCOM*, 2005.
[17] "Pyretic Tutorial," https://github.com/frenetic-lang/pyretic/wiki/Query-Policies.
[18] H. Wang, D. Zhang, and K. G. Shin, "Change-point monitoring for the detection of dos attacks," *IEEE Trans. Dependable Secur. Comput.*, 2004.
[19] B. Brodsky and B. Darkhovsky, "Nonparametric methods in change-point problems," *Kluwer Academic Publishers*, 1993.
[20] P. Winkel and N. Zhang, "Statistical development of quality in medicine," *Wiley Publishers*, 2007.
[21] AppNeta, "Tcpreplay," http://tcpreplay.synfin.net/wiki/tcpreplay.
[22] T. Benson, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 2010.
[23] Intel Corporation, "Data Plane Development Kit," http://dpdk.org/.
[24] P. Emmerich, S. Gallenmüller, D. Raumer, F. Wohlfart, and G. Carle, "Moongen: A scriptable high-speed packet generator," in *Proceedings of the 2015 ACM Conference on Internet Measurement Conference*, 2015.
[25] "LuaJIT," http://luajit.org/.
[26] Gray Mark, "INSTALL.DPDK.md: Clarify DPDK arguments," http://openvswitch.org/pipermail/dev/2015-December/063137.html.
[27] S. Ali, I. U. Haq, S. Rizvi, N. Rasheed, U. Sarfraz, S. A. Khayam, and F. Mirza, "On mitigating sampling-induced accuracy loss in traffic anomaly detection systems," *ACM SIGCOMM Computer Communication Review*, 2010.
[28] C. Sony, "Traffic data repository at the wide project," in *Proceedings of USENIX 2000 Annual Technical Conference: FREENIX Track*, 2000.

Chen, Songqing; Guo, Yang; Hao, Fang; Lakshman, T.V.; Montgomery, Douglas; Sriram, Kotikalapudi; Wang, An.
"vPROM: vSwitch Enhanced Programmable Measurement in SDN."
Paper presented at IEEE 25th International Conference on Network Protocols (ICNP), 2017, Toronto, ON, Canada. October 10, 2017 - October 13, 2017.

Testing IoT Systems

Jeff Voas
Computer Secuirty Division
NIST
Gaithersburg, USA
jeff.voas@nist.gov

Rick Kuhn
Computer Secuirty Division
NIST
Gaithersburg, USA
rkuhn@nist.gov

Phil Laplante
Great Valley School of Graduate and Professional Studies
Penn State
Malvern, USA
plaplante@psu.edu

***Abstract*— This article presents challenges and solutions to testing systems based on the underlying products and services commonly referred to as the Internet of 'things' (IoT).**

Keywords—Internet of Things, testing, Domain Range Ratio

I. INTRODUCTION

Can you test the Internet? No - it is unbounded. Can you test the Internet of Things (IoT)? Same answer.

You could test sub-nets of the IoT and other bounded components of it. The Internet and its 'things' are only boundable for mere instants in time, therefore testing is problematic. Testing systems-at-rest is easier than testing systems reorganizing themselves in real-time and at massive scale. The Internet at time zero is different than the Internet at time zero + *x*, where *x* is a millisecond.

We argue that testing the Internet and the IoT is not feasible. We further argue that we can use the concept of a Network of 'Things' (NoT) [1] to create testing schemes that are practical. This definition allows for measurement, and allows for one NoT to be compared to another. In addition, this definition allows for estimating the *testability*[1] of a specific NoT, which when said slightly differently, asks the question: is this NoT *testable*, meaning is testing even worth the effort you will put into it?

The general concept behind the term Network of 'Things' involves communication, computation, sensing, and actuation. These are simple ideas that have existed in distributed computing for years. But what makes IoT and NoT different from previous large-scale distributed computing systems is scale, heterogeneity, data integrity, sensing, and possible non-ownership of the assets in a purposed and proprietary NoT. By 'non-ownership' of assets, we include leased cloud services, leased data from vendor sensors, leased wireless communications, leased hardware and 3rd party software, and so on. This paper uses the concept of a NoT as the *entity under test.*

[1]Testability here refers to the likelihood that defects can be discovered during testing [3]; testability is clearly a function of what type of testing is occurring and how test cases are selected.

II. UNDERSTANDING A NETWORK OF 'THINGS'

To address such concerns, the National Institute of Standards (NIST) Special Publication 800-183 [1] offered a scientific foundation to describe the underpinnings of a Network of 'things.' It breaks these four activities into core distributed system components termed "primitives." The document then defines a simple class of "elements" that allow for the foreshadowing of the trustworthiness of systems built from IoT-based components, services, and commercial products. (NIST has not released a specific definition for IoT at this time).

The primitives proposed in [1] are: 1) Sensor, 2) Aggregator, 3) Communication channel, 4) eUtility, and 5) Decision trigger. Here are their descriptions from the document:

1. A *sensor* is an electronic utility that digitally measures physical properties (e.g. temperature, acceleration, weight, sound, etc.) and outputs raw data.

2. An *aggregator* is a software implementation based on mathematical function(s) that transforms/consolidates groups of raw data into intermediate data.

3. A *communication channel* is a medium by which the data is transmitted (e.g., physical via USB, wireless, wired, verbal, etc.) between sensor, aggregator, communication channel, decision trigger, or eUtility.

Kuhn, David; Laplante, Phil; Voas, Jeffrey.
"Testing IoT Systems."
Paper presented at 12th IEEE International Symposium on Service-Oriented System Engineering, Bamberg, Germany. March 26, 2018 - March 29, 2018.

4. An *eUtility* (external utility) is a software or hardware product or service, providing computing power that aggregators will likely network of 'things' have.

5. A *decision trigger* creates the final result(s) needed to satisfy the purpose, specification, and requirements of a specific network of 'things.'

III. Testability of a Network of 'Things'

A specific, purposed network of 'things' is likely to have a dynamic and rapidly changing dataflow and workflow. It will likely have numerous inputs from a variety of sources. This will, in turn, create a massive internal state space of data states created throughout the computational workflow of a NoT, and a vast number of potential interactions among components.

One way to think about the testing problems from this state-space explosion is by trying to answer the question: Do NoTs of large scale have an impact on testability? To answer that, we will first look at the Domain Range Ratio (DRR) [2][3] metric, first proposed in the 1990s by Voas and Miller.

We contend that the DRR metric addresses the inherent problem of testing networks that employ IoT-based components and services, e.g. clouds. The DRR is simply the cardinality of the set of all possible test cases for that system divided by the cardinality of the set of all possible outputs. A fundamental issue is that a particular network of 'things' is likely to process large amounts of data for the purpose of making rather limited output decisions, such as 'actuate' or 'do not actuate.' This situation makes it difficult to observe internal failures due to corrupted internal data during testing time.

For example, Fig. 1 represents a simplistic NoT purposed to buy or not buy a certain stock. (Figure 1 is not intended to represent real NoTs, but rather to highlight the primitives in [1].) This NoT has 15 sensors clustered into 3 groups, 5 aggregators, and 3 eUtilities (2 clouds, and 1 laptop). Note that Sensor 5 and Sensor 6 are blue, illustrating that they are sending out data of suspicious integrity. This NoT has 22 communication channels that carry the data that eventually gets aggregated and then fed into the NoT's decision trigger. The decision trigger is binary –a value of '1' means buy the stock, a value of '0' means do not. Stated simply, the combinatorics of faults and internal failures that can go wrong with 15 sensors, 5 aggregators, 3 eUtilities, and 22 communications channels is quite large.

Now assume that each binary value of the decision trigger variable is obtained approximately 50% of the time. Because of this minimal output space size, a fair coin toss also has a 50-50 chance of providing a correct output for any given input. Hence building a system that generates a random '1' or '0' result via a coin flip is equivalent to and cheaper than building this complex and expensive NoT. Worse, consider the scenario where '1' and '0' are not evenly distributed, e.g., the specification states that for 1 million unique test cases only 10 should produce a '1' and the other 999,990 should produce a '0'. One could build a NoT to compute this function or one could write a piece of code that just says: **for all inputs output ('0')**. This incorrect code is still 99.999 reliable, and it would be nearly impossible to discover the defect in the code with a handful of random tests sampled from the 1 million. In short, random testing here has a minimal probability of detecting this fault because each test case has a very low probability of revealing the defective logic due to the tiny output space, and its probability density function for each output. Note also that this argument likely applies to aggregators – if an aggregator is fed much sensor data and reduces that data to a single output value, in particular a binary value, this problem is the same.

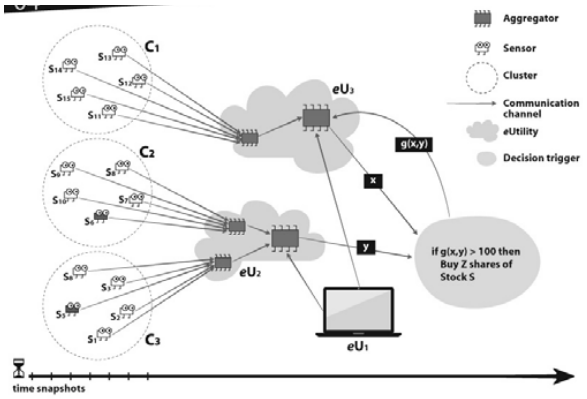


Fig. 1. A Simple NoT With a Feed-Back Loop

Furthermore, in this system, corrupted data (regardless of the reason for the corruption) can travel through any of the communication channels – it can originate from eUtilities, aggregators, sensors, and even communication channels. Given the many data-related events that happen before a decision trigger is executed, how does a system-level test of a NoT give assurance and confidence to the prior event's trustworthiness? This is the same question that resulted in the concept of software unit testing. We apply the same principles here to NoTs. We will discuss this situation further using assertions.

IV. The Oracle Problem for Networks 'Things'

The traditional software testing problem of having access to a usable oracle can be paraphrased as follows: if you do not know if an output is correct after a test is performed, what is the point of testing? Further, this problem can be subdivided into two problems: (1) a defective oracle, and (2) not having an oracle at all.

For NoTs, the oracle problem is exacerbated by this: it is unlikely that the intended functionality of a general-purpose, short-lived NoT will remain static long enough for an oracle to be built. (Hopefully for most security-critical and safety-critical properties of NoTs, this concern can be avoided.) This problem exists because NoTs offer more *extensibility* and *malleability* of the intended functionality than in previous distributed systems. For example, the sensors, eUtilities, and communication channels can all be quickly swapped out and replacements swapped in; and the algorithms and the software (in the aggregators, communication channels, and decision trigger) can be continuously tweaked when the purpose of a NoT changes.

Kuhn, David; Laplante, Phil; Voas, Jeffrey.
"Testing IoT Systems."
Paper presented at 12th IEEE International Symposium on Service-Oriented System Engineering, Bamberg, Germany. March 26, 2018 - March 29, 2018.

The ability to "modify-on-the-fly" a NoT is one of the advantages of this advancement in distributed computing and control but is problematic for testing using oracles. Hence, testing efficiency is critical. Using the Template

V. Assertions

Three conditions are necessary for software to fail: (1) a fault is executed, (2) a defective internal data state is created, and (3) the defective state propagates causing incorrect output (failure). By testing internal data states using *assertions*, we observe whether (2) occurs if the assertion is correctly coded and placed.

Assertions are internal self-tests that probe either: (1) the output of a component or function, or (2) data states that exist anytime during computation. Assertions increase testability by increasing the size of the range [4][5]. The goal is not to test the functionality of eUtilities, aggregators, sensors, and communication channels, but rather to test the data they produce.

To do this, the notion of wrappers connected to interfaces (between primitives) is possibly the easiest implementation approach for building assertions. For example, in the interface between a sensor and an aggregator, insert a wrapper on the data leaving the sensor before it rides on the communication channel, or insert a wrapper before it leaves the communication channel and enters the aggregator, or both (if there is concern something might go wrong during transmission). Tests at the interface points offer two advantages for NoTs: (1) this testing can be done even if most of the primitive 'things' are black-box entities, and (2) no access to any software or algorithms is required.

VI. Combinatorial Testing

Among the unique challenges of testing NoTs, one of the most significant is the potential for an enormous number of interactions. Instead of two, or a few, components sending and receiving data, NoTs may have 10s or 100s of nodes interacting. While internet e-commerce and information systems include thousands of nodes, interactions are typically client-server. NoTs, in contrast, may require cooperation among a much larger set of nodes to meet their design objectives. The difficulty of testing these networks has led to recognition of the need for combinatorial test methods [12][13], which are designed specifically for testing complex interactions.

Software faults may involve one or multiple factors interacting. For example, a device failure that occurs only when *pressure < 10 AND volume > 300 AND velocity = 5* (where *pressure*, *volume* and *velocity* are integer variables) is a 3-way interaction fault. Interaction faults remain dormant until the particular combination of values is encountered in practice. Combinatorial testing (CT) is an increasingly popular method for reducing the cost of finding such complex faults. The empirical basis for CT's effectiveness was shown in a series of NIST studies [6][7][8][9] that demonstrated the following: most faults involve a single parameter or two parameters; and progressively fewer interaction faults involve 3, 4, 5, and 6 parameters (a fault involving more than six parameters has not been seen) [8]. This empirical finding, referred to as the *interaction rule*, has important implications for software testing, because it means that compressing t-way combinations into a small number of tests can provide more efficient fault detection than conventional methods.

Matrices known as *covering arrays* [10] are used to produce tests covering t-way combinations of values, for some specified level of $t \geq 2$; e.g., if $t = 3$, then the covering array contains all 3-way combinations of variable values. The key property of a covering array is that it includes all *t*-way combinations of values at least once, and algorithms developed in the past 10-15 years can efficiently generate test arrays for high-order interactions, typically up to *t=6*. (In the past, nearly all combinatorial testing had been limited to pair-wise, or *t=2*.) For example, suppose we want to test a module for a lighted text display, which might be controlled using an Arduino board or similar small system. The function allows 10 effect settings for enhancing the text, each of which has two possible settings: flashing (on, off), size (large, small), three light colors that can each be on or off to produce different effects, a "glow" effect (on, off), etc. Testing all combinations would require 210, or 1024 tests. The 10 effects are labeled A through J. If we represent "on" as 1, and "off" as 0, then the array in Fig. 2 provides a compact test set that covers all 3-way combinations.

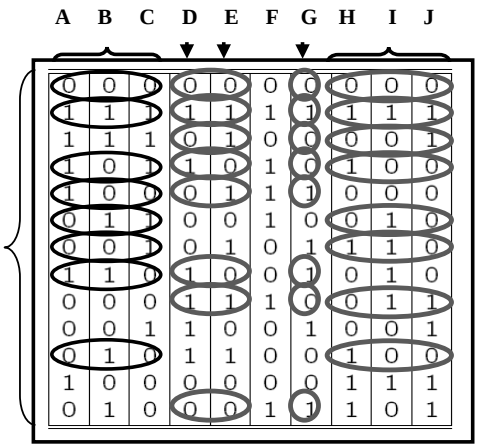


Fig. 2. 3-way covering array

Fig. 2 shows a 3-way covering array for 10 variables with two values each, where each row represents a test and each column specifies values for a variable setting. The interesting property of this array is that any three columns contain all eight possible values for three binary variables. For example, taking columns D, E, and G, we can see that all eight possible 3-way combinations (000, 001, 010, 011, 100, 101, 110, 111) occur somewhere in the three columns together. In fact, any combination of three columns chosen in any order will also contain all eight possible values. Collectively, therefore, this set of tests will exercise all 3-way combinations of input values in only 13 tests, as compared with 1,024 for exhaustive coverage. Similar arrays can be generated to cover up to all 6-way combinations. The larger the problem, the greater the improvement over exhaustive testing, because for a given

Kuhn, David; Laplante, Phil; Voas, Jeffrey.
"Testing IoT Systems."
Paper presented at 12th IEEE International Symposium on Service-Oriented System Engineering, Bamberg, Germany. March 26, 2018 - March 29, 2018.

interaction strength, the number of rows (tests) in a *t*-way covering array increases with *log n*, for *n* parameters, while the exhaustive test set size of course increases exponentially. (Note that covering arrays are not restricted to binary variables; these are used here to simplify the presentation.) For example, with 34 on-off switches, there are approximately 17 billion possible combinations, but all 3-way settings can be covered with only 33 tests, and all 4-way combinations with 85 tests.

An extensive body of empirical work shows that these methods are highly cost effective. In practical applications, combinatorial testing has been shown to provide significant advantages, with reduced cost and greater fault detection [10][11]. Fortunately, these methods can solve some of the unique problems of testing NoTs.

VII. Applying Combinatorial Testing to the Example

How can we adequately test NoTs, given the constraints identified by the DRR for these systems? Consider again the example for which 10 inputs from an input domain of 10 million produce a '1', with the rest producing '0'. For any computed result, there must be some specification that defines the conditions under which each possible result is produced. In general, these conditions will be specified by some logical predicate, especially for NoTs, where decision predicates receive inputs from various sensors and generated values elsewhere in the system. One such example is shown in figure 1. Here we have the decision trigger *"if g(x,y) > 100* then buy Z shares of stock S". In this example, an expression of two values, *x* and *y*, is used in the decision; we assume different combinations of *x* and *y* values not shown in the figure may generate different results. For decision triggers in the network, many additional expressions with different combinations of values may be used in decision triggers. It is easy to produce positive tests for this example: simply specify values to make the expression *g(x,y) > 100* true. But what if we want to ensure that "buy Z shares of stock S" action is not triggered under some other conditions? If the DRR tells us that there is only a small portion of the input space for which this action is correct, how will it be possible to test the huge portion of inputs for which the result should not be to buy Z shares of stock S?

One approach to achieving this assurance is to use the pseudo exhaustive test method described in [14], where we have a small set of possible outputs. This method produces two test arrays for each possible output, or class of outputs. One array includes tests for each condition where a particular result should be produced. For example, suppose a decision trigger is "*x+y > 200 && x >100 || x < 100 && y > 500 || y > 1000* then *R1*", where *R1* is a Boolean output corresponding to some action that the system performs. Only three positive tests are needed, one for each conjunct within the expression. A more difficult challenge is showing that the code implementing the expression does not generate the result for some combination of variables inappropriately. This expression is in disjunctive normal form, where each term contains at most two variables, so it is in 2-DNF. By generating a 2-way covering array of values for *x* and *y*, but excluding the positive cases, we have a test set with all possible 2-way combinations of values where *R1* should not be produced. Thus we address the oracle problem by verifying that results for tests in each array are equivalent, rather than specifying a set of inputs and determining the output specifically for each one. In the first array, we should see *R1* as the output for each test, and the second array we should not see *R1*. Because we verify positive and negative results for each output *Ri, (i=1,2,3,…)*, we have a sound and complete set of tests without the conventional test oracle problem of computing outputs for each set of inputs.

To see the power of this method, consider the example introduced previously, with roughly one million possible inputs, where 10 produce an output of '1' and all others produce an output of '0'. This could occur with a system of 20 Boolean parameters, for example, resulting in 220 = 1,048, 576 possible inputs. From the specification, we derive the conditions under which '1' is produced, in the form of *if-then-else* rules or a decision tree. Transforming these rules into *k*-DNF form, we produce a set of conjunctions that result in the '1' output. Suppose that the 10 conditions that result in an output of '1' contain at most three Boolean literals (e.g., *x && ~a && y*). It is easy to produce tests to verify this output for each of the 10 conditions, but how can we ensure that none of the other 1,048, 566 inputs will produce a '1' instead of the correct value of '0'? Surprisingly, we can verify this for all 3-way combinations of inputs with only 28 tests, by generating a covering array of all 3-way combinations, excluding the 10 conditions that should produce '1' [14]. The test arrays will also catch a large proportion of combinations with more than three Booleans, or we can generate arrays up to 6-way with less than 400 tests. This method is not restricted to Boolean inputs, but may be applied to complex conditionals as well, and as a result is especially well suited to testing complex conditionals in decision triggers.

Given a formal specification of the conditions for each decision trigger, the test arrays described above can be produced mechanically, but many tests will still be needed. Thus, even though a conventional test oracle is not needed (because each of the two arrays should produce the same result), the large number of tests may be prohibitive in some applications. The DRR calculation can be used to identify the most difficult to test interface points, helping to establish priorities and allocate testing resources.

VIII. Summary

We believe that because of the necessary role of decision triggers, specifically purposed NoTs have testability concerns. We explained how the testing oracle problem applies to NoTs as well as other distributed systems, and that the problem may be worse compared with other complex IT systems due to "leased assets." And finally, we described applications of combinatorial testing and the domain range ratio, and how these methods provide a practical approach to IoT testing complexities.

We hope this review of challenges and potential solutions will offer new insights into how to more efficiently test NoTs.

References

[1] J. Voas, "Networks of 'Things'", *NIST Special Publication SP 800-183* (July 2016), http://dx.doi.org/10.6028/NIST.SP.800-183.

Kuhn, David; Laplante, Phil; Voas, Jeffrey.
"Testing IoT Systems."
Paper presented at 12th IEEE International Symposium on Service-Oriented System Engineering, Bamberg, Germany. March 26, 2018 - March 29, 2018.

[2] Voas, J.M. and Miller, K.W., "Semantic metrics for software testability", The Journal of Systems and Software, vol. 20, no. 3, March 1993, pp. 207-216.

[3] J. M. Voas and K. W. Miller, "Software testability: the new verification", *IEEE Software*, vol. 12, no. 3, March 1995, pp. 17-28.

[4] J. Voas, "Software Testability Measurement for Intelligent Assertion Placement," *Software Quality Journal*, vol. 6, no. 4, December 1997, pp. 327-335.

[5] J. Voas and L. Kassab, "Using Assertions to Make Untestable Software More Testable," *Software Quality Professional*, vol. 1, no. 4, September 1999, pp. 31-40.

[6] D. R. Wallace and D. R. Kuhn "Failure Modes in Medical Device Software: an Analysis of 15 Years of Recall Data," *International Journal of Reliability, Quality, and Safety Engineering*, vol. 8, no. 4, 2001, pp. 351-371.

[7] D. R. Kuhn and M. J. Reilly, "An Investigation of the Applicability of Design of Experiments to Software Testing," *27th NASA/IEEE Software Engineering Workshop*, NASA Goddard Space Flight Center, 4-6 December, 2002, pp. 91-95 .

[8] D. R. Kuhn, D. R Wallace and A. Gallo, "Software Fault Interactions and Implications for Software Testing," *IEEE Transactions on Software Engineering*, vol 30, no. 6, 2004, pp. 418-421.

[9] D. R. Kuhn and V. Okum, "Pseudo-exhaustive testing for software," *30th Annual IEEE/NASA Software Engineering* , 2006, pp. 153-158.

[10] Y. Lei, R. Kacker, R, Kuhn, V. Okun and J Lawrence, "IPOG: a General Strategy for t-way Software Testing," *14th Annual IEEE International Conference and Workshops on the Engineering of Computer-Based Systems*, Tucson, Arizona, March 26-29, 2007, pp. 549–556.

[11] J. D. Hagar, T. L. Wissink, D. R. Kuhn, D. R. and R. N. Kacker, "Introducing combinatorial testing in a large organization," *Computer*, vol. 48, no. 4, April 2015, pp. 64-72.

[12] A.H. Patil, N. Goveas, and K. Rangarajan, "Test Suite Design Methodology Using Combinatorial Approach for Internet of Things Operating Systems," *J. Software Eng.Applications*, vol. 8, no. 7, 2015, p. 303.

[13] G. Dhadyalla, N. Kumari, and T. Snell, "Combinatorial Testing for an Automotive Hybrid Electric Vehicle Control System: A Case Study," Proc. IEEE 7th Int'l Conf. Software Testing, Verification and Validation Workshops (ICSTW 14), 2014, pp. 51–57.

[14] D. R. Kuhn, V. Hu, D. Ferraiolo, R. Kacker, R. and Y. Lei, "Pseudo-exhaustive Testing of Attribute Based Access Control Rule," *5th International Workshop on Combinatorial Testing*, 2016, pp.1-10.

Kuhn, David; Laplante, Phil; Voas, Jeffrey.
"Testing IoT Systems."
Paper presented at 12th IEEE International Symposium on Service-Oriented System Engineering, Bamberg, Germany. March 26, 2018 - March 29, 2018.

FATIGUE CRACK GROWTH RATES OF API X70 PIPELINE STEELS IN PRESSURIZED HYDROGEN GAS COMPARED WITH AN X52 PIPELINE IN HYDROGEN SERVICE*

ELIZABETH DREXLER
NIST
Boulder, CO, USA

ANDREW SLIFKA
NIST
Boulder, CO, USA

ROBERT AMARO
Colorado School of Mines
Golden, CO, USA

DAMIAN LAURIA
NIST
Boulder, CO, USA

ABSTRACT

The current ASME B31.12 code used to guide the design of hydrogen pipelines favors the use of API 5L X52, but is being modified to include higher strength steels, such as X70 to enable cost reductions without affecting safety. To provide a scientific basis for code modification, fatigue crack growth (FCG) tests were conducted on an X52 pipeline steel that is currently in service transporting hydrogen gas, as well as two X70 pipeline steels designed for natural gas. Compact tension specimens were tested in hydrogen gas pressurized to 5.5 MPa or 34 MPa. A comparison of these tests, conducted at a cyclic loading frequency of 1 Hz, shows that there is very little difference between the FCG rates of the base metal among the three steels at a given hydrogen pressure. All three metals exhibited some increase in FCG rate at a hydrogen pressure of 34 MPa compared with 5.5 MPa. Analysis of the data provide a rationale for allowing higher strength steels to be used for hydrogen gas transport. A recommendation was made to the ASME B31.12 Committee on Hydrogen Piping and Pipelines to allow higher-strength steels that is based on this and other data acquired at hydrogen pressures ≤ 21 MPa.

INTRODUCTION

Two federal agencies are tasked with ensuring that present and future hydrogen pipelines are safe and efficient. According to the US Department of Energy (DOE) [1], there are approximately 1500 miles of steel pipelines for the transportation of hydrogen gas in service today. That number is likely to increase, particularly if hydrogen fuel-cell cars become a popular alternative to gasoline-powered cars. Pipelines will be a necessary component to enable market penetration beyond the coastal US. The DOE has set a goal to reduce the cost of hydrogen delivery by the year 2020 from the production site to the point of use in consumer vehicles to <$2/gge (gallon of gasoline equivalent) for at least one delivery pathway [2]. This will help pave the way toward making hydrogen a competitive choice for powering cars and heating homes. Meanwhile, the mission of the Department of Transportation, Pipeline and Hazardous Materials Safety Administration (DOT/PHMSA) is to maintain the safety of pipelines transporting fuels in the US.

Amaro, Robert; Drexler, Elizabeth; Lauria, Damian; Slifka, Andrew; Sowards, Jeffrey.
"Fatigue Crack Growth Rates of API X70 Pipeline Steels in Pressurized Hydrogen Gas Compared with an X52 Pipeline in Hydrogen Service."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

Both agencies are working to develop the transmission infrastructure needed to support hydrogen fuel cell vehicles. Pipelines are the most cost-effective means of transporting hydrogen gas, but to attain the DOE goal for delivery cost, the expense of laying new pipelines must be further reduced. This can be achieved with higher strength steel. Fekete *et al*. [3] have described the savings generated by the use of steel with the grade API 5L X70 instead of API X52 for hydrogen pipelines. However, these savings can only be realized if the proposed material is as safe for operations as X52.

At the present time, the code used to design hydrogen pipelines is ASME B31.12 Hydrogen Piping and Pipelines [4]. This code states that API 5L X52 (PSL 2) grade steel can be used for hydrogen pipelines without additional testing. If a stronger grade of steel is desired, fracture toughness tests in pressurized hydrogen gas must be conducted. As few facilities have the capability of testing in pressurized hydrogen gas, X52 is virtually the only grade used in the US. This grade became the default choice in the code, because it exhibits minimal loss in ductility under monotonic loading in hydrogen gas.

However, it is rare for a pipeline to fail because it has exceeded its ultimate tensile strength, where the loss of ductility becomes critical. Safety factors ensure that stresses remain well below the yield strength of the steel. Rather, steel pipelines fail because fatigue cracks initiated by damage or flaws eventually propagate through the wall thickness of the pipe. If fatigue is the failure mechanism of concern, then limiting the choice of steels to X52 may not be the most effective means of designing safely operating hydrogen pipelines. For example, Cialone and Holbrook [5] found that for X42 pipeline steel there the fatigue crack growth rate (FCGR) increased by an order of magnitude for tests conducted in pressurized hydrogen and nitrogen. Other research groups have also found the FCGR of pipeline steels to increase by an order of magnitude or more when tested in pressurized hydrogen gas, as compared with those tested in air or an inert environment [6-8]. However, more data are needed to characterize the effect of strength on fatigue lifetimes in hydrogen.

In order to provide the ASME B31.12 Committee on Hydrogen Piping and Pipelines with a body of data from which to base a modification to the code, FCGR tests have been conducted that compare X52 steel from a currently operational hydrogen pipeline that was designed to the current B31.12 code, and two X70 steels from natural gas pipelines. Compact tension (CT) specimens were cyclically loaded in air and in hydrogen gas pressurized to either 5.5 MPa, a typical pressure at which to operate a hydrogen pipeline, or 34 MPa, the highest pressure currently considered.

MATERIALS AND METHODS

The base metals from two X70 pipeline steels that were designed for natural gas transmission and the modern X52 steel that is currently used in a hydrogen pipeline that went into operation in 2011 were tested to compare their FCGRs in pressurized hydrogen gas. The tests were conducted at a cyclic loading frequency of 1 Hz. Other researchers have found that in general there is an inverse relationship between the cyclic loading frequency and the hydrogen-assisted fatigue crack growth rate (HA-FCGR) for most structural alloys,

Amaro, Robert; Drexler, Elizabeth; Lauria, Damian; Slifka, Andrew; Sowards, Jeffrey.
"Fatigue Crack Growth Rates of API X70 Pipeline Steels in Pressurized Hydrogen Gas Compared with an X52 Pipeline in Hydrogen Service."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

particularly for frequencies at or above 1 Hz [9-13]. A limited number of tests were conducted at 0.1 Hz in order to determine the relationship between the FCGR for these steels and the loading frequency.

The chemical compositions of the three low-carbon, micro-alloyed steels are found in Table 1. The microstructures from near the mid-line of each steel are shown in Figure 1, and from optical microscopy were determined to be polygonal and acicular ferrite. There may be other constituents that are not resolvable without employing more advanced analytical techniques. Tensile data was acquired for each steel in air and in the transverse orientation, according to ASTM E8 [14]. The mean of those data and the dimensions of the pipes from which they came are shown in Table 2. Note that the X52 has a far higher yield strength than might be expected for an X52, although it meets the specification for API 5L X52 PSL 2, which has a minimum yield strength of 359 MPa (52 ksi) and a maximum yield strength of 531 MPa (77 ksi) [15].

The fatigue tests were conducted in accordance with ASTM E647 [16] and with a constant load ratio (R=0.5). The data generated are (increasing) stress intensity range (ΔK) and fatigue crack growth rate (da/dN); the tests were conducted in load control and regulated by the load cell located within the chamber, and the crack length was calculated from compliance, as provided by a CMOD (crack mouth opening displacement) gage located at the load line of the specimen. An internal load cell was used because it can more accurately represent the forces on the specimen(s), as the frictional forces of the seals are eliminated. The signal drift of the internal load cell in hydrogen gas up to 34 MPa results in a change in load ration of less than 2 %. In order to obtain sufficient data in a span of two years, a new apparatus was employed that permits the cyclic loading of ten specimens simultaneously within a test chamber [17]. The CT specimens were machined from the C-L orientation (see Figure 1c of ASTM E399 [18]) with a width W= 44.5 mm, a chevron notch to facilitate growth of a straight precrack, and the surface roughness Ra≤0.25 µm. The precrack was grown in air at a load ratio of R=0.1. The test chamber was purged three times with 99.9999 % helium and three times with 99.9995 % hydrogen before a final fill with the hydrogen and commencing the fatigue tests. The tests continued 24 hours/day, 7 days/week until all specimens were completed. The chamber pressure was continuously monitored and automatically maintained to ±3 % of the designated pressure.

Table 1. Chemical composition in mass percent of the steels tested. The balance is Fe.

Element	C	Mn	P	S	Si	Cu
X52	0.071	1.06	0.012	0.004	0.24	0.016
X70A	0.048	1.43	0.009	0.001	0.17	0.220
X70B	0.053	1.53	0.01	0.001	0.16	0.250

	Ni	Cr	Mo	V	Nb	Ti	Al
X52	0.016	0.033	0.003	0.004	0.026	0.038	0.017
X70A	0.14	0.240	0.005	0.004	0.054	0.027	0.015
X70B	0.14	0.230	0.003	0.004	0.054	0.024	0.012

Amaro, Robert; Drexler, Elizabeth; Lauria, Damian; Slifka, Andrew; Sowards, Jeffrey.
"Fatigue Crack Growth Rates of API X70 Pipeline Steels in Pressurized Hydrogen Gas Compared with an X52 Pipeline in Hydrogen Service."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

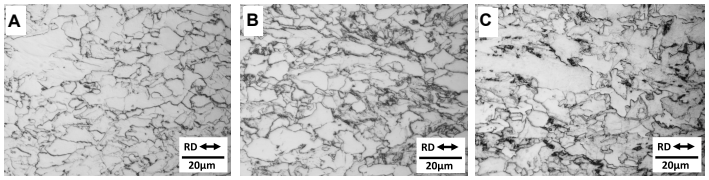


Figure 1. The microstructure from near the mid-line of the pipe through thickness for the (A) X52, (B) X70A, and (C) X70B steels.

Table 2. The tensile properties and the pipe dimensions for each of the steels reported.

Material	σ_y [MPa ± std. dev.]	σ_{UTS} [MPa ± std. dev.]	Pipe diameter [mm (in)]	Wall thickness [mm]
X52	487 ± 5	588 ± 5	508 (20)	10.6
X70A	509 ± 19	609 ± 4	914 (36)	18
X70B	553 ± 18	640 ± 9	914 (36)	22

RESULTS

The purpose of this study was to determine whether X70 can safely be used to construct pipelines for hydrogen gas transmission. Variability of the measurement on these steels in air can be found in Drexler *et al.* [17]. It is not possible to report on the variability of the measurement at any hydrogen condition because there is not sufficient data over the requisite range to allow calculations to be performed according to McKeighan et al. [19]. Uncertainty of the measurement would be expected to be much smaller than the variability, so calculating the uncertainty would not provide meaningful information.

The data at a cyclic loading frequency of 1 Hz are shown in Figure 3, and each dataset (line style) represents one to four individual specimens tested. In Figure 3A, it can be seen that for the specimens tested in hydrogen pressurized to 5.5 MPa, there is little difference between the FCGRs of steels designated as X52 and those designated as X70. The FCGRs of the steels in air (shown for comparison) are lower than those tested in hydrogen gas by as much as 20 times for the range of data tested. In pressurized hydrogen gas, subtle differences in the relative FCGR among the steels exist between low values of ΔK (<11 MPa·m^{½}) and those at higher values (>15 MPa·m^{½}). These differences are negligible when compared with the overall effect of hydrogen-assisted fatigue on the FCGR of pipeline steels.

At a hydrogen pressure of 34 MPa (Figure 3B), there is an even greater difference between the air and hydrogen data—as much as 50 times higher for the hydrogen data at a given value of ΔK. As seen in the figure, the three tests conducted in hydrogen gas on the X52 steel are visibly different. They were conducted simultaneously with the new apparatus, so the differences are not from variations in test conditions from one test to another. Furthermore, these specimens originated from a single piece of material that was removed from the pipe, and were from a similar clock position. Rather, this variability appears to be attributable to the way hydrogen interacts with microstructural features.

The data for both hydrogen pressures are shown on the same graph (Figure 4) to emphasize the effect of hydrogen test pressure on the FCGR of these steels.

Amaro, Robert; Drexler, Elizabeth; Lauria, Damian; Slifka, Andrew; Sowards, Jeffrey.
"Fatigue Crack Growth Rates of API X70 Pipeline Steels in Pressurized Hydrogen Gas Compared with an X52 Pipeline in Hydrogen Service."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

The fatigue crack grew rapidly in the higher hydrogen pressure, resulting in little data acquisition at low ΔK. Nevertheless, it is apparent that at low values of ΔK (<15 MPa·m^{½}), the FCGRs of tests conducted at 34 MPa are higher than those conducted at 5.5 MPa. Above ΔK=20 MPa·m^{½}, however, the data coalesce.

The remaining variable, cyclic loading frequency, is illustrated in Figure 5 for tests conducted at a hydrogen gas pressure of 5.5 MPa. From this graphical representation of the data, it is difficult to determine if frequency has a consistent effect on the FCGR. To clarify the effect of frequency on FCGR, data were analyzed for all materials and hydrogen gas pressures at one value of ΔK. A value of 14 MPa·m^{½} was chosen because it was the value for which the most data was available. Bar graphs showing the average value of da/dN for all the data available for that condition are shown in Figure 6. This snapshot of these limited conditions reveals that the slower cyclic loading frequency leads to slight increases in the FCGR for all conditions, except for the X52 steel when tested at a hydrogen gas pressure of 5.5 MPa (black bars). The hydrogen gas pressure (gray bars represent data acquired at a hydrogen gas pressure of 34 MPa) has a far larger effect on the FCGR than does the cyclic loading rate.

DISCUSSION

It is important to quantify the differences in the hydrogen-assisted fatigue crack growth rate (HA-FCGR) in X52 and X70 steels for two reasons. The first, as stated earlier, is that pipelines are expected to provide the means by which hydrogen fuel is transmitted between where it is produced and the end user. To accomplish this at a competitive cost, pipelines will have to be constructed of higher grade steel, so that less material can be used while still providing comparable margins of safety. Less steel lowers the cost. The impetus for the second reason can be found in the tensile data provided in Table 2. The owners of the hydrogen pipeline that provided our X52 steel, wanted and thought they were getting X52-strength steel. Instead they received material with an average yield strength closer to that of an X70 than an X52. It is not unusual for foundries to provide steel that exceeds the specified minimum yield strength (SMYS), because the API specification provides so much leeway.

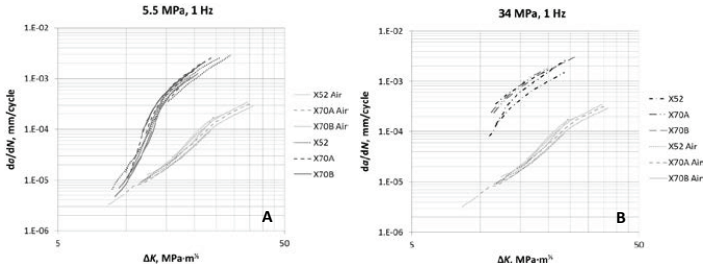


Figure 3. FCGR results from tests conducted at a cyclic loading frequency of 1 Hz and hydrogen gas pressures of (A) 5.5. MPa and (B) 34 MPa. Data collected in air are shown for comparison.

Amaro, Robert; Drexler, Elizabeth; Lauria, Damian; Slifka, Andrew; Sowards, Jeffrey.
"Fatigue Crack Growth Rates of API X70 Pipeline Steels in Pressurized Hydrogen Gas Compared with an X52 Pipeline in Hydrogen Service."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

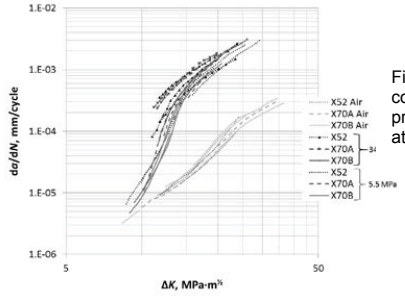


Figure 4. FCGR data comparing tests conducted in air, and at hydrogen gas pressurized to 5.5 MPa and 34 MPa at a cyclic loading frequency of 1 Hz.

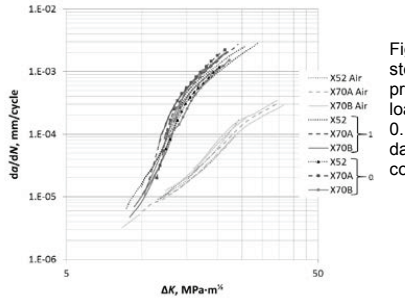


Figure 5. FCGR data for the three steels tested at a hydrogen gas pressure of 5.5 MPa and a cyclic loading frequency of 1 Hz (lines) and 0.1 Hz (lines with symbols). FCGR data acquired in air are included for comparison.

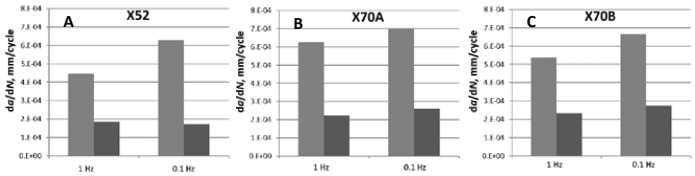


Figure 6. Bar graphs comparing the average value of the d*a*/d*N* data at ΔK = 14 MPa·m^{½} for the available data at cyclic loading frequencies of 1 Hz and 0.1 Hz for the (A) X52 steel, (B) X70A, and (C) X70B. Gray bars represent data acquired at a hydrogen gas pressure of 34 MPa and the black at 5.5 MPa.

The FCGR data generated at NIST show that all the reported materials are strongly, but comparably, affected by the presence of high-pressure hydrogen. This is observed for both hydrogen pressures and both cyclic loading rates discussed here. The findings were reported to the ASME B31.12 Committee on Hydrogen Piping and Pipelines. They concurred that, as long as pipelines operate well below the specified minimum yield strength (SMYS)—below the stresses for bursting or fracture, fatigue is the likely failure mechanism for hydrogen pipelines. (At higher operating pressures with respect to the SMYS, fracture toughness tests are still required.) Furthermore, these fatigue tests

Amaro, Robert; Drexler, Elizabeth; Lauria, Damian; Slifka, Andrew; Sowards, Jeffrey.
"Fatigue Crack Growth Rates of API X70 Pipeline Steels in Pressurized Hydrogen Gas Compared with an X52 Pipeline in Hydrogen Service."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

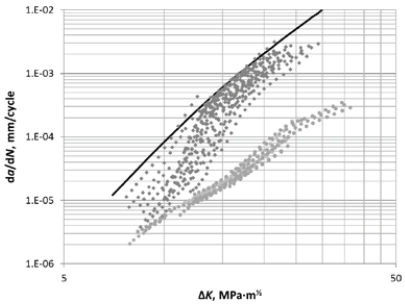


Figure 7. All of the data acquired on the fatigue crack growth rate of X52 and X70 steels in air (lt. gray circles) and in hydrogen gas (gray diamonds) pressurized to 5.5 MPa with the modeled fit of the upper bound that is now part of the ASME B31.12 code (black line).

provide the foundation upon which to modify the code. It was decided by the Committee that rather than requiring each material to be tested, an upper bound to all the available data that was acquired at hydrogen gas pressures of 21 MPa or below would be modeled [20] and that would be established as the minimum fatigue lifetime to which hydrogen pipelines will be designed. Figure 7 shows all of the NIST data acquired at hydrogen gas pressurized to 5.5 MPa throughout this test program (including an X52 pipeline steel, ca. 1964, that is not reported here) and the upper-bound fit to the data. The modification to the code has been approved by all requisite entities within ASME, and the modification will be implemented in the 2016 version of the code that is scheduled for release in February 2017.

CONCLUSIONS

Fatigue tests are a more accurate measure of how pipeline steels will perform in a pressurized hydrogen environment than tensile tests. However, sufficient FCGR data on which to base a code for designing hydrogen pipelines has not been available before now. Scores of tests were conducted at NIST on X52 steels (currently approved for use without further tests in the ASME B31.12 code) and X70 steels. Both grades of steel exhibited HA-FCGR, which accelerated crack growth up to 1 to 1.5 orders of magnitude over the FCGR in air. Since the HA-FCGRs for the two grades are comparable, X70 could be used for constructing hydrogen pipelines operating at current pressures with no loss in performance or safety when the modeled upper-bound FCGR is used. The ASME B31.12 Code on Hydrogen Piping and Pipelines has been revised and accepted to reflect this finding. Should future pipelines operate at pressures higher than 21 MPa (the maximum pressure used for the model fit for the code revision), the model will need to be modified and the code revised to reflect the higher FCGRs measured on steels tested at higher pressures, such as those reported here at 34 MPa.

Further studies on the HA-FCGR of the fusion zone and associated heat-affected zones should be conducted to elucidate whether these areas are more susceptible to degradation from hydrogen than the base metal. Even more fundamental, a general study on the interaction of hydrogen and predominant

Amaro, Robert; Drexler, Elizabeth; Lauria, Damian; Slifka, Andrew; Sowards, Jeffrey.
"Fatigue Crack Growth Rates of API X70 Pipeline Steels in Pressurized Hydrogen Gas Compared with an X52 Pipeline in Hydrogen Service."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

microstructural constituents in ferritic steels is needed. With that data, a fully-predictive physics-based model can be developed.

ACKNOWLEDGEMENTS

We thank Nik Hrabe for providing the microstructure images, James Merritt and the US Department of Transportation for their support for this work through agreement DTPH5615X00004, and Louis Hayden and the members of the ASME B31.12 Committee on Hydrogen Piping and Pipelines for their guidance.

REFERENCES

1. Energy.gov. July 12, 2016]; Available from: http://energy.gov/eere/fuelcells/hydrogen-pipelines.
2. US Department of Energy, E.E.R.E., *Multi-Year Research, Development, and Demonstration Plan: Planned program activities for 2011-2020*, F.C.T. Office, Editor. 2012.
3. Fekete, J.R., Sowards, J. W., Amaro, R. L., *Economic impact of applying high strength steels in hydrogen gas pipelines.* International Journal of Hydrogen Energy, 2015. **40**(33): p. 10547-10558.
4. *ASME B31.12-2011, "Hydrogen Piping and Pipelines"*. 2012, American Society of Mechanical Engineers: New York, NY. p. 258.
5. Cialone, H.J., Holbrook, J. H. *Microstructural and Fractographic Features of Hydrogen-Accelerated Fatigue-Crack Growth in Steels*. in *Welding, Failure Analysis, and Metallography*. 1987. Denver, CO, USA: ASM.
6. San Marchi, C., Stalheim, D., G., Somerday, B., P., Boggess, T., Nibur, K., A., Jansto, S., *Fracture Resistance and Fatigue Crack Growth of X80 Pipeline Steel in Gaseous Hydrogen*, in *ASME Pressure Vessels & Piping Division/ K-PVP 2011 Conference*. 2011, ASME: Baltimore, Maryland, USA. p. 9.
7. Nelson, H., G. *Hydrogen-Induced Slow Crack Growth of a Plain Carbon Pipeline Steel Under Conditions of Cyclic Loading*. in *Effect of Hydrogen Behavior of Materials: Proceedings of the International Conference*. 1976. Lake Moran, WY: Metall Soc of AIME, New York, NY.
8. Suresh, S., Ritchie, R. O., *Mechanistic dissimilarities between environmentally influenced fatigue-crack propagation at near-threshold and higher growth rates in lower strength steels.* Metal Science, 1982. **16**(11): p. 529-538.
9. Walter, R.J., Chandler, W. T. *Cyclic-Load Crack Growth in ASME SA-105 Grade II Steel in High-Pressure Hydrogen at Ambient Temperature*. in *Effect of Hydrogen Behavior of Materials: Proceedings of the International Conference*. 1976. Lake Moran, WY: Metall Soc of AIME, New York, NY.
10. Yoshioka, S., Kumasawa, M., Demizu, M., Inoue, A., , *Fatigue Crack Growth Behavior in Hydrogen Gas Environment*, in *Third International Conference on Fatigue and Fatigue Thresholds*. June 28- July 3, 1987: Charlottesville, VA.
11. Nelson, H.G. *On the mechanism of hydrogen-enhanced crack growth in ferritic steels*. in *Proceedings of the Second International Conference on Mechanical Behavior of Materials*. 1978. Boston, MA, August 1976.
12. Johnson, H.H. *Hydrogen brittleness in hydrogen and hydrogen-oxygen gas mixtures*. in *Stress Corrosion Cracking and Hydrogen Embrittlement of Iron Base Alloys, June 12-16, 1973*. 1977. Unieux Firminy, France.
13. Nibur, K., and Somerday, BP, *Fracture and fatigue test methods in hydrogen gas*, in *Gaseous hydrogen embrittlement of materials in energy technologies, Volume 1: The problem, its characterisation and effects on particular alloy classes*, R. Gangloff, and Somerday, BP, Editor. 2012, Woodhead Publishing: Cambridge, England. p. 195-236.
14. *ASTM Standard E8/E8M-09 "Standard Test Method for Tension testing of Metallic Materials"*. 2009: West Conshohocken, PA.
15. *API SPEC 5L, "Specification for Line Pipe, Forty-fourth Edition"*. 2007, American Petroleum Institute: Washington, DC. p. 154.
16. *ASTM Standard E 647-11 "Test Method for Measurement of Fatigue Crack Growth Rates"*. 2011, ASTM International. p. 46.

Amaro, Robert; Drexler, Elizabeth; Lauria, Damian; Slifka, Andrew; Sowards, Jeffrey.
"Fatigue Crack Growth Rates of API X70 Pipeline Steels in Pressurized Hydrogen Gas Compared with an X52 Pipeline in Hydrogen Service."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

17. Drexler, E.S., McColskey, J. D., Dvorak, M., Rustagi, N., Lauria, D. S., and Slifka, A. J. , *Apparatus for simultaneous fatigue testing of multiple compact tension specimens in air and controlled (harsh) environments.* Experimental Techniques, 2014. **40**(1): p. 429-439.
18. *ASTM Standard E399-12, "Standard Test Method for Linear-Elastic Plane-Strain Fracture Toughness KIc of Metallic Materials"*. 2012, ASTM: West Conshohocken, PA. p. 33.
19. McKeighan, P.C., Feiger, J.H., and McKnight, D.H., *Round Robin Test Program and Results for Fatigue Crack Growth Measurement in Support of ASTM Standard E647: Final Report*. 2008, Southwest Research Institute.
20. Amaro, R.L., Drexler, E. S., Slifka, A. J., *Development of an Engineering-Based Hydrogen-Assisted Fatigue Crack Growth Design Methodology for Code Implementation*, in *ASME 2014 Pressure Vessel and Piping Conference*. 2014: Anaheim, CA, USA.

Amaro, Robert; Drexler, Elizabeth; Lauria, Damian; Slifka, Andrew; Sowards, Jeffrey.
"Fatigue Crack Growth Rates of API X70 Pipeline Steels in Pressurized Hydrogen Gas Compared with an X52 Pipeline in Hydrogen Service."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

COMPUTATIONAL MODELING OF HYDROGEN-ASSISTED FATIGUE CRACK GROWTH IN PIPELINE STEELS*

D. T. O'CONNOR
National Institute of Standards and Technology, Boulder CO, USA
A. J. SLIFKA
National Institute of Standards and Technology, Golden CO, USA
B. E. LONG
Colorado School of Mines, Golden CO, USA
E. S. DREXLER
National Institute of Standards and Technology, Boulder CO, USA
R. L. AMARO
Colorado School of Mines, Golden CO, USA

ABSTRACT

In this work we further develop a model to predict hydrogen-assisted fatigue crack growth in steel pipelines and pressure vessels. This model is implemented by finite element code, which uses an elastic-plastic constitutive model in conjunction with a hydrogen diffusion model to predict the deformation and concentration of hydrogen around a fatigue crack tip. The hydrogen concentration around the crack tip is used to inform our fatigue crack growth model and account for the effect of hydrogen embrittlement. We first use our model to predict the fatigue crack growth of X100 pipeline steel at different levels of applied hydrogen pressure. The simulated results are within a factor of ± 2 of the experimental X100 results.

INTRODUCTION

Hydrogen is expected to play a key role in transitioning the United States' energy and transportation sectors away from fossil fuels towards a more sustainable and climate-friendly alternative. While not an energy source, per se, hydrogen is seen as an energy carrier. In which case, hydrogen is used in conjunction with a catalyst to produce energy. Hydrogen may also be used to store energy from green energy-producing alternatives such as solar, wind, wave, and so on that produce energy regardless of demand. In order for hydrogen to see widespread use as an energy carrier, hydrogen must be safely and efficiently transported from the source to the location of end-use.

Hydrogen fuel cell vehicles are currently being manufactured and sold by Toyota, Hyundai, and Honda. Mercedes-Benz, Lexus, and Nissan have concept and evaluation vehicles currently in testing [1]. Unfortunately, there are currently only 29 hydrogen fueling stations in the United States [2]. The vast majority of those are in Northern and Southern California. The infrastructure required to transport hydrogen across the United States efficiently does not currently exist. Steel pipeline is thought to be the best method to transport

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Computational Modeling of Hydrogen-Assisted Fatigue Crack Growth in Pipeline Steels."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

hydrogen long-distances. While the United States currently has on the order of 305,000 miles of natural gas pipeline [3] there is currently only approximately 700 miles to 1500 miles of hydrogen-dedicated pipeline for hydrogen delivery [4, 5]. A primary barrier to the use of steel pipeline to transport hydrogen is hydrogen's deleterious effects on steels monotonic and fatigue deformation response [6, 7]. The ASME B31.12 Committee on Hydrogen Piping and Pipelines is supporting the design, engineering, and installation of hydrogen-dedicated pipeline by creating and updating the B31.12 code based upon current hydrogen-assisted fatigue crack growth (HA-FCG) data collected at laboratories, such as the National Institute of Standards and Technology (NIST) and Sandia National Laboratories (SNL) [8-10] The data collected has been used to inform a phenomenological HA-FCG model for pipeline steels that predicts cycles to failure for known operating conditions. The model was based initially upon closed-form solutions for the crack-tip deformation response, the hydrogen diffusion within the material, and the coupling of the two [11, 12]. This work details the implementation of these key aspects of the HA-FCG model into the finite element code ABAQUS[1]. The results from the combination of deformation and hydrogen diffusion in ABAQUS are then coupled with the phenomenological HA-FCG model to predict the crack-growth response of X100 compact tension (CT) specimens tested in 1.72 MPa, 6.89 MPa, and 20.68 MPa gaseous hydrogen (250 psi, 1000 psi, and 3000 psi gaseous hydrogen).

Hydrogen-Assisted Fatigue Crack Growth Model

The existing phenomenological HA-FCG model, calibrated to X100 pipeline steel, is detailed in [11, 12]. The form of the model is as follows:

$$\frac{da}{dN_{\text{total}}} = \frac{da}{dN_{\text{fatigue}}} + \delta\left(P_{\text{H}} - P_{\text{H}_{\text{th}}}\right)\frac{da}{dN_{\text{H}}}, \qquad \text{EQ. 1}$$

where, a is the crack length, N is the number of cycles, and P is the hydrogen pressure. From left to right, the total fatigue crack growth is calculated as a summation of the fatigue crack growth resulting from fatigue only and the hydrogen-assisted fatigue crack growth. The HA-FCG term has the form:

$$\frac{da}{dN_{\text{H}}} = \left[\left(\frac{da}{dN_{P_{\text{H}}}}\right)^{-1} + \left(\frac{da}{dN_{\Delta K}}\right)^{-1}\right]^{-1}, \qquad \text{EQ. 2}$$

where the overall contribution of HA-FCG is modeled as a competition between a hydrogen-pressure-dominated component, $\frac{da}{dN_{P_{\text{H}}}}$, and a component dominated by hydrogen-assistance from the crack-extension driving force , $\frac{da}{dN_{\Delta K}}$. This competition is a result of two independent damage mechanisms. When the crack extension per cycle is on the order of the fatigue process zone (FPZ) size, the crack growth rate is dominated by the accumulated damage of the FPZ and the increased hydrogen concentration within the FPZ. However, when the crack extension per cycle extends far beyond the FPZ, P_{H}, the crack growth rate is dominated less by the effects within the FPZ and more by the far-field crack driving force, ΔK. The hydrogen-pressure-dominated FCG term is defined as

[1] Identified for clarity only, no endorsement by NIST is implied

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Computational Modeling of Hydrogen-Assisted Fatigue Crack Growth in Pipeline Steels."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

$$\frac{da}{dN_{P_H}} = a1\Delta K^{B1}\left(P_H^{m1} exp^{\left(\frac{-Q+V\sigma_h}{RT}\right)}\right)^{d1}, \quad \text{EQ. 3}$$

and is referred to as transient HA-FCG, where ΔK is the stress intensity range, Q is the activation energy for hydrogen diffusion; V is the partial molar volume of hydrogen in the metal; R is the universal gas constant; T the absolute temperature; σ_h is the hydrostatic stress at a critical distance in front of the crack tip; P_H is the ambient hydrogen pressure; and *a1*, *B1*, *m1*, and *d1* are fitting parameters. The component dominated by hydrogen-assistance from the crack-extension driving force is defined as

$$\frac{da}{dN_{\Delta K}} = a2\Delta K^{B2}\left(P_H^{m2} exp^{\left(\frac{-Q+V\sigma_h}{RT}\right)}\right)^{d2}, \quad \text{EQ. 4}$$

and is referred to as the steady-state HA-FCG, where *a2*, *B2*, *m2*, and *d2* are fitting parameters. Explicit details of the model justification, constants, parameters, calibration, etc. may be found in [11]. One will note that Eqs. 3 and 4 employ closed-form solutions for the stress-free hydrogen concentration within the material, P_H^{m2}, the stress field at a crack tip, σ_h, and the stress-assisted hydrogen concentration near the crack tip, $P_H^{m2} exp^{\left(\frac{-Q+V\sigma_h}{RT}\right)}$.

ABAQUS Implementation

Ultimately, the HA-FCG model defined above is to be implemented in a physics-based format that predicts the FCG, based upon microstructure-specific material responses, such as hydrogen diffusivity, hydrogen-dislocation interactions, microstructure-specific deformation, etc. A first step towards this aim is to determine the stress-free and stress-assisted hydrogen concentration within the material, both near the crack tip and far field, by use of the finite element code ABAQUS. For this purpose, Eq. 3 is then replaced by

$$\frac{da}{dN_{P_H}} = a1\Delta K^{B1}(C_L)^{d1}, \quad \text{EQ. 5}$$

where C_L is the spatial- and time-dependent lattice hydrogen concentration determined from ABAQUS (defined in EQ. 10 below). Equation 4 is also replaced with

$$\frac{da}{dN_{\Delta K}} = a2\Delta K^{B2}(C_L)^{d2}, \quad \text{EQ. 6}$$

and *a2*, *B2*, *m2*, and *d2* are fitting parameters [12].

The finite-element implementation to determine C_L requires an understanding of the elastic-plastic deformation response of the material. This modeling effort employs the Ramberg-Osgood (RO) elastic-plastic constitutive model

$$\frac{\varepsilon}{\varepsilon_0} = \frac{\sigma}{\sigma_0} + \alpha\left(\frac{\sigma}{\sigma_0}\right)^n, \quad \text{EQ. 7}$$

where ε is the total strain, σ is the total stress, ε_0 and σ_0 are the strain and stress at yielding, respectively, and α and n are constants. The Ramberg-Osgood deformation model is implemented in ABAQUS with the existing cyclic plasticity model incorporating linear kinematic hardening.

Hydrogen transport is modeled in ABAQUS by use of a new user-defined subroutine, H-diff, which is built upon the structure of the existing ABAQUS

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Computational Modeling of Hydrogen-Assisted Fatigue Crack Growth in Pipeline Steels."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

subroutine UMATHT. The subroutine H-diff explicitly calculates the hydrogen concentration, C_L, the plastic-hardening curve, and the hydrostatic stress by use of EQ. 7 at the integration points as a function of time. The hydrogen concentration is calculated via the hydrogen transport equation of [13] and modified by [14]

$$\frac{D}{D_{\text{eff}}}\frac{\partial C_L}{\partial t} = D\nabla^2 C_{\text{L}} - \nabla\cdot\left(\frac{DV_{\text{H}}}{3RT}C_{\text{L}}\nabla\sigma_{\text{h}}\right) - \left(\sum_j \eta^j \theta_{\text{T}}^j \frac{\partial N_{\text{T}}^j}{\partial \varepsilon^{\text{p}}}\right)\frac{\partial \varepsilon^{\text{p}}}{\partial t}. \quad \text{EQ. 8}$$

The variables in Eq. 8 are as follows: D is the hydrogen diffusion coefficient, D_{eff} is the effective diffusion coefficient, C_{L} is the hydrogen concentration in the normal interstitial lattice site (NILS), V_{H} is the partial molar volume of hydrogen, R is the universal gas constant, T is the absolute temperature, σ_{h} is the hydrostatic stress, η is the number of trapping sites per trap type j, θ_{T} is the trap site occupancy for a given trap site j, N_{T} is the trap-site density for a given trap site j, ε^{p} is the equivalent plastic strain and the ∇ is the mathematical vector differential operator. Equation 8 is an extension of Fick's law that incorporates the influence of both hydrostatic stress and plastic strain on hydrogen transport. While Eq. 8 may be used for three types of traps, i.e. carbides, grain boundaries and dislocations, this model implementation is only concerned with the so-called weak traps. In this case, Eq. 8 is used to model hydrogen diffusion resulting only from NILS and dislocations. Oriani's theory [15] provides the relationship between the trap-site occupancies, θ_{T}^j, and the lattice-site occupancies, θ_L, as

$$\frac{\theta_{\text{T}}^j}{1-\theta_{\text{T}}^j} = \frac{\theta_{\text{L}}}{1-\theta_{\text{L}}} exp\left(\frac{W_{\text{B}}^j}{RT}\right), \quad \text{EQ. 9}$$

where W_{B}^j is the trap binding energy for the trap of interest, dislocations in this case. The hydrogen concentration in the NILS and the trap sites is given by EQ. 10 and EQ. 11, respectively.

$$C_{\text{L}} = \beta N_{\text{L}}\theta_{\text{L}} \quad \text{EQ. 10}$$

$$C_{\text{T}}^j = \eta^j N_{\text{T}}^j \theta_{\text{T}}^j \quad \text{EQ. 11}$$

Literature values are used for the number of interstitial sites per atom, β, the number of solvent atoms per unit volume, N_{L}, and the number of trap sites per trap type, η^j. The trap densities for a given trap type, N_{T}^j, taken here as the trap density for only dislocations, is solved as a function of equivalent plastic strain, per the work of [16]. The relationship is given in EQ. 12

$$N_{\text{T}}^{\text{dislocations}} = \frac{10^{23.26-2.33exp(-5.5\varepsilon^p)}}{N_A}, \quad \text{EQ. 12}$$

where N_A is Avogadro's constant. Finally, the effective hydrogen diffusion is calculated by use of

$$\frac{D}{D_{\text{eff}}} = 1 + \sum_j \frac{\partial C_{\text{T}}^j}{\partial C_{\text{L}}}. \quad \text{EQ 13.}$$

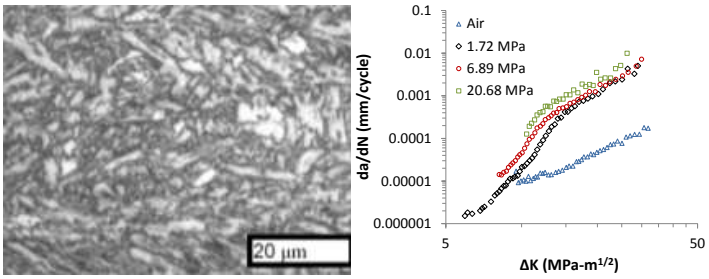
Equations 8 through 13 are solved for each integration point at each time step within the new user-defined material model H-Diff.

HA-FCG Calibration to an API-5L X100 Pipeline Steel

An API-5L X100 pipeline steel has been characterized at NIST and was used as the model material for the original HA-FCG implementation [11]. HA-FCG tests were conducted on compact tension specimens in the transverse-

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Computational Modeling of Hydrogen-Assisted Fatigue Crack Growth in Pipeline Steels."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

longitudinal (TL) orientation at a load ratio of R=0.5 (were R=K_{min}/K_{max} in this case) in various hydrogen pressures. The material's chemical composition is provided in Table 1. The material's microstructure is shown in Fig. 1a and can be characterized by having an average grain size on the order of 1 micron.



(a) **(b)**

Figure 1: (a) Image of X100 microstructure, (b) HA-FCG of X100 steel as a function of environment, R=0.5.

The monotonic test results for this material, Ramberg-Osgood fit parameters, as a function of air and varying hydrogen pressures, along with its HA-FCG response are detailed in [11, 12] and provided here for completeness. Full experimental details for collection of the data in Table 2 and Figure 1b are provided in [11, 17].

Table 1: Chemical composition of API steels tested, mass %.

	Al	C	Co	Cr	Cu	Fe	Mn	Mo
X100	0.012	0.064	0.003	0.023	0.28	96.90	1.87	0.23
	N	Nb	Ni	P	Si	Ti	V	
X100	0.003	0.017	0.47	0.009	0.099	0.017	0.002	

Table 2: Monotonic test results and R-O fitting parameters for X100 [12].

	H_2 Pressure	σ_0	ε_0	n	α	E
Material	MPa	MPa	-	-	-	GPa
X100	AIR	693.01	0.0032	13.48	0.92	214.14
Longitudinal	5.5	700.37	0.0032	13.39	1.01	219.17
	13.8	700.90	0.0032	13.78	1.11	218.89
	27.6	708.86	0.0031	13.56	1.03	229.61
	68.9	714.01	0.0033	14.34	0.97	215.74
X100	AIR	804.47	0.0035	17.18	2.97	229.58
Transverse	13.8	810.23	0.0035	15.33	3.52	230.52
X52	AIR	442.21	0.0021	11.74	3.10	212.42

One will note from Table 2 that for a given orientation with respect to the rolling direction (longitudinal or transverse), the X100 tested has relatively stable yielding and post-yielding response as a function of environment, as indicated by the R-O parameters. Perhaps not surprising given the microstructural texture shown in Fig. 1a, the material exhibits noticeable orientation anisotropy in both

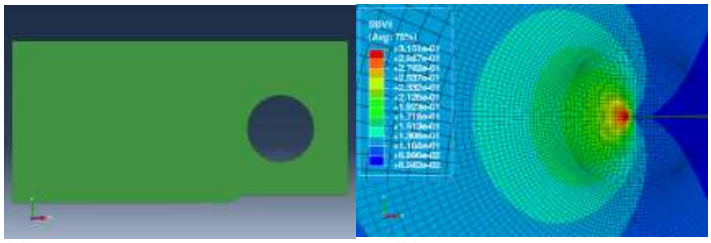
Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Computational Modeling of Hydrogen-Assisted Fatigue Crack Growth in Pipeline Steels."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

its yielding and post-yielding behavior. Figure 1b depicts the FCG results of the X100 in both air and high-pressure hydrogen. Figure 1b indicates that the X100 material tested is susceptible to changes in hydrogen pressure, with increased FCG rates at higher pressures.

To determine the hydrogen lattice concentration, C_L, a one-half symmetry CT specimen was modeled in ABAQUS and is shown in Fig. 2a. The symmetry plane was created along the presumed crack path at the geometric center line. Given that the CT specimens were tested in the TL orientation, the X100 material properties from the transverse orientation were used in this study. Specimens having various crack lengths were created. The crack lengths and applied loads were calculated by use of ASTM E647 to ensure resulting ΔK values of 7 MPa-m$^{0.5}$, 9 MPa-m$^{0.5}$, 11 MPa-m$^{0.5}$, 13 MPa-m$^{0.5}$, and 15 MPa-m$^{0.5}$. Simulations were then run at the five ΔK values for air, and hydrogen gas pressures of 1.72 MPa, 6.89 MPa, and 20.68 MPa. The ambient hydrogen pressure was implemented in ABAQUS by way of the calculated chemical potential. The full model parameters required for the hydrogen diffusion study in ABAQUS are taken from [14] and are given in Table 3. Figure 2b provides the resulting imagery of the lattice hydrogen concentration at a crack tip of a CT specimen experiencing ΔK=15 MPa-m$^{0.5}$ in 6.89 MPa gaseous hydrogen.

Table 3: Hydrogen diffusion material parameters.

D (m^2/s)	V_H (m^3/mol)	N_L (mol/m^3)	W_B (J/mol)	β	η
1.28 x 10^{-8}	2.0 x 10^{-6}	8.46 x 10^{-28}	6.1 x 10^4	1	1



(a) **(b)**

Figure 2: (a) Symmetry CT specimen geometry; (b) Predicted hydrogen concetration at the crack tip for ΔK=15 MPa-m$^{0.5}$, R=0.5, and a hydrogen pressure of 6.89 MPa.

Although no data currently exists to calibrate the lattice hydrogen concentration, the results provided in Fig. 2b are as expected [13]. The predicted fatigue crack growth for the twenty combinations of ΔK and environments of interest were calculated by use of the lattice hydrogen concentrations predicted from ABAQUS in conjunction with Eqs. 5 and 6. The FCG predictions are shown with the experimental data in Fig. 3. The results depicted in Fig. 3 indicate that the model implementation performs very well at predicting the HA-FCG for a

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Computational Modeling of Hydrogen-Assisted Fatigue Crack Growth in Pipeline Steels."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

CT-specimen of X100 for varying ΔK and hydrogen pressures with predicted values within a factor of ± 2 of the experimental results.

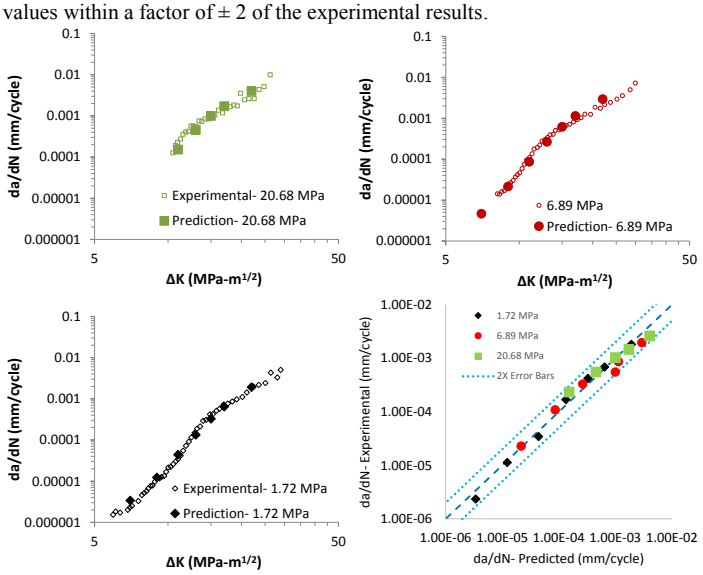


Figure 3: Experimental and predicted HA-FCG for an X100 steel at various pressures of hydroegn gas.

CONCLUSIONS

A new ABAQUS user-defined material model for stress- and plastic strain-assisted hydrogen diffusion, has been created by use of the existing ABAQUS UMATHT. The new hydrogen diffusion model has been coupled with the elastic-plastic material response to predict predict hydrogen-free deformation, deformation in the presence of hydrogen, and the lattice hydrogen concentration of an API-5L X100 material. Furthermore, the physics-based model, which combines deformation and the hydrogen diffusion, has been coupled with an existing phenomenological HA-FCG model. The modeling results accurately predict HA-FCG within a factor of ± 2.

REFERENCES

1. *U.S. Depratment of Energy, Energy Eficiency & Renewable Energy, Fuel Cell Links- Vehicles and Manufacturers*. 2016 7/24/2016]; Available from: www.fueleconomy.gov/feg/fcv_links.shtml.
2. *U.S. Depratment of Energy, Energy Eficiency & Renewable Energy, Alternative Fuels Data Center- Alternative Fueling Station Locator*. 2016 7/24/2016]; Available from: www.afdc.energy.gov/locator/stations/.

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Computational Modeling of Hydrogen-Assisted Fatigue Crack Growth in Pipeline Steels."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

3. *United States Energy Information Administration, About U.S. Natural Gas Pipelines*. 2016 7/24/2016]; Available from: www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/ngpipeline/index.html.
4. *U.S. Depratment of Energy, Energy Eficiency & Renewable Energy, Alternative Fuels Data Center, Hydrogen Production and Distribution*. 2016 7/24/2016]; Available from: www.afdc.energy.gov/fuels/hydrogen_production.html#distribution.
5. *U.S. Office of Energy Efficiency & Renewable Energy, ENERGY.gov, Hydrogen Pipelines*. 2016; Available from: http://energy.gov/eere/fuelcells/hydrogen-pipelines.
6. Somerday, B., P., *Technical Reference on Hydrogen Compatibility of Materials- Plain Carbon Ferritic Steels: C-Mn Alloys (Code 1100)*, 2008, Sandia National Laboratories, Livermore, CA. p. 32.
7. Nanninga, N.E., Levy, Y. S., Drexler, E. S., Condon, R. T., Stevenson, A. E., Slifka, A. J., *Comparison of hydrogen embrittlement in three pipeline steels in high pressure gaseous hydrogen environments*. Corrosion Science, 2012. **59**(0): p. 1-9.
8. Slifka, A.J., Drexler, E. S., Nanninga, N. E., Levy, Y. S., McColskey, J. D., Amaro, R. L., Stevenson, A. E., *Fatigue crack growth of two pipeline steels in a pressurized hydrogen environment.* Corrosion Science, 2014. **78**(0): p. 313-321.
9. Drexler, E.S., Slifka, A. J., Amaro, R. L., Barbosa, N., Lauria, D. S., Hayden, L. E., Stalheim, D. G., *Fatigue crack growth rates of API X70 pipeline steel in a pressurized hydrogen gas environment.* Fatigue & Fracture of Engineering Materials & Structures, 2014. **37**(5): p. 517-525.
10. San Marchi, C., Somerday, B. P., Nibur, K. A., Stalheim, D. G., Boggess, T., Jansto, S., *Fracture and Fatigue of Commercial Grade API Pipeline Steels in Gaseous Hydrogen*, in *ASEM 2010 Pressure Vesselss and Piping Conference*2010: Bellevue, WA, USA.
11. Amaro, R.L., Rustagi, N., Findley, K. O., Drexler, E. S., Slifka, A. J., *Modeling the fatigue crack growth of X100 pipeline steel in gaseous hydrogen.* International Journal of Fatigue, 2014. **59**(0): p. 262-271.
12. Amaro, R.L., Drexler, E. S., Slifka, A.J., *Fatigue crack growth modeling of pipeline steels in high pressure gaseous hydrogen.* International Journal of Fatigue, 2014. **62**(0): p. 249-257.
13. Sofronis, P., McMeeking, R. M., *Numerical analysis of hydrogen transport near a blunting crack tip.* Journal of the Mechanics and Physics of Solids, 1989. **37**(3): p. 317-350.
14. Novak, P., Yuan, R., Somerday, B. P., Sofronis, P., Ritchie, R. O., *A statistical, physical-based, micro-mechanical model of hydrogen-induced intergranular fracture in steel.* Journal of the Mechanics and Physics of Solids, 2010. **58**(2): p. 206-226.
15. Oriani, R.A., *The diffusion and trapping of hydrogen in steel.* Acta Metallurgica, 1970. **18**(1): p. 147-157.
16. Kumnick, A.J., Johnson, H. H., *Deep trapping states for hydrogen in deformed iron.* Acta Metallurgica, 1980. **28**(1): p. 33-39.
17. Nanninga, N.E., Levy, Y. S., Drexler, E. S., Condon, R. T., Stevenson, A. E., Slifka, A. J., *Comparison of hydrogen embrittlement in three pipeline steels in high pressure gaseous hydrogen environments*. Corrosion Science, 2012. **59**: p. 1-9.

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Computational Modeling of Hydrogen-Assisted Fatigue Crack Growth in Pipeline Steels."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

APPLICATION OF A MODEL OF HYDROGEN-ASSISTED FATIGUE CRACK GROWTH IN 4130 STEEL*

R. L. AMARO
Colorado School of Mines
Golden, Colorado USA

B. E. LONG
Colorado School of Mines
Golden, Colorado USA

A. J. SLIFKA
National Institute of Standards and Technology
Boulder, Colorado USA

E. S. DREXLER
National Institute of Standards and Technology
Boulder, Colorado USA

D. T. O'CONNOR
National Institute of Standards and Technology
Boulder, Colorado USA

ABSTRACT

In this work, we applied a finite element model to predict the cyclic lifetime of 4130 steel cylinders under the influence of hydrogen. This example is used to demonstrate the efficacy of a fatigue crack growth (FCG) model we have developed. The model was designed to be robust and incorporate features of stress-assisted hydrogen diffusion, large-scale plasticity, hydrogen gas pressure, loading frequency, and effects of microstructure. The model was calibrated to the 4130 steel material by use of tensile tests and experimental FCG results of a compact tension specimen. We then used the model to predict the hydrogen-assisted FCG rate and cycle life of a pressurized cylinder with a deliberate initial thumbnail crack. The results showed good correlation to the cyclic lifetime results of 4130 pressurized cylinders found in the literature.

INTRODUCTION

If hydrogen is to be used as an energy carrier to provide an alternative to fossil fuels, a vast network of pipelines is required to transport the hydrogen across the country. Furthermore, truck-mounted and loose pressure vessels will likely continue to be employed as short-distance hydrogen transportation and storage solutions. Steel pipelines are likely the most economical means of long-distance hydrogen transportation. The most common materials used in natural gas pipelines are API-5L grade steels that have a specified minimum yield stresses between 52 ksi (358 MPa) and 80 ksi (551 MPa). These steels carry the designation of API-5L X52, X65, X70, X80, etc. Future hydrogen-specific pipe installations may include the use of X100 and X120 pipeline steels. Pressure vessels for hydrogen use are commonly produced with ASTM SA/A516 or AISI 413X steels, where the X is replaced with a 0 or 5 depending upon carbon content. While the difference in geometries and boundary conditions between pipes and the cylinder portion of pressure vessels may be easily represented, the steels used comprise vastly different microstructures. Although the differing microstructures

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Application of a Model of Hydrogen-Assisted Fatigue Crack Growth in 4130 Steel."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

may or may not yield significantly different deformation responses under normal operating conditions, they do likely produce varying hydrogen diffusivities [1-4] and, therefore, have the potential for significantly different fatigue and fracture responses in the presence of hydrogen.

The pipeline and infrastructure required to transport hydrogen across the United States does not currently exist. Given that steel pipelines have for a long time been the best method to transport fuels long distances, the United States will need a hydrogen transmission network similar to that of the current natural gas transmission network in size. The United States currently has approximately 305,000 miles of natural gas pipelines [5] and only one half of one percent of that quantity is hydrogen-dedicated pipeline [6, 7]. A primary barrier to the use of carbon steel to transport hydrogen, whether by pipeline or pressure vessel, is the deleterious effect that hydrogen has upon the deformation, fracture, and fatigue response of the steel [8, 9]. Design and engineering codes for hydrogen transmission via pipeline [10] and distribution via piping [11], as well as pressure vessels [12], have been developed to ensure safe and effective use of these steels for hydrogen service. The newest versions of the codes that support the transmission and distribution of hydrogen and hydrogen-bearing gasses are being informed by the recent experimental results from laboratories such as the National Institute of Standards and Technology (NIST) and Sandia National Laboratories (SNL) [13-15]. It is nearly universally understood that microstructure-specific physics-based models are required to inform the large-scale infrastructure expansion required to meet the needs of a future network for hydrogen transmission.

A phenomenological hydrogen-assisted fatigue crack growth (HA-FCG) model that predicts cycles to failure for given material-specific calibration parameters and known initial and boundary conditions has been created and used to inform the forthcoming 2016 version of the ASME B31.12 Hydrogen Piping and Pipeline code (to be released in 2017). The current HA-FCG model is based upon the understanding that the crack-growth response of a steel results from an interaction between competing damage mechanisms and was initially implemented with closed-form solutions to the crack stress response and the stress-assisted hydrogen diffusion [16, 17]. In our current effort, the elastic-plastic deformation response coupled with the hydrogen diffusion model is implemented in the finite element (FE) program ABAQUS[1]. The predicted hydrogen concentration, as a function of elastic and plastic deformation and gas pressure, is coupled with the existing phenomenological framework to predict the HA-FCG response of the steel. The strength of this modeling framework is that one can predict a steel's HA-FCG response, for any geometry that can be modeled, based upon laboratory results from simple specimens, e.g. compact tension (CT) specimens. This work details the use of the modeling framework to predict cycles to failure for pipes and pressure vessels that have thumbnail-shaped internal cracks. Ultimately, the model implementation will be updated to include

[1] Identified for clarity only, no endorsement by NIST is implied

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Application of a Model of Hydrogen-Assisted Fatigue Crack Growth in 4130 Steel."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

microstructure-specific steel domains and their associated deformation and hydrogen diffusion properties.

Coupled Hydrogen-Assisted Fatigue Crack Growth Model

The phenomenological HA-FCG model, calibrated to X100 pipeline steel is detailed in [16, 17]. The model has also been partially calibrated to X52 and X70 steels [18]. The model has the following functional form:

$$\frac{da}{dN_{\text{total}}} = \frac{da}{dN_{\text{fatigue}}} + \delta\left(P_{\text{H}} - P_{\text{H}_{\text{th}}}\right)\frac{da}{dN_{\text{H}}}, \quad \text{EQ. 1}$$

where, a is the crack length, N is the number of cycles, and P is the hydrogen pressure. The model predicts the total fatigue crack growth as the sum of the fatigue crack growth resulting from fatigue only and the HA-FCG. HA-FCG is predicted by

$$\frac{da}{dN_{\text{H}}} = \left[\left(\frac{da}{dN_{\text{P}_{\text{H}}}}\right)^{-1} + \left(\frac{da}{dN_{\Delta K}}\right)^{-1}\right]^{-1}. \quad \text{EQ. 2}$$

Equation 2 predicts that the HA-FCG results from a competitive interaction between a hydrogen-pressure-dominated component, $\frac{da}{dN_{\text{P}_{\text{H}}}}$, and a component dominated by hydrogen-assistance from the crack-extension driving force, $\frac{da}{dN_{\Delta K}}$. The hydrogen-pressure-dominated FCG term is defined as

$$\frac{da}{dN_{\text{P}_{\text{H}}}} = a1\Delta K^{B1}(C_{\text{L}})^{d1}, \quad \text{EQ. 3}$$

where ΔK is the stress intensity range, C_{L} is the spatial and time-dependent lattice hydrogen concentration determined from ABAQUS, and $a1$, $B1$, and $d1$ are fitting parameters. The component dominated by hydrogen-assistance from the crack-extension driving force is defined as

$$\frac{da}{dN_{\Delta\text{K}}} = a2\Delta K^{B2}(C_{\text{L}})^{d2}, \quad \text{EQ. 4}$$

and $a2$, $B2$, and $d2$ are fitting parameters.

The lattice hydrogen concentration, C_{L}, is predicted by use of a new user-defined material (UMAT) model that we developed in ABAQUS called H-diff. The UMAT H-diff is based upon the architecture of the existing high-temperature material model UMATHT as in [19]. In order to account for hydrogen trapping by dislocations, H-diff utilizes the extended Fick's law provided in [20]. The elastic-plastic material response is predicted by the Ramberg-Osgood (RO) constitutive model by use of the ABAQUS "Deformation Plasticity" parameters. The R-O model,

$$\frac{\varepsilon}{\varepsilon_0} = \frac{\sigma}{\sigma_0} + \alpha\left(\frac{\sigma}{\sigma_0}\right)^{n}, \quad \text{EQ. 5}$$

calculates the total strain, ε, as a function of the total stress, σ, or vice versa. The parameters ε_0 and σ_0 are the strain and stress at yielding, respectively, and α and n are constants. The hydrogen concentration is calculated via the hydrogen transport equation of [21] and modified by [20]

$$\frac{D}{D_{\text{eff}}}\frac{\partial C_{\text{L}}}{\partial t} = D\nabla^2 C_{\text{L}} - \nabla\cdot\left(\frac{DV_{\text{H}}}{3RT}C_{\text{L}}\nabla\sigma_{\text{h}}\right) - \left(\sum_j \eta^{\text{j}}\theta_T^j\frac{\partial N_{\text{T}}^j}{\partial\varepsilon^{\text{p}}}\right)\frac{\partial\varepsilon^{\text{p}}}{\partial t}. \quad \text{EQ. 6}$$

The parameters that are required in order to solve for the hydrogen concentration are defined as follows: D is the hydrogen diffusion coefficient, D_{eff} is the effective

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Application of a Model of Hydrogen-Assisted Fatigue Crack Growth in 4130 Steel."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

diffusion coefficient, C_L is the hydrogen concentration in the normal interstitial lattice site (NILS), V_H is the partial molar volume of hydrogen, R is the universal gas constant, T is the absolute temperature, σ_h is the hydrostatic stress, η is the number of trapping sites per trap type j, θ_T is the trap site occupancy for a given trap site j, N_T is the trap-site density for a given trap site j, ε^p is the equivalent plastic strain and the ∇ is the mathematical vector differential operator. Equation 6 is an extension of Fick's law that incorporates the influence of both hydrostatic stress and plastic strain on hydrogen transport. The hydrogen concentration equation has the capability to solve for the hydrogen concentration in the lattice and three separate trap sites. This implementation is concerned with only the weakly-trapped hydrogen, in which case the model only determines the hydrogen diffusion resulting from NILS and dislocations. Oriani's theory [22] provides the relationship between the trap-site occupancies, θ_T^j, and the lattice-site occupancies, θ_L, as

$$\frac{\theta_T^j}{1-\theta_T^j} = \frac{\theta_L}{1-\theta_L} exp\left(\frac{W_B^j}{RT}\right), \quad \text{EQ. 7}$$

where W_B^j is the trap binding energy for the trap of interest (dislocations in this case). The hydrogen concentration in the NILS and the trap sites is given by EQ. 8 and EQ. 9, respectively.

$$C_L = \beta N_L \theta_L \quad \text{EQ. 8}$$

$$C_T^j = \eta^j N_T^j \theta_T^j \quad \text{EQ. 9}$$

Literature values are used for the number of interstitial sites per atom, β, the number of solvent atoms per unit volume, N_L, and the number trap sites per trap type, η^j. The trap densities for a given trap type, N_T^j, taken here as only the trap density for dislocations, is solved as a function of equivalent plastic strain per the work of [23]. The relationship is given in EQ. 10

$$N_T^{\text{dislocations}} = \frac{10^{23.26-2.33exp(-5.5\varepsilon^p)}}{N_A}, \quad \text{EQ. 10}$$

where N_A is Avogadro's constant. Finally, the effective hydrogen diffusion is calculated by use of

$$\frac{D}{D_{\text{eff}}} = 1 + \sum_j \frac{\partial C_T^j}{\partial C_L}. \quad \text{EQ. 11}$$

Equations 6 through 11 are solved at each integration point and at each time step within the new user-defined material model H-Diff.

HA-FCG Calibration to a 4130 Pressure Vessel Steel

The data in [24] are leveraged here to calibrate the elastic-plastic constitutive behavior, the HA-FCG response and the diffusivity parameters. The R-O parameters and the hydrogen diffusion parameters are provided in Tables 1 and 2, respectively. The experimental HA-FCG results of the 4130 steel [25] are provided in Fig. 1(a).

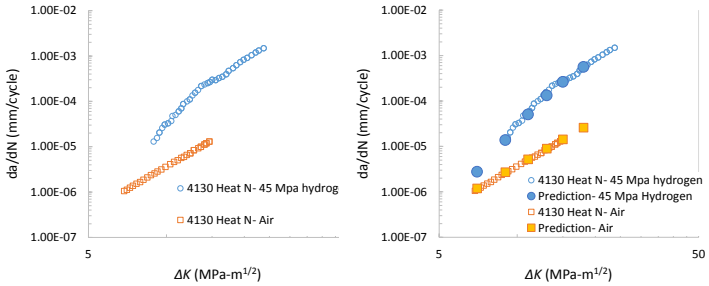
Table 1: Ramberg-Osgood parameters for 4130 steel

Material	ε_0	σ_0 (MPa)	α	n
4130	0.0034	762	0.58	20

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Application of a Model of Hydrogen-Assisted Fatigue Crack Growth in 4130 Steel."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

Table 2: Model input data for hydrogen diffusion

D (m^2/s)	V_H (m^3/mol)	N_L (mol/m^3)	W_B (J/mol)	β	η
1.5×10^{-8}	2.0×10^{-6}	8.46×10^{5}	6.1×10^{4}	1	1

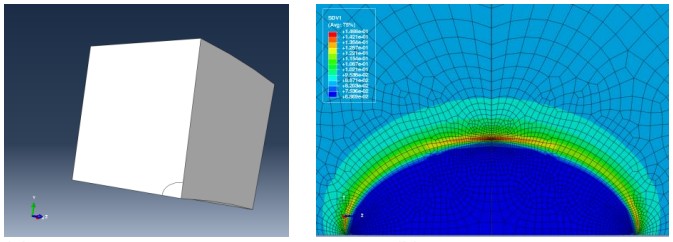


(a) **(b)**

Figure 1: (a) Experimental HA-FCG results for 4130 steel in air and 45 MPa gaseous hydrogen [25], (b) experimental and predicted HA-FCG results for 4130 steel in air and 45 MPa hydrogen gas.

A one-half symmetry CT specimen was modeled in ABAQUS to elucidate the predictive capabilities of the model for HA-FCG of 4130 steel. The coupled deformation-hydrogen diffusion parameters are provided in Table 2. Figure 1(b) shows the predicted HA-FCG results of the 4130 CT specimen in 45 MPa gaseous hydrogen, a frequency of 1 Hz, and a load ratio of R=0.1. The fatigue crack growth values were predicted at ΔK values of 7 MPa-$m^{0.5}$, 9 MPa-$m^{0.5}$, 11 MPa-$m^{0.5}$, 13 MPa-$m^{0.5}$, and 15 MPa-$m^{0.5}$. The predicted data is overlaid upon the experimental data of [25]. Analysis of the predictions in Fig. 1(a) indicate that the coupled model predicts HA-FCG within a factor of ± 1.5 and is deemed to be sufficiently calibrated to HA-FCG of 4130 steel. Of interest, then, is to use the model implementation on realistic geometries. To do so, this work leverages the results of [26], in which 4130 pressure vessels with engineered flaws of known size, semi-elliptical shape with nominal root radius 0.5 mm and an aspect ratio (a/2c) of 1/3 aligned along the length of the cylinder, are cycled from 3.5 MPa (500 psi) to 43.8 MPa (6350 psi) until failure. A finite element model of the pressure vessel, which incorporates extensive symmetry boundary conditions, was created in ABAQUS and is shown in Fig. 2a.

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Application of a Model of Hydrogen-Assisted Fatigue Crack Growth in 4130 Steel."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.



(a) **(b)**

Figure 2: (a) Pressure vessel FE model with thumbnail-shaped crack, (b) predicted hydrogen concentration at the crack tip resulting from 43.8 MPa internal hydrogen pressure. Thumbnail-shaped crack has root radius of 0.5 mm in both images.

The UMAT H-diff was then used to determine the hydrogen concentration at the crack tip as a function of the internal hydrogen pressure, 3.5 MPa (500 psi) and 43.8 MPa (6350 psi), and the elastic-plastic deformation response resulting from each loading condition. Model predictions for the hydrogen coverage are provided in Fig. 2b. Coupling the hydrogen-concentration prediction from ABAQUS and the phenomenological HA-FCG model described above, predictions of cycles to failure have been generated, based the size of the initial thumbnail-shaped crack. The model predictions, shown as a blue line, are compared to the experimental results of [26], in Fig. 3. The model accurately predicts the cycles to failure of the 4130 pressure vessels tested in [25] within a factor of 2 and is currently being updated for use with other pipeline and pressure vessel steels of interest.

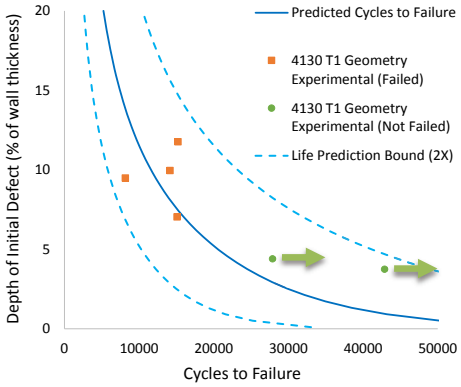


Figure 3: Cycles to failure for internal, thumbnail-shaped cracks in pressure vessels. HA-FCG predictions and experimental results as a function of initial crack size.

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Application of a Model of Hydrogen-Assisted Fatigue Crack Growth in 4130 Steel."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

CONCLUSIONS

A coupled HA-FCG model framework has been implemented that accurately predicts the cycles to failure of laboratory specimens and realistic pipe and pressure vessel geometries. The current model implementation requires minimal calibration to steels of interest (Tables 1 and 2). The strength of this current model implementation is in its ability to predict HA-FCG for many grades of steel and geometries of interest.

REFERENCES

1. Park, G.T., Koh, S. U., Jung, H. G., Kim, K. Y., *Effect of microstructure on the hydrogen trapping efficiency and hydrogen induced cracking of linepipe steel.* Corrosion Science, 2008. **50**(7): p. 1865-1871.
2. Luppo, M.I., Ovejero-Garcia, J., *The influence of microstructure on the trapping and diffusion of hydrogen in a low carbon steel.* Corrosion Science, 1991. **32**(10): p. 1125-1136.
3. Chan, S.L.I., *Hydrogen Trapping Ability of Steels with Different Microstructures.* Journal of the Chinese Institute of Engineers, 1999. **22**(1): p. 43-53.
4. Dong, C.F., Liu, Z. Y., Li, X. G., Cheng, Y. F., *Effects of hydrogen-charging on the susceptibility of X100 pipeline steel to hydrogen-induced cracking.* International Journal of Hydrogen Energy, 2009. **34**(24): p. 9879-9884.
5. *United States Energy Information Administration, About U.S. Natural Gas Pipelines*. 2016 7/24/2016]; Available from: www.eia.gov/pub/oil_gas/natural_gas/analysis_publications/ngpipeline/index.html.
6. *U.S. Department of Energy, Energy Eficiency & Renewable Energy, Alternative Fuels Data Center, Hydrogen Production and Distribution*. 2016 7/24/2016]; Available from: www.afdc.energy.gov/fuels/hydrogen_production.html#distribution.
7. *U.S. Office of Energy Efficiency & Renewable Energy, ENERGY.gov, Hydrogen Pipelines*. 2016; Available from: http://energy.gov/eere/fuelcells/hydrogen-pipelines.
8. Somerday, B., P., *Technical Reference on Hydrogen Compatibility of Materials- Plain Carbon Ferritic Steels: C-Mn Alloys (Code 1100)*. 2008, Sandia National Laboratories, Livermore, CA. p. 32.
9. Nanninga, N.E., Levy, Y. S., Drexler, E. S., Condon, R. T., Stevenson, A. E., Slifka, A. J., *Comparison of hydrogen embrittlement in three pipeline steels in high pressure gaseous hydrogen environments.* Corrosion Science, 2012. **59**(0): p. 1-9.
10. *B31.12 Hydrogen Piping and Pipelines*. 2014, American Society of Mechanical Engineers.
11. *ANSI/CSA CHMC 1-2014 Test Methods for Evaluating Material Compatibility in Compressed hydrogen Applications- Metals*. 2014, CSA Group: Ontario, Canada.
12. *ISO 11114-4:2005 Transportable Gas Cylinders- Compatibility of Cylinder and Valve Materials and Gas Contents*. 2005, ISO: Switzerland.
13. Slifka, A.J., Drexler, E. S., Nanninga, N. E., Levy, Y. S., McColskey, J. D., Amaro, R. L., Stevenson, A. E., *Fatigue crack growth of two pipeline steels in a pressurized hydrogen environment.* Corrosion Science, 2014. **78**(0): p. 313-321.
14. Drexler, E.S., Slifka, A. J., Amaro, R. L., Barbosa, N., Lauria, D. S., Hayden, L. E., Stalheim, D. G., *Fatigue crack growth rates of API X70 pipeline steel in a pressurized hydrogen gas environment.* Fatigue & Fracture of Engineering Materials & Structures, 2014. **37**(5): p. 517-525.
15. San Marchi, C., and Somerday, B.P. *Technical reference on hydrogen compatibility of materials: Plain carbon ferritic steels: C-Mn alloys (code 1100)*. 2010 October 2010 [cited 2012 January 3, 2012]; Available from: http://www.sandia.gov/matlsTechRef/.
16. Amaro, R.L., Rustagi, N., Findley, K. O., Drexler, E. S., Slifka, A. J., *Modeling the fatigue crack growth of X100 pipeline steel in gaseous hydrogen.* International Journal of Fatigue, 2014. **59**(0): p. 262-271.

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Application of a Model of Hydrogen-Assisted Fatigue Crack Growth in 4130 Steel."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

17. Amaro, R.L., Drexler, E. S., Slifka, A.J., *Fatigue crack growth modeling of pipeline steels in high pressure gaseous hydrogen.* International Journal of Fatigue, 2014. **62**: p. 249-257.
18. Amaro, R.L., Drexler, E. S., Slifka, A. J. *Development of an Engineering-Based Hydrogen-Assisted Fatigue Crack Growth Design Methodology for Code Implementation*. in *ASME 2014 Pressure Vessel and Piping Conference*. 2015. Anaheim, CA.
19. Moriconi, C., G. Henaff, and D. Halm, *Influence of hydrogen coverage on the parameters of a cohesive zone model dedicated to fatigue crack propagation.* Procedia Engineering, 2011. **10**(0): p. 2657-2662.
20. Novak, P., Yuan, R., Somerday, B. P., Sofronis, P., Ritchie, R. O., *A statistical, physical-based, micro-mechanical model of hydrogen-induced intergranular fracture in steel.* Journal of the Mechanics and Physics of Solids, 2010. **58**(2): p. 206-226.
21. Sofronis, P., McMeeking, R. M., *Numerical analysis of hydrogen transport near a blunting crack tip.* Journal of the Mechanics and Physics of Solids, 1989. **37**(3): p. 317-350.
22. Oriani, R.A., *The diffusion and trapping of hydrogen in steel.* Acta Metallurgica, 1970. **18**(1): p. 147-157.
23. Kumnick, A.J., Johnson, H. H., *Deep trapping states for hydrogen in deformed iron.* Acta Metallurgica, 1980. **28**(1): p. 33-39.
24. Boller, C., Seeger, T., *Materials Data for Cyclic Loading. Part B: Low-Alloy Steels*. Materials Science Monographs, 42B. 1987, Amsterdam, The Netherlands: Elsevier Science Publishers.
25. San Marchi, C., Dedrick, D. E., Van Blarigan, P., Somerday, B. P., Nibur, K. A., *Pressure Cycling of Type 1 Pressure Vessels with gaseous Hydrogen*, in *Fourth International Conference on Hydrogen Safety*. 2011: San Francisco, CA.
26. San Marchi, C., Harris, A., Yip, M., Somerday, B. P., Nibur, K. A. *Pressure Cycling of Steel Pressure Vessels with Gaseous Hydrogen*. in *ASME 2012 Pressure Vessels & Piping Conference, PVP2012-78709*. 2012. Toronto, Ontario, Canada: American Society of Mechanical Engineers.

Amaro, Robert; Drexler, Elizabeth; Long, Benjamin; O'Connor, Devin; Slifka, Andrew.
"Application of a Model of Hydrogen-Assisted Fatigue Crack Growth in 4130 Steel."
Paper presented at International Hydrogen Conference 2016, Moran, WY, United States. September 11, 2016 - September 14, 2016.

IN SITU NEUTRON TRANSMISSION BRAGG EDGE MEASUREMENTS OF STRAIN FIELDS NEAR FATIGUE CRACKS GROWN IN AIR AND IN HYDROGEN

MATTHEW CONNOLLY
National Institute of Standards and Technology
Boulder, CO, USA

PETER BRADLEY
National Institute of Standards and Technology
Boulder, CO, USA

ANDREW SLIFKA
National Institute of Standards and Technology
Boulder, CO, USA

ELIZABETH DREXLER
National Institute of Standards and Technology
Boulder, CO, USA

ABSTRACT

In situ transmission Bragg edge measurements of the strain fields in an X70 steel were performed near fatigue cracks grown in air and in hydrogen. Through the use of a novel test chamber which is capable of pressurization with hydrogen gas, and amenable to neutron-scattering measurements, fatigue cracks were grown in X70 steel specimens at the NG-6 neutron-imaging beam line at the NIST Center for Neutron Research. The measured strain fields are presented and discussed in the context of proposed mechanisms of hydrogen-assisted fatigue crack growth.

INTRODUCTION

Pipelines are the most likely means of transporting gaseous hydrogen to support clean power generation and clean transportation[1]. Although steels are a cost-effective solution for the construction of hydrogen gas pipelines, their fatigue and fracture properties are adversely affected by the presence of gaseous hydrogen[2-4]. The corrosive effect of hydrogen on steels manifests in fatigue crack growth rates (FCGRs) which are one to two orders of magnitude faster when grown in H_2 compared with those grown in air. Although the embrittlement effect has been observed since 1875[5], the exact mechanism (or mechanisms) that dominates remains to be elucidated. Until recently, most hydrogen embrittlement research comprised anecdotal material-specific observations, and concentrated on urgent technical problems[6]. Determination of the mechanism(s) is necessary to develop designs for high-functioning hydrogen transportation and storage applications, which may, in turn, provide insight into hydrogen effects in other material classes.

The National Institute of Standards and Technology (NIST) has an ongoing program to generate hydrogen-assisted FCGR (HA-FCGR) data[7]. The laboratory at NIST is one of only a few capable of performing FCGR measurements in gaseous hydrogen — the same failure mode that would be expected in service. The program has recently begun expanding to complement

Bradley, Peter; Connolly, Matthew; Drexler, Elizabeth; Slifka, Andrew.
"In Situ Neutron Transmission Bragg Edge Measurements of Strain Fields Near Fatigue Cracks Grown in Air and in Hydrogen."
Paper presented at International Hydrogen Conference 2016, Jackson Hole, WY, United States. September 26, 2016 - September 29, 2016.

in-air and HA-FCGR data with physics-based modeling and a scientific corroboration of the prevailing mechanisms. Ultimately, the goal is to use the data collected at NIST to create and calibrate a physics-based predictive model for the damage and deformation response of pipeline steels in gaseous hydrogen.

There exist several proposed mechanisms of embrittlement including hydrogen-enhanced decohesion (HEDE) and hydrogen-enhanced localized plasticity (HELP)[8-9]. In the HEDE mechanism, interstitial accumulation of hydrogen at locations of high triaxial stresses leads to the weakening of Fe-Fe bonds once the hydrogen concentration reaches a critical concentration. In the HELP mechanism, the introduction of hydrogen gas creates areas of extended dislocations in the Fe lattice and enhances dislocation mobility in the steel framework. A full quantification of the elastic and plastic deformation as a function of stress and hydrogen concentration is not yet determined.

Neutron-diffraction measurements of strain in steel are readily available to study elastic lattice deformation leading to HA-FCGR. In these measurements, a spatial mapping of the atomic lattice spacing is produced. With an appropriate measurement of an unstressed lattice spacing, the measured lattice spacing during mechanical loading can then determine the elastic lattice strain. However, the determination of strain fields near fatigue cracks grown in H_2 is challenging experimentally, because of the rapid diffusion of hydrogen from the steels. Even after extended exposure to H_2, in-air FCGRs are observed in steels once the specimen is removed from the H_2 [10]. In order to fully understand the HA-FCGR mechanism, it is necessary to perform any measurements in situ. To achieve this, a test chamber has been developed that can hold moderate gas pressure (3.4 MPa, 500 psi), and has the capability of mechanical loading of steel specimens for neutron and synchrotron x-ray scattering measurements. The chamber has been designed to be nearly transparent to neutron radiation, ideal for diffraction and radiography. The chamber is compatible with load frames available at the user facilities at Argonne National Laboratory Advanced Photon Source, the NIST Center for Neutron Research, and Oak Ridge National Laboratory Spallation Neutron Source.

In this paper, we present neutron Transmission Bragg Edge Spectroscopy (TBES) measurements of the strain fields around crack tips grown via fatigue in air and in a hydrogen environment. Drastic differences in both magnitude and spatial extent of the crack tip strain fields grown in each condition are demonstrated.

MATERIALS

The material used for this study was an X70 pipeline steel. The material was chosen due to its heavy use in pipelines as well as the abundance of in-air and hydrogen-assisted FCGR data on the material. Table 1 shows the chemical composition of the material used in this study; the balance is Fe. Table 2 shows the tensile properties of the material used in this study. From optical microscopy, the X70 steel was determined to be polygonal ferrite and either acicular ferrite or bainite. There may be other constitutents that are not resolvable without employing more advanced analytical techniques.

Bradley, Peter; Connolly, Matthew; Drexler, Elizabeth; Slifka, Andrew.
"In Situ Neutron Transmission Bragg Edge Measurements of Strain Fields Near Fatigue Cracks Grown in Air and in Hydrogen."
Paper presented at International Hydrogen Conference 2016, Jackson Hole, WY, United States. September 26, 2016 - September 29, 2016.

Table 1: Chemical compositions of the X70 pipeline steel, in mass percent.

	C	Mn	P	S	Si	Cu	Ni
Mass %	0.048	1.43	0.009	0.001	0.17	0.220	0.014
	Cr	**Mo**	**V**	**Nb**	**Ti**	**Al**	**Fe**
Mass %	0.240	0.005	0.004	0.054	0.027	0.015	Balance

Table 2: Tensile properties of the X70 pipeline steel, measured in the transverse orientation.

Yield Strength (MPa)	Ult. Tensile Strength (MPa)
509	609

METHODS

Fatigue Crack Growth Rate Measurements

Measurement of the FCGR was performed according to ASTM E647[11] for compact tension (C(T)) specimens (length W = 26.67 mm and thickness B = 3 mm) with a Crack Mouth Opening Diplacement (CMOD) gauge attached to the load line. All C(T) specimens were fatigue pre-cracked in air to obtain a sharp initial crack. The pre-crack length for all specimens was approximately 10 mm. All FCGR measurements were performed at a load ratio R = 0.5 and maximum load Pmax = 1.7 kN. A cycling frequency f = 1 Hz was used for the H2 test and a cycling frequency of f = 10 Hz was used for the air test. Research-grade (99.9995 % pure) H was used for testing. Analysis of the test gas indicated O and H2O concentrations below the detection limits of 0.5 ppm for O and 1 ppm for H2O. Figure 1 shows the FCGR for the X70 steel in air and H2.

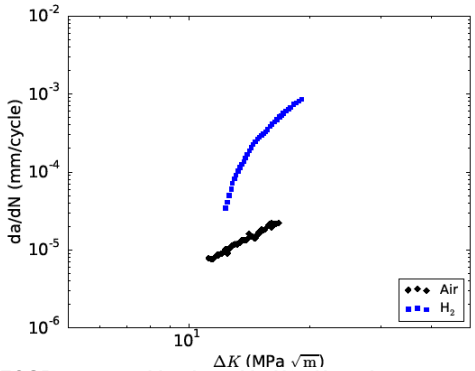


Figure 1: FCGR measured in air and in H2. Note the presence of a "knee", or a sharp increase in FCGR in the H2 data at ~ ΔK = 12 MPa m1/2.

Bradley, Peter; Connolly, Matthew; Drexler, Elizabeth; Slifka, Andrew.
"In Situ Neutron Transmission Bragg Edge Measurements of Strain Fields Near Fatigue Cracks Grown in Air and in Hydrogen."
Paper presented at International Hydrogen Conference 2016, Jackson Hole, WY, United States. September 26, 2016 - September 29, 2016.

Gas Pressure Chamber

Figure 1 shows a diagram of the test chamber utilized for this experiment. To be sufficiently transparent to neutrons and x-rays, the chamber was constructed of the aluminum alloy 6061-T6. The wall thickness of the thin portion of the chamber is 3.175 mm, which allows ≈ 95% of incident neutrons to be transmitted through the chamber.

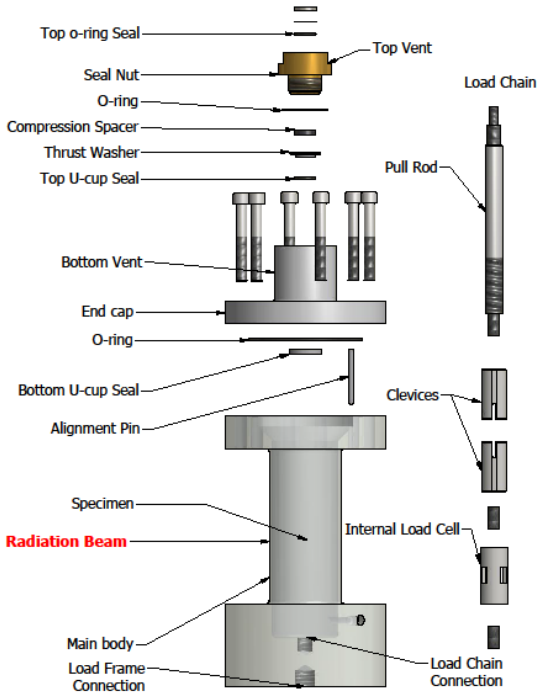


Figure 2: Schematic of the test chamber.

Strain Measurements

TBES measurements were performed at the NIST Center for Neutron Research (NCNR) NG-6 beamline[12]. This beamline uses a pyrolytic graphite double monochromator system to vary the incident neutron wavelength. A LiF pixelated detector plate was used [13]. The detector plate is 28 cm × 28 cm in area, and the pixel size is 50 µm × 50 µm. For strain measurements, a C(T) specimen with length W = 26.67 mm and thickness B = 6 mm was used. Fatigue cracks were grown with a load ratio R = 0.5, maximum load Pmax = 3.4 kN, and

Bradley, Peter; Connolly, Matthew; Drexler, Elizabeth; Slifka, Andrew.
"In Situ Neutron Transmission Bragg Edge Measurements of Strain Fields Near Fatigue Cracks Grown in Air and in Hydrogen."
Paper presented at International Hydrogen Conference 2016, Jackson Hole, WY, United States. September 26, 2016 - September 29, 2016.

a cycling frequency f =0.033 Hz. A larger Pmax was used for the strain measurements compared to the FCGR measurements because of the larger specimen thickness. The loading frequency was limited because of the constraints of a stepper motor on the load frame. Separate, identically prepared specimens were used for the in-air and in- H2 measurements. Each fatigue crack was cyclically loaded for at least N = 3000 cycles to ensure fresh crack growth in each environment. The specimens were then held at a given load for the TBES measurements. TBES employs Bragg's law[14-15]:

$$\lambda = 2\, d_{hkl} sin\theta \qquad \textbf{(1)}$$

where λ is the wavelength of incident radiation, dhkl is the lattice spacing corresponding to reflection from a particular crystallographic (hkl) plane, and θ is half of the scattering angle. For a given (hkl) reflection, the Bragg angle increases with λ until the back-scattering condition, θ = π. For larger wavelengths, λ > 2dhkl, Eqn. 1 is no longer satisfied for any θ, and therefore no scattering from that particular crystallographic plane occurs, and the transmission through the sample will increase sharply (so-called "Bragg Edge", see Fig. 3). Thus, by locating the wavelength in which a Bragg Edge occurs, the material lattice spacing can be determined. In practice, the sharp increase in transmitted intensity is smoothed by the finite wavelength resolution of the instrument. Close to the Bragg Edge, the transmitted intensity is accurately expressed as the convolution of a step function, which represents the Bragg edge, and a Gaussian function, which represents the instrument wavelength broadening. The convolution of a step and Gaussian function can be written as a complimentary error function,

$$Tr(\lambda) = A + C\ Erfc(\lambda - d_{hkl}/2\sigma) \qquad \textbf{(2)}$$

where A is a parameter related to the background intensity, C is related to the neutron scattering cross section and the thickness of the specimen, and σ is related to the full-width at half-maximum of the distribution of neutron wavelengths from instrument broadening. Figure 3 shows a sample of the transmission acquired over a 2x2 pixel area (100 μm x 100 μm), as well as the fit to Eqn. 2. Thus, a measurement of the transmission of neutrons, through a material and incident on each detector pixel, provides a route to a fast, accurate determination of the lattice spacing, d, of the material with spatial resolution governed by the pixel size. Shifts in this measured lattice spacing with respect to a reference, unstrained lattice spacing, d_0, provide a measure of the material's strain, given by

$$\varepsilon = \frac{d - d_0}{d_0} \qquad \textbf{(3)}$$

Figure 4 shows radiographs of the experimental setup using incident neutron wavelength above and below the Bragg edge.

Bradley, Peter; Connolly, Matthew; Drexler, Elizabeth; Slifka, Andrew.
"In Situ Neutron Transmission Bragg Edge Measurements of Strain Fields Near Fatigue Cracks Grown in Air and in Hydrogen."
Paper presented at International Hydrogen Conference 2016, Jackson Hole, WY, United States. September 26, 2016 - September 29, 2016.

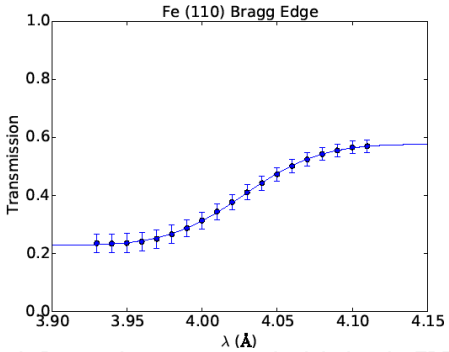


Figure 3:Sample Bragg edge spectra acquired during the TBES measurement, with fit function according Eqn 2. The transmission was averaged over a 2x2 pixel area to achieve the counting statistics indicated by the error bars.

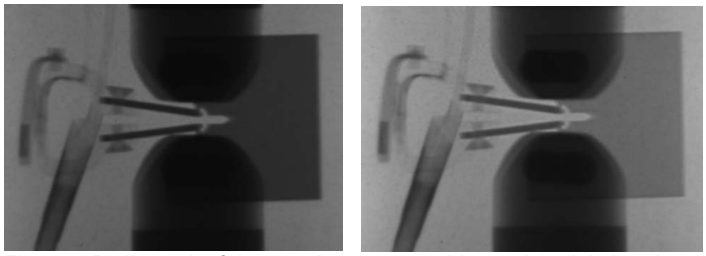


Figure 4: Radiograph of the experiment setup with wavelength below the Bragg edge (left) and above the Bragg edge (right). On the left side of each image are a clip gauge and wires associated with the load cell. In the top and bottom center the two dark areas are the two clevises, which hold the C(T) specimen during the test. The transmission through the specimen is noticeably larger for the radiograph above the Bragg edge.

RESULTS AND DISCUSSION

Contour images of the measured strain fields in air and in H2 for loading P=5.15 kN are shown in Fig. 5. The contour images show a larger compressive crack tip strain for the crack grown in H2 as compared to air. Because a tensile load is applied in the plane of the specimen, the through-thickness strain component observed via TBES is compressive. The results of this study appear to suggest that the effect of hydrogen is to enhance the crack-tip strain for a given applied load beyond that observed in air.

Bradley, Peter; Connolly, Matthew; Drexler, Elizabeth; Slifka, Andrew.
"In Situ Neutron Transmission Bragg Edge Measurements of Strain Fields Near Fatigue Cracks Grown in Air and in Hydrogen."
Paper presented at International Hydrogen Conference 2016, Jackson Hole, WY, United States. September 26, 2016 - September 29, 2016.

This result is consistent with the HEDE mechanism; as H absorbs into the material, the presence of H in the lattice decreases the Fe-Fe interaction energy, leading to larger elastic strains for a given applied load. The HEDE mechanism predicts both an intragranular decohesion, between the Fe atoms within the lattice, as well as an intergranular decohesion, between grains in the material. However, it should be emphasized that the strain measurements presented here provide only a measurement of the elastic lattice strain. Because lattice strain is measured, hydrogen-induced intergranular decohesion is not indicated in these measurements.

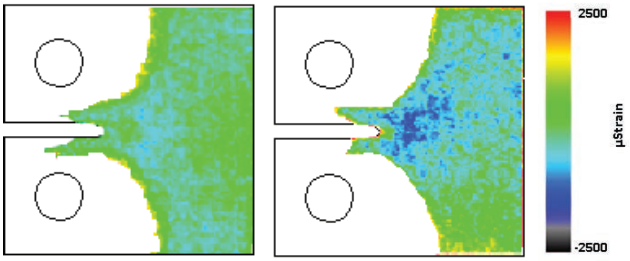


Figure 5: Measured strain fields for the cracks grown in air (left) and in H2 (right).

However, the results are consistent with an intragranular decohesion due to the presence of H2. Further, because only elastic strains are measured without any information on plasticity (ie. dislocation pileups), this measurement cannot provide evidence for, or against, the HELP mechanism. It has been argued that the increase in dislocation pileup due to the HELP mechanism likely aides in providing the large hydrogen concentrations necessary for inter- and intra-granular decohesion in the HEDE mechanism[16]. These measurements are primarily susceptible to effects of the HEDE mechanism, it is therefore likely for concurrent mechanisms to be acting.

The power in the measurements presented here is the possibility to quantify the extent of the enhanced elastic strain field ahead of the crack tip. For distances ahead of the crack tip in which the material conforms to Linear Elastic Fracture Mechanics (LEFM), the crack tip stress intensity factor, K, completely defines the stress state ahead of the crack tip. In order to achieve the enhanced elastic strain field in H2 shown here, the stress state must be characterized by a larger K than observed in air. Because ΔK is the independent variable in FCGR measurements, this quantification is crucial to understand the differences in air and H2 environment FCGR, as presented in Fig. 2. Attempts are underway to quantify these differences for a range of loads, crack lengths, and gas pressures.

CONCLUSIONS

Strain fields near fatigue crack tips that were grown in air and in H2 are presented. In H2, the magnitude and spatial extent of the strain field was enhanced

Bradley, Peter; Connolly, Matthew; Drexler, Elizabeth; Slifka, Andrew.
"In Situ Neutron Transmission Bragg Edge Measurements of Strain Fields Near Fatigue Cracks Grown in Air and in Hydrogen."
Paper presented at International Hydrogen Conference 2016, Jackson Hole, WY, United States. September 26, 2016 - September 29, 2016.

compared with the strain field in air. The results presented are consistent with a presumed intragranular HEDE mechanism contributing to the deformation. In order to fully elucidate the range over which the HEDE mechanism is exhibited, it is crucial to determine the crack-tip strain field over a larger range of applied K. In addition, the strain field must be measured as a function of H pressure, mechanical load, and crack length to provide the physics-based models with sufficient data to be fully predictive. Future measurements are planned to acquire this information, including synchrotron x-ray measurements, which will increase the spatial resolution of the measurements by a factor of ≈ 5. Further, the combination of elastic strain measurements with a quantification of the plasticity via the dislocation density, which can be acquired through synchroton x-ray measurements, will further elucidate the mechanisms of HA-FCG.

REFERENCES

[1] Fekete, J. R., Sowards, J. W., & Amaro, R. L. (2015). Economic impact of applying high strength steels in hydrogen gas pipelines. *International Journal of Hydrogen Energy*, *40*(33), 10547-10558.

[2] Slifka, A. J., Drexler, E. S., Nanninga, N. E., Levy, Y. S., McColskey, J. D., Amaro, R. L., & Stevenson, A. E. (2014). Fatigue crack growth of two pipeline steels in a pressurized hydrogen environment. *Corrosion Science*, *78*, 313-321.

[3] Nanninga, N. E., Levy, Y. S., Drexler, E. S., Condon, R. T., Stevenson, A. E., & Slifka, A. J. (2012). Comparison of hydrogen embrittlement in three pipeline steels in high pressure gaseous hydrogen environments. *Corrosion Science*, *59*, 1-9.

[4] Amaro, R. L., Rustagi, N., Findley, K. O., Drexler, E. S., & Slifka, A. J. (2014). Modeling the fatigue crack growth of X100 pipeline steel in gaseous hydrogen. *International Journal of Fatigue*, *59*, 262-271.

[5] Johnson, W. H. (1874). On some remarkable changes produced in iron and steel by the action of hydrogen and acids. *Proceedings of the Royal Society of London*, *23*(156-163), 168-179.

[6] Barnoush, A. (2011). Hydrogen embrittlement. *Saarland University*.

[7] Slifka, A. J., Drexler, E. S., Amaro, R., Lauria, D., Fekete, J., & Eason, K. R. (2013). Materials testing in hydrogen gas at NIST, Boulder. Retrieved August 14, 2016, from http://www.sandia.gov/matlsTechRef/advmat_presentations/Slifka_NIST.pdf

[8] Du, Y. A., Ismer, L., Rogal, J., Hickel, T., Neugebauer, J., & Drautz, R. (2011). First-principles study on the interaction of H interstitials with grain boundaries in α-and γ-Fe. *Physical Review B*, *84*(14), 144121.

[9] Moody, N. R., Thompson, A. W., Ricker, R. E., Was, G. S., & Jones, R. H. (2003). Hydrogen effects on material behavior and corrosion deformation interactions. *TMS, Warrendale, PA*.

[10] Darcis, P. P., McColskey, J. D., Lasseigne, A. N., & Siewert, T. A. (2009). Hydrogen effects on fatigue crack growth rate in high strength pipeline steel. In*Effects of Hydrogen on Materials: Proceedings of the 2008 International Hydrogen Conference, September 7-10, 2008, Jackson Lake Lodge, Grand Teton National Park, Wyoming, USA* (p. 381). ASM International.

[11] ASTM International. (2011). *Standard test method for measurement of fatigue crack growth rates*. ASTM International.

[12] Hussey, D. S., Jacobson, D. L., Arif, M., Huffman, P. R., Williams, R. E., & Cook, J. C. (2005). New neutron imaging facility at the NIST. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *542*(1), 9-15.

[13] Faenov, A., Matsubayashi, M., Pikuz, T., Fukuda, Y., Kando, M., Yasuda, R., ... & Kodama, R. (2015). Using LiF crystals for high-performance neutron imaging with micron-scale resolution. *High Power Laser Science and Engineering*, *3*, e27.

Bradley, Peter; Connolly, Matthew; Drexler, Elizabeth; Slifka, Andrew.
"In Situ Neutron Transmission Bragg Edge Measurements of Strain Fields Near Fatigue Cracks Grown in Air and in Hydrogen."
Paper presented at International Hydrogen Conference 2016, Jackson Hole, WY, United States. September 26, 2016 - September 29, 2016.

[14] Tremsin, A. S., McPhate, J. B., Vallerga, J. V., Siegmund, O. H. W., Feller, W. B., Bilheux, H. Z., ... & Penumadu, D. (2010). Transmission Bragg edge spectroscopy measurements at ORNL spallation neutron source. In *Journal of Physics: Conference Series* (Vol. 251, No. 1, p. 012069). IOP Publishing.

[15] Santisteban, J. R., Edwards, L., Fitzpatrick, M. E., Steuwer, A., Withers, P. J., Daymond, M. R. & Schooneveld, E. M. (2002). Strain imaging by Bragg edge neutron transmission. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *481*(1), 765-768.

[16] Novak, P., Yuan, R., Somerday, B. P., Sofronis, P., & Ritchie, R. O. (2010). A statistical, physical-based, micro-mechanical model of hydrogen-induced intergranular fracture in steel. *Journal of the Mechanics and Physics of Solids*, *58*(2), 206-226.

Bradley, Peter; Connolly, Matthew; Drexler, Elizabeth; Slifka, Andrew.
"In Situ Neutron Transmission Bragg Edge Measurements of Strain Fields Near Fatigue Cracks Grown in Air and in Hydrogen."
Paper presented at International Hydrogen Conference 2016, Jackson Hole, WY, United States. September 26, 2016 - September 29, 2016.

Surface Integral Equation Formulation of Electromagnetic Scattering for Cloaking Applications

Alex J. Yuffa
National Institute of Standards and Technology, Boulder, CO 80305

We investigate the role of the electric field and its normal derivative in three-dimensional electromagnetic scattering theory. In particular, we present an alternative integral equation formulation that uses the electric field and its normal derivative as the boundary unknowns. In particular, we extend a traditional formulation that is used in two-dimensional scattering theory to three-dimensions. Our formalism may be advantageous in some avant-garde applications such as cloaking (invisibility) devices that have fascinated minds for decades, if not centuries. It is well-known that the boundary conditions have an extensive effect on the solution, and thus it is not surprising that unorthodox boundary conditions provide the most elegant means for achieving a solution. Indeed, for cloaking applications, such unorthodox boundary conditions on the normal components of the electromagnetic field or its derivative, rather than on the tangential components, provide the required solution. In particular, it has been shown that the vanishing of these normal components is required to achieve some cloaking aspects (R. Weder, J. Phys. A: Math. Theor., 41, 415401, 2008 and I.V. Lindell and A.H. Sihvola, IEEE Trans. Antennas Propag., 58, 1128–1135, 2010).

The presentation begins with a derivation of the continuity conditions for the electric field and its normal derivative across an interface. From these conditions, we uncover several intriguing relationships involving closed surface integrals of the field and/or its normal derivative. For example, we show that

$$\int_{\Sigma} \boldsymbol{N} \cdot \frac{\partial \boldsymbol{E}}{\partial N} \, \mathrm{d}S = 0$$

for closed surfaces with a zero mean curvature. We end the presentation with a physical interpretation of the surface integrals appearing in the above discussed integral equation formulation. In order not to obscure the physical/geometric awareness, the presentation is given from a tensor-calculus perspective.

Yuffa, Alexey.
"Surface Integral Equation Formulation of Electromagnetic Scattering for Cloaking Applications."
Paper presented at USNC-URSI National Radio Science Meeting, Boulder, CO, United States. January 4, 2017 - January 6, 2017.

Preprint: Software Quality, Reliability and Security Companion (QRS-C), 2017 IEEE International Conference on, pp. 61-68. IEEE, 2017

Combinatorial and MC/DC Coverage Levels of Random Testing

Sergiy Vilkomir, Aparna Alluri
Department of Computer Science
East Carolina University
Greenville, NC 27858, USA
vilkomirs@ecu.edu

D. Richard Kuhn, Raghu N. Kacker
Information Technology Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899, USA
{kuhn, raghu.kacker}@nist.gov

***Abstract*—Software testing criteria differ in effectiveness, numbers of required test cases, and a process of test generation. Specific criteria are often compared with random testing as a simplest basic approach and, in some cases, random testing shows a surprisingly high level of effectiveness. One of the reasons is that any random test set has a specific level of coverage according to any coverage criterion. Numerical evaluation of coverage levels of random testing according various coverage criteria is interesting research task important for understanding relationship between different testing approaches. In this paper we experimentally evaluate coverage levels of random testing for two criteria – MC/DC and combinatorial *t*-way testing. The results could be used for selection optimal methods for practical testing, and development of new testing methods based on integration of existing approaches.**

Keywords—random testing; combinatorial testing; MC/DC; pairwise; coverage

I. INTRODUCTION

Testing coverage criteria are widely used in software testing. According to ISO/IEC/IEEE 29119-1:2013 Standard [1], test coverage is "a degree, expressed as a percentage, to which specified test coverage items have been exercised by a test case or test cases". Simple examples include statement coverage, path coverage, and branch coverage. A more sophisticated example is *t*-way combinatorial coverage, which requires every possible combination of values of *t* parameters be included in some test case in the test suite [2, 3]. In the most basic form for *t*=2, this criterion is known as pairwise testing and requires all possible pairs of values be covered [4]. One of the strongest coverage criteria for testing logical predicates is Modified Condition/Decision Coverage (MC/DC) [5], which requires every condition in a decision be covered, and "each condition has been shown to independently affect the decision's outcome" [6]. MCDC coverage subsumes branch coverage, and in turn, statement coverage of code.

Because MCDC is such a strong criterion, the US Federal Aviation Administration (FAA) has for many years required MCDC testing for Level A (life critical) software aboard commercial aircraft [7, 8], but it is rarely used outside of the aerospace industry. One of the most significant barriers to wider use of MCDC is its expense. While testing for consumer grade software is roughly the same as the cost of producing the code, in the aviation industry spending could be seven times more on verification than development [9]. To encourage wider use of MCDC beyond the aerospace industry, the cost of its application will need to be reduced.

Testing criteria differ in effectiveness, numbers of required test cases, and a process of test generation. Empirical evaluation and experimental comparison of testing criteria started in 1980's [10, 11] but this is still an important research direction [12, 13, 14]. Specific criteria are often compared with random testing as a simplest basic approach [15, 16, 17, 18].

In some cases, random testing shows a surprisingly high level of effectiveness. For example, for testing logical expressions, *t*-way testing have an advantage over random testing but this benefit is not significant [18, 19]. One of the reasons is that any random (or any other) test set has a specific level of coverage according to any coverage criterion. Thus, random testing has a certain level of pairwise coverage, etc. This level is never 100% but could be quite high. Numerical evaluation of coverage levels of random testing according various coverage criteria is interesting and important research task. It could be useful for understanding relationship between different testing approaches, selection optimal methods for practical testing, and development of new testing methods based on integration of existing approaches.

In this paper we experimentally evaluate coverage levels of random testing for two criteria – MC/DC and *t*-way testing. The paper is structured as follows: Section II considers a general problem of the integration of testing techniques to reduce the number of test cases and increase the code coverage and testing effectiveness. As a part of this problem, the more specific research question about evaluation of the MC/DC and combinatorial coverage levels of random testing is discussed. The methods and scope of the investigation are described in Section III. Section IV provides experimental results on the MC/DC coverage level of random testing and Section V provides similar results for combinatorial coverage levels. Conclusions and directions for future work are presented in Section VI.

II. RESEARCH QUESTION

Testing according to any coverage criterion, as well as other testing approaches, has some benefits and some challenges. For example, MC/DC has good effectiveness for testing logical expressions but the process of test generation to satisfy MC/DC is quite complicated. Random testing could be effective but usually only for a large number of test cases.

Alluri, Aparna; Kacker, Raghu; Kuhn, David; Vilkomir, Sergiy.
"Combinatorial and MC/DC Coverage Levels of Random Testing."
Paper presented at IEEE International Conference on Software Quality Reliability and Security, Prague, Czech Republic. July 25, 2017 - July 29, 2017.

Combinatorial criteria are very effective in many situations but not for testing logical expressions.

Taking this into consideration, a natural idea is trying to combine different testing approaches to use advantages of each of them. The purpose of such combination is to minimize a number of test cases, maximize effectiveness of testing, and simplify test generation as much as possible. Researches in this direction include combining functional and structural testing [20, 21], model-based and combinatorial testing [22], model-based and search-based testing [23] and more.

Some preliminary research suggests that using combinatorial testing in conjunction with model-based approaches can significantly reduce the cost of achieving MCDC coverage [24]. Part of the savings occurs because tests based on *t*-way covering arrays will necessarily also cover some proportion of MCDC and branch coverage. For example, a test set covering all *2*-way combinations of binary variable settings will ensure that branch predicates containing only two binary variables will be instantiated in all possible ways. But *2*-way (pairwise) test sets will also cover some proportion of *t*-way combinations for all $t > 2$, up to n, where n is the number of variables. Similarly, tests with random values will also cover a significant proportion of *t*-way combinations.

Even when two criteria are completely independent and use different principles, a test set satisfied the first criterion has some level of coverage according to the second criterion and opposite. For example, 100% MC/DC test set has some pairwise coverage and the test set provided 100% pairwise coverage also provides a certain percentage of MC/DC coverage. This fact is well-known and there are some initial results on such relationships (i.e., for pairwise and MC/DC testing [14]) but this area still requires more numerical evaluations based on empirical investigations.

Because of its simplicity, random testing is a good candidate for combining with coverage criteria. For this purpose it is necessary to understand how good random testing is in providing coverage for different criteria and how level of coverage changes depending on the number of test cases in the random test set. We consider two specific research questions:

- RQ1: What is the level of MC/DC coverage of random test sets of different sizes?
- RQ2: What is the level of *t*-way coverage of random test sets of different sizes?

Answers on these questions do not solve the problem of combining different testing criteria but are necessary and important steps in this direction.

III. Methods and Scope of Investigation

A. Used tools

To measure MC/DC coverage, we used CodeCover and Testwell CTC++ tools. CodeCover is an open-source white-box testing tool developed at the University of Stuttgart, Germany in 2007 [25, 26]. CodeCover measures several types of coverage including term coverage (subsumes MC/DC) and supports several programming languages including Java and C. Testwell CTC++ is a code coverage and dynamic analysis tool for C and C++ code but also supports Java, and C#. The tool has been developed by Testwell Ltd company (Finland) [27] and since 2013 is owned by Verifysoft Technology GmbH [28]. As well as CodeCover, CTC++ provides measurements of several types of coverage including MC/DC. A sample report produced by Testwell CTC++ is presented in Fig. 1.

We used two tools to evaluate MC/DC levels because results of such evaluation for the same software and the same test sets often are different for different tools. The main reason for this is that there is no commonly accepted definition of an incomplete (not 100%) MC/DC coverage. Different principles are used in different tools. Some tools evaluate coverage separately for each logical condition and calculate an average value then. Other tools create complete MC/DC test sets, compare them with actually used test sets, and evaluate percentage of coverage based on this.

```
HITS/TRUE      FALSE    LINE DESCRIPTION
=============================================================

        2                6 FUNCTION main()
        1          1     10   if (myfile . is_open ( ))
        8          1     12     while (getline ( myfile , line ))
       24          8     19       while (iss >> val)
                         22       }+
        3          5     26       if (( a && b ) || ( a && c ))
        2                26         1: (T && T) || (_ && _)
        1                26         2: (T && F) || (T && T)
        0                26         3: (F && _) || (T && T)
                   1     26         4: (T && F) || (T && F)
                   0     26         5: (T && F) || (F && _)
                   0     26         6: (F && _) || (T && F)
                   4     26         7: (F && _) || (F && _)
                         26         MC/DC (cond 1): 1 + 7, 1 - 6
                         26         MC/DC (cond 2): 1 + 4, 1 - 5
                     -   26         MC/DC (cond 3): 2 - 5, 3 - 7
                         26         MC/DC (cond 4): 2 + 4, 3 - 6
                         26       }+
                         27       else
                         27       }+
                         28     }+
                         30   }+
                         31   else
                         31   }+
        2                32   return 0
                         33 }

***TER  93 % ( 13/ 14) of FUNCTION main()
      100 % ( 21/ 21) statement
```

Fig. 1. Testwell CTC++ sample report.

It is not our task in this paper to judge different approaches to MC/DC evaluation. There are some justifications for all of them. However, we provide the results of MC/DC evaluation from two tools and compare them in Section 4.

To measure *t*-way coverage, we used the Combinatorial Coverage Measurement (CCM) tool developed by National Institute of Standards and Technology (NIST) and the Centro Nacional de Metrologia of Mexico [29, 30]. CCM can analyze existing levels of *t*-way (t=2…6) coverage for any test set. Fig. 2 shows CCM tool user interface. The tool displays a graph showing the coverage for the given tests. Additional tests can be added to increase the coverage. The percentage of coverage can be viewed in the "Results" tab.

B. Main steps and the scope of the investigation

The general organization of experimental evaluation is illustrated in Fig. 3.

Alluri, Aparna; Kacker, Raghu; Kuhn, David; Vilkomir, Sergiy.
"Combinatorial and MC/DC Coverage Levels of Random Testing."
Paper presented at IEEE International Conference on Software Quality Reliability and Security, Prague, Czech Republic. July 25, 2017 - July 29, 2017.

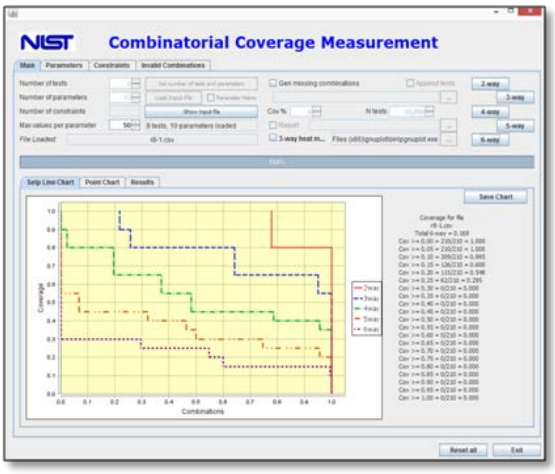


Fig. 2. CCM tool user interface.

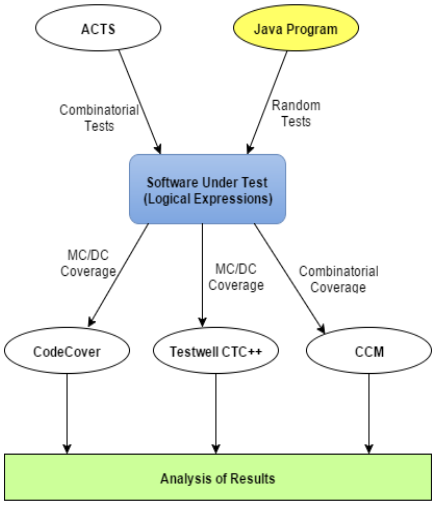


Fig. 3. Organization of experimental evaluation.

The investigation included the following main steps:

Step 1. Generation of logical expressions of different sizes (i.e., different numbers of logical variables in expressions). The sets of expressions were put into software programs which are without real functionality but used only for testing purposes to evaluate coverage levels. A total of 100 logical expressions were generated for testing. 50 expressions were generated from 10 fixed input variables and another 50 from 20 fixed input variables. Both sets contain 25 simple and 25 complex expressions of different sizes: 25 expressions of size 3, 15 of size 4, 5 of size 5, 2 of size 6, and 1 of each size 7, 8, and 9. Some examples of the generated expressions from 10 variables are presented in Table I. The sizes of expressions were chosen in such a way that they reflect situations in real software i.e. more expressions of small size and less of large size. The used proportion of different sizes approximately corresponds with data on expression sizes reported at [31, 32].

TABLE I. EXAMPLES OF LOGICAL EXPRESSIONS

	Simple expressions	**Complex expressions**
Size 3	d∨f∨c	(i∨!d) ∧ (!i∨!e)
Size 4	(c∨f) ∧ (i∨g)	(f∧ (a∨h)) ∨ (!h∧e∨!f)
Size 5	(c∧e) ∨ (!a∧f∨d)	(!h∨!f) ∧ (!c∨h∨e∧a) ∧ (f∨!e)

Step 2. Generation of random test sets of different sizes, where the size of a random test set is the number of test cases in it. To generate test cases and check that all test cases in a test set are different, we create a simple software program which used Java random number generator. 42 random test sets of different sizes were generated for 10 variables and 42 similar sets for 20 variables. The sizes of the generated random test sets were 1, 2, 3,… 25, 30, 40,…100, 200,….900, 1024. For each size, 3 test sets were generated, so the coverage for any random test size was the average of the three coverage values.

Step 3. Evaluation of MC/DC coverage for random test sets using the CodeCover tool.

Step 4. Evaluation of MC/DC coverage for random test sets using the Testwell CTC++ tool.

Step 5. Evaluation of *t*-way coverage for random test sets using the CCM tool.

Step 6. Analysis of experimental data.

The total numbers of logical expressions and test sets are summarized in Table II.

TABLE II. SCOPE OF EXPERIMENTAL TESTING

Number of		
Logical Variables	**Logical Expressions of mixed sizes**	**Random Test Sets**
10	50	126
20	50	126
Total	100	252

IV. MC/DC LEVEL OF RANDOM TESTING

All test inputs in generated random test sets have values T (True) or F (False). The example of the test set of size 5 as one of 42 generated random test sets is presented in Table III. Detailed data on MC/DC coverage by CoderCover and CTC++ tools for all random test sets are given in Table IV.

TABLE III. EXAMPLE OF THE RANDOM TEST SET OF SIZE 5

	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**
1	T	T	F	F	T	F	F	F	T	F
2	T	F	T	F	F	T	T	F	F	F
3	T	F	T	F	T	T	T	T	T	F
4	T	T	F	F	F	T	F	F	F	F
5	F	T	F	F	F	T	T	T	T	F

The obtained results showed that MC/DC coverage demonstrated a fast growing trend when the number of random test cases increased. This trend was shown both by CoderCover (Fig. 4) and CTC++ (Fig. 5) tools for the simple

Alluri, Aparna; Kacker, Raghu; Kuhn, David; Vilkomir, Sergiy.
"Combinatorial and MC/DC Coverage Levels of Random Testing."
Paper presented at IEEE International Conference on Software Quality Reliability and Security, Prague, Czech Republic. July 25, 2017 - July 29, 2017.

and complex expressions. For the simple expressions, MC/DC coverage reached 99% level for 55 random test cases and complete 100% MC/DC coverage was achieved started from 100 tests. The results by CoderCover and CTC++ for simple expressions were similar and close to each other (Fig. 6).

For complex expressions, more test cases were required to reach maximum MC/DC coverage. The coverage was very close to maximum from 200 test cases and maximum MC/DC coverage required approximately 400 random test cases. However, it is necessary to mention two distinctive features of estimation of MC/DC coverage for complex expressions:

- The MC/DC levels reported by two tools were significantly different (Fig. 7).
- The maximum of MC/DC coverage did not reach 100%. Even for exhaustive testing (all 1024 possible test cases for 10 variables), the maximum level of MC/DC coverage was 93% according to CodeCover and only 77% according to CTC++.

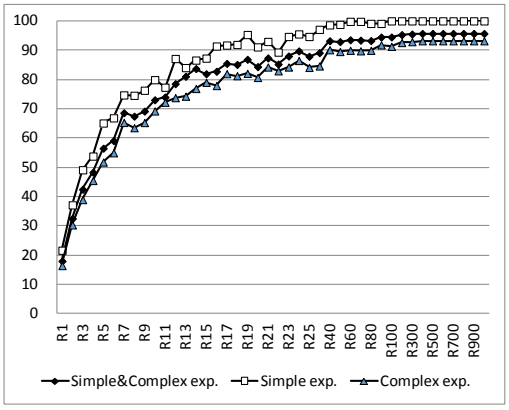


Fig. 4. MC/DC coverage by CodeCover for Random tests (10 variables).

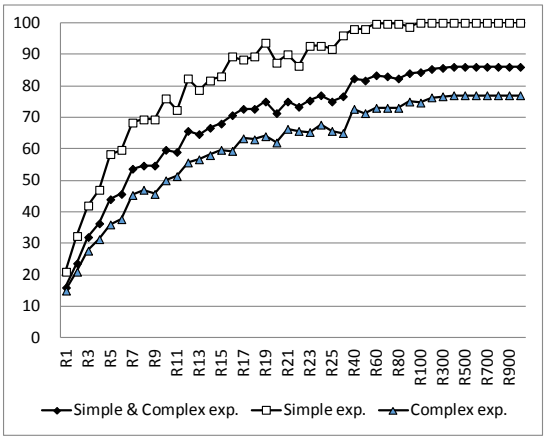


Fig. 5. MC/DC coverage by Testwell CTC++ for Random tests (10 variables.).

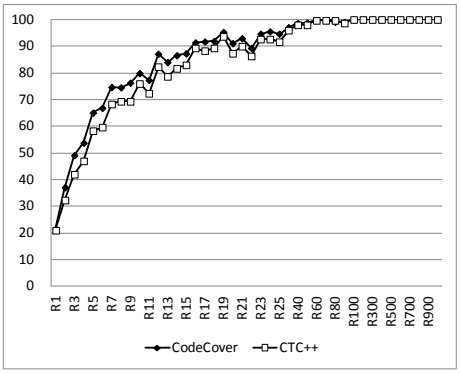


Fig. 6. Comparison of CodeCover and Testwell CTC++ results for Simple Expressions (10 variables).

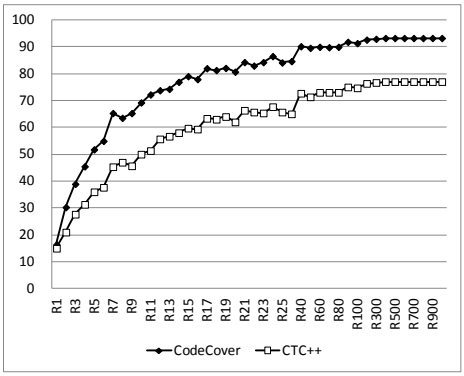


Fig. 7. Comparison of CodeCover and Testwell CTC++ results for Complex Expressions (10 variables).

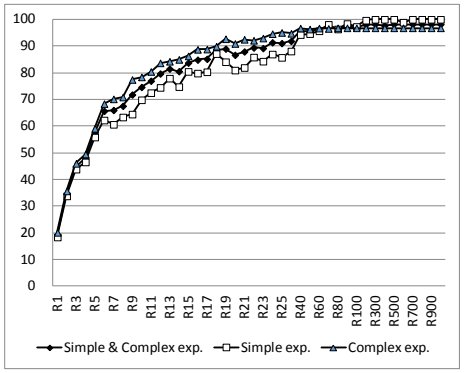


Fig. 8. MC/DC coverage by CodeCover for Random tests (20 variables).

The similar situation is not only for CodeCover and CTC++ but also for other tools provided MC/DC coverage, for example, Kalimetrix Logiscope [33] or TESSY [34]. The reason, as it was mentioned in Section 3.A, could be that different tools could use different principles for coverage evaluation. They also can evaluate different types of MC/DC, for example, Unique-Cause MC/DC vs. Masking MC/DC.

Alluri, Aparna; Kacker, Raghu; Kuhn, David; Vilkomir, Sergiy.
"Combinatorial and MC/DC Coverage Levels of Random Testing."
Paper presented at IEEE International Conference on Software Quality Reliability and Security, Prague, Czech Republic. July 25, 2017 - July 29, 2017.

Other factors that affect evaluation of coverage are how tools treat short-circuit Boolean expressions and multiple occurrences of logical variables in expressions. Thus, in our case, CTC++ considers multiple occurrences of the same variable as different variables and requires test cases with different values of the first and second occurrences of the variable, that is impossible. Under this definition used, complex expressions never can have 100% MC/DC coverage. In contrast with CTC++, CodeCover considers multiple occurrences as the same variable so 100% MC/DC coverage for complex expressions is possible. However, when the short-circuit operator is used for evaluation of expression values, CodeCover considers some variables as uncovered even if test cases provide necessary coverage.

TABLE IV. MC/DC COVERAGE FOR RANDOM TESTS (10 VARIABLES)

Random Test Set Size	MC/DC Coverage for Simple and Complex Expressions		MC/DC Coverage for Simple Expressions		MC/DC Coverage for Complex Expressions	
	CodeCover	Testwell CTC++	CodeCover	Testwell CTC++	CodeCover	Testwell CTC++
1	17.97	16.00	21.63	21.00	16.40	15.00
2	32.50	23.67	37.13	32.33	30.30	21.00
3	42.47	32.00	49.13	42.00	38.93	27.67
4	48.30	36.33	53.80	47.00	45.47	31.33
5	56.50	44.00	65.10	58.33	51.77	36.00
6	59.10	45.67	66.87	59.67	54.93	37.67
7	68.60	53.67	74.70	68.33	65.30	45.33
8	67.43	54.67	74.57	69.33	63.50	47.00
9	69.17	54.67	76.30	69.33	65.27	45.67
10	73.03	59.67	79.90	76.00	69.23	50.00
11	73.97	59.00	77.33	72.33	72.20	51.33
12	78.53	65.67	87.13	82.33	73.77	55.67
13	81.07	64.67	84.03	78.67	74.27	56.67
14	83.70	66.67	86.60	81.67	76.90	58.00
15	81.93	68.00	87.27	83.00	79.00	59.67
16	82.77	70.67	91.40	89.33	77.90	59.33
17	85.47	72.67	91.73	88.33	81.97	63.33
18	85.10	72.67	91.93	89.33	81.27	63.00
19	86.87	75.00	95.33	93.67	82.07	64.00
20	84.37	71.33	91.10	87.33	80.67	62.00
21	87.33	75.00	92.97	90.00	84.23	66.33
22	85.23	73.33	89.33	86.33	82.93	65.67
23	88.00	75.33	94.67	92.67	84.23	65.33
24	89.73	77.00	95.53	92.67	86.47	67.67
25	87.90	75.00	94.67	91.67	84.10	65.67
30	89.13	76.67	97.10	96.00	84.60	65.00
40	93.20	82.33	98.63	98.00	90.13	72.67
50	92.90	81.67	98.80	98.00	89.53	71.33
60	93.53	83.33	99.83	99.67	89.93	73.00
70	93.40	83.00	99.83	99.67	89.77	73.00
80	93.27	82.33	99.13	99.67	89.93	73.00
90	94.47	84.00	99.13	98.67	91.80	75.00
100	94.50	84.33	100.00	100.00	91.30	74.67
200	95.30	85.33	100.00	100.00	92.60	76.33
300	95.50	85.67	100.00	100.00	92.90	76.67
400	95.70	86.00	100.00	100.00	93.20	77.00
500	95.70	86.00	100.00	100.00	93.20	77.00
600	95.70	86.00	100.00	100.00	93.20	77.00
700	95.70	86.00	100.00	100.00	93.20	77.00
800	95.70	86.00	100.00	100.00	93.20	77.00
900	95.70	86.00	100.00	100.00	93.20	77.00
1024	95.70	86.00	100.00	100.00	93.20	77.00

To investigate how the total number of logical variables in software affects MC/DC coverage of random testing, we repeated testing with 20 variables instead of 10 variables. The sizes of logical expressions remained the same (from 3 to 9) but variables for each expression were selected from the set of 20 variables. Each random test case had also 20 input values though not all of them where used for each expression. Detailed data for this testing are presented in Table V and Fig. 8 and 9.

Some conclusions were similar for 10 and 20 variables:

- The levels of MC/DC coverage for simple expressions by CodeCover and CTC++ were close each other (Fig.10).
- The levels of MC/DC coverage for complex expressions by CodeCover and CTC++ were significantly different (Fig. 11) and CodeCover reported higher levels of coverage.

Alluri, Aparna; Kacker, Raghu; Kuhn, David; Vilkomir, Sergiy.
"Combinatorial and MC/DC Coverage Levels of Random Testing."
Paper presented at IEEE International Conference on Software Quality Reliability and Security, Prague, Czech Republic. July 25, 2017 - July 29, 2017.

TABLE V. MC/DC COVERAGE FOR RANDOM TESTS (20 VARIABLES)

Random Test Set Size	MC/DC Coverage for Simple and Complex Expressions		MC/DC Coverage for Simple Expressions		MC/DC Coverage for Complex Expressions	
	CodeCover	Testwell CTC++	CodeCover	Testwell CTC++	CodeCover	Testwell CTC++
1	19.10	16.00	18.40	20.00	20.13	17.00
2	34.60	26.67	33.80	31.67	35.67	26.00
3	44.87	34.00	43.83	38.67	46.00	33.33
4	48.03	39.00	46.53	43.33	49.50	37.33
5	57.57	43.67	55.87	49.00	59.13	42.33
6	65.67	51.00	62.23	53.67	68.40	50.67
7	66.00	53.33	60.63	54.33	70.17	53.67
8	67.57	54.67	63.37	57.33	70.87	54.33
9	71.80	57.67	64.47	57.33	77.43	59.33
10	74.63	62.67	69.83	64.33	78.40	63.00
11	76.97	63.67	72.50	65.67	80.43	63.67
12	79.67	67.33	74.47	68.33	83.67	68.33
13	81.50	68.33	77.93	75.33	84.27	68.00
14	80.53	68.00	74.77	68.67	84.97	68.33
15	83.73	72.00	80.47	74.67	86.27	71.67
16	84.87	74.00	79.83	75.00	88.73	74.67
17	85.17	74.00	80.33	75.00	88.83	74.33
18	88.67	78.67	87.13	83.00	89.90	76.33
19	89.00	79.00	84.10	80.33	92.70	78.67
20	86.67	75.67	80.97	76.67	90.93	76.33
21	87.93	77.33	81.93	77.00	92.50	78.33
22	89.40	79.67	85.83	81.33	92.13	78.67
23	89.20	79.33	84.30	79.00	92.97	80.00
24	91.37	81.00	86.97	81.33	94.60	81.33
25	91.07	82.00	85.73	81.67	95.07	82.67
30	91.87	82.67	88.10	86.67	94.73	82.00
40	95.67	88.00	94.27	91.67	96.70	85.00
50	95.57	88.33	94.63	93.00	96.33	85.00
60	96.27	89.00	95.70	94.33	96.70	85.33
70	97.23	90.67	98.10	97.33	96.57	85.67
80	96.57	90.00	96.37	95.33	96.80	85.67
90	97.50	91.33	98.40	97.67	96.80	86.00
100	97.00	90.67	97.30	96.00	96.80	86.00
200	98.07	92.00	99.67	99.33	96.80	86.00
300	98.20	92.00	100.00	100.00	96.80	86.00
400	97.70	92.00	100.00	100.00	96.80	86.00
500	98.20	92.00	100.00	100.00	96.80	86.00
600	98.20	92.00	98.90	98.33	96.80	86.00
700	98.20	92.00	100.00	100.00	96.80	86.00
800	98.20	92.00	100.00	100.00	96.80	86.00
900	98.20	92.00	100.00	100.00	96.80	86.00
1024	98.20	92.00	100.00	100.00	96.80	86.00

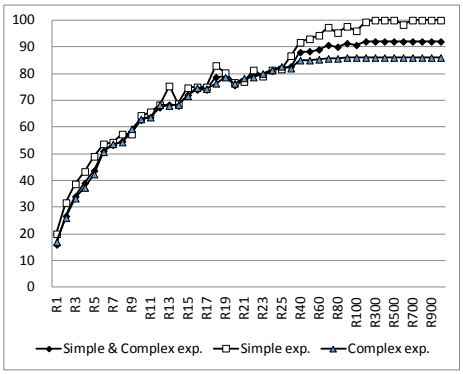


Fig. 9. MC/DC coverage by Testwell CTC++ for Random tests (20 variables).

- Maximal level of MC/DC coverage for complex expressions did not reach 100%.

However, numerical data were slightly different for 20 variables vs. 10 variables:

- For the simple expressions, MC/DC coverage reached 99% level for 200 random test cases and complete 100% MC/DC coverage was achieved started from 400 tests.
- For the complex expressions, the maximal possible levels of MC/DC coverage was 96.8% by CodeCover and 86% by CTC++.
- For the complex expressions, the maximal levels of MC/DC coverage were reached after 100 random test cases.

Alluri, Aparna; Kacker, Raghu; Kuhn, David; Vilkomir, Sergiy.
"Combinatorial and MC/DC Coverage Levels of Random Testing."
Paper presented at IEEE International Conference on Software Quality Reliability and Security, Prague, Czech Republic. July 25, 2017 - July 29, 2017.

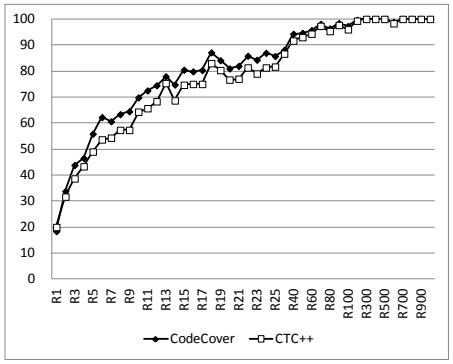


Fig. 10. Comparison of CodeCover and Testwell CTC++ results for Simple Expressions (20 variables).

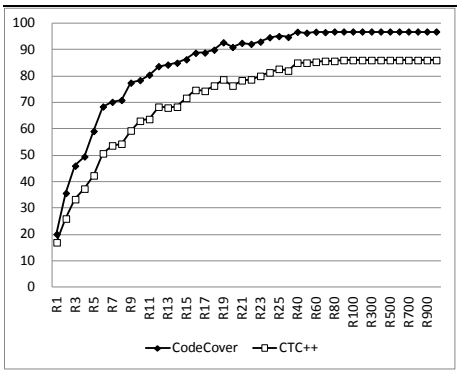


Fig. 11. Comparison of CodeCover and Testwell CTC++ results for Complex Expressions (20 variables).

In general, random test cases achieved a high level of MC/DC coverage very fast when the number of test increased. The precise data are given above but, very approximately, around 100 random tests provided high MC/DC coverage in all cases. Of course, this number is much higher than the amount of MC/DC test cases for one expression which is *n+1* for expressions size *n*, i.e., maximum 10 tests for expressions size 9. However, the number of different MC/DC tests necessary for all expressions together can be close to these 100 tests. At the same time, the process of MC/DC test generation is much harder comparing with random testing and should been done separately for each expression. It makes random testing a good basis for development new approaches to achieve MC/DC coverage.

V. COMBINATORIAL COVERAGE LEVEL OF RANDOM TESTING

In contrast to MC/DC, combinatorial coverage level does not depend on logical expressions in software and depends only on input variables.

In this section we evaluate *t*-way coverage of random test cases for *t*=2…6. Similar to Section IV, we consider this coverage for 10 input variables (Table VII) and 20 input variables (Table VIII). We use the same random test cases as in Section IV with sizes from 2 to 1024.

To understand how high the combinatorial coverage level of random test cases is, it is necessary to random sets of the same sizes as *t*-way sets. The sizes of combinatorial test sets for *t*=2…6 and for *n*=10 and *n*=20 are presented in Table VI. They are not optimal (minimal) but are close to optimal values and reflect sizes of combinatorial test sets generated by ACTS tool.

TABLE VI. SIZES OF COMBINATORIAL TEST SETS

Number of logical variables	2-way	3-way	4-way	5-way	6-way
10	10	20	44	93	178
20	12	27	66	165	375

TABLE VII. COMBINATORIAL COVERAGE FOR RANDOM TESTS (10 VARIABLES)

Random Test Set Size	2-way	3-way	4-way	5-way	6-way
2	46.85	24.51	12.44	6.25	3.12
3	60.55	34.27	18.08	9.25	4.67
4	68.33	41.24	22.63	11.86	6.08
5	76.85	49.51	28.12	14.96	7.70
6	83.52	55.49	32.11	17.31	9.01
7	85.74	59.23	35.48	19.58	10.33
8	85.92	63.96	39.61	22.20	11.78
9	92.78	70.48	44.58	25.28	13.47
10	96.85	76.70	49.29	27.94	14.86
11	94.44	76.60	51.57	30.10	16.20
12	98.52	81.39	54.71	32.17	17.52
13	97.78	83.40	57.68	34.31	18.76
14	97.78	83.16	58.00	35.03	19.48
15	99.26	87.95	63.13	38.37	21.18
16	98.15	86.84	63.43	39.53	22.30
17	99.44	91.42	69.30	43.66	24.43
18	99.44	90.73	68.26	43.27	24.63
19	99.81	94.10	73.06	46.64	26.40
20	98.89	91.18	70.29	45.43	26.21
21	99.81	94.69	75.81	50.24	29.09
22	99.81	95.69	77.41	51.41	29.87
23	100.00	95.49	77.80	52.35	30.76
24	100.00	96.84	79.70	54.14	32.02
25	100.00	97.51	81.58	56.31	33.43
30	100.00	98.23	86.07	61.76	37.85
40	100.00	99.51	92.74	72.51	47.37
50	100.00	99.96	97.12	81.48	55.84
60	100.00	100.00	96.32	84.80	60.76
70	100.00	100.00	99.01	93.29	75.71
80	100.00	100.00	99.53	93.01	73.16
90	100.00	100.00	99.65	94.43	76.68
100	100.00	100.00	99.87	96.60	81.34
200	100.00	100.00	100.00	99.96	97.42
300	100.00	100.00	100.00	100.00	99.61
400	100.00	100.00	100.00	100.00	99.98
500	100.00	100.00	100.00	100.00	100.00
600	100.00	100.00	100.00	100.00	100.00
700	100.00	100.00	100.00	100.00	100.00
800	100.00	100.00	100.00	100.00	100.00
900	100.00	100.00	100.00	100.00	100.00
1024	100.00	100.00	100.00	100.00	100.00

As it is possible to conclude from Tables VII and VIII, the level of combinatorial coverage of random test cases is quite high. Thus, in all situations for any *n* and *t*, this level for

Alluri, Aparna; Kacker, Raghu; Kuhn, David; Vilkomir, Sergiy.
"Combinatorial and MC/DC Coverage Levels of Random Testing."
Paper presented at IEEE International Conference on Software Quality Reliability and Security, Prague, Czech Republic. July 25, 2017 - July 29, 2017.

random test sets of the same size as combinatorial test sets is around 90-97%.

The level of combinatorial coverage grows very fast when the number of random tests increases. However, after 90% the increase becomes significantly slow. To reach 100% of combinatorial coverage, significantly more random tests are required comparing with the combinatorial test sets. Thus, 23 (for n=10) and 24 (for n=20) random test cases are necessary to archive 100% pairwise coverage comparing with 10 and 12 test cases in pairwise test sets. The similar situations are for t-way coverage as it is possible to see from Tables VII and VIII.

TABLE VIII. COMBINATORIAL COVERAGE FOR RANDOM TESTS (20 VARIABLES)

Random Test Set Size	2-way	3-way	4-way	5-way	6-way
2	46.57	24.43	12.41	6.23	3.12
3	58.34	33.41	17.80	9.16	4.64
4	64.02	38.91	21.67	11.54	6.01
5	77.19	48.80	27.50	14.62	7.54
6	84.61	56.70	32.82	17.60	9.10
7	88.03	61.58	36.68	20.02	10.46
8	90.48	65.84	40.40	22.46	11.86
9	92.19	68.71	43.16	24.45	13.07
10	94.38	74.22	48.49	28.14	15.29
11	95.97	77.20	50.84	29.45	15.89
12	96.88	80.29	54.19	31.79	17.26
13	97.24	81.22	55.66	33.22	18.27
14	98.64	84.71	59.52	35.94	19.83
15	98.64	85.55	61.20	37.54	20.94
16	99.47	90.09	66.24	40.86	22.71
17	99.65	89.75	66.19	41.38	23.33
18	99.69	91.83	69.41	43.92	24.87
19	99.56	91.38	69.98	45.04	25.82
20	99.56	92.38	71.31	46.10	26.54
21	99.83	94.10	73.94	48.33	27.96
22	99.96	95.95	77.69	51.54	29.88
23	99.91	95.73	78.03	52.40	30.72
24	100.00	96.79	80.17	54.36	31.96
25	99.96	96.99	81.05	55.43	32.81
30	100.00	98.69	86.76	62.43	38.18
40	100.00	99.60	92.40	71.66	46.51
50	100.00	99.89	96.10	79.74	54.68
60	100.00	99.94	97.82	84.91	60.86
70	100.00	99.99	98.86	88.99	66.54
80	100.00	100.00	99.52	92.51	72.08
90	100.00	100.00	99.64	93.97	75.40
100	100.00	100.00	99.84	95.59	78.80
200	100.00	100.00	100.00	99.83	95.64
300	100.00	100.00	100.00	99.99	99.10
400	100.00	100.00	100.00	100.00	99.79
500	100.00	100.00	100.00	100.00	99.96
600	100.00	100.00	100.00	100.00	99.99
700	100.00	100.00	100.00	100.00	100.00
800	100.00	100.00	100.00	100.00	100.00
900	100.00	100.00	100.00	100.00	100.00
1024	100.00	100.00	100.00	100.00	100.00

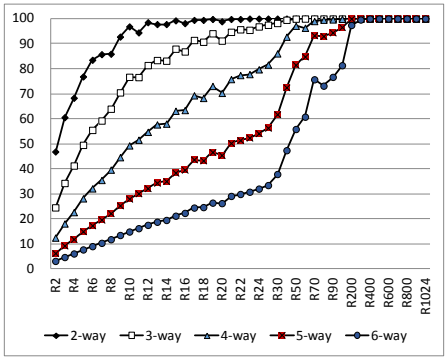


Fig. 12. Combinatorial coverage for Random tests (10 variables).

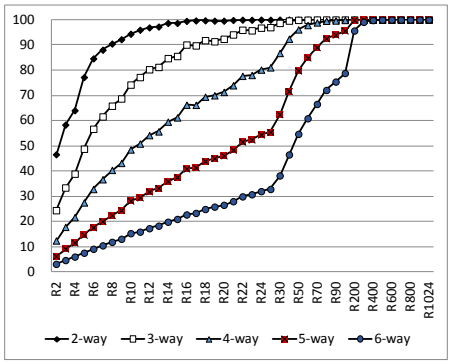


Fig. 13. Combinatorial coverage for Random tests (20 variables).

VI. CONCLUSIONS AND FUTURE WORK

This paper evaluates the ability of random testing to provide coverage according to MC/DC and t-way testing criteria. One hundred logical expressions of different sizes were generated. We also generated 252 random test sets with from 2 to 1024 test cases in a set. These sets were used to test a software program with logical expressions and the levels of coverage were evaluated using the structural coverage tools CodeCover, Testwell CTC++, and CCM, which measures the coverage of input value combinations in a test suite.

Our experiments show that for simple expressions, when the number of random test cases increased, they quickly reach a high level of coverage both for MC/DC and 2-way. The close to 100% level of MC/DC coverage is achieved after approximately 100 random tests in many cases, but full coverage may require doubling test set size. Complex expressions required a much larger test set size for MC/DC coverage on the order of 90%, again doubling the number of tests for full coverage. The paper provides detailed data for different types of expressions and different testing tools.

An unexpected result of the study was that structural coverage tools differ in their definition of partial MC/DC coverage, resulting in significant variation in coverage calculations. The primary standard for MC/DC coverage,

Alluri, Aparna; Kacker, Raghu; Kuhn, David; Vilkomir, Sergiy.
"Combinatorial and MC/DC Coverage Levels of Random Testing."
Paper presented at IEEE International Conference on Software Quality Reliability and Security, Prague, Czech Republic. July 25, 2017 - July 29, 2017.

RTCA DO-178B, requires full coverage, so partial coverage numbers are generally not used. In cases where knowledge of less than complete MC/DC coverage is of interest, a consistent definition of partial coverage will need to be specified.

Random test sets of the same sizes as *t*-way sets provide the 90-97% level of combinatorial coverage. However, much more random tests are required to reach 100% coverage. The paper provides detailed data for different numbers of input variables, different types of expressions, and *t*-way coverage for *t*=2…6.

The obtained results can help for integration random testing with other approaches (in particular, MC/DC and combinatorial testing) to increase of effectiveness of testing software with complex logical structure.

ACKNOWLEDGMENT

This work was performed under the following financial assistance award 70NANB15H217 from the U.S. Department of Commerce, National Institute of Standards and Technology.

Disclaimer: *Products may be identified in this document, but identification does not imply recommendation or endorsement by NIST, nor that the products identified are necessarily the best available for the purpose.*

REFERENCES

[1] International Standard ISO/IEC/IEEE 29119-1:2013 "Software and systems engineering — Software testing — Part 1: Concepts and definitions," 2013.

[2] M. Grindal, J. Offutt, S. Andler, "Combination Testing Strategies: a Survey," Software Testing, Verification and Reliability, Vol. 15, No. 3, pp. 167-199, 2005.

[3] D. R. Kuhn, R. Kacker, and Y. Lei, Introduction to Combinatorial Testing, Chapman and Hall/CRC, 2013, 341 pages.

[4] D. R. Kuhn, R. Kacker, Y. Lei, and J. Hunter, "Combinatorial software testing, " IEEE Computer, vol. 42, no. 8, August 2009.

[5] RTCA/DO-178B, "Software Considerations in Airborne Systems and Equipment Certification," RTCA, Washington D.C., USA, 1992.

[6] J. Chilenski and S. Miller, "Applicability of Modified Condition/Decision Coverage to software testing," Software Engineering Journal, September 1994, pp. 193-200.

[7] RTCA/DO-178B "Software Considerations in Airborne Systems and Equipment Certification," Radio Technical Commission for Aeronautics, 1992.

[8] RTCA/DO-178C "Software Considerations in Airborne Systems and Equipment Certification," Radio Technical Commission for Aeronautics, 2012.

[9] Y. Moy, E. Ledinot, H. Delseny, V. Wiels, and B. Monate, "Testing or Formal Verification: DO-178C Alternatives and Industrial Experience," IEEE Software, May/June 2013, Volume 30, Issue 3, pp. 50-57.

[10] M. R. Girgis and M. R. Woodward, "An experimental comparison of the error exposing ability of program testing criteria," Proceedings of the Workshop on Software Testing, pp. 64-73. IEEE Computer Society Press, July 1986.

[11] V. Basili and R. Selby, "Comparing the Effectiveness of Software Testing Strategies," IEEE Trans. Softw. Eng. SE-13, (December 1987), pp. 1278–1296.

[12] C.-A. Sun, Y. Zai, and H. Liu, "Evaluating and comparing fault-based testing strategies for general boolean specifcations: A series of experiments," The Computer Journal, 2015, Volume 58, Issue 5, pp. 1199-1213.

[13] P.S. Kochhar, F. Thung, and D. Lo, "Code coverage and test suite effectiveness: Empirical study with real bugs in large systems," Proceedings of the IEEE 22nd International Conference on Software Analysis, Evolution and Reengineering (SANER), Montreal, Canada, 2-6 March 2015, pp. 560-564.

[14] S. Vilkomir and D. Anderson, "Relationship between pair-wise and MC/DC testing: Initial experimental results," Proceedings of the IEEE 8th International Conference on Software Testing, Verification and Validation Workshops (ICSTW 2015), 13-17 April 2015, Graz, Austria.

[15] P. Thevenod-Fosse, H. Waeselynck, and Y. Crouzet, "An experimental study on software structural testing: deterministic versus random input generation," Proceedings of the 21st International Symposium on Fault-Tolerant Computing (FTCS 91), IEEE Press, Jun. 1991, pp. 410-417.

[16] M. A. Vouk, K-C. Tai, and A. Paradkar, "Empirical studies of predicate-based software testing," Proceedings of the 5th International Symposium on Software Reliability Engineering, 1994.

[17] T. Y. Chen, F. C. Kuo, H. Liu, and W. E. Wong, "Code coverage of adaptive random testing," IEEE Transactions on Reliability, 2013, 62(1), pp. 226-237.

[18] S. Vilkomir, O. Starov, and R. Bhambroo, "Evaluation of t-way Approach for Testing Logical Expressions in Software," Proceedings of the IEEE Sixth International Conference on Software Testing, Verification and Validation Workshops (ICSTW 2013), 18–20 March 2013, Luxembourg, pp. 249-256.

[19] W. Ballance, S. Vilkomir, and W. Jenkins, "Effectiveness of Pair-wise Testing for Software with Boolean Inputs," Proceedings of the Fifth International Conference on Software Testing, Verification and Validation (ICST 2012), April 17-21, 2012, Workshop on Combinatorial Testing (CT-2012), Montreal, Canada, pp. 580-585.

[20] S. Liu and Y. Chen, "A relation-based method combining functional and structural testing for test case generation," Journal of Systems and Software 81.2 (2008), pp. 234-248.

[21] C. Pfaller and M. Pister, "Combining Structural and Functional Test Case Generation," Proceedings of the Software Engineering Conference (SE08) , Munich, February 2008., pp. 229-241.

[22] C. D. Nguyen, A. Marchetto, and P. Tonella, "Combining model-based and combinatorial testing for effective test case generation," Proceedings of the 2012 International Symposium on Software Testing and Analysis, pp. 100-110. ACM, 2012.

[23] E. P. Enoiu, K. Doganay, M. Bohlin, D. Sundmark, and P. Pettersson, "MOS: an integrated model-based and search-based testing tool for function block diagrams," Proceedings of the 1st International Workshop on Combining Modelling and Search-Based Software Engineering, pp. 55-60. IEEE Press, 2013.

[24] R. Bartholomew, "An industry proof-of-concept demonstration of automated combinatorial test," Proceedings of the 8th International Workshop on Automation of Software Test (AST'13), May 18-19, 2013, San Francisco, CA, USA, pp. 118-124.

[25] "CodeCover," http://codecover.org

[26] R. Schmidberger, "Well-Defined Coverage Metrics for the Glass Box Test," In Testing Software and Systems, pp. 113-128. Springer Berlin Heidelberg, 2014.

[27] Testwell, "Testwell CTC++. Test Coverage Analyzer for C/C++," http://www.testwell.fi/ctcdesc.html

[28] Verifysoft Technology GmbH, "Testwell CTC++ Test Coverage Analyser," http://www.verifysoft.com/en_ctcpp.html

[29] National Institute of Standards and Technology (NIST), "Combinatorial Coverage Measurement Tool, User Guide, January 30, 2011," http://csrc.nist.gov/groups/SNS/acts/documents/ComCoverage110130.pdf

[30] I. D. Mendoza, D. R. Kuhn, R. N. Kacker, and Y. Lei, "CCM: A Tool for Measuring Combinatorial Coverage of System State Space," Proceedings of the ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM 2013), 10-11 Oct. 2013, Baltimore, Maryland, USA, p. 291.

[31] J. Chilenski, "An investigation of three forms of the modified condition decision coverage (MCDC) criterion," Tech. Report DOT/FAA/AR-01/18, FAA, 2001.

[32] V. Durelli, et al., "What to expect of predicates: An empirical analysis of predicates in real world programs," Journal of Systems and Software 113, 2016, pp. 324-336.

Alluri, Aparna; Kacker, Raghu; Kuhn, David; Vilkomir, Sergiy.
"Combinatorial and MC/DC Coverage Levels of Random Testing."
Paper presented at IEEE International Conference on Software Quality Reliability and Security, Prague, Czech Republic. July 25, 2017 - July 29, 2017.

[33] Kalimetrix, "Logiscope TestChecker," http://www.kalimetrix.com/logiscope/testchecker

[34] Razorcat, "Automated testing of embedded software," http://www.razorcat.eu/tessy.html

Alluri, Aparna; Kacker, Raghu; Kuhn, David; Vilkomir, Sergiy.
"Combinatorial and MC/DC Coverage Levels of Random Testing."
Paper presented at IEEE International Conference on Software Quality Reliability and Security, Prague, Czech Republic. July 25, 2017 - July 29, 2017.

MAC-Layer Coexistence Analysis of LTE and WLAN Systems Via Listen-Before-Talk

Yao Ma and Daniel G. Kuester
Communications Technology Laboratory, National Institute of Standards and Technology
325 Broadway, Boulder, Colorado, USA

***Abstract*— With the congestion and scarcity of available spectrum resources, spectrum sharing between long-term evolution (LTE) and the IEEE 802.11 (aka. WLAN) systems is an ongoing research topic. Considering the LTE license assisted access (LAA) with the listen before talk (LBT) procedure, recent research efforts try to evaluate the performance in several LTE-LBT and WLAN coexistence scenarios. However, the available approaches have not adequately modeled and analyzed general case of LBT (such as Category 4), and the case when there are more than two types of transmissions. In this paper, to fill this technical gap, we implement a systematic modelling and analysis of the media access control (MAC) layer coexisting performance of LTE-LBT and WLAN systems. We consider the coexistence scenario of multiple LTE downlink with multiple WLAN uplink and downlink transmissions. We develop analytical results on time-efficiency throughput, transmission and collision probabilities of LTE and WLAN nodes, and then generalize the result to multiple types of transmissions (e.g., more than three types). To validate the analysis, we implement LBT and WLAN MAC algorithm programming and extensive simulations, which confirm the accuracy of our analysis. Our result shows that replacing WLAN stations with LTE transmitters may, in some cases, significantly degrade the overall throughput, depending on the original efficiencies of WLAN systems and channel access schemes. To address this, we propose a 4-way handshaking channel access scheme for LTE-LBT, which can significantly improve the coexistence performance. These results put new insight into relationship between coexistence performance and MAC parameters of LTE-LBT and WLAN systems, and may aid in the design and optimization of coexistence systems.**

I. Introduction

To enable the coexistence between cellular and IEEE 802.11 wireless local area network (WLAN) systems in the industrial, scientific, and medical (ISM) radio band, the long-term evolution (LTE) with license assisted access (LAA) has been proposed and studied by the 3rd Generation Partnership Project (3GPP), industry, and research community [1]–[7]. Constructive coexistence between unlicensed LTE with WLAN relies on proper medium access control (MAC) layer channel access coordination between the heterogeneous systems. To avoid destructive collision between LTE and other radio systems, the 3GPP LAA has proposed features such as dynamic carrier selection, discontinuous transmission with limited maximum transmission duration, transmit power control, and listen before talk (LBT) [4], [5], [7].

The LAA has defined 4 categories of LBT schemes [5], [6]. Category 1 has no LBT, and category 2 defines LBT without random back-off, related to the frame-based (FBE) sensing and access. Category 3 LBT includes random back-off with fixed size of contention window (CW), and category 4 LBT uses variable size of CW with multiple-stage random back-off [5]. Among them, Categories 3 and 4 LTE-LBT schemes implement load-based coexistence, and have attracted major interest from the telecom industry and research community. Various coexistence settings based on LTE-LBT and WLAN transmissions have been intensively evaluated, and abundant experimental and field test results are reported in [5]–[7].

Recently, some analytical approaches for the performance evaluation of LTE-LBT and WLAN coexistence systems have been developed, see e.g. [8], [9]. In [8], coexistence of WLAN access points (APs) and LTE-LBT downlink transmissions is modelled and analyzed. The LTE-LBT transmission uses higher data rate when there is no collision, and a low rate when a collision is detected. In another LTE-LBT scheme [9], LTE base stations (BSs) share unlicensed spectrum with WLAN APs in downlink transmissions, and the overall system throughput is examined when some APs are replaced by an equal number of BSs. The work in [8], [9] evaluate only Category-3 LBT schemes, and analytical curves are provided as numerical result (there is no evidence that simulation results are provided in [8], [9]).

Some recent work [10]–[12] try to optimize the coexistence performance under various fairness constraints. In [10], the authors propose an LBT scheme that has short initial clear channel assessment (CCA) duration, which controls the access priority of LTE and WLAN systems. They develop optimization method to maximize the LTE-LAA system performance under constraint of fair channel sharing with the WLAN system. A new LBT protocol was developed in [11] to maximize the LTE and WLAN overall normalized channel rate by adjusting the LTE transmission durations, under a constraint of protection to WLAN service. In [12], an adaptive LBT scheme was developed to satisfy the quality of service (QoS) requirement of LTE users under condition that the resulting collision probability of WLAN users is constrained.

While the above mentioned [8]–[12] and other works provide new performance evaluation and optimization methods on LTE-LAA and WLAN coexistence, the general case of Category-4 LTE-LBT based coexistence has not been clearly modelled and well analyzed. Furthermore, in most of the available works, only two types of transmissions were considered, namely WLAN downlink and LTE LBT-based downlink.

Kuester, Daniel; Ma, Yao.
"MAC-Layer Coexistence Analysis of LTE and WLAN Systems Via Listen-Before-Talk."
Paper presented at Same as above, Las Vegas, NV, United States. January 8, 2017 - January 11, 2017.

However, future coexistence applications will likely involve three or more types of transmissions, and related performance analysis is lacking.

In this paper, we study the coexistence of multiple types of LTE-LBT and WLAN transmissions with different parameters, and provide a stationary steady-state analysis method of channel access probability, collision probability, and throughput, assuming saturated transmissions. Many recent works claim that the overall throughput performance is typically enhanced by replacing some WLAN APs with LTE nodes [5], [7], [9]. Here, we try to examine the conditions when the overall performance improves or degrades, supported by simulation.

The contribution of this paper is highlighted as follows:

- Following Category-4 LBT given by recent 3GPP LTE-LAA document [5], we provide new performance analysis that is valid for multiple types of LTE and WLAN transmissions, and generalize the analysis to more than 3 types of transmissions.
- We implement computer programming and extensive simulation to evaluate the MAC-layer coexistence of LTE-LBT and WLAN systems, and validate our analytical results.
- Our results show that the coexistence sum throughput may either improve or degrade when WLAN APs are replaced by LTE nodes, depending on system parameters and efficiency of the original WLAN system. Also, we propose a 4-way handshaking access scheme for LTE LBT, which is shown to provide significantly enhanced coexistence performance.

Finally, our new analytical result can be utilized for optimization design of general LBT schemes in coexistence scenario.

The remainder of this paper is organized as follows: Section II describes the LTE and WLAN coexistence system model, and Section III presents the performance analysis for Category-4 LBT assuming 3 types of transmissions. Technical discussions are provided in Section IV, and numerical results and conclusions are given in Sections V and VI, respectively.

II. System Model

Suppose several LTE small cells with N_L enhanced NodeB's (eNBs) coexist with multiple WLAN networks. The LTE-LAA utilizes carrier aggregation (CA), with primary traffic on licensed channels and secondary traffic on unlicensed channels. It shares the unlicensed channels with WLAN transmissions, and hence the LBT is important for system coexistence. Note that the LTE and WLAN aggregation (LWA) is another spectrum sharing candidate technique, which relies on dual connectivity of licensed LTE and unlicensed WLAN systems, and allows the LTE traffic to be carried by WLAN transmissions. The study of the LWA method is outside the scope of this paper.

Here, we consider the case that LTE LAA utilizes only unlicensed spectrum and shares it with incumbent WLAN users. In this model, the LTE-LBT downlink transmissions coexist with incumbent WLAN downlink and uplink transmissions in the same spectrum band, and all of them have

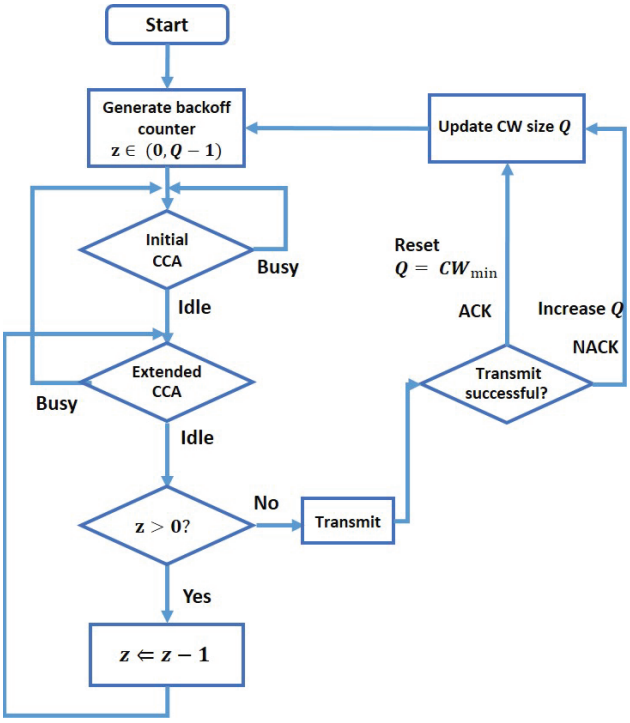


Fig. 1: Flow diagram of LTE downlink LAA LBT Category-4 procedure (adopted from [5], [6] with modifications).

saturated incoming traffic. In WLAN enhanced distributed channel access (EDCA), based on QoS requirement, different transmissions may belong to different access categories in terms of arbitration inter-frame space (AIFS), transmission duration, and initial CW size and backoff cut-off stage [18], [19]. In the model considered here, we assume that WLAN uplink and downlink transmissions have different CW sizes. We propose an LTE downlink LAA Category-4 LBT scheme adopted from [5] and [6], and show it in Fig. 1. It is modified from Fig. 7.2.1.6.1 of [5] in that we remove the buffer idle states, and explicitly show the relation between CW size Q and the receiver acknowledgement. It is also related to the Category-3 LBT figure given in [6] which has fixed CW size. In comparison with [5] and [6], we made one additional major change that the order of blocks "extended CCA" and "$z > 0$" is switched in Fig. 1. This change removes the channel access priority of an LAA node which just finishes a transmission, and makes it different from the case of transmission of a WLAN node.

The initial CCA in Fig. 1 has dual purposes: it implements initial channel sensing of duration T_{iCCA} to avoid collision with other transmissions, and it is implemented after a busy channel (either successful transmission or collision), for a duration called defer period T_{Defer} [5], [6]. When the counter z is reduced to 0, transmission starts. The CW size is adjusted based on ACK and NACK responses via hybrid automatic repeat request (HARQ) from the receiver. We assume that after NACK, the CW size is doubled until the cutoff stage

Kuester, Daniel; Ma, Yao.
"MAC-Layer Coexistence Analysis of LTE and WLAN Systems Via Listen-Before-Talk."
Paper presented at Same as above, Las Vegas, NV, United States. January 8, 2017 - January 11, 2017.

is reached, similarly to the WLAN distributed coordination function (DCF) procedure.

III. Performance Analysis

A. Markov Model of Category-4 LBT Procedure

Based on Fig. 1, we provide a Markov chain-based modelling of the LTE-LBT process in Fig. 2, which has maximum backoff stage R_L (also called cutoff stage). In each stage, the counter decrements when the channel is sensed idle for an extended CCA duration, denoted as T_{eCCA}. When the backoff counter is reduced to 0, an LTE downlink transmission starts. The $P_{f,L}$ is the probability that the LTE eNB experiences collision when it senses the channel idle and transmits. When we consider only the 0th backoff stage (with fixed CW size), this Markov chain state diagram models a Category-3 LBT process.

It appears that the Markov model in Fig. 2 is similar to the model given in [13], [15]. They are related and yet have some differences. First, after active transmission, the LTE node has to wait for channel idle duration $T_{\text{iCCA}}+T_{\text{eCCA}}$, before counter decrement. We assume that $T_{\text{Defer}} = T_{\text{iCCA}} = T_{\text{DIFS}}$, where T_{DIFS} is the WLAN DCF inter-frame space (DIFS) duration. Define δ_L as an LBT idle slot (or eCCA) duration. In Fig. 1, we let the LTE node with a successful transmission to wait for an additional δ_L, so that it does not have priority over other competing stations. This is in contrast to the WLAN transmission case [15]. The difference is reflected in the stage 0 backoff counter in Fig. 2, where in [15] the backoff CW size after successful transmission is shorter than the CW size after a collision. Second, MAC parameters of LTE LBT procedure may be different from those of WLAN. The transmission packet duration of LTE is assumed to be larger than the WLAN packet duration in this paper. To facilitate the coexistence analysis, however, we assume that $\delta_L = \delta_W$, where δ_W is the backoff idle slot duration of WLAN system.

B. Stationary Probability and Throughput Analysis

Based on the LTE-LBT Markov model in Fig. 2, a performance analysis of LTE and WLAN coexistence is provided next. The conditional state transition probabilities are given by:

$$P(0,k|j,0) = (1-P_{f,L})/Z_0, \quad j \in [0, R_L-1] \quad (1)$$

$$P(0,k|R_L,0) = P_{f,L}/Z_0, \quad (2)$$

$$P(j,k|j,k+1) = 1, \quad j \in [0,R_L], k \in [0, Z_j-2] \quad (3)$$

where Z_j is the CW size at backoff stage j, $j = 0, 1, \ldots, R_L$, $k \in (0, Z_j-1)$ is the backoff counter value, and R_L is the cutoff stage for LTE Category-4 LBT.

We define $\pi_{j,k}$ as the stationary probability of state (j,k). Using π as stationary probability of Markov state is consistent with the notations used in the literature [17], [18]. From eqs. (1)–(3) it follows that:

$$\pi_{j,0} = p_{f,L}^j \pi_{0,0}, \quad j \in [0, R_L],$$

$$\pi_{j,k} = \frac{Z_j - k}{Z_j}\pi_{j,0} \quad k \in [0, Z_j-1]; j \in [0, R_L].$$

Using the fact that total probability of all states is 1, we have $\sum_{j=0}^{R_L}\sum_{k=0}^{Z_j-1}\pi_{j,k} = 1$, and obtain that

$$\pi_{0,0} = \left[0.5\sum_{j=0}^{R_L} p_{f,L}^j (1+Z_j)\right]^{-1}.$$

The transmission probability of each LTE station is given by:

$$\tau_L = \sum_{j=0}^{R_L}\pi_{j,0} = \pi_{0,0}\frac{1-p_{f,L}^{R_L+1}}{1-p_{f,L}} = \frac{2(1-p_{f,L}^{R_L+1})}{(1-p_{f,L})\sum_{j=0}^{R_L} p_{f,L}^j (1+Z_j)}. \quad (4)$$

The joint stationary probability distribution of LTE and WLAN networks based on Markov process is provided next. We assume that LTE and WLAN systems have the same slot duration δ in the counter idle slots, which is the same as that considered in [8], [9]. We define the failed transmission probabilities (due to packet collisions) of LTE, and WLAN downlink and uplink systems as $p_{f,L}$, p_{f,W_D}, and p_{f,W_U}; corresponding transmission (or channel access) probabilities as τ_L, τ_{W_D}, and τ_{W_U}; the initial CW sizes as Z_0, $W_{0,D}$, and $W_{0,U}$; and the cutoff stages as R_L, R_D and R_U, respectively.

Based on [15], the transmission probabilities of a WLAN node in downlink and uplink transmissions are given by:

$$\tau_{W_D} = \frac{1}{1+\frac{1-p_{f,W_D}}{2(1-p_{f,W_D}^{R_D+1})}\left[\sum_{j=0}^{R_D} p_{f,W_D}^j (2^j W_{0,D}-1) - (1-p_{f,W_D}^{R_D+1})\right]}, \quad (5)$$

$$\tau_{W_U} = \frac{1}{1+\frac{1-p_{f,W_U}}{2(1-p_{f,W_U}^{R_U+1})}\left[\sum_{j=0}^{R_U} p_{f,W_U}^j (2^j W_{0,U}-1) - (1-p_{f,W_D}^{R_U+1})\right]}. \quad (6)$$

The probabilities of failed transmissions in LTE and WLAN systems are derived as:

$$p_{f,W_D} = 1-(1-\tau_{W_D})^{n_{W_D}-1}(1-\tau_{W_U})^{n_{W_U}}(1-\tau_L)^{n_L}, \quad (7)$$

$$p_{f,W_U} = 1-(1-\tau_{W_U})^{n_{W_U}-1}(1-\tau_{W_D})^{n_{W_D}}(1-\tau_L)^{n_L}, \quad (8)$$

$$p_{f,L} = 1-(1-\tau_L)^{n_L-1}(1-\tau_{W_D})^{n_{W_D}}(1-\tau_{W_U})^{n_{W_U}}, \quad (9)$$

where $\tau_L, \tau_{W_D}, \tau_{W_U}$ are given by (4), (5), and (6), respectively. Eqs. (4)-(9) model 6 unknowns in 6 equalities, and they can be solved by using an iterative numerical search method. Numerical evaluation shows that the iteration was stable and accurate for the considered range of parameters. The iterative search method for solving the transmission and collision probabilities has been used popularly in the literature, for example in [8], [9], [13]–[15].

Kuester, Daniel; Ma, Yao.
"MAC-Layer Coexistence Analysis of LTE and WLAN Systems Via Listen-Before-Talk."
Paper presented at Same as above, Las Vegas, NV, United States. January 8, 2017 - January 11, 2017.

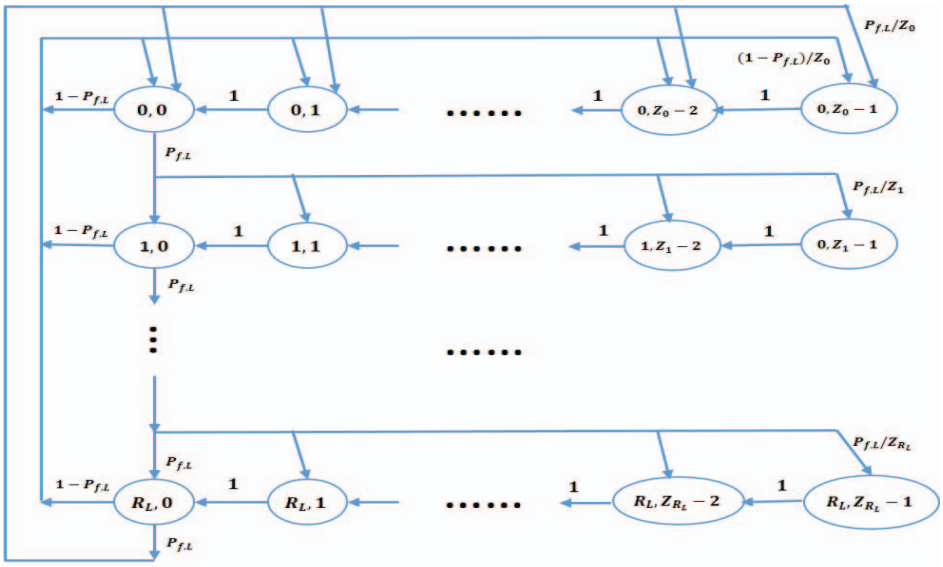


Fig. 2: Markov model of LTE-LAA LBT category 4 procedure.

Let P_{s,W_D}, P_{s,W_U}, and $P_{s,L}$ be the successful transmission probabilities, T_{P,W_D}, T_{P,W_U}, $T_{P,L}$ be the corresponding payload durations, and P_{b,W_D}, P_{b,W_U}, and $P_{b,L}$ be the probabilities of busy states (node transmitting), for WLAN downlink, uplink and LTE systems, respectively.

The MAC-layer sum throughput (normalized successful transmission time-duration) for WLAN downlink, uplink and LTE transmissions are given by:

$$S_{W_D} = P_{s,W_D}(1-P_{b,W_U})(1-P_{b,L})\overline{T}_{P,W_D}/T_{\text{ave}}, \quad (10)$$
$$S_{W_U} = P_{s,W_U}(1-P_{b,W_D})(1-P_{b,L})\overline{T}_{P,W_U}/T_{\text{ave}}, \quad (11)$$
$$S_L = P_{s,L}(1-P_{b,W_D})(1-P_{b,W_U})T_{P,L}/T_{\text{ave}}, \quad (12)$$

where $\overline{T}_{P,W_D} = T_{P,W_D}W_{0,D}/(W_{0,D}-1)$, $\overline{T}_{P,W_U} = T_{P,W_U}W_{0,U}/(W_{0,U}-1)$, respectively. Furthermore, we have

$$P_{s,W_D} = n_{W_D}\tau_{W_D}(1-\tau_{W_D})^{n_{W_D}-1}, \quad (13)$$
$$P_{s,W_U} = n_{W_U}\tau_{W_U}(1-\tau_{W_U})^{n_{W_U}-1}, \quad (14)$$
$$P_{s,L} = n_L\tau_L(1-\tau_L)^{n_L-1}, \quad (15)$$
$$P_{b,W_D} = 1-(1-\tau_{W_D})^{n_{W_D}}, \quad (16)$$
$$P_{b,W_U} = 1-(1-\tau_{W_U})^{n_{W_U}}, \quad (17)$$
$$P_{b,L} = 1-(1-\tau_L)^{n_L}. \quad (18)$$

In (10)–(12), T_{ave} is the average time duration spent when a packet is sent successfully from either WLAN or LTE system. The T_{ave} is given by

$$\begin{aligned} T_{\text{ave}} &= (1-P_{b,W_D})(1-P_{b,W_U})(1-P_{b,L})\delta \\ &+ P_{s,W_D}(1-P_{b,W_U})(1-P_{b,L})\overline{T}_{s,W_D} \\ &+ P_{s,W_U}(1-P_{b,W_D})(1-P_{b,L})\overline{T}_{s,W_U} \\ &+ P_{s,L}(1-P_{b,W_D})(1-P_{b,W_U})T_{s,L} \\ &+ P_{c,WW}(1-P_{b,L})T_{c,W} \\ &+ P_{c,LL}(1-P_{b,W})T_{c,L} + P_{c,WL}T_{c,M}, \end{aligned} \quad (19)$$

where $\overline{T}_{s,W_D} = \delta + T_{s,W_D}\frac{W_{0,D}}{W_{0,D}-1}$ and $\overline{T}_{s,W_U} = \delta + T_{s,W_U}\frac{W_{0,U}}{W_{0,U}-1}$ are the effective transmission durations of WLAN downlink and uplink, respectively, following a method in [15]. In (19), T_{s,W_D}, T_{s,W_U} and $T_{s,L}$ (or T_{c,W_D}, T_{c,W_U} and $T_{c,L}$) are durations of channel busy condition caused by one successful transmission (or collision), respectively. The $\delta = \delta_L = \delta_{W_D} = \delta_{W_U}$, where $\delta_L, \delta_{W_D}, \delta_{W_U}$ are idle slot durations of the three types of transmissions. Furthermore, $T_{c,W} = \max(T_{c,W_D}, T_{c,W_U})$, and $T_{c,M} = \max(T_{c,W}, T_{c,L})$.

The $P_{c,WW}$, $P_{c,LL}$, $P_{c,WL}$ refer to WLAN intra-system collision probability, LTE intra-system collision probability, and WLAN LTE inter-system collision probability, respectively:

$$P_{c,WW} = P_{b,W} - P_{s,W}, \quad (20)$$
$$P_{c,LL} = P_{b,L} - P_{s,L}, \quad (21)$$
$$P_{c,WL} = P_{b,W}P_{b,L}, \quad (22)$$

where

$$P_{b,W} = 1-(1-\tau_{W_D})^{n_{W_D}}(1-\tau_{W_U})^{n_{W_U}}, \quad (23)$$
$$\begin{aligned} P_{s,W} &= P_{s,W_D}(1-P_{b,W_U}) + P_{s,W_U}(1-P_{b,W_D}) \\ &= n_{W_D}\tau_{W_D}(1-\tau_{W_D})^{n_{W_D}-1}(1-\tau_{W_U})^{n_{W_U}} \\ &+ n_{W_U}\tau_{W_U}(1-\tau_{W_U})^{n_{W_U}-1}(1-\tau_{W_D})^{n_{W_D}}. \end{aligned} \quad (24)$$

The first term on the right hand side (RHS) of (19) gives the average idle duration, the 2nd to 4th terms on the RHS provide the average successful transmission durations of the three types of transmissions, and the 5th to 7th terms provide the average collision durations, respectively. Our approach shown in (19) is concise and flexible in computing the throughput statistics. It decouples the idle, successful transmission, and collision events of each type of transmission, and then groups the events (or related probabilities) properly to compute the statistics.

By substituting (13)–(24) into (10)–(12), the average

throughput (time efficiency) of WLAN and LTE systems can be evaluated. This method is more concise and flexible than those developed in [8], [9], and easily scalable to the case of multiple coexisting types of transmissions (even more than three).

IV. Technical Extension and Discussions

A. Generalization to Coexistence of Multiple Systems

In future coexistence applications, the number of types of transmissions may be larger than three. It is of practical interest to develop a general coexistence performance analysis method valid for multiple types of transmissions, as shown next. Assume N types of transmissions, and each transmission involves n_i nodes (or links), for $i = 1, \ldots, N$. Define δ, $T_{s,i}$, $T_{c,i}$, respectively, as the durations of counter idle slot, successful transmission, and collision; and $P_{f,i}$, $P_{b,i}$, $P_{s,i}$, respectively, as the probabilities of a failed transmission, channel busy (at the counter decrement phase), and successful transmission, for a node in transmission type i.

The steps of analytical evaluation are described below. First, construct the Markov chain based on the MAC layer scheme, and find the channel access probability of transmission type i ($i = 1, \ldots, N$), given by

$$\tau_i = g_i(P_{f,i}, Z_i, R_i) \tag{25}$$

where Z_i, R_i are the initial CW size and cutoff stage of type i, respectively, and g_i is the mapping function based on the MAC scheme of type i. For example, τ_i in (25) is provided by eqs. (4), (5), (6), for LTE-LBT and WLAN downlink and uplink systems, respectively. Next,

$$P_{f,i} = 1 - (1-\tau_i)^{n_i - 1} \prod_{\substack{j \neq i \\ j=1}}^{N} (1-\tau_j)^{n_j}. \tag{26}$$

For N types of transmissions, eqs. (25) and (26) involves $2N$ equations and $2N$ unknowns $\{f_i, \tau_i\}_{i=1,\ldots,N}$, which can be solved uniquely by iterative numerical research techniques.

The sum throughput of transmission type i is derived as

$$\begin{aligned} S_i &= P_{s,i} \prod_{\substack{j \neq i \\ j=1}}^{N} (1-P_{b,j}) T_{s,i}/T_{\text{ave}} \\ &= n_i \tau_i (1-\tau_i)^{n_i-1} \prod_{\substack{j \neq i \\ j=1}}^{N} (1-\tau_j)^{n_j} T_{s,i}/T_{\text{ave}} \end{aligned} \tag{27}$$

where T_{ave} is the average duration spent for each transmission type to send one payload successfully. Its formula is given by (28), shown on the top of the next page.

The first, second and third terms in (28) refer to durations for events of channel idle, successful transmission, and collision within one transmission type, respectively. The fourth and last terms refer to durations of collisions of two different types of transmissions, and simultaneous collisions of all types of transmissions, respectively. The skipped terms in (28) are durations of collisions among three types to $(N-1)$ types, whose notations are obvious and omitted here for the brevity of presentation.

B. LTE LBT Hand-Shaking Schemes

To evaluate the LTE transmission and collision durations $T_{s,L}$ and $T_{c,L}$ needed in (19), we propose two candidate hand-shaking schemes. The first one is similar to the WLAN basic access scheme, where the LTE transmission phase includes a 2-way handshaking: downlink data transmission and uplink ACK/NACK response. In the first setting (LBT basic channel access),

$$T_{s,L} = T_{P,L} + T_{\text{Defer}}, \tag{29}$$

$$T_{c,L} = T_{s,L}, \tag{30}$$

where the T_{Defer} (= T_{DIFS}) is the required defer period. Based on the method in Fig. 1, LTE transmitter's priority over the non-transmitting stations is removed. Notice that in DCF we have

$$T_{s,W} = T_{P,W} + T_{\text{SIFS}} + T_{\text{ACK}} + T_{\text{DIFS}}, \tag{31}$$

where T_{SIFS} and T_{ACK} are the short inter-frame space (SIFS) and ACK packet durations, respectively. In comparison, $T_{s,L}$ does not have terms T_{SIFS} and T_{ACK}. We explain it as follows: It is stated in [5] that the downlink HARQ-ACK timing rules from Release-12 carrier aggregation (CA) can be reused at least for DL-only LAA transmissions. This means that the delay between LTE transmission and the ACK/NACK response is on the order of several milliseconds (ms), much larger than the WLAN DCF SIFS (which in DCF is the delay between transmission and ACK/NACK). The milliseconds level of ACK/NACK delay involved in LTE-LAA does not block the other nodes for accessing the channel, and thus it is not counted in $T_{s,L}$. In the first scheme, when the LTE collision happens, the data packets sent in transmission duration are lost and the cost of collision may be high.

To improve the MAC efficiency, as the second setting, we propose an RTS/CTS-type scheme for LTE-LBT. Its transmitting phase includes 4-way handshaking: downlink RTS, uplink CTS, downlink data packet, and uplink ACK/NACK response. The time statistics are given by

$$\begin{aligned} T_{s,L} &= T_{L,\text{RTS}} + T_{L,\text{SIFS}} + T_{L,\text{CTS}} + T_{L,\text{SIFS}} + T_{P,L} \\ &\quad + T_{\text{Defer}} \end{aligned} \tag{32}$$

$$T_{c,L} = T_{L,\text{RTS}} + T_{L,\text{SIFS}} + T_{L,\text{CTS}} + T_{\text{Defer}}, \tag{33}$$

where $T_{L,\text{RTS}}$ and $T_{L,\text{CTS}}$ are the durations of RTS and CTS packets, respectively; $T_{L,\text{SIFS}}$ is the short handshaking delay (or LTE SIFS) between uplink transmission and downlink response, and we assume $T_{L,\text{SIFS}} = T_{\text{SIFS}}$.

The 2-way handshaking basic access scheme (the first setting) proposed above is consistent with the LAA LBT and HARQ-ACK guidelines provided in [5]. To our knowledge, the second setting – RTS/CTS-type LBT scheme has not been discussed in [5]. This scheme is proposed here for both theoretical and practical interest: it provides substantial performance enhancement because it can reduce the time-

$$T_{\text{ave}} = \prod_{i=1}^{N}(1-P_{b,i})\delta + \sum_{i=1}^{N} P_{s,i}\left(\prod_{\substack{j=1\\ j\neq i}}^{N}(1-P_{b,i})\right)T_{s,i} + \sum_{i=1}^{N}(P_{b,i}-P_{s,i})\left(\prod_{\substack{j=1\\ j\neq i}}^{N}(1-P_{b,i})\right)T_{c,i}$$
$$+ \sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} P_{b,i}P_{b,j}\left(\prod_{\substack{l=1\\ l\neq i,j}}^{N}(1-P_{b,l})\right)\max(T_{c,i},T_{c,j}) + \cdots + \left(\prod_{i=1}^{N} P_{b,l}\right)\max(\{T_{c,i}\}_{i=1,\ldots,N}). \quad (28)$$

efficiency loss caused by collisions. Performance of both schemes will be evaluated in Section V.

V. Numerical Results

In this section, we provide both analytical and simulation results of the WLAN downlink and uplink transmissions in coexistence with LTE-LBT downlink transmission. We implement computer programming based on modified DCF MAC algorithms given in [15], and the Category-4 LBT algorithm described in Fig. 1 adopted from [5]. In our simulation, we define three global events: channel idle, successful transmission, and collision; and four local events for each WLAN and LTE node: channel idle, counter freezing (due to channel being busy), successful transmission, and collision. The simulation results were obtained by running for 1×10^5 time slots on each parameter setting. In each slot, each node updates its local event and a global tracker updates the global event. Finally, based on accumulated numbers of events, the statistics of time-efficiency throughput and channel access and collision probabilities are computed for each node. Every analytical curve shown in this section is accompanied by a simulation curve and validated.

The parameters used for analysis and simulation are listed in Table I, where the WLAN parameters were adopted from [7], [9], [15], [16]. The $T_{s,W}$ and $T_{c,W}$ can be computed from the parameters in Table I using a method in [13], [15]. Here, channel bit rate (CBR) = 100 mega-bits per second (Mbps), assumed to be the same for both LTE and WLAN systems. Saturated traffic is assumed for all nodes. The spectrum sensing (aka. CCA) in both WLAN and LTE-LBT systems is assumed to be perfect (no hidden node problem, no false alarm or miss detection). We study the effects of basic access and RTS/CTS schemes for the WLAN system, and basic access and 4-way handshaking schemes for the LTE LBT.

First, we study the effect of the LTE-LBT cutoff stage setup, and let R_L increase from 0 to 8, when $n_L = n_{W_D} = 4$, $n_{W_U} = 20$, $R_D = R_U = 6$, $W_{0,D} = Z_L = 16$, and $W_{0,U} = 80$, with WLAN RTS/CTS access. Note that the case of R_L = 0 models the Category-3 LBT, and $R_L = 1, \ldots, 8$ models Category-4 LBT. Here, we set $W_{0,U} = 5W_{0,D}$ so that WLAN downlink has higher priority to access the channel than the uplink. The results on throughput and transmission probabilities are presented in Figs. 3 and 4, respectively. The analytical and simulation results are in close agreement. The results show that when R_L increases, each LTE eNB node decreases its own throughput and transmission probability, but the LTE and WLAN sum throughput increases.

TABLE I: LTE and WLAN Parameters Used for Simulation

LTE parameters

Parameter	Value
Packet payload duration $T_{P,L}$	2 ms
$T_{L,\text{RTS}}$ (= $T_{L,\text{CTS}}$)	10 μ s
$T_{L,\text{SIFS}}$	16 μ s
LBT defer period: T_{Defer} (=T_{DIFS})	34 μ s
LBT eCCA period: T_{eCCA} (=δ_W)	9 μ s

WLAN parameters

Parameter	Value
Packet payload duration:	1 ms
MAC and PHY headers	272 and 128 bits
T_{SIFS}	16 μ s
T_{DIFS}	34 μ s
Idle slot duration δ_W	9 μ s
Downlink: $W_{0,D}$, R_D	16, 6
Uplink: $W_{0,U}$, R_U	80, 6

Next, we check the effect of replacing some WLAN APs with an equal number of LTE eNBs, on the throughput of LTE and WLAN systems, respectively. In the first setting, we let $n_{W_U} = 20$, $n_L + n_{W_D} = 8$, and $Z_L = W_{0,D} = 16$, and show throughput results in Fig. 5 assuming WLAN basic access, and in Fig. 6 with WLAN RTS/CTS access, respectively. The LTE basic access scheme is assumed. As n_L increases from 0 to 8 (with $n_L + n_{W_D} = 8$), the overall throughput of three types of transmissions increases from about 70 % to 74 % with WLAN basic access (constructive coexistence), but decreases from about 88 % to 78 % with WLAN RTS/CTS access (non-constructive coexistence). This demonstrates that the coexistence results critically depend on whether the original WLAN system has low efficiency (70 %, basic access) or high efficiency (88 %, RTS/CTS access).

In the second setting, we compare two cases: WLAN only with $n_{W_D} = 8, n_L = 0$; and four WLAN APs are replaced by LTE eNBs ($n_{W_D} = 4, n_L = 4$). We study the effect of CW size Z_0, and effect of LTE-LBT access methods (basic vs. RTS/CTS) on the throughput. We consider WLAN RTS/CTS

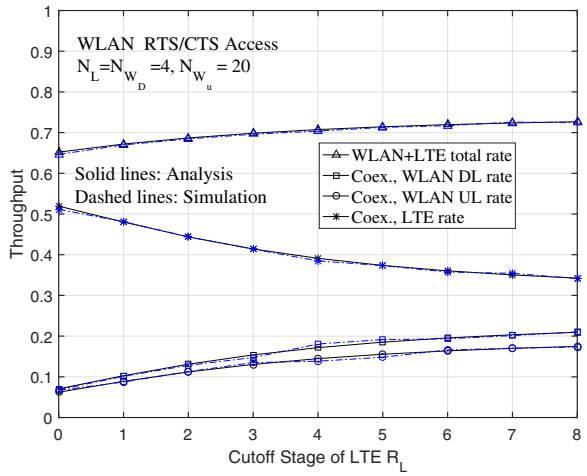


Fig. 3: Throughput of LTE and WLAN systems, when $n_L = n_{W_D} = 4$, $n_{W_U} = 20$, $R_D = R_U = 6$, $W_{0,D} = Z_L = 16$, and $W_{0,U} = 80$, with WLAN RTS/CTS access.

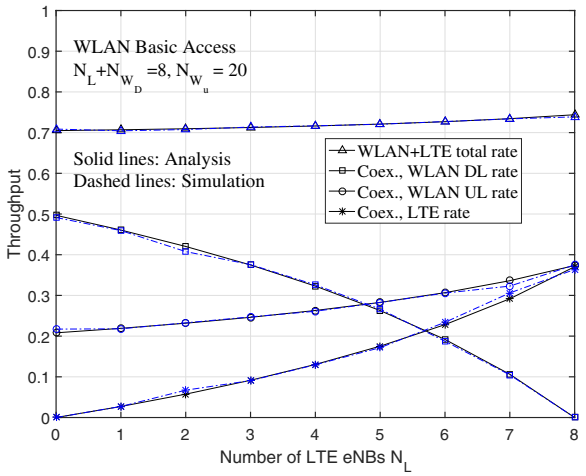


Fig. 5: Throughput of LTE and WLAN systems, when $n_L + n_{W_D} = 8$, $n_{W_U} = 20$, $R_D = R_U = R_L = 6$, $W_{0,D} = Z_L = 16$, and $W_{0,U} = 80$, with WLAN basic access.

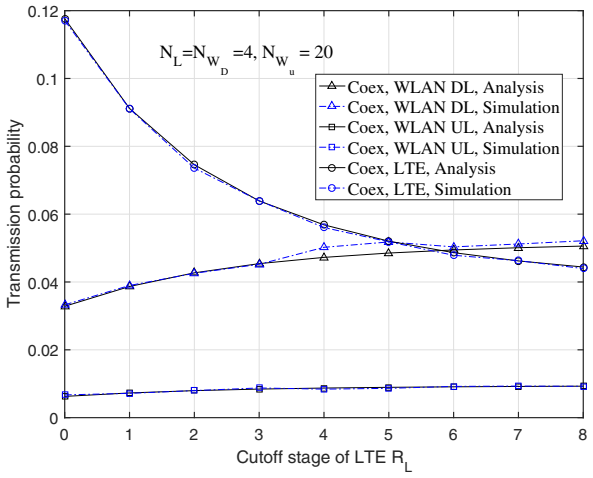


Fig. 4: Transmission (channel access) probabilities of LTE and WLAN systems, respectively.

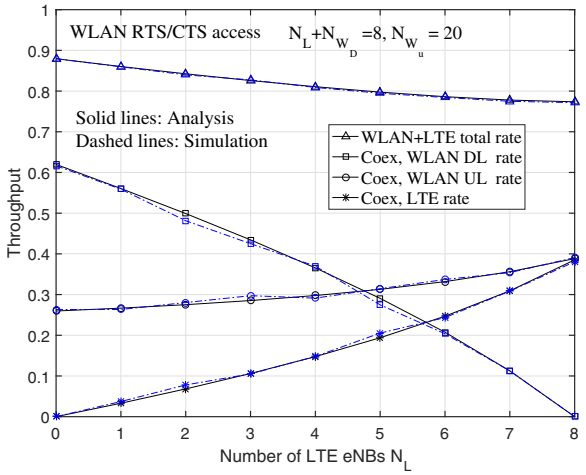


Fig. 6: Throughput of LTE and WLAN systems, with WLAN RTS/CTS access.

access, when $n_{W_U} = 20$, $R_U = R_D = R_L = 6$, $W_{0,D} = 16$, and $W_{0,U} = 80$. The results are provided in Fig. 7 assuming LTE basic access and in Fig. 8 assuming the proposed LTE-LBT 4-way handshaking channel access scheme (aka, LTE RTS/CTS-type access), respectively. It is observed that with LTE basic access scheme the coexistence sum throughput increases from about 70 % to 84 %, but is constantly worse than that of the original WLAN system (about 88 %). Fortunately, by using the LTE-LBT with our proposed RTS/CTS access, Fig. 8 shows that the coexistence sum rate (about 90 % $\sim$ 92 %) is consistently better than that of the original WLAN system. AT $Z_0 \approx 16$, the WLAN and LTE systems have approximately equal normalized throughput.

VI. Conclusion

In this paper, we have studied MAC-layer performance of LTE-LBT downlink coexisting with WLAN downlink and uplink transmissions. We have provided a flexible analytical approach to evaluate the transmission probability, collision probability and time-efficiency throughput. Since future coexistence scenarios will likely involve multiple systems, to support the related performance analysis, we have generalized the analysis result to multiple transmission types (more than three). We have implemented LTE and WLAN MAC algorithm programming and extensive computer simulation, and simulation results have verified the validity of our analysis result. Effects of the LBT cutoff stage, CW sizes, and handshaking access schemes have been evaluated. Regarding the fair coexistence between LTE and WLAN systems, our work shows

Kuester, Daniel; Ma, Yao.
"MAC-Layer Coexistence Analysis of LTE and WLAN Systems Via Listen-Before-Talk."
Paper presented at Same as above, Las Vegas, NV, United States. January 8, 2017 - January 11, 2017.

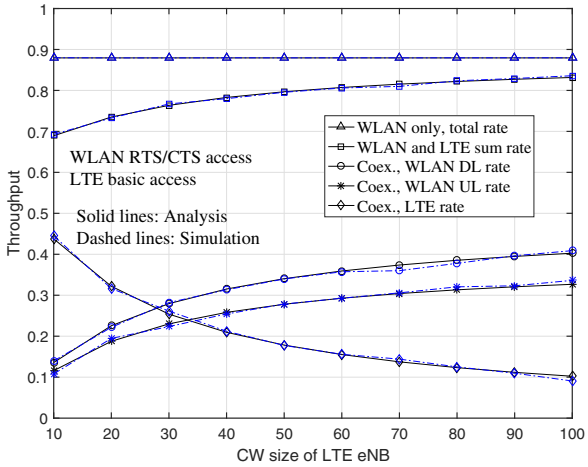


Fig. 7: Throughput of LTE-LBT and WLAN systems, when $R_L = 6$ with WLAN RTS/CTS access.

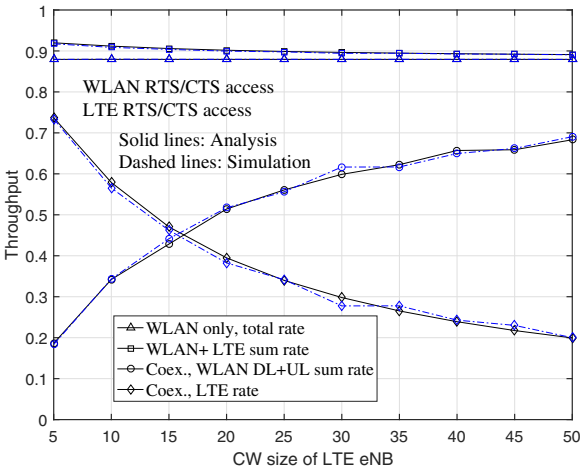


Fig. 8: Throughput of LTE-LBT and WLAN systems, when $R_L = 6$ and both WLAN and LTE systems use RTS/CTS access schemes.

that when the WLAN system uses RTS/CTS with saturated downlink and uplink traffic, replacing WLAN nodes with an equal number of LTE-LBT nodes may substantially reduce the overall throughput. To enhance the overall coexistence throughput, we have proposed an RTS/CTS type LBT channel access scheme and illustrated its competitive performance. In summary, the coexistence results depend heavily on the proper selection of handshaking schemes (basic or RTS/CTS access) and MAC parameters of both LTE-LBT and WLAN systems. Our results demonstrate that achieving satisfactory coexistence of LTE and WLAN systems is a non-trivial task, and additional research is needed. Perfect channel sensing is considered in this paper. In future work, the effect of channel sensing threshold and sensing errors will be studied, and measurement and testing procedure will be implemented to further validate the analytical and simulation results.

ACKNOWLEDGMENT

The authors thank Jason Coder, Bill Young, and Adam Wunderlich at the NIST for some helpful discussions and feedback during the preparation of this paper.

REFERENCES

[1] R. Zhang, M. Wang, L. X. Cai, Z. Zheng, and X. Shen, "LTE-unlicensed: the future of spectrum aggregation for cellular networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 150–159, Jun. 2015.

[2] A. Al-Dulaimi, S. Al-Rubaye, N. Qiang, and E. Sousa, "5G communications race: pursuit of more capacity triggers LTE in unlicensed band," *IEEE Veh. Technol. Mag.*, vol. 10, no. 1, pp. 43–51, Mar. 2015.

[3] F. M. Abinader, et. al. "Enabling the coexistence of LTE and Wi-Fi in unlicensed bands," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 54–61, Nov. 2014.

[4] A. Mukherjee et al., "Licensed-assisted access LTE: coexistence with IEEE 802.11 and the evolution toward 5G," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 50-57, Jun. 2016.

[5] 3GPP TSG RAN, "Study On Licensed-Assisted Access To Unlicensed Spectrum", 3GPP TR 36.889 V13.0.0, Jun. 2015.

[6] Ericsson, "Discussion on LBT protocols," 3GPP Tech. Rep. R1-151996, Apr. 2015.

[7] LTE-U Forum, "Coexistence study for LTE-U SDL", LTE-U Technical Report, V1.0, Feb. 2015.

[8] C. Chen, R. Ratasuk, and A. Ghosh, "Downlink performance analysis of LTE and WiFi coexistence in unlicensed bands with a simple listen-before-talk scheme," *Proc. IEEE VTC*, pp. 1–5, May 2015.

[9] Y. Song, K. W. Sung, and Y. Han, "Coexistence of Wi-Fi and cellular with listen-before-talk in unlicensed spectrum," *IEEE Commun. Lett.*, vol. 20, no. 1, pp. 161–164, Jan. 2016.

[10] V. Valls, A. Garcia-Saavedra, X. Costa and D. J. Leith, "Maximizing LTE capacity in unlicensed bands (LTE-U/LAA) while fairly coexisting with 802.11 WLANs," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1219–1222, Jun. 2016.

[11] S. Han, Y. C. Liang, Q. Chen and B. H. Soong, "Licensed-assisted access for LTE in unlicensed spectrum: A MAC protocol design," *Proc. IEEE ICC*, pp. 1–6, Kuala Lumpur, Malaysia, 2016.

[12] R. Yin, G. Yu, A. Maaref, and G. Y. Li, "A framework for co-channel interference and collision probability tradeoff in LTE licensed-assisted access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6078–6090, Sept. 2016.

[13] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[14] H. F. Chuan and J. W. Tantra,"Comments on IEEE 802.11 saturation throughput analysis with freezing of backoff counters," *IEEE Commun. Lett.*, vol.9, no.2, pp.130–132, Feb. 2005.

[15] I. Tinnirello, G. Bianchi, and X. Yang, "Refinements on IEEE 802.11 distributed coordination function modeling approaches," *IEEE Trans. Veh. Technol.*, vol.59, no.3, pp.1055–1067, Mar. 2010.

[16] E. H. Ong, et. al. "IEEE 802.11ac: Enhancements for very high throughput WLANs," *Proc. IEEE PIMRC*, pp. 849-853, 11–14 Sept. 2011.

[17] L. Dai and X. Sun, "A unified analysis of IEEE 802.11 DCF networks: stability, throughput, and delay," *IEEE Trans. Mobile Computing*, vol.12, no.8, pp.1558–1572, Aug. 2013.

[18] Y. Gao, X. Sun and L. Dai, "IEEE 802.11e EDCA networks: modeling, differentiation and optimization," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3863–3879, July 2014.

[19] IEEE WG802.11 Wireless LAN Working Group, *IEEE Std. 802.11e-2005 Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements*, Sep. 2005.

Predicting Face Recognition Performance in Unconstrained Environments

P. Jonathon Phillips

Amy N. Yates
National Institute of Standards and Technology
Gaithersburg, MD, USA

J. Ross Beveridge
Department of Computer Science, Colorado State University
Fort Collins, CO, USA

Geof Givens
Givens Statistical Solutions, LLC
Fort Collins, CO, USA

Abstract

While face recognition algorithms perform under many different unconstrained conditions, predicting this performance is not possible when a new location is introduced. Analyzing the impostor distribution of the videos of the Point-and-Shoot Challenge (PaSC) as well as its relationship to the genuine match distribution, we show that there is large variation in the false accept rate over the impostor distribution, demonstrate there is a correlation between changes in the verification and false accept rates over factor, and using this, present a method for predicting the performance of an algorithm using only unlabeled data for a new location.

1. Introduction

Face recognition algorithms operate under a variety of unconstrained conditions, and performance varies substantially across locations, i.e., the physical location. Given videos in a new location, how well will an algorithm perform? Without explicitly testing the new location, a common method is to use the overall performance of the system on previously known locations. We present a way to better model the performance without needing to identify and label individuals in the videos.

Are there any locations that are "easy" (high verification rate and low false accept rate)? It is commonly believed that there exist locations that are easy as well as ones that are "hard." From our analysis, we show that such locations do not necessarily exist.

On the Point-and-Shoot Face Recognition Challenge (PaSC) [2], Lee et al. [10] found that verification rate (VR) varies across locations. The effect of factors on the genuine match distribution has been studied [1], [7]. However, there has not been as much research on the impostor distribution. O'Toole et al. [13] found that performance changes when the impostor distribution is restricted to people of the same gender or race. Several researchers have focused on the effects of pose, expression, and illumination [8], [6].

Extending the work of Lee et al. [10] , we investigate how the false accept rate (FAR) varies across locations in the PaSC dataset. The algorithms in this study are from the Face and Gesture 2015 Person Recognition Evaluation [3]. In our analysis, we include video-based factors which are automatically computed [10]. We also analyze the relationship between the genuine match distribution and impostor distribution. Using this analysis, we demonstrate it is possible to predict the performance of an algorithm in a new location based solely on unlabeled data acquired from the new location.

Novel contributions in this paper are:

- We show that when a threshold is set so that the global FAR is fixed, there is a large variation in the false accept rates over the locations.
- We show that with this fixed threshold, changes in verification and face accept are correlated across locations.
- This correlation allows us to predict the verification rate for new locations using a regression model.

2. PaSC Challenge and Data Set

To investigate the false accept rate across the impostor distribution, we needed a data set that documented many factors about the videos themselves, especially with the location of videos systematically varied. The Point-and-Shoot Face Recognition Challenge (PaSC) was designed to advance the development of face recognition algorithms on videos taken with digital point and shoot cameras, particularly for handheld cameras found in cell phones; full details of the protocol can be found in [2]. What follows is a brief summarization of the relevant details.

Beveridge, J.; Givens, Geof; Phillips, P; Yates, Amy.
"Predicting Face Recognition Performance in Unconstrained Environments."
Paper presented at IEEE Computer Society Workshop on Biometrics 2017, Honolulu, HI, United States. July 21, 2017 - July 21, 2017.

2.1. Data Set

In our analysis, we focus on the effect of location and sensor. This is possible because videos in the PaSC are taken from six locations with six sensors, five of those being handheld.

In the video portion of the PaSC, 2802 videos of 265 subjects were taken over 7 different weeks at the University of Notre Dame in the spring semester of 2011. The videos show people carrying out tasks rather than looking into a camera. Collection was carried out according to a plan–a script–in which generally a person entered a scene, approached some designated spot, carried out an action, and then left the scene. The videos typically begin as the person is moving into the scene and terminate as the person is leaving.

Each subject is present in videos for at least four of the weeks, implying the differences in weeks' performances is not due to the subjects. Video length ranges roughly between 50 and 400 frames with most videos containing between 200 and 250 frames, and the resolutions ranged between 640×480 to 1280×720.

There were six different locations with six different sensors. Five of the sensors were handheld, and these varied by week. Additionally, data was collected by a tripod-mounted sensor, and this sensor filmed the same actions at the same location and time as the handheld sensor of the week.

Figure 1. Sampled portions of video frames from PaSC videos indicating some of the situations that make recognition challenging. Courtesy of Beveridge et al. [4].

Figure 1 shows a sample of frames from PaSC videos from different locations. Characterizing the videos are four primary factors: location, action being performed, video camera (sensor), and person in the video (subject).

2.2. Location Factor

One aspect the design of this data set allows us to analyze is how an algorithm performs when restricted to pairs of videos from certain locations. During each week, the videos were collected with a new combination of location and action taking place, for example picking up a newspaper in an office. No combination of location and action was repeated on subsequent weeks. Table 1 shows a summary of the location, handheld camera, and action combinations.[1]

Table 1. Location, camera, and action combinations. The abbreviations for the location is in the right column.

Sensor	Location	Action	Abbrev.
Flip Mino F360B	canopy	golf swing	Ca
Kodak Zi8	canopy	bag toss	Ca
Samsung M. CAM	office	pickup newspaper	Pa
Sanyo Xacti	lab 1	write on easel	Ea
Sanyo Xacti	lawn	blow bubbles	Bu
Nexus Phone	hallway	ball toss	Ba
Kodak Zi8	lab 2	pickup phone	Ph

Each location and action combination was captured on a specific week by two different cameras, one being handheld. Consequently, each video depicts a single subject at a certain location doing a specific action captured by one particular sensor, e.g. for a specific subject, there is exactly one video depicting the subject on the lawn blowing bubbles captured by a Sanyo Xacti. There is also a video of the subject blowing bubbles on the lawn captured by the tripod-mounted sensor, a Panasonic HD700. From Table 1, it is clear that the handheld sensors are confounded with the locations and actions.

In the findings below, the influence that location, camera, and action combinations (called the location factor for simplicity) exert over performance is strong, and the abbreviations introduced in Table 1 will be used when reporting results. Therefore here, briefly, is a bit more information about each. The **canopy** (Ca) was a white pop-up material structure setup outside in bad weather. Two actions were carried out on different days. The first was swinging a golf club, and the second was tossing a bean bag. The **office** (Pa) was a large well-lit room where a subject picked up and looked at a newspaper. In **Lab 1** (Ea) each subject wrote on a large floor standing easel set out in a large open lab space. The **lawn** (Bu) was an open grassy area in a plaza with bright sun. Subjects approached a table and blew bubbles. The **hallway** (Ba) was an interior space of an older building with relatively dark stone walls where subjects threw a toy basketball. In **lab 2** (Ph) a subject picked up a phone in a relatively cluttered lab area.

As videos are compared in pairs, the location factor is defined by location-pairs, i.e. the locations of the videos for a given pair. In total, there are 22 location-pairs. For 6 pairings the videos are from the same location and collected in the same week; these only include impostor pairs, i.e. pairs

[1]The identification of any commercial product or trade name does not imply endorsement or recommendation by NIST.

Beveridge, J.; Givens, Geof; Phillips, P; Yates, Amy.
"Predicting Face Recognition Performance in Unconstrained Environments."
Paper presented at IEEE Computer Society Workshop on Biometrics 2017, Honolulu, HI, United States. July 21, 2017 - July 21, 2017.

of videos of different people. However, we focus mainly on cross-week comparisons, i.e. video-pairs in which the weeks of capture are different. There are 16 cross-week location-pairs. In 15 of the cross-week location-pairs, the videos were collected at different locations from different weeks, but for one pair, the videos were collected at the same location (canopy) on different weeks.

2.3. Video-Based Factors

Location, action, and sensor are not the only factors effecting performance. Another class of factors effecting performance comes directly from the videos themselves; that is, these factors, called video-based factors, are dependent on the video from which they are estimated. As we show later in Section 6, video-based factors can encode properties of a location-pair. For our work, we measure this encoding by looking at aggregate statistics of video-based factors from all video-pairs of the location-pair.

We consider three video-based factors: face size, face confidence, and yaw. Estimated by the Pittsburgh Pattern Recognition (PittPatt) face recognition SDK 5.2.2, face size is the number of pixels between the eyes, face confidence is PittPatt's self-assessment of how certain the algorithm was in detecting the true face, and yaw is the measurement of how far the face was turned to the left or right.

The real-valued factors are converted to levels by ordering video-pairs from smallest to largest factor value and then dividing them into n equal sized bins. The result is n levels ranging from smallest to largest factor value. The PittPatt SDK 5.2.2 software estimated these factors for the frames of the videos, and the generalizations to videos and video-pairs follow the methods of Lee et al. [10].

3. Algorithms

Our analysis is performed on the four top performers in the Face and Gesture 2015 Person Recognition Evaluation [3]. The algorithms were developed independently by four different research groups from four different countries on four different continents. Each algorithm is very different in how it computes a similarity score (the degree of similarity between two faces in two videos). This independence provides evidence that our conclusion will generalize to algorithms not included in this study.

The Chinese Academy of Science (CAS) algorithm uses two convolutional neural networks, one for larger and one for smaller faces [9].

The Stevens Institute of Technology (SIT) algorithm combines scale-invariant feature transform (SIFT) features with a probabilistic modeling procedures and principal component analysis based dimensionality reduction process [11], [12].

The University of Ljubljana (Ljub) algorithm combines four feature types with a probabilistic principal component analysis [15].

The Univeristy of Technology, Sydney, (UTS) algorithm uses three-dimensional face pose normalization and face descriptors [5].

4. Measuring Performance

Our results are reported on participants in the Face and Gesture 2015 Person Recognition Evaluation [3], and in this competition, the participants followed the PaSC protocol. In the protocol for the PaSC, algorithms are given two videos and then return a number measuring the degree of similarity between the subjects in the pair of videos. Hence, in calculating and predicting performance, we compare videos in pairs.

In measuring performance, we are observing how often an algorithm correctly declares the same person to be in two videos. We are also interested in how often the algorithm incorrectly believes two different people from videos are the same person. However, we are not interested in the overall performance of the algorithm. Instead, we are more interested in how the performance changes over levels of a factor. Later in this paper, for a set of videos of a factor-level, we are predicting how well an algorithm will correctly match videos of the same person (marginal VR). In our prediction, we use how often the algorithm incorrectly declared different people to be the same (marginal FAR). We then compare our predicted performance to the actual observed performance.

The focus of analysis in this paper is on performance when comparing videos for a factor-level. Presented with two faces from videos x and y, an algorithm A returns a similarity score, $s_A(x, y)$, for video-pair (x, y). The similarity score denotes how similar the faces are estimated to be; a higher similarity score indicates a higher likelihood of the two faces belonging to the same subject.

To make a decision, a threshold τ_g is set so that every video-pair score at least as large τ_g is declared a match and every score below the threshold is considered a non-match. We divide the set of videos into two sets: the set of video-pairs that are genuine matches and the set of video-pairs that are impostors. With the threshold τ_g, we calculate the verification rate $\mathrm{VR}(\tau_g)$ as the ratio of genuine matched video-pairs correctly identified as a match and the false accept rate $\mathrm{FAR}(\tau_g)$ as the ratio of impostor video-pairs incorrectly identified as a match.

Generally, the threshold τ_g is set to specify the FAR at a certain instance. In our paper, we select τ_g for each algorithm so that globally $\mathrm{FAR}(\tau_g) = 0.10$. For PaSC, the standard for reporting VR is $\mathrm{FAR} = 0.01$. However, we shifted the threshold to have enough false matches for analysis.

Nonetheless, the analysis in this paper is not focused on the overall performance over the set of all video-pairs.

Beveridge, J.; Givens, Geof; Phillips, P; Yates, Amy.
"Predicting Face Recognition Performance in Unconstrained Environments."
Paper presented at IEEE Computer Society Workshop on Biometrics 2017, Honolulu, HI, United States. July 21, 2017 - July 21, 2017.

Rather, for this paper, as previously mentioned, the analysis is centered on performance when comparing video-pairs of factor levels such as locations. For the marginal verification and false accept rates for factor F_i, a threshold τ_g is set so that globally $\mathrm{FAR}(\tau_g) = 0.10$, and with this threshold, the verification rate $\mathrm{VR}(F_i, \tau_g)$ and false accept rate $\mathrm{FAR}(F_i, \tau_g)$ are then calculated only on the video-pairs in F_i.

5. Imposter-Pair Analysis For Location-Pairs

In this section, we picked the threshold τ_g so that the global FAR = 0.10 and then investigated the impostor distribution over the different location-pairs, calculating the marginal false accept rates over the location-pairs using the threshold τ_g. We showed that there is large variation in the false accept rate. Then we showed that, keeping the threshold τ_g constant, the changes in the verification and false accept rates over the location-pairs are correlated.

5.1. Range of Marginal FARs over Location-Pairs

It is well known that location significantly effects algorithm performance. The design of the PaSC data set enabled us to characterize the impact of location on performance. Previous studies have investigated the effect of location on verification rates [1], [10]. We proceed by examining the effect of location on the FAR and then look at the relationship between FAR and VR.

Since comparisons are between two videos, we look at performance for location-pairs. For the four algorithms in our study, we computed the FAR for the 22 location-pairs as described in Section 4. Figure 2 demonstrates how location factors effect FAR (upper graph) and VR (lower) for the four algorithms on handheld video-pairs when the global FAR is set to 0.10. Along the horizontal axes are the pairs of locations described in Section 2.2. All 22 pairs are present in the upper graph, but only the 16 cross-week pairs are present in the lower graph because the same-week comparisons only contain impostor pairs. The vertical axes show the marginal FAR and VR values, respectively, using a τ_g that corresponds to a global FAR of 0.10. The location pairs are ordered by the mean rate over all the algorithms for both graphs. In the top graph, all location pairs to the left of the vertical line (from pairs Ba-Ca to CaDW-CaDW) are cross-week pairs; CaDW signifies canopy videos taken in different weeks. All pairs to the right consist of video-pairs taken in the same week.

The principal finding is that location exerts a dramatic influence over the impostor distribution and hence the marginal FAR. For handheld video-pairs, Algorithm Ljub has the greatest range in FAR from 0.01 to 0.42, and CAS has the smallest range from 0.05 to 0.27; for tripod video-pairs, Ljub still has the greatest range in FAR from 0.02 to 0.39, and CAS has the smallest range from 0.03 to 0.22. Table 2 shows the ranges for the cross-week location-pairs over both sets of video-pairs. For the handheld video-pairs, the FAR for the four algorithms CAS, UTS, Ljub, and SIT varies by a factor of 3.6, 7.33, 21, and 11.5, respectively. For the tripod video-pairs, the FAR for the algorithms CAS, UTS, Ljub, and SIT varies by a factor of 4.33, 7.67, 9, and 7, respectively. Prior work has already suggested the importance of location [1], [10]; this is the first clear evidence of how significantly it effects the impostor distribution.

Table 2. The cross-week ranges of location-pair marginal $\mathrm{FAR}(L_i, \tau_g)$ location-pairs over both sets of video-pairs with a threshold τ_g set so that global FAR = 0.10.

Algorithm	Handheld	Tripod
CAS	$0.05 - 0.18$	$0.03 - 0.13$
UTS	$0.03 - 0.22$	$0.03 - 0.23$
Ljub	$0.01 - 0.21$	$0.02 - 0.18$
SIT	$0.02 - 0.23$	$0.03 - 0.21$

A related finding is the importance of the cross-week versus same-week distinction. For both sets of video-pairs, the mean cross-week marginal FAR averaged over the algorithms was 0.09 compared to 0.21 for same-week pairs. A recent related result on still face image by Sgori et al. [14] also showed higher FAR values for same day image-pairs compared to different day image-pairs. One important conclusion is that the presence of impostor pairs in a data set taken at the same time biases upward the expected FAR for the data set as a whole.

5.2. Do VR and FAR Track Together?

We will now look at the relationship between the location-pair FARs and VRs for the cross-week pairs. Scatterplots in Figure 3 relate marginal VR to marginal FAR, described in Section 4, for the 16 cross-week location-pairs over the different sensor-pairs. The horizontal axis is the FAR on a log-scale, and the vertical axis is the VR on a linear scale. The points represent location-pairs over different sensor-pairs, and the line is a linear regressor. For all four algorithms, the regression line suggests a linear relationship between $\log(\mathrm{FAR})$ and VR. In other words, a location-pair that has a higher marginal VR will likely have a higher marginal FAR. Unfortunately, this linear relationship suggests that finding a location-pair that is easier than others is unlikely. We say a location-pair is easier if it has both a higher VR and a lower FAR than other pairs.

6. Imposter-Pair Analysis For Video-Based Factors

In this section, we investigated the impostor distribution over the different video-based factors and showed that there is large variation in the false accept rate. Then we

Beveridge, J.; Givens, Geof; Phillips, P; Yates, Amy.
"Predicting Face Recognition Performance in Unconstrained Environments."
Paper presented at IEEE Computer Society Workshop on Biometrics 2017, Honolulu, HI, United States. July 21, 2017 - July 21, 2017.

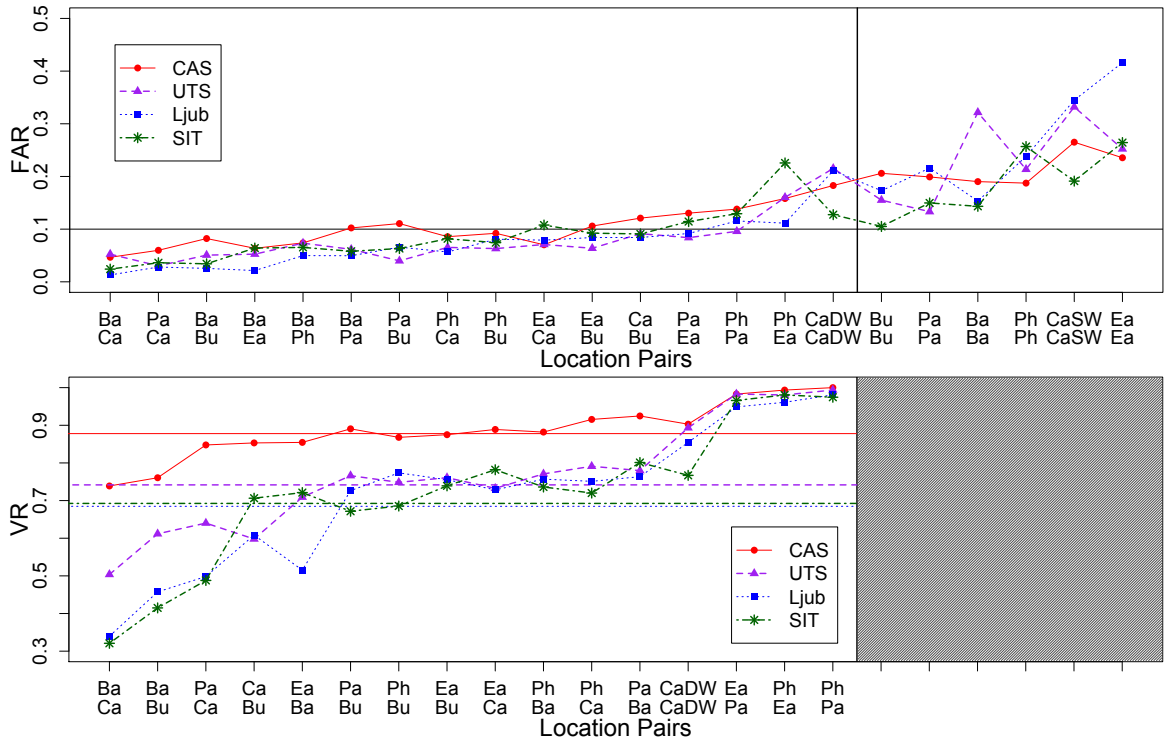


Figure 2. With a threshold set τ_g so that the global FAR is fixed, these graphs show the marginal $\mathrm{FAR}(L_i, \tau_g)$ and $\mathrm{VR}(L_i, \tau_g)$ of each location-pair on handheld video-pairs for each algorithm–ordered by the mean rate over all the algorithms. The top graph is on $\mathrm{FAR}(L_i, \tau_g)$. The horizontal line corresponds to the global FAR = 0.10, and the vertical line between pairs CaDW-CaDW and Bu-Bu separates the pairs into cross-week (left) and same week. The bottom graph is on $\mathrm{VR}(L_i, \tau_g)$. The horizontal lines correspond to the global $\mathrm{VR}(\tau_g)$ for each algorithm when the global FAR = 0.10. There are no same-week pairs for matches.

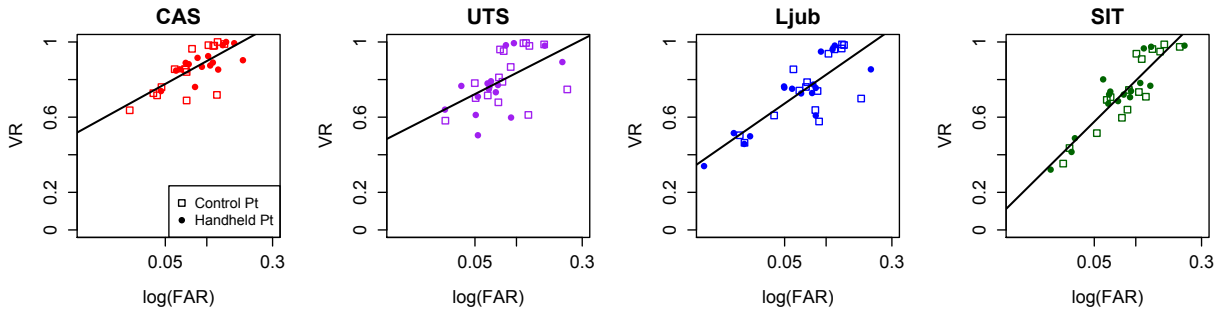


Figure 3. Scatterplots of $\mathrm{VR}(L_i, \tau_g)$ vs $\log(\mathrm{FAR}(L_i, \tau_g))$ of location-pairs over different sensor-pairs with a threshold τ_g set to that global FAR = 0.10.

showed that changes in the verification and false accept rates over the location-pairs are correlated and interact with the location-pairs.

The impact of image- and video-based factors on verification rates has been extensively studied; however, their impact on the FAR has not been examined. We first look at the relationship between FAR and VR for three video-based factors and then investigate if there is an interaction between location-pairs and the video-based factors.

Figure 4 shows the trade-off between FAR and VR for face size. The procedure described at the end of Section 2.3 for creating factor levels through sorting and binning was

Beveridge, J.; Givens, Geof; Phillips, P; Yates, Amy.
"Predicting Face Recognition Performance in Unconstrained Environments."
Paper presented at IEEE Computer Society Workshop on Biometrics 2017, Honolulu, HI, United States. July 21, 2017 - July 21, 2017.

used to create 10 face size factor levels: smallest faces to largest faces. Each point in Figure 4 is plotted according to the average marginal VR and FAR for all those video-pairs at one face size level. A trend similar to that seen for location factors is evident, changes in face size associated with higher marginal VR correlate with higher marginal FAR. There is a similar relationship for yaw and face size.

Figure 5 highlights the interactions between location and video factors for Algorithm Ljub. Like the scatterplots in Figure 3, each point corresponds to a location-pair and sensor-pair. Unlike in Figure 3, in Figure 5 circle size varies and is proportional the mean video factor for a location-pair. For the yaw-factor, all the circles are about the same size, which means that yaw does not interact with the location-pair. In contrast, a clear interaction effect between location and face size is evident: location-pairs with smaller VR and FAR tend to have small circle sizes and hence smaller mean face sizes. Figure 5 also suggests some interaction between location and face confidence.

This analysis was repeated for Algorithms SIT, UTS, and CAS, and the conclusions were the same. Across all four algorithms for all three video factors, we saw a trade-off between VR and FAR for different levels of each factor. Further analysis suggested an interaction between location and both face size and face confidence with face size having a larger interaction.

7. Predicting Performance

7.1. Models

With a new, previously unseen, location being compared to a known location, how well can performance (marginal VR) be predicted? We know that there is a wide range of potential marginal VR. Figure 3 illustrates this, showing scatterplots of VR vs $\log(\mathrm{FAR})$ of location-pairs over different sensor-pairs. Recall that additionally, a linear regressor is fit to the points for each algorithm. Observe the ranges of the marginal VR for the location-pairs of the four algorithms. For the algorithm SIT, the range is from 0.32 to 0.99 when the threshold τ_g is picked to set the global FAR to 0.10, described in Section 4.

What if, instead of one new location, two locations are new and compared against each other? How well can we accurately predict performance of this entirely new pair? Is it even possible to predict the performance with the same technique used when only one location is new? Which factors should be included in a model?

We started with a very simple model. As explained below, Linear Model 1 uses only the FAR of a location-pair to predict what the observed VR will be. Simply knowing how many false positives are in the set of video-pairs for a location-pair can indicate how well the algorithm will perform for those video-pairs. Additionally knowing some more information on the video-pairs, i.e. the video-based factors from Section 2.3, a better prediction can be made using Linear Model 2.

In Figure 3, a simple linear regressor is fit solely to the marginal verification and false accept rates of the location-pairs. The linear regressor is given by

$$\mathrm{VR} = \alpha + \beta \log(\mathrm{FAR}). \tag{1}$$

This is Linear Model 1.

Video-based factors are not incorporated into Linear Model 1. However, as we noted earlier, there is interaction between location and two video-based factors. There is interaction between location and face size, there is less interaction between location and face confidence, but there is no interaction seen between location and yaw.

To find a second model that utilizes video-based factors, we removed each location and partitioned the subjects into training and testing sets. On the remaining video-pairs that had both subjects in the training set, we fit models on the marginal VR using marginal FAR as well video-based factors from Section 2.3 and any relevant two-way interaction terms for each location-pair; we only kept terms that were significant ($p < 0.05$).

Many models resulted, and they performed robustly the same across the algorithms indicating that specifically which terms are in the model is not highly significant. With a set of second models being robustly the same in terms of prediction performance, we chose for Linear Model 2 to be given by

$$\mathrm{VR} = \alpha + \beta_1 \log(\mathrm{FAR}) + \beta_2\, \mathrm{Yaw} + \beta_3\, \mathrm{FC} + \beta_4\, \mathrm{Yaw} * \log(\mathrm{FAR}) \tag{2}$$

where Yaw is the mean yaw and FC stands for the mean face confidence for the video-pairs of the location-pair. We use these models in the method described below in Section 7.2 for predicting performance.

7.2. Prediction Procedure

In order to predict how well a set of videos of a location-pair might perform, we do the following. There are sixteen cross-week location-pairs over different sensor-pairs. For each location-pair L_i, one of the locations is randomly dropped. There will be no location-pair (no video) containing the dropped location; this location will be new. On the video-pairs of the remaining cross-week location-pairs, the subjects are partitioned into two sets: training and testing. Only video-pairs with both subjects in the training set are used.

With the video-pairs of the training set subjects, the global threshold τ_g is set so that the global FAR is 0.10. The global VR is calculated over all video-pairs in the training set using τ_g; this is denoted as VR_g. For the extant location-pairs, none of which use the new location, the marginal val-

Beveridge, J.; Givens, Geof; Phillips, P; Yates, Amy.
"Predicting Face Recognition Performance in Unconstrained Environments."
Paper presented at IEEE Computer Society Workshop on Biometrics 2017, Honolulu, HI, United States. July 21, 2017 - July 21, 2017.

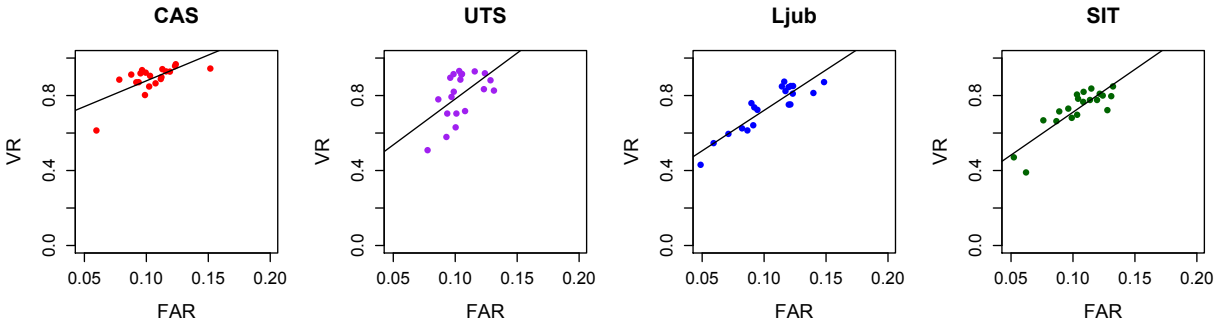


Figure 4. Scatterplots of $\mathrm{VR}(F_i, \tau_g)$ vs $\mathrm{FAR}(F_i, \tau_g)$ for Face Size over different sensor-pairs, divided into 10 bins, fitted with a linear regressor for each algorithm. Thresholds τ_g set to global FAR = 0.10.

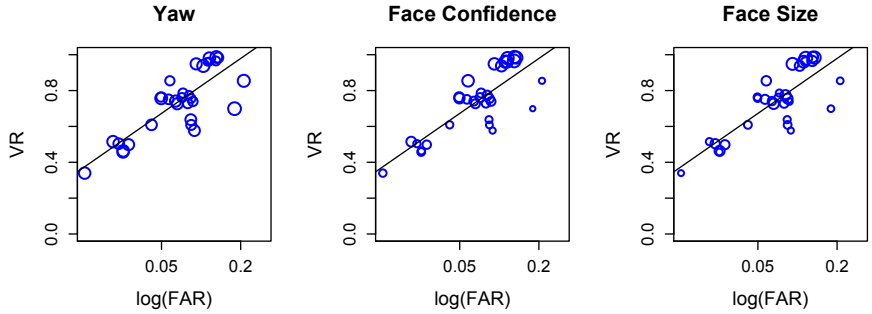


Figure 5. With threshold τ_g set for a global FAR = 0.10, interactions between Algorithm Ljub location-pairs from Figure 3 and each of the three video-based factors: yaw, face confidence, and face size. Each panel looks at the interaction for the factor in its title. The size of each circle is proportional to the mean of the factor for each location-pair.

ues are calculated over the different sensor-pairs, and these are used to fit the regression models from Section 7.1.

Using τ_g and the method described in Section 4, the observed marginal VRs of the location-pair L_i are calculated over sensor-pairs; we denote this by vr_i. Furthermore, the marginal FARs, far_i, are also calculated. With the marginal values, a regression line can predict the observed verification rate. This predicted VR is $\widehat{\mathrm{vr}}_i = f(\mathrm{far}_i)$ where the function f is Linear Model 1 (eq. 1) or Linear Model 2 (eq. 2).

The root mean square error (RMSE) is used to determine the standard deviation between the predicted VR and the observed VR (vr_i). When using the global rate, VR_g, to predict the observed VR, the RMSE is denoted by $\mathcal{G}$. When using the VR predicted by a regression line, $\widehat{\mathrm{vr}}_i$, the RMSE is denoted by $\mathcal{E}$. Equations 3 and 4 formally express the definitions, respectively.

$$\mathcal{G} = \sqrt{\frac{\sum_{i=1}^{n} (\mathrm{VR}_g - \mathrm{vr}_i)^2}{n}} \tag{3}$$

$$\mathcal{E} = \sqrt{\frac{\sum_{i=1}^{n} (\widehat{\mathrm{vr}}_i - \mathrm{vr}_i)^2}{n}} \tag{4}$$

8. Results of Prediction

Are these models better than using the global VR? In order to test the models from Section 7.1, we implemented the procedure from Section 7.2 100 times with one location being new for each location-pair. Then, in order to test if the method was valid for two new, unseen locations, we ran the procedure another 100 times, but this time, both locations of a location-pair were new.

After 100 iterations of the Section 7.2 process, Figure 6(a) displays the mean RMSEs, equations 3 and 4, of predicting the observed VR with the previous global VR, with the VR produced from Linear Model 1, and with the VR produced from Linear Model 2 over all location-pairs and sensor-pairs. The bars extend one standard deviation. For Algorithms Ljub and SIT, the mean RMSEs from forecasting using Linear Model 1 are much lower than using the global VR, which are over 0.21. For the algorithms CAS

Beveridge, J.; Givens, Geof; Phillips, P; Yates, Amy.
"Predicting Face Recognition Performance in Unconstrained Environments."
Paper presented at IEEE Computer Society Workshop on Biometrics 2017, Honolulu, HI, United States. July 21, 2017 - July 21, 2017.

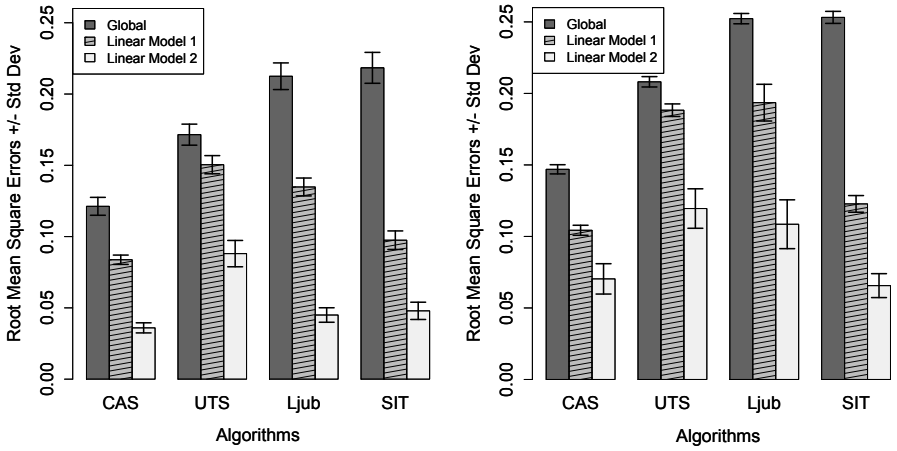


Figure 6. Bar plots of the mean RMSEs with standard deviation bars. In (a), one location is new. In (b), two locations are new.

and UTS, using Linear Model 1 is still better than using the global VR, which have mean RMSEs over 0.12, but the gap is not as large as it is for the other two algorithms.

The second linear model predicts the observed VR even better than the first linear model. The RMSEs from Linear Model 2 are much smaller than those from Linear Model 1 and definitely from those using the global VR. In fact, the means from Linear Model 2 are below 0.05 across three of the algorithms: CAS, Ljub, and SIT. The mean RMSE of Algorithm UTS is under 0.09, which is much smaller than it was from using the global VR or Linear Model 1 VR.

After 100 iterations, Figure 6(b) displays the mean RMSEs of predicting the observed VR with the global VR, with the VR produced from Linear Model 1, and with the VR produced from Linear Model 2 as in Figure 6(a), but in Figure 6(b), instead of one location being new, now both locations are new. Again, in general forecasting with Linear Model 1 is better than simply using the global VR. Using the global VR, Algorithm CAS has a mean RMSE of 0.15, and UTS has a mean RMSE of over 0.20. Algorithms Ljub and SIT have mean RMSEs over 0.25. For the algorithm SIT, the mean RMSE of Linear Model 1 less than half the mean RMSE of the global VR prediction. For Algorithms CAS, UTS, and Ljub, Linear Model 1 is still better than the previous global VR, but the differences are not as large as it is for SIT.

The second linear model still does even better than the first. There is a little more variability than before, but that is not surprising as now both locations are new. The mean RMSEs are under 0.12 for Algorithms UTS and Ljub, and the mean RMSEs are below 0.08 for Algorithms CAS and SIT.

9. Conclusion

We have shown that it is possible to predict the performance of an algorithm on unseen videos at a new location. We demonstrated that using the previously-known global VR is not a very good estimate; there is a lot of variability in marginal VR across location-pairs. We presented two models for predicting the marginal VR of a new location. The first model uses only the marginal FAR, and the second uses the marginal FAR as well as two video-based factors: yaw and face confidence. Both methods are better than simply using the previous global VR, but the second model came the closest to predicting the observed VR. Given two new locations, the second model is much better than using the global VR. The algorithms on which we tested were from four different groups on four different continents, implying that our results will generalize well.

To develop these models, we looked at the effect of location-camera-action (simply called location) and video factors on the FAR. Surprisingly, for location and video-based factors there was a clear relationship between VR and FAR. For these factors, one level is not better than another; there is a trade-off between VR and FAR. An increase (resp. decrease) in the FAR results in an increase (resp. decrease) in the VR. Our results illuminate a path for better understanding the performance of face recognition algorithms in unconstrained scenarios. The results underscore a need to better control a tendency of current algorithms to increase impostor scores in favorable settings as defined by higher genuine match scores. These results also establish a foundation for better modeling of distributional changes conditioned on measurable, knowable, attributes of target application locations, and thus bring us closer to the goal of predicting performance on unseen videos at new locations.

Beveridge, J.; Givens, Geof; Phillips, P; Yates, Amy.
"Predicting Face Recognition Performance in Unconstrained Environments."
Paper presented at IEEE Computer Society Workshop on Biometrics 2017, Honolulu, HI, United States. July 21, 2017 - July 21, 2017.

References

[1] J. R. Beveridge, G. H. Givens, P. J. Phillips, B. A. Draper, and Y. M. Lui. Focus on quality, predicting FRVT 2006 performance. In *Proceeding of the Eighth International Conference on Automatic Face and Gesture Recognition*, 2008.

[2] J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Givens, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In *IEEE Conference on Biometrics: Theory, Applications and Systems*, 2013.

[3] J. R. Beveridge, H. Zhang, B. A. Draper, P. J. Flynn, Z. Feng, P. Huber, J. Kittler, Z. Huang, S. Li, Y. Li, M. Kan, R. Wang, S. Shan, X. Chen, H. Li, G. Hua, V. Štruc, J. Križaj, C. Ding, D. Tao, and P. J. Phillips. Report on the FG 2015 video person recognition evaluation. In *Proceedings Eleventh IEEE International Conference on Automatic Face and Gesture Recognition*, 2015.

[4] J. R. Beveridge, H. Zhang, P. Flynn, Y. Lee, V. E. Liong, J. Lu, M. Angeloni, T. Pereira, H. Li, G. Hua, V. Struc, J. Krizaj, and P. J. Phillips. The IJCB 2014 PaSC video face and person recognition competition. In *Proceedings of the International Joint Conference on BIometrics*, 2014.

[5] C. Ding, C. Xu, and D. Tao. Multi-task pose-invariant face recognition. *IEEE Transactions on Image Processing*, 24(3):980–993, March 2015.

[6] A. Dutta, R. N. J. Veldhuis, and L. J. Spreeuwers. Predicting face recognition performance using image quality. *CoRR*, abs/1510.07119, 2015.

[7] G. H. Givens, J. R. Beveridge, P. J. Phillips, B. A. Draper, Y. M. Lui, and D. S. Bolme. Introduction to face recognition and evaluation of algorithm performance. *Computational Statistics and Data Analysis*, 67:236–247, 2013.

[8] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807 – 813, 2010.

[9] Z. Huang, R. Wang, S. Shan, and X. Chen. Hybrid Euclidean-and-Riemannian Metric Learning for Image Set Classification. In *Proceedings of the 12th Asian Conference on Computer Vision (ACCV 2014)*, Singapore, November 2014.

[10] Y. Lee, P. J. Phillips, J. J. Filliben, J. R. Beveridge, and H. Zhang. Generalizing face quality and factor measures to video. In *International Joint Conference on Biometrics (IJCB)*, 2014.

[11] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3499–3506. IEEE, 2013.

[12] H. Li, G. Hua, X. Shen, Z. Lin, and J. Brandt. Eigen-Pep for Video Face Recognition. In *Proceedings of the 12th Asian Conference on Computer Vision (ACCV 2014)*, 2104.

[13] A. J. O'Toole, P. J. Phillips, X. An, and J. Dunlop. Demographic effects on estimates of automatic face recognition performance. *Image and Vision Computing*, 30:169–176, 2012.

[14] A. Sgroi, K. W. Bowyer, P. Flynn, and P. J. Phillips. SNoW: understanding the causes of strong, neutral, and weak face impostor pairs. In *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013.

[15] V. Štruc, J. Križaj, and S. Dobrišek. Modest face recognition. In *3rd International Workshop on Biometrics and Forensics (IWBF 2015)*, pages 1–6, March 2015.

Beveridge, J.; Givens, Geof; Phillips, P; Yates, Amy.
"Predicting Face Recognition Performance in Unconstrained Environments."
Paper presented at IEEE Computer Society Workshop on Biometrics 2017, Honolulu, HI, United States. July 21, 2017 - July 21, 2017.

Equivalent Isotropic Response as a Surrogate for Incident Field Strength

Daniel G. Kuester, Duncan A. McGillivray, John Ladbury, Adam Wunderlich,
Ari Feldman, William F. Young, and Sheryl Genco

Communications Technology Laboratory
National Institute of Standards and Technology
Boulder, CO, USA
daniel.kuester@nist.gov, duncan.a.mcgillivray@nist.gov, john.ladbury@nist.gov, adam.wunderlich@nist.gov,
ari.feldman@nist.gov, william.young@nist.gov, sheryl.genco@nist.gov

***Abstract*—The strength of an electromagnetic plane wave incident in the free field can be characterized in terms of power output by an idealized isotropic antenna probe. We refer to the parameter as equivalent isotropic incident power (EIIP), though it lacks an accepted name. This parameter has begun to enter use in various industry standards, technical reports, and peer-reviewed papers. To our knowledge, however, it has not been defined or studied in detail by prior work. We start to address this gap here with a proposed a definition, physical interpretation, and comparison to field strength.**

I. Introduction

Stakeholders in various spectrum sharing scenarios are increasingly asked to specify and test impacts of new systems upon incumbent spectrum users. In these problems, plane waves from multiple radiators impinge upon each receive system with different angles, frequencies, and waveforms. The strength of the wave incident from each radiator upon each receiver must be understood clearly in order to enable direct comparison or combination among simulations, tests, and analytical models. The ideal parameter fits into 1) established terminology, 2) radiated "black box" testing of receiver systems with integrated antennas, and 3) simple, direct application to link budgeting.

A parameter that is an alternative to incident field strength has quietly entered use for this purpose [1]–[7]. The idea is to characterize plane wave strength in terms of the output power response of a hypothetical isotropic probe antenna. It is the complement to EIRP on the receive side of the Friis equation.

We summarize here this "equivalent isotropic" receive parameter, which we call EIIP. We propose an explicit physical and mathematical definition, offer some interpretation of the parameter, and discuss its relationship with the standardized antenna terminology.

II. Definitions

Terminology standardized in [8] includes the well-known effective isotropic radiated power (EIRP) as

$$\text{EIRP} = P_t G_t. \tag{1}$$

Here, G_t is the transmit antenna absolute gain (polarization losses are not included – there is no standardized "partial EIRP"). An interpretation of EIRP is: "the power absorbed by a lossless isotropic antenna that excites far-field plane waves equivalent to the transmit system along a free space path."

The parameter finds use in regulation, system models, and tests for which internal "subsystem" parameters are not known. Like gain, an EIRP value could be specified as a pattern plot, a value at some specified transmit antenna orientation like boresight, or an implied maximum value.

On the other side of the link, the receiving antenna is impinged by an incident electric field with magnitude $|E_r|$, polarized with the transmit antenna. The receive antenna outputs available power P_r, depending on the partial gain of its antenna $G_r e_p$, where G_r is receive antenna absolute gain and e_p is the link polarization efficiency. Consider the following definition of "equivalent isotropic *incident* power" to relate these parameters:

$$\text{EIIP} = \frac{P_r}{G_r e_{pr}} = \frac{|E_r|^2}{\eta_0}\frac{\lambda_0^2}{4\pi}, \tag{2}$$

where wavelength λ_0 and $\eta_0 = \sqrt{\mu_0/\epsilon_0} \approx 377\,\Omega$. The name is meant to emphasize the nature of the parameter — a surrogate for incident field strength and complement to EIRP. A physical interpretation of EIIP is: "the output power available from an isotropic antenna impinged upon by a plane wave with field strength $|E_r|$ and $e_{pr} = 1$." The EIIP does not vary with receive antenna orientation because the reference antenna is defined as isotropic — the factor $1/(e_{pr}G_r)$ cancels the orientation dependence in P_r.

III. EIIP in Link Analysis and Testing

a) Reference Polarization: We decompose the complete Friis link polarization as follows:

$$e_p = |\hat{\rho}_r \cdot \hat{\rho}_t^*|^2 = |\hat{\rho}_r \cdot \hat{\rho}_{\text{ref}}^*|^2 |\hat{\rho}_{\text{ref}} \cdot \hat{\rho}_t^*|^2 = e_{pr} e_{pt}, \tag{3}$$

since dot products commute and $|\hat{\rho}_{\text{ref}} \cdot \hat{\rho}_{\text{ref}}^*| = 1$.

The reference polarization efficiencies for the transmitter and receiver (e_{pt} and e_{pr}) are determined by the corresponding antenna polarization vectors ($\hat{\rho}_t$ and $\hat{\rho}_r$) and some specified

Feldman, Ari; Genco, Sheryl; Kuester, Daniel; Ladbury, John; McGillivray, Duncan; Wunderlich, Adam; Young, William.
"Equivalent Isotropic Response as a Surrogate for Incident Field Strength."
Paper presented at 2017 IEEE International Symposium on Antennas and Propagation, San Diego, CA, United States. July 9, 2017 - July 14, 2017.

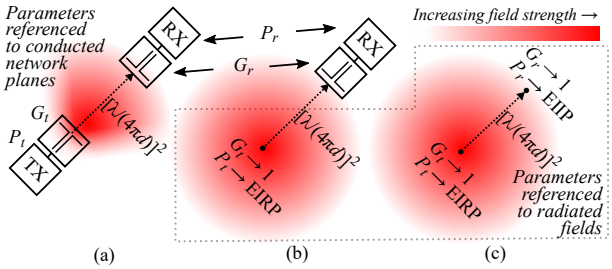


Fig. 1. An actual transmit system (a) and the "equivalent" radiator (b)-(c) excite plane waves with strength that is equal only along the dotted lines. The available output power from the actual receive antenna (a)-(b) is P_r; the output power response of an isotropic antenna probe (c) given an equivalent incident plane wave is EIIP.

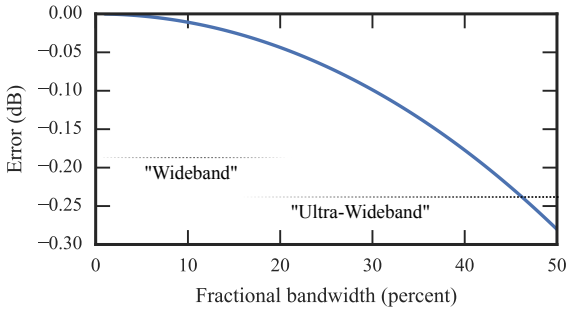


Fig. 2. Error in converting RMS EIIP to mean-squared field strength with the approximation (6) for band-limited white Gaussian noise signals.

reference polarization vector ($\hat{\rho}_{\text{ref}}$). The $\hat{\rho}_{\text{ref}}$ can be chosen arbitrarily to suit an application, but needs to be specified (as with partial gain parameters).

b) "Equivalent Isotropic" Friis Transmission Equation: The EIRP and EIIP definitions in (1) and (2) can substitute directly into the Friis transmission equation, as illustrated in Fig. 1. The relationship between equivalent isotropic parameters in free space is

$$\text{EIIP} = \text{EIRP}\left(\frac{\lambda_0}{4\pi d}\right)^2 e_{pt}, \tag{4}$$

where d is the separation distance between antennas. Equation (4) quantifies radiated field strength excited by the transmit antenna only along the path between the antennas. Like incident field strength, it does not depend on parameters defined inside the receive system (such as P_r, e_{pr}, and G_r), making it a "black box" characterization.

c) Units: Labeling EIRP values with power units is standard practice. The same can apply to EIIP. A potential source of confusion, however, is that these parameters do not correspond with any measurable conducted power. One approach to emphasize this distinction could be to borrow the "i" from the "dBi" of antenna gain: "dBWi" or "dBmi" for EIRP, or "dBW/i" or "dBm/i" for EIIP.

d) Receiving System Response to EIIP: If a characterized receiving system is excited at some known EIIP level, then

$$P_r\,(\text{dBm}) = \text{EIIP}\,(\text{dBm/i}) + G_r\,(\text{dBi}) + e_{pr}\,(\text{dB}) + e_m\,(\text{dB}). \tag{5}$$

This equation is a means to determine received power in wireless link budgets from 1) internal receive parameters G_r, e_{pr}, and matching efficiency e_m, and 2) the incident plane wave strength via EIIP. The P_r result is subject to the usual far-field link estimation constraints and has the expected orientation dependence via G_r and e_{pr}.

e) Modulated Field Approximation: If the incident field is modulated, its power spectral density is distributed across a range of frequencies, not the single tone implied by (2). An approximate relationship in terms of RMS power is

$$\text{EIIP} \approx \frac{\text{E}\left[|E_r(t)|^2\right]}{\eta_0}\frac{\lambda_c^2}{4\pi}, \tag{6}$$

by substitution into (2). Now λ_c is the wavelength at the modulation center frequency, and $|E_r|^2$ is the expected value of $|E_r(t)|^2$ ("mean-squared" field strength). The approximation error in (6) depends on the power spectral density function of the field modulation. For the special case of band-limited white Guassian noise, this error is shown in Fig. 2.

IV. Conclusion

The EIIP parameter has the properties we sought in the introduction. Further, the definition of (2) means that an EIIP can be computed from measurements of either antenna gain and RF power or field strength, producing a derived metrology quantity. Errors caused by under-defined probe response to modulation present issues of definitional uncertainty.

Still, there are caveats to use of EIIP. Analytical conversion between field strength and EIIP involving ultra-wideband modulated waveforms need to be treated carefully. Few commercial field strength probes are specified for use with modern communications waveforms, so the most appropriate class of test equipment to measure these quantities is not clear. The implications of the parameter's use in realistic propagation conditions are also unclear. Application of EIIP in these problem areas could benefit from more research in the future.

References

[1] "User Equipment (UE) Conformance Specification for UE Positioning; Part 1: Conformance Test Specification," 3GPP, Tech. Rep. TS 37.571-1, 2016.

[2] "Measurements of radiated performance for MIMO and multi-antenna reception for HSPA and LTE terminals," 3GPP, Tech. Rep. TR 37.976 v11.0.0, 2012. [Online]. Available: http://www.3gpp.org/DynaReport/37976.htm

[3] P. V. Nikitin and K. V. S. Rao, "Antennas and Propagation in UHF RFID Systems," in *2008 IEEE Int. Conf. RFID*, 2008, pp. 277–288.

[4] S. Saunders and A. Aragón-Zavala, *Antennas and Propagation for Wireless Communication Systems*, 2nd ed. John Wiley & Sons, 2007.

[5] "High Rate 60 GHz PHY, MAC and HDMI PAL," Ecma International, Geneva, CH, Tech. Rep. ECMA Std. 387, dec 2008.

[6] A. Saakian, *Radio Wave Propagation Fundamentals*. Artech House, 2011.

[7] "Standard for Air Interface for Broadband Wireless Access Systems," IEEE, Tech. Rep. Std. 802.16n-2013, 2013.

[8] "IEEE Standard for Definitions of Terms for Antennas," IEEE Standards Association, Tech. Rep. Std. 145-2013, 2013.

Feldman, Ari; Genco, Sheryl; Kuester, Daniel; Ladbury, John; McGillivray, Duncan; Wunderlich, Adam; Young, William.
"Equivalent Isotropic Response as a Surrogate for Incident Field Strength."
Paper presented at 2017 IEEE International Symposium on Antennas and Propagation, San Diego, CA, United States. July 9, 2017 - July 14, 2017.

Coexistence Analysis of LTE and WLAN Systems With Heterogenous Backoff Slot Durations

Yao Ma, Daniel G. Kuester, Jason Coder, and William Young
Communications Technology Laboratory, National Institute of Standards and Technology
325 Broadway, Boulder, Colorado, USA

***Abstract*— To enable constructive coexistence with wireless local area networks (WLANs), unlicensed long-term evolution (LTE) systems use listen before talk (LBT) as a major candidate technique. The LBT has a flexible backoff idle slot duration, which can be significantly larger than the WLAN counterpart. To our knowledge, however, available analytical results on the LTE and WLAN coexistence have considered only identical idle backoff slot durations. There is a formidable technical difficulty to coexistence analysis for different backoff slot durations. In this paper, we develop a new technical approach to address this open issue. First, we point out an LBT backoff slot jamming effect, and propose a modified LBT backoff scheme to address this problem. Second, for our proposed LBT scheme, we develop a new analytical framework to address system interactions with non-equal backoff slot durations, model the LTE backoff process as super-counters, and provide a thorough analysis on the throughput, backoff counter hold time, and successful transmission probabilities of LTE-LBT and WLAN systems. Finally, we program the algorithms and use computer simulation to validate the analysis. This result fills a major gap and provides practical value for LTE-LBT and WLAN coexistence performance analysis with heterogeneous sensing and backoff slot durations.**

Index Terms: LTE; WLAN; Wireless System Coexistence; CSMA/CA; MAC-layer Performance Analysis.

I. Introduction

With the congestion and scarcity of available spectrum resources, spectrum sharing between long-term evolution license assisted access (LTE-LAA) and the IEEE 802.11 wireless local area network (WLAN) systems is a major ongoing research topic [1]–[6]. The 3rd Generation Partnership Project (3GPP) proposes to use listen before talk (LBT) to enable constructive coexistence between LAA and WLAN systems. The 3GPP LAA has defined 4 categories of LBT schemes [4], [5]. Category 3 and 4 LBT are system-load based sensing schemes, and have attracted significant interest. Various coexistence settings based on LTE-LAA and WLAN transmissions have been intensively evaluated, and experimental and field test results are reported in [4]–[6]. The WLAN uses carrier sense multiple access with collision avoidance (CSMA/CA) in the medium access control (MAC) layer, and load-based LBT uses a similar CSMA/CA method. However, due to the sensing reliability and other system requirement, the sensing (backoff slot) duration in the LBT Category 3 can be significantly larger than its counterpart in the WLAN [4].

Recently, some analytical approaches for the evaluation of LTE-LAA and WLAN coexistence systems have been developed, see e.g., [8]–[10]. Furthermore, optimization methods of the LAA and WLAN coexistence systems have been studied under various fairness constraints in [11]–[13].

However, to our knowledge, available analytical results are only valid when the backoff slot durations in different systems (such as the LAA and WLAN) are identical. The WLAN backoff (or idle/empty) slot duration includes clear channel assessment (CCA) time, and LAA backoff idle slot duration is equal to extended CCA (eCCA) time. In the current 3GPP development documents [4], [5], the LAA eCCA slot duration may be 20 μs or even larger, while the WLAN backoff slot duration (which includes CCA time) is 9 μs for several popular physical layer specifications [7]. Sensing performance of the LBT is closely related to eCCA sensing duration – a larger eCCA duration (aka. backoff duration) causes a better signal to noise ratio (SNR) for signal detection, but a slower backoff process, and vice versa. Robust and reliable detection of WLAN signals at LTE nodes, especially in multiparty fading channels, requires a reasonably large channel sensing duration (such as during eCCA). The channel sensing (and backoff slot) durations in different CSMA/CA-based systems are typically not identical, such as IEEE 802.15.4, IEEE 802.11, and LTE-LAA systems. Hence, analyzing the case of heterogeneous backoff slot durations will have important theoretical and practical value, useful for future coexistence applications of heterogeneous systems.

Available methods face formidable challenges to address the case of non-equal idle backoff slot durations. The Bianchi-proposed Markov chain method is a popular approach for CSMA/CA MAC-layer performance analysis [14], [15], and has been extended in [8]–[10] for coexistence analysis. However, this method is not flexible enough to model very complex coexisting behaviors in both the backoff phase and transmission phase, experienced in non-identical slot durations between coexistence systems. Recently, another method on WLAN MAC-layer performance analysis is provided in [16]–[18]. This method is more flexible than Bianchi's framework in that it explicitly models the backoff counter hold time, and uses a different set of statistics to compute the MAC-layer throughput. However, this method is based on assumption of identical backoff idle slot durations among all transmitting nodes.

Coexistence analysis between IEEE 802.15.4 and IEEE 802.11 WLAN systems has recently been implemented in [19], where the 802.15.4 devices are assumed to have a backoff slot duration three times as large as the counterpart WLAN nodes. However, besides the differences in the MAC protocols

Coder, Jason; Kuester, Daniel; Ma, Yao; Young, William.
"Coexistence Analysis of LTE and WLAN Systems With Heterogenous Backoff Slot Durations."
Paper presented at IEEE International Conference on Communications (ICC) 2017, Paris, France. May 21, 2017 - May 25, 2017.

between LTE-LAA and the 802.15.4, the 802.15.4 device does not have the backoff slot frozen effect as the LAA node. Thus, the problem at hand is more difficult to solve.

In this paper, we model and solve this challenging problem. The contributions are highlighted as follows:

- We show that with heterogeneous backoff durations between LAA and WLAN systems, there is a previously-unknown backoff slot jamming effect to LAA nodes. We then propose a MAC scheme to avoid this negative effect.
- We develop a novel analysis tool to model the non-identical backoff slots, such as LTE super counters and weighted probability transition paths, to model interaction between LTE and WLAN nodes. Then we provide analytical results on the counter hold time, successful transmission probability, and throughput.
- We program the algorithms and implement extensive simulation to validate our analytical results on the co-existence performance.

This new technique fills a major gap in coexistence analysis of LTE-LAA and WLAN systems, and can be extended to the analysis of other CSMA/CA based heterogeneous wireless systems. The technical insight and method provided by this work may be used for optimization of coexisting systems.

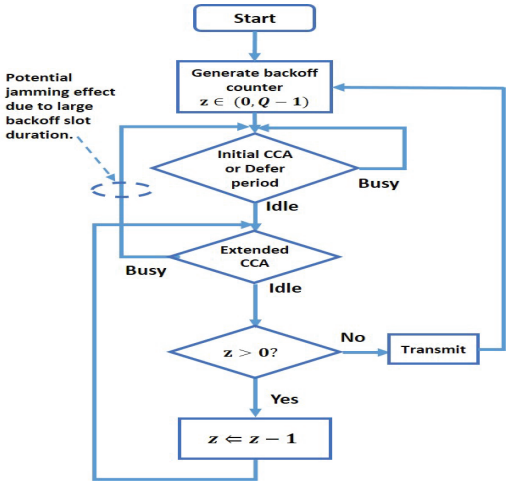


Fig. 1: Flow diagram of LTE downlink LAA LBT Category-3 procedure, adopted from [4], [5] with major revision. We mark the backoff slot jamming effect assuming that the LAA has backoff slot duration substantially larger than that of the WLAN system.

II. System Model

Here, we consider the case that LTE LAA utilizes only unlicensed spectrum for the downlink and shares it with incumbent WLAN users. The processing flow of LAA Category 3 LBT scheme is shown in Fig. 1, adopted from [4], [5]. In comparison with [4], [5], we switched the order of the blocks "$z > 0$" and "extended CCA". This revision lets the transmitter which finishes one transmission to wait for an eCCA period, in addition to initial CCA (or extended defer period), before a backoff counter reduction. This change is significant in that it makes sure that after a channel busy period, the active transmitter which finishes its transmission opportunity (TXOP) does not have more priority in next channel access than the competing stations.

Define $N_s = \delta_L/\delta_W$, where δ_L and δ_W are the backoff idle slot durations for LAA and WLAN, respectively. To facilitate smooth coexistence, we assume $T_{\text{DIFS}} = T_{\text{Defer}}$, where T_{DIFS} and T_{Defer} are WLAN distributed coordination function interframe spacing (DIFS) and LAA eCCA defer durations, respectively. In the LAA backoff counter reduction scheme, shown in Fig. 1 (and those in [4], [5]), by default, an LAA counter reduction is permitted in either of the two cases: 1) when the channel becomes idle for $T_{\text{DIFS}} + \delta_L$ right after channel busy state; 2) the channel becomes idle for δ_L right after previous counter reduction.

We point out that when $N_s > 1$, this LBT scheme can cause a slot jamming effect disadvantageous to the LAA station, not investigated in the available literature.

This slot-jamming effect is illustrated in Fig. 2 for the row "LAA states (default)" in the eCCA duration, assuming $N_s = 2$. In detail, an LAA counter reduction takes a longer duration ($N_s\delta_W$) than a WLAN counter reduction (δ_W), and before it reaches a slot boundary, a WLAN counter may first reduce to zero and begin transmission. After the channel busy state is over, the LAA node has to reset the counter value to the state before the WLAN transmission: that is, the reduction can be jammed if there are frequent WLAN transmissions (when $N_s > 1$); please refer to WLAN slots 4-6 in Fig. 2. Though the jamming does not happen in these slots, it can happen if any WLAN node reduces its counter to 0 from slot 5 to 6. Based on this observation, the jamming effect is due to that a WLAN node always has higher counter reduction opportunity in both cases 1 and 2 discussed above.

To address this problem, we propose a modified LAA Counter Reduction Scheme, shown next.

Proposed LAA Counter Reduction Scheme

1) Draw counter value $Z \in (0, Z_0 - 1)$, where Z_0 is the LAA initial contention window (CW) size. Wait until the channel is idle for initial CCA (iCCA) duration. If $Z = 0$, the LAA node transmits; otherwise, it goes to backoff stage.
2) Decrease counter Z by 1 in either of the following two channel idle cases:
   Case 1: Right after a channel busy state, if channel becomes idle for $T_{\text{DIFS}} + \delta_W$ (use $\delta_L = \delta_W$);
   Case 2: After the previous counter reduction, channel is idle again for $\delta_L = N_s\delta_W$.
3) If Z is reduced to zero, starts transmission. Restart from Step 1).

In our proposed LBT MAC scheme, in Case 1, LAA and WLAN nodes have equal priority in reducing their counter values. After an LAA counter reduction, if the idle period continues, then we still set $\delta_L = N_s\delta_W$, which enables an adequate slot period for channel sensing. The state transition and counter reduction for the proposed scheme is given by the row "LAA states (our proposed)" in Fig. 2. During WLAN

Coder, Jason; Kuester, Daniel; Ma, Yao; Young, William.
"Coexistence Analysis of LTE and WLAN Systems With Heterogenous Backoff Slot Durations."
Paper presented at IEEE International Conference on Communications (ICC) 2017, Paris, France. May 21, 2017 - May 25, 2017.

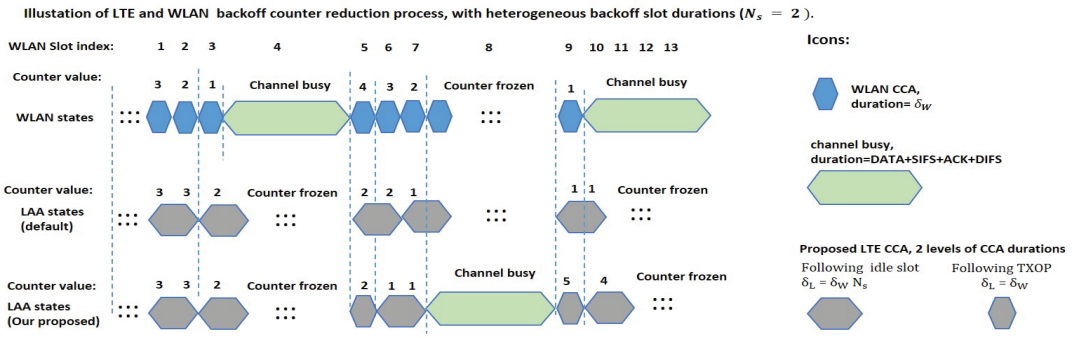


Fig. 2: Flow diagram of LTE and WLAN backoff counter reduction and transmission process, when the LAA has backoff slot duration twice as large as that of the WLAN system.

slot 5 to 6, the LAA idle slot is reduced from $N_s\delta_W$ to δ_W, providing equal counter reduction opportunity for all LAA and WLAN nodes after a channel busy state is over. In WLAN slot indexes 5 and 9 of Fig. 2, after the WLAN and LAA transmissions (channel busy), their counter values (4 and 5 respectively) were randomly generated based on their initial CW sizes. Furthermore, ACK and SIFS refer to acknowledgement signal duration and short interframe spacing, respectively.

Our scheme has two advantages: 1) It mostly avoids the slot jamming effect; 2) It causes only negligible impact on channel sensing accuracy of channel idle state in case 1, because although the total idle duration used for channel detection is reduced to $T_{\text{DIFS}} + \delta_W$ from $T_{\text{DIFS}} + N_s\delta_W$ (case 1), it is typically larger than $N_s\delta_W$ (case 2).

III. Performance Analysis

We developed a new Markov chain method to model the LAA CW countdown process, and its interactions with WLAN transmissions. The model is shown in Fig. 3. The basic backoff-and-transmission state transition model is shown in Fig. 3(a), and its equivalent expanded model for $N_s > 1$ is described in Fig. 3.(b). Based on the LTE-LAA Markov model in Fig. 3, a performance analysis of LTE and WLAN coexistence is provided next. In this section, we use subscripts L, W, i, S, C, p to denote LAA, WLAN, idle, successful transmission, collision, and payload, respectively. The MAC throughput of an LAA node and a WLAN node are, respectively, given by

$$S_L = \pi_{S,L} T_{p,L} / T_{\text{ave},L} \quad (1)$$

$$S_W = \pi_{S,W} T_{p,W} / T_{\text{ave},W}, \quad (2)$$

where $T_{p,L}$ and $T_{p,W}$ are payload durations, $\pi_{S,L}$ and $\pi_{S,W}$ are the probabilities of successful transmissions, and $T_{\text{ave},L}$ and $T_{\text{ave},W}$ are the average total durations caused by one successful transmission, in LAA and WLAN systems, respectively. Define $\pi_{F,L}$ and $\pi_{R,L}$ as probabilities for failed transmission and backoff stage, respectively. Based on the model in Fig. 3.(a), we have $\pi_{S,L} = 0.5P_{t,L}$, $\pi_{F,L} = 0.5(1 - P_{t,L})$, and $\pi_{R,L} = 0.5$, where $P_{t,L}$ is probability of successful transmission conditioned on that an LAA transmission starts.

Suppose that a WLAN node has cutoff stage M, with maximum CW size W_m at stage m, for $m = 0, 1, \ldots, M$. For a WLAN node, define $\pi_{F,W,m}$ and $\pi_{R,W,m}$ as probabilities for failed transmission and backoff, at stage m, respectively. Below, we use a method similar to that in [16], with one major difference that once the transmission at cut-off stage fails, the counter is reset to the initial stage $(m = 0)$ immediately. By using the equality

$$\pi_{S,W} + \sum_{m=0}^{M} [\pi_{R,W,m} + \pi_{F,W,m}] = 1, \quad (3)$$

we can solve for the state probabilities as: $\pi_{S,W} = P_{t,W}/2$,

$$\pi_{R,W,0} = \frac{0.5P_{t,W}}{1 - (1 - P_{t,W})^{M+1}} \quad (4)$$

$$\pi_{R,W,m} = \pi_{R,W,0}(1 - P_{t,W})^m \quad (5)$$

$$\pi_{F,W,m} = \pi_{R,W,0}(1 - P_{t,W})^{m+1} \quad (6)$$

for $m = 0, \ldots, M$, where $P_{t,W}$ is the successful transmission probability given that a WLAN node transmission starts. Also,

$$T_{\text{ave},L} = \pi_{S,L} T_{S,L} + \pi_{F,L} T_{C,L} + 0.5 T_{R,L}$$

$$T_{\text{ave},W} = \pi_{S,W} T_{S,W} + \sum_{m=0}^{M} [\pi_{F,W,m} T_{C,W} + \pi_{R,W,m} T_{R,W,m}],$$

where the $T_{S,L}$ ($T_{S,W}$) and $T_{C,L}$ (and $T_{C,W}$) are the channel busy durations due to successful transmission and collision, for LAA (and WLAN), respectively. The $T_{R,L}$ is the LTE counter hold time per transmission, and $T_{R,W,m}$ (and $P_{R,W,m}$) is the WLAN counter hold time (and probability) in backoff stage m, $m = 0, \ldots, M$.

Conditioned on a counter reduction, the transmission probability for LAA Category-3 node is derived as

$$\tau_L = 2/(1 + Z_0), \quad (7)$$

where Z_0 is the initial CW size of the LAA node. Define $\tilde{\pi}_{R_m} = \pi_{R,W,m} T_{R,W,m} / T_{\text{ave},W}$ as the normalized duration in backoff stage m. The transmission probability for a WLAN

Coder, Jason; Kuester, Daniel; Ma, Yao; Young, William.
"Coexistence Analysis of LTE and WLAN Systems With Heterogenous Backoff Slot Durations."
Paper presented at IEEE International Conference on Communications (ICC) 2017, Paris, France. May 21, 2017 - May 25, 2017.

node is obtained as

$$\tau_W = 1 - \frac{\sum_{m=0}^{M} \tilde{\pi}_{R_m}(1 - 2/(1+W_m))}{\sum_{m=0}^{M} \tilde{\pi}_{R_m}}. \quad (8)$$

It easily follows that

$$\begin{aligned} T_{R,L} &= \frac{Z_0 - 1}{2} T_{L,0} \\ T_{R,W,m} &= \frac{W_m - 1}{2} T_{W,0}, \end{aligned}$$

where $T_{L,0}$ and $T_{W,0}$ are the hold-time per counter reduction at LAA and WLAN nodes, respectively.

To compute MAC-layer throughput, we still need to find $T_{L,0}$, $P_{t,L}$ for the LAA, and $T_{W,0}$ and $P_{t,W}$ for the WLAN. Refer to Figs. 3 and 4: Even for the case of equal LTE and WLAN slot duration ($N_s = 1$), this model is different from those of available approaches [8]–[10], [14], [16]–[18]. For the case of $N_s > 1$, the difference is more significant.

To illustrate our method, we show the case of $N_s = 1$ first, and then we develop more details for the case of $N_s > 1$.

A. Equal Slot Duration ($N_s = 1$)

When $N_s = 1$, it follows that

$$\begin{aligned} P_{t,L} &= (1-\tau_W)^{n_W}(1-\tau_L)^{n_L - 1} \\ P_{t,W} &= (1-\tau_W)^{n_W - 1}(1-\tau_L)^{n_L}, \end{aligned}$$

where τ_W and τ_L are the transmitting (channel access) probabilities of WLAN and LAA systems, given by (8) and (7), respectively. Let P and $\hat{P}$ denote probabilities observed by a node when observing its own system (e.g., state of LAA system observed by an LAA node), and the other system (e.g., state of LAA system observed by a WLAN node), respectively. For example $P_{i,L} = (1-\tau_L)^{n_L - 1}$, $P_{i,W} = (1-\tau_W)^{n_W - 1}$, but $\hat{P}_{i,L} = (1-\tau_L)^{n_L}$, and $\hat{P}_{i,W} = (1-\tau_W)^{n_W}$.

Refer to Fig. 4: The feedforward path of $1 - \hat{P}_{i,W}P_{i,L}$ consists of 5 sub-events: LAA successful transmission (with probability $P_{S,L}$), LAA intra-system signal collision ($P_{C,L}$), WLAN successful transmission ($\hat{P}_{S,W}$), WLAN intra-system signal collision ($\hat{P}_{C,W}$), and LAA-WLAN inter-system signal collision (with probability $(1-\hat{P}_{i,W})(1-P_{i,L})$). When $N_s = 1$, we obtain an average counter hold time (per counter reduction) for an LAA node as

$$\begin{aligned} T_{L,0} &= P_{i,L}\hat{P}_{i,W}\delta_W + (P_{S,L}T_{S,L} + P_{C,L}T_{C,L})\hat{P}_{i,W} \\ &+ (\hat{P}_{S,W}T_{S,W} + \hat{P}_{C,W}T_{C,W})P_{i,L} \\ &+ (1-\hat{P}_{i,W})(1-P_{i,L})T_{C,M}, \end{aligned} \quad (9)$$

where $T_{C,M} = \max(T_{C,W}, T_{C,L})$, $\hat{P}_{S,W} = n_W\tau_W(1-\tau_W)^{n_W-1}$, $\hat{P}_{C,W} = 1 - \hat{P}_{i,W} - \hat{P}_{S,W}$,

$$P_{S,L} = \begin{cases} (n_L - 1)\tau_L(1-\tau_L)^{n_L - 2}, & \text{when } n_L \geq 2; \\ 0, & \text{when } n_L \leq 1, \end{cases}$$

and $P_{C,L} = 1 - P_{i,L} - P_{S,L}$.

Similarly, we have the average counter hold time (per counter reduction) for a WLAN node as

$$\begin{aligned} T_{W,0} &= \hat{P}_{i,L}P_{i,W}\delta_W + (\hat{P}_{S,L}T_{S,L} + \hat{P}_{C,L}T_{C,L})P_{i,W} \\ &+ (P_{S,W}T_{S,W} + P_{C,W}T_{C,W})\hat{P}_{i,L} \\ &+ (1-P_{i,W})(1-\hat{P}_{i,L})T_{C,M}, \end{aligned} \quad (10)$$

where $\hat{P}_{S,L} = n_L\tau_L(1-\tau_L)^{n_L-1}$, $\hat{P}_{C,L} = 1 - \hat{P}_{i,L} - \hat{P}_{S,L}$,

$$P_{S,W} = \begin{cases} (n_W - 1)\tau_W(1-\tau_W)^{n_W-2}, & \text{when } n_W \geq 2; \\ 0, & \text{when } n_W \leq 1, \end{cases}$$

and $P_{C,W} = 1 - P_{i,W} - P_{S,W}$.

Based on the above results, the throughput of the LAA and WLAN nodes in the coexistence case with $N_s = 1$ can be readily evaluated.

B. Non-Equal Slot Durations ($N_s > 1$)

We need to consider two cases for the LAA backoff counter reduction:

1) channel is idle for $T_{\text{DIFS}} + \delta_W$ following a transmission (channel busy); and
2) channel is idle for $\delta_L = N_s\delta_W$ right after a previous counter reduction.

We model the transition paths between the two cases during an LAA counter reduction in Fig. 5. Define the probabilities of cases 1 and 2 as $\Pr(C_1)$ and $\Pr(C_2)$, and the transition probability from case n_1 to case n_2 as $\Pr(C_{n_2}|C_{n_1})$, for $n_1, n_2 \in (1, 2)$. For example, $\Pr(C_1|C_1)$ is the sum of all the probability paths from Case 1 (on the right side in Fig. 5) to Case 1 (on the left side), and $\Pr(C_1|C_1) = 1 - \hat{P}_{i,W}P_{i,L}$.

From Fig. 5, it follows that

$$\begin{aligned} \Pr(C_1) &= \Pr(C_1)(1 - P_{i,L}\hat{P}_{i,W}) + [\Pr(C_1) + \Pr(C_2)] \\ &\quad \cdot P_{i,L}\hat{P}_{i,W}[1 - P_{i,L}\hat{P}_{i,W}^{N_s}] \quad (11) \\ \Pr(C_2) &= [\Pr(C_1) + \Pr(C_2)]P_{i,L}\hat{P}_{i,W}\hat{P}_{i,W}^{N_s - 1}. \quad (12) \end{aligned}$$

We can verify that equations (11) and (12) are equivalent, as expected. To determine $\Pr(C_1)$ and $\Pr(C_2)$, we need one more equality. The sum probability of all the counter states within one counter reduction in Fig. 5 equals unity. Thus,

$$\begin{aligned} &\Pr(C_1) + [\Pr(C_1) + \Pr(C_2)]P_{i,L}\hat{P}_{i,W} \\ &\quad \cdot(1 + \hat{P}_{i,W} + \ldots + \hat{P}_{i,W}^{N_s - 1}) = 1. \quad (13) \end{aligned}$$

Based on (12) and (13), we derive:

$$\begin{aligned} \Pr(C_2) &= \left(\frac{1 - P_{i,L}\hat{P}_{i,W}^{N_s}}{P_{i,L}\hat{P}_{i,W}^{N_s}} + \frac{1 - \hat{P}_{i,W}^{N_s}}{\hat{P}_{i,W}^{N_s - 1} - \hat{P}_{i,W}^{N_s}} \right)^{-1} \\ \Pr(C_1) &= \Pr(C_2)\frac{1 - P_{i,L}\hat{P}_{i,W}^{N_s}}{P_{i,L}\hat{P}_{i,W}^{N_s}}. \end{aligned}$$

When $N_s = 1$, (11) and (12) reduce to

$$\begin{aligned} \Pr(C_1|N_s = 1) &= (1 - P_{i,L}\hat{P}_{i,W}) \quad (14) \\ \Pr(C_2|N_s = 1) &= P_{i,L}\hat{P}_{i,W}, \quad (15) \end{aligned}$$

as expected. This means that when $N_s = 1$, Case 1 corresponds to a channel busy event, which is always followed by DIFS and idle slot δ_W, and Case 2 corresponds to a channel idle event, where all LAA and WLAN nodes stay idle.

Successful transmission probabilities

Define $\Pr(\text{WTx})$ as the probability that only the WLAN node has transmit opportunity, and $\Pr(\text{JTx})$ as the probability that all LTE and WLAN nodes have transmit opportunity, respectively, from WLAN's observation. To compute $P_{t,W}$, refer to Fig. 5 again. $\Pr(\text{WTx})$ is the sum probability the

Coder, Jason; Kuester, Daniel; Ma, Yao; Young, William.
"Coexistence Analysis of LTE and WLAN Systems With Heterogenous Backoff Slot Durations."
Paper presented at IEEE International Conference on Communications (ICC) 2017, Paris, France. May 21, 2017 - May 25, 2017.

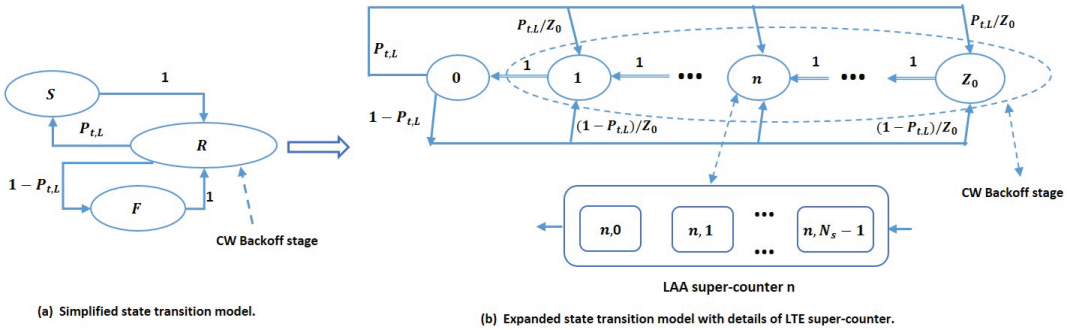


Fig. 3: Our proposed Markov model for the LTE-LAA LBT category 3 procedure in coexistence with WLAN.

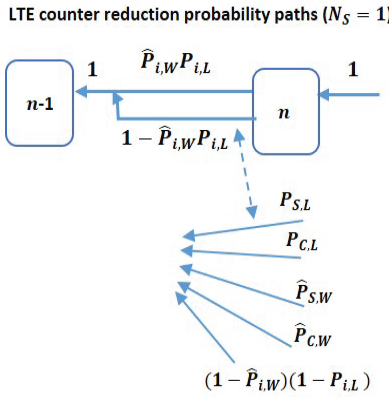


Fig. 4: Illustration of Markov model for the LAA counter reduction when $N_s = 1$.

$N_s - 1$ subcells in the right side of the super-counter. When $N_s \geq 2$, we have $\Pr(\text{JTx}) = 1 - \Pr(\text{WTx})$, and

$$\begin{aligned}\Pr(\text{WTx}) &= [\Pr(\tilde{C}_1) + \Pr(\tilde{C}_2)]\hat{P}_{i,L}P_{i,W} \\ &\quad \cdot(1 + P_{i,W} + \ldots + P_{i,W}^{N_s-2}) \end{aligned} \quad (16)$$

where $\Pr(\tilde{C}_1)$ and $\Pr(\tilde{C}_2)$ are obtained from $\Pr(C_1)$ and $\Pr(C_2)$ by replacing $P_{i,L}$ and $\hat{P}_{i,W}$ with $\hat{P}_{i,L}$ and $P_{i,W}$ therein, respectively. With probability $\Pr(\text{WTx})$, all LAA nodes stay silent. Thus, the successful transmission probability of a WLAN node ($P_{t,W}$) is given by

$$\begin{aligned}P_{t,W} &= \Pr(\text{WTx})(1-\tau_W)^{n_W-1} \\ &+ \Pr(\text{JTx})(1-\tau_W)^{n_W-1}(1-\tau_L)^{n_L}. \end{aligned} \quad (17)$$

We define successful transmission probability of an LAA node based on each counter reduction (which happens in Cases 1 and 2), then

$$P_{t,L} = (1-\tau_L)^{n_L-1}(1-\tau_W)^{n_W}, \quad (18)$$

which is independent of N_s. This is because each LAA node can transmit only upon the two channel idle cases.

Average counter hold time for LAA and WLAN nodes

TABLE I: Probability and duration pairs to compute LAA counter hold time.

Probability ($P_{n,L}$)	Duration ($T_{n,L}$)
$\Pr(C_1)(1 - P_{i,L}\hat{P}_{i,W})$	$\overline{T}_{L,W}$
$\Pr(C_1, C_2)(1 - \hat{P}_{i,W})$	$\overline{T}_W$
$\Pr(C_1, C_2)(1 - \hat{P}_{i,W})\hat{P}_{i,W}$	$\overline{T}_W + \delta_W$
$\ldots$	$\ldots$
$\Pr(C_1, C_2)(1 - \hat{P}_{i,W})\hat{P}_{i,W}^{N_s-2}$	$\overline{T}_W + (N_s - 2)\delta_W$
$\Pr(C_1, C_2)(1 - \hat{P}_{i,W}P_{i,L})\hat{P}_{i,W}^{N_s-1}$	$\overline{T}_{L,W} + (N_s - 1)\delta_W$
$\Pr(C_1, C_2)P_{i,L}\hat{P}_{i,W}^{N_s}$	$N_s\delta_W$

Refer to Fig. 5 again. The average hold time for an LAA node $T_{L,0}$ is obtained by summing the duration of each path from the start states to the end states, weighted by the path probability. The probability and duration pairs of each path is listed in Table I. In Table I, $\overline{T}_W$ is the average channel busy duration when any one or more WLAN nodes transmit, and $\overline{T}_{L,W}$ is the average channel busy duration when any one or more of the LAA and WLAN nodes transmit. They are given by

$$\begin{aligned}\overline{T}_W &= \frac{1}{(1-\hat{P}_{i,W})}[\hat{P}_{S,W}T_{S,W} + \hat{P}_{C,W}T_{C,W}] \\ \overline{T}_{L,W} &= \frac{1}{(1-\hat{P}_{i,W}P_{i,L})}[(\hat{P}_{S,W}T_{S,W} + \hat{P}_{C,W}T_{C,W})P_{i,L} \\ &+ (P_{S,L}T_{S,L} + P_{C,L}T_{C,L})\hat{P}_{i,W} \\ &+ (1-\hat{P}_{i,W})(1-P_{i,L})T_{C,M}]. \end{aligned}$$

In Table I, the first item is for the direct path through the regular counter on the top side, from Case 1 to Case 1 which is a channel busy event. The 2nd to $(N_s + 1)$th terms are for the paths through the super-counter on the bottom side from both Cases 1 and 2 to Case 1, which are channel busy events. The final term ($(N_s + 2)$th term) is for the 0th subcell in the super-counter, from Case 2 to Case 2. In Table I,

$$\Pr(C_1, C_2) = [\Pr(C_1) + \Pr(C_2)]\hat{P}_{i,W}P_{i,L}, \quad (19)$$

which corresponds to the path from Cases 1 and 2 in slot $n+1$ to the super-counter in slot n. Finally, $T_{L,0}$ can be computed

Coder, Jason; Kuester, Daniel; Ma, Yao; Young, William.
"Coexistence Analysis of LTE and WLAN Systems With Heterogenous Backoff Slot Durations."
Paper presented at IEEE International Conference on Communications (ICC) 2017, Paris, France. May 21, 2017 - May 25, 2017.

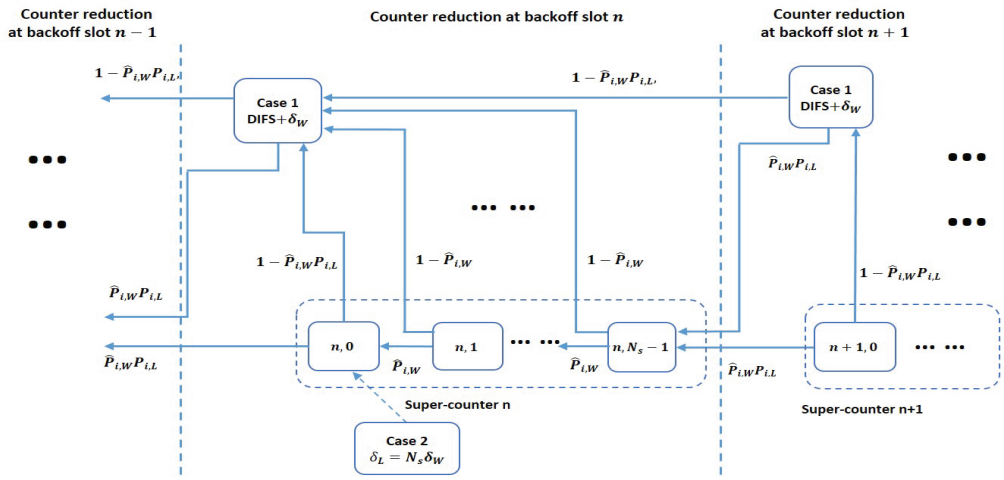


Fig. 5: Flow diagram for LAA counter backoff probability paths.

by summing up all the probability-weighted durations in Table I, that is

$$T_{L,0} = \frac{1}{\Pr(C_1)+\Pr(C_2)} \sum_{n=1}^{N_s+2} P_{n,L} T_{n,L}, \quad (20)$$

where the normalization by factor $\Pr(C_1)+\Pr(C_2)$ is used, because an LAA node transmits only upon the two idle cases with sum probability $\Pr(C_1)+\Pr(C_2)$.

When $\overline{T}_W \gg \delta_W$ and $\overline{T}_{L,W} \gg \delta_W$, we obtain an approximate formula for $T_{L,0}$:

$$\begin{aligned} T_{L,0} &\simeq \frac{\Pr(C_1)(1-P_{i,L}\hat{P}_{i,W})}{\Pr(C_1)+\Pr(C_2)} \overline{T}_{L,W} + \hat{P}_{i,W} P_{i,L} \\ &\times [(1-\hat{P}_{i,W}^{N_s-1})\overline{T}_W + \hat{P}_{i,W}^{N_s-1}(1-P_{i,L}\hat{P}_{i,W})\overline{T}_{L,W}]. \end{aligned}$$

By use of the concept of joint transmission and WLAN-only transmission, average hold time $T_{W,0}$ for a WLAN node is derived as

$$T_{W,0} \simeq \Pr(\text{WTx})\tilde{T}_W + \Pr(\text{JTx})\tilde{T}_{W,L}, \quad (21)$$

where

$$\begin{aligned} \tilde{T}_W &= P_{S,W}T_{S,W} + P_{C,W}T_{C,W} + P_{i,W}\delta_W \\ \tilde{T}_{W,L} &= P_{i,W}\hat{P}_{i,L}\delta_W + (P_{S,W}T_{S,W} + P_{C,W}T_{C,W})\hat{P}_{i,L} \\ &+ (\hat{P}_{S,L}T_{S,L} + \hat{P}_{C,L}T_{C,L})P_{i,W} \\ &+ (1-P_{i,W})(1-\hat{P}_{i,L})T_{C,M}. \end{aligned}$$

Based on results above, the coexistence performance of LAA and WLAN with non-equal idle slot durations can be readily evaluated.

IV. Numerical Results

In this section, we provide both analytical and simulation results of the coexistence behavior of LTE-LAA links with WLAN links. The proposed LBT backoff counter reduction scheme is used. The simulation results were obtained by running for 10^5 time slots on each parameter setting. The parameters used for analysis and simulation are listed in Table II, where the WLAN parameters were adopted from [6], [9], [15], with basic access scheme. We assume that the WLAN and LAA systems have channels fully overlapped at the 5 GHz industrial, scientific, and medical (ISM) band. When the transmission time efficiency is 100%, the upper bound for the physical layer channel bit rate (CBR) is set to 100 Mega bits per second (Mbps) for both the LAA and WLAN systems.

TABLE II: LTE and WLAN Parameters in Simulation.

LTE parameters

Parameter	Value
Payload duration per transmission	2 ms
$T_{L,\text{SIFS}}$	16 μs
LBT defer period: T_{Defer} (=T_{DIFS})	34 μs
LBT eCCA period: T_{eCCA} (=$N_s\delta_W$)	$N_s \times 9$ μs
Initial CW size Z_0	8

WLAN parameters

Parameter	Value
Payload duration per transmission	1 ms
MAC and PHY headers	272 and 128 bits
T_{SIFS}	16 μs
T_{DIFS}	34 μs
Idle slot duration δ_W	9 μs
Initial CW size W_0	16

We show the average counter hold durations for the LAA and WLAN systems in Fig. 6, and the throughput results in Fig. 7, respectively, assuming $N_s = 3$, and $n_L + n_W$ changes from 4 to 28. We observe that the analytical and simulation results match very well. From Fig. 6, as total number of links $n_W + n_L$ increases, the gap between counter hold durations among LAA and WLAN nodes decreases, and this corresponds

Coder, Jason; Kuester, Daniel; Ma, Yao; Young, William.
"Coexistence Analysis of LTE and WLAN Systems With Heterogenous Backoff Slot Durations."
Paper presented at IEEE International Conference on Communications (ICC) 2017, Paris, France. May 21, 2017 - May 25, 2017.

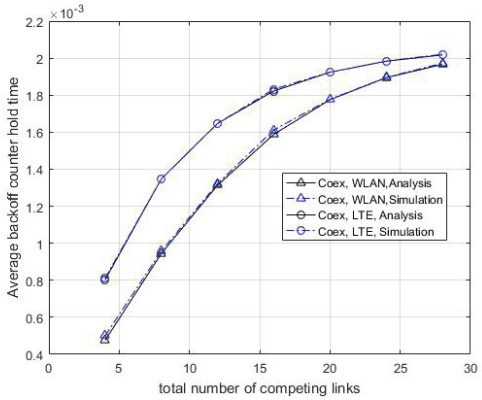


Fig. 6: Counter hold times of LTE and WLAN systems, when $N_s = 3$, $M = 3$, and $n_L + n_W$ changes from 4 to 28.

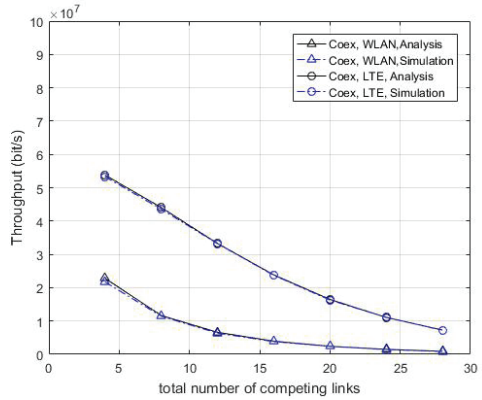


Fig. 7: Throughput of LTE and WLAN systems, when $N_s = 3$, $M = 3$, and $n_L + n_W$ changes from 4 to 28.

to a better access fairness for LAA nodes based on our proposed LBT scheme. For results not shown here, when the original LBT is used, this gap is significantly larger and LAA nodes experience a backoff jamming effect. The related analysis and simulation result is not shown here due to space limitation. From Fig. 7, we observe that throughput of LAA and WLAN systems decrease with number of links. The LAA system has larger throughput because it uses half the CW size ($Z_0 = W_0/2$), although its idle slot duration is 3 times that of the WLAN system.

V. Conclusion

In this paper, we have studied the impact of heterogeneous backoff slot durations on the MAC-layer performance of LTE-LAA coexisting with WLAN transmissions. We first pointed out a slot-jamming effect due to difference in backoff idle slot durations, and proposed an LBT slot backoff scheme to avoid this problem. To evaluate the coexistence performance, we have developed a novel Markov chain approach with several new features to capture the complicated coexistence behaviors caused by different backoff slot durations. Then, we provided analytical results on the backoff counter hold time, successful transmission probability and throughput. We have implemented LTE and WLAN MAC scheme programming and extensive computer simulation, which have verified our analysis results. The new analytical tool can be leveraged for CSMA parameter optimization in coexisting systems, and provide theoretical support for related measurement and experiment. In future work, our method can be extended to coexistence of other CSMA/CA based wireless systems, and the effects of various fading channel models will be studied.

References

[1] R. Zhang, M. Wang, L. X. Cai, Z. Zheng, and X. Shen, "LTE-unlicensed: the future of spectrum aggregation for cellular networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 150–159, Jun. 2015.

[2] F. M. Abinader, et. al. "Enabling the coexistence of LTE and Wi-Fi in unlicensed bands," *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 54–61, Nov. 2014.

[3] A. Mukherjee et al., "Licensed-assisted access LTE: coexistence with IEEE 802.11 and the evolution toward 5G," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 50-57, Jun. 2016.

[4] 3GPP TSG RAN, "Study On Licensed-Assisted Access To Unlicensed Spectrum", 3GPP TR 36.889 V13.0.0, Jun. 2015.

[5] Ericsson, "Discussion on LBT protocols," 3GPP Tech. Rep. R1-151996, Apr. 2015.

[6] LTE-U Forum, "Coexistence study for LTE-U SDL", LTE-U Technical Report, V1.0, Feb. 2015.

[7] IEEE LAN/MAN Standards Committee, IEEE Std 802.11-2012, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Feb. 2012.

[8] C. Chen, R. Ratasuk, and A. Ghosh, "Downlink performance analysis of LTE and WiFi coexistence in unlicensed bands with a simple listen-before-talk scheme," *Proc. IEEE VTC*, pp. 1–5, May 2015.

[9] Y. Song, K. W. Sung, and Y. Han, "Coexistence of Wi-Fi and cellular with listen-before-talk in unlicensed spectrum," *IEEE Commun. Lett.*, vol. 20, no. 1, pp. 161–164, Jan. 2016.

[10] Y. Ma and D. G. Kuester, "MAC-Layer Coexistence Analysis of LTE and WLAN Systems Via Listen-Before-Talk," Proc. IEEE CCNC, Las Vegas, USA, Jan. 2017.

[11] V. Valls, A. Garcia-Saavedra, X. Costa and D. J. Leith, "Maximizing LTE capacity in unlicensed bands (LTE-U/LAA) while fairly coexisting with 802.11 WLANs," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1219–1222, Jun. 2016.

[12] S. Han, Y. C. Liang, Q. Chen and B. H. Soong, "Licensed-assisted access for LTE in unlicensed spectrum: A MAC protocol design," Proc. IEEE ICC, pp. 1–6, Kuala Lumpur, Malaysia, 2016.

[13] R. Yin, G. Yu, A. Maaref, and G. Y. Li, "A framework for co-channel interference and collision probability tradeoff in LTE licensed-assisted access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6078–6090, Sept. 2016.

[14] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[15] I. Tinnirello, G. Bianchi, and X. Yang, "Refinements on IEEE 802.11 distributed coordination function modeling approaches," *IEEE Trans. Veh. Technol.*, vol.59, no.3, pp.1055–1067, Mar. 2010.

[16] L. Dai and X. Sun, "A unified analysis of IEEE 802.11 DCF networks: stability, throughput, and delay," *IEEE Trans. Mobile Computing*, vol.12, no.8, pp.1558–1572, Aug. 2013.

[17] Y. Gao, X. Sun and L. Dai, "Throughput Optimization of Heterogeneous IEEE 802.11 DCF Networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 398–411, Jan. 2013.

[18] Y. Gao, X. Sun and L. Dai, "IEEE 802.11e Std EDCA networks: modeling, differentiation and optimization," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3863–3879, July 2014.

[19] W. Zhang, M. A. Suresh, R. Stoleru and H. Chenji, "On modeling the coexistence of 802.11 and 802.15.4 networks for performance tuning," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, pp. 5855-5866, Oct. 2014.

Coder, Jason; Kuester, Daniel; Ma, Yao; Young, William.
"Coexistence Analysis of LTE and WLAN Systems With Heterogenous Backoff Slot Durations."
Paper presented at IEEE International Conference on Communications (ICC) 2017, Paris, France. May 21, 2017 - May 25, 2017.

The Iterated Random Function Problem*,**

Ritam Bhaumik[1], Nilanjan Datta[2], Avijit Dutta[1], Nicky Mouha[3,4], and Mridul Nandi[1]

[1] Indian Statistical Institute, Kolkata, India.
[2] Indian Institute of Technology, Kharagpur, India.
[3] National Institute of Standards and Technology, Gaithersburg, MD, USA.
[4] Project-team SECRET, Inria, France.
bhaumik.ritam@gmail.com, nilanjan_isi_jrf@yahoo.com, avirocks.dutta13@gmail.com, nicky@mouha.be, mridul.nandi@gmail.com

Abstract. At CRYPTO 2015, Minaud and Seurin introduced and studied the *iterated random permutation* problem, which is to distinguish the r-th iterate of a random permutation from a random permutation. In this paper, we study the closely related *iterated random function* problem, and prove the first almost-tight bound in the adaptive setting. More specifically, we prove that the advantage to distinguish the r-th iterate of a random function from a random function using q queries is bounded by $O(q^2r(\log r)^3/N)$, where N is the size of the domain. In previous work, the best known bound was $O(q^2r^2/N)$, obtained as a direct result of interpreting the iterated random function problem as a special case of CBC-MAC based on a random function. For the iterated random function problem, the best known attack has an advantage of $\Omega(q^2r/N)$, showing that our security bound is tight up to a factor of $(\log r)^3$.

Keywords: Iterated random function, random function, pseudorandom function, password hashing, Patarin, H-coefficient technique, provable security.

1 Introduction

Take any n-bit hash function h. Assuming that this hash function can be modelled as a random function, the probability that the outputs of h collide given $q \ll 2^{n/2}$ distinct inputs is about $q^2/2^n$: the well-known birthday attack.

Now let us consider another hash function g, defined as the r-th iterate of h, i.e. $g(m) = h(h(\ldots h(m)))$, where h is applied r times. For the same number of queries $q \ll 2^{n/2}$, the birthday attack has about an r times higher probability to succeed for g than for h (see e.g. Preneel and van Oorschot [17, Lemma 2]).

* Certain algorithms and commercial products are identified in this paper to foster understanding. Such identification does not imply recommendation or endorsement by NIST, nor does it imply that the algorithms or products identified are necessarily the best available for the purpose.

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security, ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

Iteration is of fundamental importance in many cryptographic constructions. For example, a "possibly weak" function may be iterated to improve its resistance against various cryptanalysis attacks, or a password hashing function may be iterated to slow down dictionary attacks. But quite surprisingly, the security of iterating a random function is not yet a well-understood problem.

In the aforementioned (non-adaptive) birthday attack, the distinguishing advantage between a random function and an iterated random function increases by about a factor r. But what happens if we consider adaptive collision-finding attacks as well? Or in general, what if we want to consider any adaptive attack, not necessarily a collision-finding attack? Could there be more efficient attacks that have not yet been discovered?

Recently at CRYPTO 2015, Minaud and Seurin [14] put this possibility to rest for the iterated random permutation problem. They proved that the advantage to distinguish an iterated random permutation from a random permutation using q queries is bounded by $O(qr/N)$, where N is the size of the domain, and showed that their bound is almost tight by providing a matching attack.

In this paper, we will do the same for the iterated random function problem. Whereas the best bound in previous work is $O(q^2r^2/N)$, we will prove a bound of $O(q^2r(\log r)^3/N)$, where log is the logarithm to the base e. Our bound is tight up to a factor of about $(\log r)^3$, and thereby rules out the possibility of better attacks.

Note. We will focus on asymptotic bounds for large r, as this is parameter range where large improvements over the currently best-known bounds can be achieved. Although our bounds hold for any $r \geq 2$, we will apply generous relaxations to derive an easy-to-see bound that only improves the currently-known bounds for larger, but nevertheless practically-relevant values of r. Also, we will only consider the iteration of a uniformly random function in an information-theoretic setting. A simple hybrid argument can be used to extend this result to the pseudorandom function (prf) advantage in a computational setting, as shown by Minaud and Seurin [14, Theorem 1] for the iterated random permutation problem.

Applications. In spite of the frequent use of iterated random functions in practice, this paper is the first to study this problem without relying on the trivial CBC-MAC bound. The most obvious application of iterated random functions is in password hashing, where a hash function is iterated in order to slow down brute force attacks. This idea is used in PKCS #5's PBKDF1 and PBKDF2. In typical password-based key derivation functions, the iteration count is often quite high, ranging from several hundreds of thousands [8], to even ten million [19], as suggested by NIST for critical keys. To analyse the effect of iteration in these constructions, it is common to model the secret low-entropy password as a random-but-known key [10], or even an adversarially-chosen input [20]. But also small values of r, such as $r = 2$, appear in practical applications. In the book "Practical cryptography" [12], Ferguson and Schneier suggest to use SHA-256(SHA-256(m)) to avoid length-extension attacks. They

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

use this construction in their RSA encryption implementation, as well as in their Fortuna random number generator. Interestingly, about 2^{64} evaluations of SHA-256(SHA-256(m)) are performed *every second* as part of bitcoin mining [21].

Related Work. The security of an iterated random function was first analysed by Yao and Yin [22,23], when they analysed the security of the password-based key derivation functions PBKDF1 and PBKDF2. Their work is parallel to that of Wagner and Goldberg [20], who analysed the security of an iterated random permutation in the context of the Unix password hashing algorithm. Bellare et al. [4] extended these results, and also pointed out some problems in the proofs of Yao and Yin.

As Wagner and Goldberg explain in [20], it is possible to interpret the iterated random permutation problem as a special case of CBC-MAC where the iteration count r equals the number of message blocks, and all message blocks except for the first one are all-zero. The same holds for the iterated random function problem, except that a random function instead of a random permutation is used inside the CBC-MAC construction.

A first proof of the security of CBC-MAC was given by Bellare et al. in [1,2]. For CBC-MAC with a random function, they prove that the advantage of an information-theoretic adversary that makes at most q queries is upper bounded by $1.5r^2q^2/N$. Using the well-known prp-prf switching lemma [5], they derive from this an upper bound of $2r^2q^2/N$ for CBC-MAC with a random permutation. The simplicity of CBC-MAC makes it a good test case for various proof techniques. Of particular interest is the short proof of CBC-MAC by Bernstein [7]. For a more detailed proof using the same technique, we refer to Nandi [15].

In [3], Bellare et al. proved a security bound that is linear in r, instead of quadratic in r as in previous proofs. They point out that their analysis only applies to CBC-MAC with a random permutation, and not with a random function: such a bound is ruled out by an attack by Berke [6]. However, Berke's attack cannot be translated to the iterated random function problem, as the number of message blocks for each of the queries in the attack is not constant.

The iterated random function problem is similar to the nested iterated (NI) construction that Gaži et al. [13] analysed at CRYPTO 2014. However, the analysis of the NI construction critically relies on the use of two *different* random functions, or more precisely on the use of a pseudo-random function (prf) with two different keys. Our analysis applies to the case where only *one* random function is iterated. As we will show, the iterated random function problem will require a more complicated analysis of collision probabilities, in order to avoid ending up with a bound that is quadratic in r.

Main Results. The main results of this paper are the proofs of two theorems. Theorem 1 bounds the success probability of a common class of collision adversaries, and Theorem 2 bounds the advantage of distinguishing an iterated random function from a random function. In these theorems, the function $\phi(q, r)$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

is defined as

$$\phi(q,r) := 2\left(\frac{q^2\sqrt{r}}{N}\right) + 2\sqrt{\frac{q^2 r \log r}{N}} + 16\left(\frac{q^2 r \log r}{N}\right)^2 + 49(\log r)^2\left(\frac{q^2 r \log r}{N}\right).$$

Theorem 1. *Let f be a random function, and let $\mathcal{A}$ be a collision-finding adversary that makes q queries to f^r as follows: every query is either chosen from a set (of size $m \leq q$) of predetermined points, or is the response of a previous query. Under the assumption that $N \log r > 90$, the following bound holds for the success probability $\mathsf{cp}^r[q]$ of $\mathcal{A}$:*

$$\mathsf{cp}^r[q](\mathcal{A}) \leq \phi(q,r).$$

Theorem 2. *Let f be a random function, and let $\mathcal{A}$ be an adversary trying to distinguish f^r from f through q queries. Then, under the assumption that $N \log r > 90$, we have*

$$\mathbf{Adv}_{f,f^r}(q) \leq \frac{q^2 r}{N} + \frac{2q^2}{N} + \phi(q,r).$$

A Note on the Setting. We should point out that our results are in an indistinguishability setting. Our goal is to distinguish, in a black-box way, between an iterated random function and a random function. In the indifferentiability setting, the adversary also has access to the underlying random function, or to a simulator that tries to mimic its behaviour. Dodis et al. [11] proved that indifferentiability for an iterated random function holds only with poor concrete security bounds, as they provide a lower bound on the complexity of any successful simulator.

Outline. Notation and preliminaries are introduced in Sect. 2. We study the probabilities to find various types of collisions in a random function in Sect. 3. These results are used in Sect. 4 to bound the probabilities of single-trail attacks and two-trail collision attacks, and eventually to also bound a more general collision attack on an iterated random function. The advantage of distinguishing an iterated random function from a random function is bounded in Sect. 5. For readability, we defer the technical proof of Lemma 7 of Sect. 4 to Sect. 6. We conclude the paper in Sect. 7. The proofs of the lemmas of Sect. 2 are given in the App. A.

2 Notation and Preliminaries

In this section, we will state some simple lemmas without proof. The proofs of these lemmas can be found in App. A.

Functions. Let $f : \mathcal{D} \to \mathcal{D}$ be a function over a domain $\mathcal{D}$ of size N. A collision for a function f is defined as a pair $(x, x') \in \mathcal{D}$ with $x \neq x'$ such that $f(x) = f(x')$. A three-way collision is a triple (x, x', x'') such that $f(x) = f(x') = f(x'')$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security, ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

for distinct x, x' and x''. For a positive integer r, the r-th iterate f^r of a function f is defined inductively as follows:

$$f^1 = f,$$
$$f^r = f \circ f^{r-1}, r > 1.$$

By convention, let f^0 be the identity function. In the remainder of this paper, we will assume that $r \geq 2$. Let a random function denote a function that is drawn uniformly at random from the set of all functions of the same domain and range.

Falling Factorial Powers and the β Function. We use the falling factorial powers notation, where for a non-negative integer $i \leq N$, $N^{\underline{i}}$ is defined as

$$N^{\underline{i}} := \frac{N!}{(N-i)!} = N(N-1)\cdots(N-i+1). \tag{1}$$

Note that $N^{\underline{i}}$ denotes the number of permutations of N items taken i at a time, or the number of ways to choose a sample of size i without replacement from a population of size N. When $i > N$, we define $N^{\underline{i}} := 0$. We also define a function $\beta(i)$ that we will frequently encounter:

$$\beta(i) := \frac{N^{\underline{i}}}{N^i}. \tag{2}$$

Again, we define $\beta(i) := 0$ for $i > N$. We derive below a simple bound on $\beta(i)$.

Lemma 1. *Let $\alpha > 0$ be a real number. Then, for $i \geq \sqrt{2\alpha N} + 1$, we have*

$$\beta(i) \leq e^{-\alpha}.$$

Partial Sums of the Harmonic Series. The divergent infinite series

$$\sum_{i=1}^{\infty} \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$

is known as the harmonic series. We will be interested in partial sums of the series of the form

$$\sum_{i=a+1}^{b} \frac{1}{i} = \frac{1}{a+1} + \frac{1}{a+2} + \cdots + \frac{1}{b-1} + \frac{1}{b}.$$

We will use the following simple bound for this sum. Throughout this paper, let log denote the natural logarithm, that is the logarithm to the base e.

Lemma 2. *For any two positive integers a and b with $b \geq a$,*

$$\sum_{i=a+1}^{b} \frac{1}{i} \leq \log\left(\frac{b}{a}\right)$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

Counting Divisors. For a positive integer a and an integer b we use the notation $a|b$ to denote a *divides* b, i.e., $ak = b$ for some integer k. We write $a \nmid b$ when a does not divide b. The number of divisors of b is denoted $\mathsf{d}(b)$. We will use the following simple bound on $\mathsf{d}(b)$.

Lemma 3. *For any positive integer b,*

$$\mathsf{d}(b) < 2\sqrt{b}.$$

The σ Function. The function $\sigma(b)$ defined as

$$\sigma(b) := \sum_{a|b} a$$

denotes the sum of the divisors of b. We will use the following simple lemma about $\sigma(b)$.

Lemma 4. *For any positive integer b,*

$$\sum_{a|b} \frac{b}{a} = \sigma(b).$$

A simple bound on $\sigma(b)$ can be obtained as follows.

Lemma 5. *For any positive integer $b \geq 2$,*

$$\sigma(b) < 3b \log b.$$

3 Random Function Collisions

In this section, we look at different approaches to find collisions on a random function f. We will bound their success probabilities, and use them in Sect. 4 to get bounds on the success probabilities of collision attacks on an iterated random function f^r.

3.1 Single-Trail Attack

Single-Trail Attack. Let $[q]$ denote the set $\{1, \ldots, q\}$. The single-trail attack works by starting with an arbitrary initial point x and producing a *trail* of points, hoping to find a collision. A trail is uniquely defined by q queries $f^{i-1}(x)$ for $i \in [q]$, where the i-th query $f^{i-1}(x)$ has response $f^i(x)$. We assume that the attack does not stop when a collision is found, but makes q queries and then checks for collisions. If a collision is found, it will appear as a rho-shaped trail, as illustrated in Fig. 1. Therefore, a collision obtained through a single-trail attack will be called a ρ-collision.

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

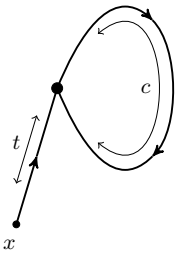


Fig. 1. Single-trail attack starting from x, resulting in a ρ collision with tail length t and cycle length c. We call the probability of this collision $\mathsf{cp}_\rho(t, c)$.

TERMINOLOGY. Suppose the q-query single-trail attack finds a collision. For some t, c, suppose it takes $t + c$ queries to find this collision, so that

$$f^{t+c}(x) = f^t(x),$$

i.e., the output of the $(t+c)$-th query is identical to the output of the t-th query. Then, t is called the tail length of the ρ-collision, and c is called the cycle length. For fixed t, c, we want to bound the probability that a q-query single-trail attack gives a ρ-collision on f with tail length t and cycle length c. Call this probability $\mathsf{cp}_\rho[q](t, c)$.

BOUNDING $\mathsf{cp}_\rho[q](t, c)$. To get a ρ-collision on f with tail length t and cycle length c, we need to call f at $t + c$ distinct values. Thus, if $q < t + c$, $\mathsf{cp}_\rho[q](t, c) = 0$. So suppose $q \geq t + c$. Out of these $t + c$ calls to f, the first $t + c - 1$ give distinct outputs, and the last coincides with the t-th output. Thus, the number of different ways this can happen is $N^{\underline{t+c-1}}$, out of the total N^{t+c} possible outcomes for the $t + c$ calls to f. Thus,

$$\mathsf{cp}_\rho[q](t, c) = \frac{N^{\underline{t+c-1}}}{N^{t+c}} = \frac{\beta(t + c - 1)}{N}.$$

This is just a function of t and c (since the queries made after the collision is found are of no consequence), so we will use the simpler notation $\mathsf{cp}_\rho(t, c)$, with the implicit assumption that $q \geq t + c$. For a fixed real $\alpha > 0$, when $t + c \geq \sqrt{2\alpha N} + 2$, Lemma 1 gives us the bound

$$\mathsf{cp}_\rho(t, c) \leq \frac{e^{-\alpha}}{N}. \tag{3}$$

When $t + c < \sqrt{2\alpha N} + 2$, we will simply use the bound

$$\mathsf{cp}_\rho(t, c) \leq \frac{1}{N}. \tag{4}$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

3.2 Two-Trail Attack

Two-Trail Attack. In the two-trail attack, we start with two different points x_1 and x_2, and produce two trails: the trail $f^{i-1}(x_1)$ for $i \in [q_1]$, and the trail $f^{i-1}(x_2)$ for $i \in [q_2]$, hoping to find a collision. In total $q_1 + q_2$ queries are made, where the i-th query for $i \in [q_1]$ is $f^{i-1}(x_1)$, with response $f^i(x_1)$, and the $(q_1 + i)$-th query for $i \in [q_2]$ is $f^{i-1}(x_2)$, with response $f^i(x_2)$. If a collision is found, the two trails will form a lambda shape, as illustrated in Fig. 2. Therefore, a collision obtained through a two-trail attack will be called a λ-collision.

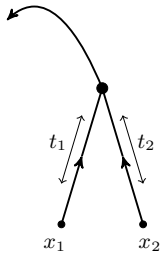


Fig. 2. Two-trail attack starting from x_1 and x_2, resulting in a λ-collision with foot lengths t_1 and t_2, respectively. We call the probability of this collision $\mathsf{cp}_\lambda(t_1, t_2)$.

Terminology. Suppose the (q_1, q_2)-query two-trail attack finds a λ-collision, regardless of whether a ρ-collisions has occurred on either trail. Suppose that a λ-collision is found after making t_1 queries along the first trail and t_2 queries along the second, i.e.,

$$f^{t_1}(x_1) = f^{t_2}(x_2).$$

t_1 and t_2 are called the foot lengths of the λ-collision. For fixed t_1, t_2, we want to bound the probability that a (q_1, q_2)-query two-trail attack finds a λ-collision with foot lengths t_1 and t_2. Denote this probability as $\mathsf{cp}_\lambda[q_1, q_2](t_1, t_2)$.

Bounding $\mathsf{cp}_\lambda[q_1, q_2](t_1, t_2)$. To get a λ-collision on f with foot lengths t_1 and t_2, we need to call f at t_1 distinct values on the first trail and t_2 distinct values on the second trail. Thus, if $q_1 < t_1$ or $q_2 < t_2$, $\mathsf{cp}_\lambda[q_1, q_2](t_1, t_2) = 0$. So we assume $q_1 \geq t_1$ and $q_2 \geq t_2$. Out of these $t_1 + t_2$ queries, the first $t_1 - 1$ in one trail and the first $t_2 - 1$ in the other trail give distinct outputs, and the last calls on the two trails coincide on a value distinct from all the earlier ones, i.e., the $t_1 + t_2$ calls lead to $t_1 + t_2 - 1$ distinct outputs, and one collision. Thus, the number of different ways this can happen is $N^{\underline{t_1+t_2-1}}$, out of the total $N^{t_1+t_2}$ possible outcomes for the $t_1 + t_2$ calls to f. Thus,

$$\mathsf{cp}_\lambda[q_1, q_2](t_1, t_2) = \frac{N^{\underline{t_1+t_2-1}}}{N^{t_1+t_2}} = \frac{\beta(t_1 + t_2 - 1)}{N}.$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

Again, this is only a function of t_1 and t_2 (since the queries made after the collision is found are of no consequence), so we will use the simpler notation $\mathsf{cp}_\lambda(t_1, t_2)$, with the implicit assumption that $q_1 \geq t_1$ and $q_2 \geq t_2$. For our purposes it will be enough to use the bound

$$\mathsf{cp}_\lambda(t_1, t_2) \leq \frac{1}{N}. \tag{5}$$

3.3 A $\lambda\rho$-Double-Collision on a Two-Trail Attack

When a two-trail attack leads to two collisions, a double-collision is said to occur. In Sect. 4, in addition to the above bounds, we also need a bound on the probability of two closely related double-collisions. We deal with a $\lambda\rho$-double-collision in this section, and a ρ'-double-collision in the next. A $\lambda\rho$-double-collision takes place when a two-trail attack leads to a λ-collision, and then the combined trail becomes the tail of a ρ-collision, as shown in Fig 3.[5]

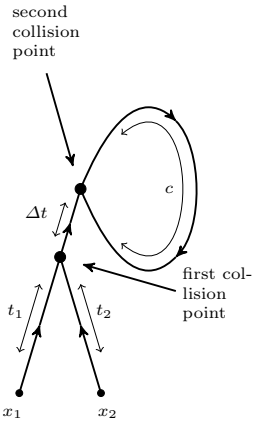


Fig. 3. Two-trail attack starting from x_1 and x_2, resulting in a $\lambda\rho$-collision. First, there is a λ-collision with foot lengths t_1 and t_2, respectively. Then, the combined trail continues for Δt queries, and completes a cycle of length c, after which a ρ-collision occurs. We call the probability of this double-collision $\mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c)$.

Terminology. We assign four parameters to this collision: the foot lengths t_1 and t_2 of the λ, the intervening length Δt between the two collisions, and the cycle length c of the ρ. Note that Δt can be seen as the tail length of the ρ-collision if we imagine it to have resulted from a single-trail attack beginning

[5] Note that we only call it a double-collision if both trails continue up to the point of second collision.

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

at the point of the λ-collision. For fixed $t_1, t_2, \Delta t, c$ we want to find the probability that a (q_1, q_2)-query two-trail attack finds a $\lambda\rho$-double-collision with foot lengths t_1 and t_2, intervening length Δt and cycle length c. Call this probability $\mathsf{cp}_{\lambda\rho}[q_1, q_2](t_1, t_2, \Delta t, c)$.

Bounding $\mathsf{cp}_{\lambda\rho}[q_1, q_2](t_1, t_2, \Delta t, c)$. To get a λ-collision on f with foot lengths t_1 and t_2, we need to call f at t_1 distinct values on the first trail, and t_2 distinct values on the second trail; and to get a ρ-collision on f with tail length Δt and cycle length c, we need to call f at Δt common values on each trail, and a further c points on the first trail; this adds up to $t_1 + t_2 + \Delta t + c$ distinct values in all. Thus, when $q_1 < t_1 + \Delta t + c$ or $q_2 < t_2 + \Delta t$, $\mathsf{cp}_{\lambda\rho}[q_1, q_2](t_1, t_2, \Delta t, c) = 0$. So we assume $q_1 \geq t_1 + \Delta t + c$ and $q_2 \geq t_2 + \Delta t$. These $t_1 + t_2 + \Delta t + c$ calls lead to $t_1 + t_2 + \Delta t + c - 2$ distinct outputs, and two collisions. Thus, the number of different ways this can happen is $N^{\underline{t_1+t_2+\Delta t+c-2}}$, out of the total $N^{t_1+t_2+\Delta t+c}$ possible outcomes for the $t_1 + t_2 + \Delta t + c$ calls to f. Thus,

$$\mathsf{cp}_{\lambda\rho}[q_1, q_2](t_1, t_2, \Delta t, c) = \frac{N^{\underline{t_1+t_2+\Delta t+c-2}}}{N^{t_1+t_2+\Delta t+c}} = \frac{\beta(t_1 + t_2 + \Delta t + c - 2)}{N^2}.$$

As before, this is only a function of $t_1, t_2, \Delta t$ and c (since the queries made after the ρ collision is found are of no consequence), so we use the simpler notation $\mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c)$, with the implicit assumption that $q_1 \geq t_1 + \Delta t + c$ and $q_2 \geq t_2 + \Delta t$. For a fixed real $\alpha > 0$, when $t_1 + t_2 + \Delta t + c \geq \sqrt{2\alpha N} + 3$, Lemma 1 gives us the bound

$$\mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c) \leq \frac{e^{-\alpha}}{N^2}. \tag{6}$$

When $t_1 + t_2 + \Delta t + c < \sqrt{2\alpha N} + 3$, we will simply use the bound

$$\mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c) \leq \frac{1}{N^2}. \tag{7}$$

3.4 A ρ'-Double-Collision on a Two-Trail Attack

A ρ'-double-collision takes place when a two-trail attack leads to a ρ with two tails. This is shown in Figure 4. We will allow $\Delta t = 0$, in which case a three-way collision occurs.

Terminology. As before, we assign four parameters to this collision: the tail lengths t_1 and t_2 of the ρ, the intervening length Δt between the two collisions, and the cycle length c of the ρ. For fixed $t_1, t_2, \Delta t, c$ we want to find the probability that a two-trail attack with sufficiently many queries finds a ρ'-double-collision with tail lengths t_1 and t_2, intervening length Δt, and cycle length c. Call this probability $\mathsf{cp}_{\rho'}[q_1, q_2](t_1, t_2, \Delta t, c)$.

Bounding $\mathsf{cp}_{\rho'}[q_1, q_2](t_1, t_2, \Delta t, c)$. The bounding of $\mathsf{cp}_{\rho'}[q_1, q_2](t_1, t_2, \Delta t, c)$ is almost identical to that of $\mathsf{cp}_{\lambda\rho}[q_1, q_2](t_1, t_2, \Delta t, c)$. To get a ρ'-double-collision with tail lengths t_1 and t_2, intervening length Δt, and cycle length c, we need to

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

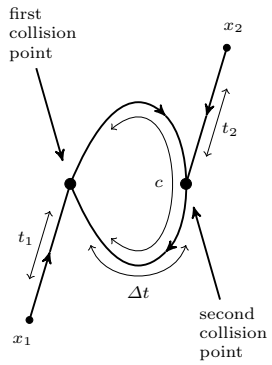


Fig. 4. Two-trail attack starting from x_1 and x_2, resulting in a ρ'-collision with tail lengths t_1 and t_2, intervening length Δt, and cycle length c. We will allow $\Delta t = 0$, in which case a three-way collision occurs. We call the probability of this double-collision $\mathsf{cp}_{\rho'}(t_1, t_2, \Delta t, c)$.

call f at $t_1+c-\Delta t$ distinct values on the first trail, t_2 distinct values on the second trail, and Δt common values on each trail, resulting in calls at t_1+t_2+c distinct values in all. Thus, when $q_1 < t_1+c$ or $q_2 < t_2+\Delta t$, $\mathsf{cp}_{\rho'}[q_1, q_2](t_1, t_2, \Delta t, c) = 0$. So we assume $q_1 \geq t_1 + c$ and $q_2 \geq t_2 + \Delta t$. These $t_1 + t_2 + c$ calls lead to t_1+t_2+c-2 distinct outputs. Thus, the number of different ways this can happen is $N^{\underline{t_1+t_2+c-2}}$, out of the total $N^{t_1+t_2+c}$ possible outcomes for the $t_1 + t_2 + c$ calls to f. Thus,

$$\mathsf{cp}_{\rho'}[q_1, q_2](t_1, t_2, \Delta t, c) = \frac{N^{\underline{t_1+t_2+c-2}}}{N^{t_1+t_2+c}} = \frac{\beta(t_1 + t_2 + c - 2)}{N^2}.$$

As before, this is only a function of $t_1, t_2, \Delta t$ and c (since the queries made after the ρ collision is found are of no consequence), so we use the simpler notation $\mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c)$, with the implicit assumption that $q_1 \geq t_1 + \Delta t + c$ and $q_2 \geq t_1 + \Delta t$. Recalling that

$$\mathsf{cp}_{\lambda\rho}(t_1, t_2, 0, c) = \frac{\beta(t_1 + t_2 + c - 2)}{N^2},$$

we conclude that

$$\mathsf{cp}_{\rho'}(t_1, t_2, \Delta t, c) = \mathsf{cp}_{\lambda\rho}(t_1, t_2, 0, c). \tag{8}$$

4 Iterated Random Function Collisions

In this section we revisit the two types of collision attacks described in Sect. 3, and analyse their success probabilities when applied to f^r. The main proof in this paper relies heavily on the results obtained in this section.

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

A Cautionary Note. At first glance, this section may appear to be similarly organised as Sect. 3. It is important to keep in mind that we are now interested in something entirely different. In Sect. 3, we looked at the probabilities of specific ρ- and λ-collisions with fixed parameters. In this section, instead, we focus on the probabilities that single-trail attacks and two-trail attacks of some specified number of queries succeed in finding collisions on f^r. By reducing these collisions to collisions on f, we can use the union bound on the bounds obtained in Sect. 3 to get the desired bounds. To distinguish from the collision probabilities on f, which we denoted $\mathsf{cp}[\cdot]$, we now use the notation $\mathsf{cp}^r[\cdot]$ for the collision probabilities on f^r.

4.1 Single-Trail Attack

We want to bound the probability that a q-query single-trail attack finds a collision on f^r. Call this probability $\mathsf{cp}^r_\rho[q]$.

Reducing to Collision on f. Suppose the q-query single-trail attack finds a ρ-collision on f^r with tail length t' and cycle length c'. Observe that this collision necessarily arises out of a ρ-collision on f, with tail length t and cycle length c for some t, c. This can happen in two ways:

– Direct Collision. This happens when r divides c. Then, define k such that rk is the first multiple of r that is not less than t, i.e.,

$$k := \left\lceil \frac{t}{r} \right\rceil,$$

then $rk + c$ is also a multiple of r, and since $f^{t+c}(x) = f^t(x)$, and $rk \geq t$, we also have

$$f^{rk+c}(x) = f^{rk}(x).$$

Writing

$$k' = \frac{c}{r},$$

we have

$$(f^r)^{k+k'}(x) = (f^r)^k(x),$$

our ρ-collision on f^r. Note that according to this notation,

$$t' = k = \left\lceil \frac{t}{r} \right\rceil, c' = k' = \frac{c}{r}.$$

Loosely speaking, in a direct collision, the first collision on f arrives *in phase* with r, i.e.,

$$t = t + c \bmod r,$$

so that this first collision on f leads immediately to a collision on f^r at the next multiple of r.

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

– Delayed Collision. A *delayed collision* occurs when r does not divide c, i.e., the first collision arrives *out of phase*. Then we need to keep cycling about the ρ of f till the phase is adjusted, and only then we arrive at the next multiple of r and find a collision on f^r. Suppose it cycles around η times. For the phase to be adjusted, $c\eta$ should be a multiple of r. The smallest value of η that satisfies this is

$$\eta = \frac{r}{d}$$

where $d = \mathsf{gcd}(c, r)$ is the greatest common divisor of c and r. Let $k = \lceil \frac{t}{r} \rceil$ as before, and let

$$k' = \frac{c}{d}.$$

As before, since we have $f^{t+c\eta}(x) = f^t(x)$, and $rk \geq t$, we have

$$f^{rk+c\eta}(x) = f^{rk}(x),$$

which gives us the ρ-collision

$$(f^r)^{k+k'}(x) = (f^r)^k(x),$$

as before. Again, according to this notation,

$$t' = k = \left\lceil \frac{t}{r} \right\rceil, c' = k' = \frac{c}{d}.$$

Required Conditions. Observing that a direct collision can be seen as a special case of delayed collision, where $d = \mathsf{gcd}(c, r) = r$, we can summarise the above as follows: a ρ-collision on f with tail length t and cycle length c *eventually* leads to a ρ-collision on f^r with tail length t' and cycle length c' where

$$t' = k = \left\lceil \frac{t}{r} \right\rceil, c' = k' = \frac{c}{d},$$

with $d = \mathsf{gcd}(c, r)$ as before. Thus, for a ρ-collision on f to result in a ρ-collision on f^r, the only required condition is that q is sufficiently large, i.e.,

$$t' + c' \leq q.$$

In terms of t and c, this becomes

$$\left\lceil \frac{t}{r} \right\rceil + \frac{c}{d} \leq q.$$

Recall that we are trying to bound the probability $\mathsf{cp}^r_\rho[q]$ of finding a ρ-collision on f^r in q queries. This is equivalent to the probability of finding a ρ-collision on f with the parameters t and c satisfying the above condition. Recall that in Sect. 3, we bounded this probability for a fixed (t, c), which we called $\mathsf{cp}_\rho(t, c)$. We can now use the union bound to get a bound on $\mathsf{cp}^r_\rho[q]$.

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

Using the Union Bound on $\mathsf{cp}^r_\rho[q]$. Let $\mathcal{S}$ be the set of (t, c) values that satisfy the requirement

$$\left\lceil \frac{t}{r} \right\rceil + \frac{c}{\mathsf{gcd}(c, r)} \leq q.$$

For a fixed $\alpha > 0$, we can split $\mathcal{S}$ into two parts:

$$\mathcal{S}^+[\alpha] := \left\{ (t, c) \in \mathcal{S} \mid t + c \geq \sqrt{2\alpha N} + 2 \right\},$$
$$\mathcal{S}^-[\alpha] := \left\{ (t, c) \in \mathcal{S} \mid t + c < \sqrt{2\alpha N} + 2 \right\}.$$

Applying the union bound with bounds (3) and (4) obtained for $\mathsf{cp}_\rho(t, c)$ gives

$$\begin{aligned}
\mathsf{cp}^r_\rho[q] &\leq \sum_{\mathcal{S}} \mathsf{cp}_\rho(t, c) \\
&= \sum_{\mathcal{S}^+[\alpha]} \mathsf{cp}_\rho(t, c) + \sum_{\mathcal{S}^-[\alpha]} \mathsf{cp}_\rho(t, c) \\
&\leq \sum_{\mathcal{S}^+[\alpha]} \frac{e^{-\alpha}}{N} + \sum_{\mathcal{S}^-[\alpha]} \frac{1}{N} \\
&= \#\mathcal{S}^+[\alpha] \cdot \frac{e^{-\alpha}}{N} + \#\mathcal{S}^-[\alpha] \cdot \frac{1}{N} \qquad (9)
\end{aligned}$$

Bounding $\#\mathcal{S}^-[\alpha]$. We observe that whenever $(t, c) \in \mathcal{S}^-[\alpha]$,

$$t < \sqrt{2\alpha N} + 2,$$

and

$$c < q \cdot \mathsf{gcd}(c, r).$$

If we count the number of (t, c) satisfying these conditions, it will give us an upper bound on $\#\mathcal{S}^-[\alpha]$. There are at most $\sqrt{2\alpha N} + 2$ values of t satisfying $t < \sqrt{2\alpha N} + 2$. For a fixed $d = \mathsf{gcd}(c, r)$, c has to be a multiple of d not exceeding qd. The number of such values of c is q. Since d must be a factor of r, we get the total number of values of c satisfying $c < q \cdot \mathsf{gcd}(c, r)$ to be at most $q \cdot \mathsf{d}(r)$. Putting it all together we get

$$\#\mathcal{S}^-[\alpha] \leq (\sqrt{2\alpha N} + 2) \cdot q \cdot \mathsf{d}(r). \qquad (10)$$

Bounding $\#\mathcal{S}^+[\alpha]$. For $(t, c) \in \mathcal{S}^+[\alpha]$, it will be enough for our purposes to consider the bounds

$$t \leq qr,$$

and

$$c < q \cdot \mathsf{gcd}(c, r).$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security, ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

Using the same reasoning as before, the number of values of c that satisfy $c < q \cdot \mathsf{gcd}(c, r)$ is at most $q \cdot \mathsf{d}(r)$. For t there are now at most qr values. Thus, we obtain the bound

$$\#\mathcal{S}^+[\alpha] \leq q^2 r \cdot \mathsf{d}(r). \tag{11}$$

Final Bound for $\mathrm{CP}^r_\rho[q]$. We can now plug (10) and (11) into (9):

$$\begin{aligned}\mathsf{cp}^r_\rho[q] &\leq \#\mathcal{S}^+[\alpha] \cdot \frac{e^{-\alpha}}{N} + \#\mathcal{S}^-[\alpha] \cdot \frac{1}{N} \\ &\leq q^2 r \cdot \mathsf{d}(r) \cdot \frac{e^{-\alpha}}{N} + (\sqrt{2\alpha N} + 2) \cdot q \cdot \mathsf{d}(r) \cdot \frac{1}{N}\end{aligned}$$

for any real $\alpha > 0$. We will simplify it by plugging in a suitable value of α.

Simplifying the Bound. We know from Lemma 3 that

$$\mathsf{d}(r) < 2\sqrt{r}.$$

We put $\alpha = \log r$. Then we have

$$\sqrt{2\alpha N} = \sqrt{2N \log r},$$

and

$$e^{-\alpha} = \frac{1}{r}.$$

When $N \log r \geq 16$, we have

$$\begin{aligned}\sqrt{2\alpha N} + 2 &= \sqrt{2N \log r} + 2 \\ &= 2\sqrt{N \log r} - \left[(2 - \sqrt{2}) \cdot \sqrt{N \log r} - 2\right] \\ &\leq 2\sqrt{N \log r} - \left[(2 - \sqrt{2}) \cdot 4 - 2\right] \\ &= 2\sqrt{N \log r} - \left[6 - \sqrt{2}\right] \\ &< 2\sqrt{N \log r}.\end{aligned}$$

Thus,

$$\mathsf{cp}^r_\rho[q] \leq 2 \cdot \left(\frac{q^2\sqrt{r}}{N}\right) + 2 \cdot \sqrt{\frac{q^2 r \log r}{N}}.$$

This gives us a bound for the success probability of a q-query single-trail attack on f^r. We state the result as a lemma.

Lemma 6. *Under the assumption that $N \log r \geq 16$, we have*

$$\mathsf{cp}^r_\rho[q] \leq 2 \cdot \left(\frac{q^2\sqrt{r}}{N}\right) + 2 \cdot \sqrt{\frac{q^2 r \log r}{N}}.$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

4.2 Two-Trail Attack

We want to bound the probability that a (q_1, q_2)-query two-trail attack finds a λ-collision on f^r. Call this probability $\mathsf{cp}^r_\lambda[q_1, q_2]$.

Reducing to Collision on f. Suppose the (q_1, q_2)-query two-trail attack finds a λ-collision on f^r with foot lengths t'_1 and t'_2. As in the case of the ρ-collision on f^r, this can only arise from a λ-collision on f, say with foot lengths t_1 and t_2, which can again happen in two ways:

– Direct Collision. A direct collision takes place when the two f-trails collide in phase, i.e.,

$$t_1 = t_2 \bmod r.$$

When this happens, the two trails continue till the next multiple of r, where they give a λ-collision on f^r. This collision takes place at

$$t'_1 = \left\lceil \frac{t_1}{r} \right\rceil, t'_2 = \left\lceil \frac{t_2}{r} \right\rceil.$$

– Delayed Collision. A delayed collision takes place when the two f-trails collide out of phase, i.e.,

$$t_1 \neq t_2 \bmod r.$$

If one of the trails results in a ρ-collision on f^r, this implies that a successful single-trail attack has been carried out on f^r. Here, we will only focus on the scenario where a λ-collision on f^r can still happen. But then one of the two f-trails must have entered into a cycle, otherwise both f-trails will remain out of phase. This can only happen in one of two ways:

- After the λ-collision on f, the combined trail forms the tail of a ρ collision on f, that is, they form a $\lambda\rho$-collision on f as in Fig. 3. One of the trails, say the one from x_1, cycles around the ρ enough number of times to adjust the phase, and then the two f-trails continue to the next multiple of r, giving a λ-collision on f^r;[6]
- After the λ-collision on f, one of the two f-trails, say the one from x_1, continues and collides with the trail from x_2, that is, they form a ρ'-collision on f as in Fig. 4. When $\Delta t = 0$, a three-way collision on f occurs. The trail from x_1 cycles around the ρ enough number of times to adjust the phase, giving a λ-collision on f^r.

In our calculations, we assume that it is the trail from x_1 that cycles multiple times, while the one from x_2 waits for the collision on f^r to happen. We obtain a bound which is symmetric over q_1 and q_2, and thus also holds for the case when the two trails reverse roles. Let τ_1 and τ_2 be the respective

[6] This is indeed a (delayed) λ-collision on f^r: from the point of view of f^r, neither of the two trails could be seen to enter into a cycle.

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

lengths of the two trails till *the point of waiting*, i.e., the point of ρ-collision of the trail from x_1. Calling Δt the distance between the two collision points, we simply have

$$\tau_1 = t_1 + \Delta t, \tau_2 = t_2 + \Delta t$$

for the $\lambda\rho$-collision, and

$$\tau_1 = t_1, \tau_2 = t_2 + \Delta t$$

for the ρ'-collision. Let the cycle length of this ρ be c (note that its tail length is τ_1 with respect to this trail). Suppose this trail cycles η times about the ρ in order to adjust the phase difference. Then η is the smallest number that satisfies

$$\tau_1 + c\eta = \tau_2 \bmod r.$$

Suppose k is such that

$$\tau_1 + c\eta = \tau_2 + rk.$$

Also, let

$$k_2 = \left\lceil \frac{\tau_2}{r} \right\rceil.$$

From our definition of τ_1 and τ_2, we have that

$$f^{\tau_1}(x_1) = f^{\tau_2}(x_2),$$

and from the ρ-collision $f^{\tau_1+c}(x_1) = f^{\tau_1}(x_1)$, it follows that

$$f^{\tau_1+c\eta}(x_1) = f^{\tau_1}(x_1).$$

From these two we get

$$f^{\tau_1+c\eta}(x_1) = f^{\tau_2}(x_2).$$

From the definition of k we have

$$f^{\tau_2+rk}(x_1) = f^{\tau_2}(x_2).$$

Continuing on to rk_2, we get a λ-collision on f^r as

$$(f^r)^{k+k_2}(x_1) = (f^r)^{k_2}(x_2).$$

According to this notation we have a λ-collision on f^r with foot lengths t'_1 and t'_2, such that

$$t'_1 = k + k_2 = \left\lceil \frac{\tau_1 + c\eta}{r} \right\rceil, t'_2 = k_2 = \left\lceil \frac{\tau_2}{r} \right\rceil.$$

When this comes from a $\lambda\rho$-collision, we have

$$t'_1 = \left\lceil \frac{t_1 + \Delta t + c\eta}{r} \right\rceil, t'_2 = \left\lceil \frac{t_2 + \Delta t}{r} \right\rceil.$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

When this comes from a ρ'-collision, we have

$$t'_1 = \left\lceil \frac{t_1 + c\eta}{r} \right\rceil, t'_2 = \left\lceil \frac{t_2 + \Delta t}{r} \right\rceil.$$

We will treat these two cases separately, even though they are closely related.

Required Conditions. Again, we observe that the direct collision is a special case of the delayed collision with $\Delta t = 0$ and $\eta = 0$. *However, there is an important difference.* For the delayed λ-collision, we require *two* collisions on f, unlike all other collisions we have seen so far, which need only one. This case corresponds to the $\lambda\rho$-double-collision and the ρ'-double-collision from Sect. 3, and requires some special treatment, as we will see in the course of our calculations. The condition needed here is that both trails continue long enough for the collision to happen, i.e.,

$$t'_1 \leq q_1, t'_2 \leq q_2.$$

In terms of $t_1, t_2, \Delta t, c, \eta$, this translates to

$$\left\lceil \frac{t_1 + \Delta t + c\eta}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2 + \Delta t}{r} \right\rceil \leq q_2$$

for the $\lambda\rho$-double-collision and

$$\left\lceil \frac{t_1 + c\eta}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2 + \Delta t}{r} \right\rceil \leq q_2$$

for the ρ'-double-collision. Recall that we are trying to calculate $\mathsf{cp}^r_\lambda[q_1, q_2]$, the probability of getting a λ-collision on f^r with a (q_1, q_2)-query two-trail attack starting from x_1 and x_2. Based on our observations above, this can happen in two ways:

- A Direct λ-collision on f. This is the direct collision scenario, where the collision is in phase. The foot lengths t_1 and t_2 have the constraints

$$\left\lceil \frac{t_1}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2}{r} \right\rceil \leq q_2, t_1 = t_2 \bmod r.$$

  For fixed t_1, t_2, we recall that the probability of this collision is $\mathsf{cp}_\lambda(t_1, t_2)$.

- A $\lambda\rho$-double-collision on f. This is the first case of the delayed collision scenario, where the collision is out of phase. Here, t_1 and t_2 are the foot lengths of the λ, Δt is the distance between the two collision points, c is the cycle length of the ρ, and η is the number of cycles necessary around the ρ. Recall that one of the trails circles around the ρ, while the other waits for the λ-collision on f^r to happen. We continue with our assumption that the one from x_1 does the cycling and the one from x_2 waits, since we will eventually count over all pairs of trails. Now $t_1, t_2, \Delta t, c, \eta$ have the constraints

$$\left\lceil \frac{t_1 + \Delta t + c\eta}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2 + \Delta t}{r} \right\rceil \leq q_2, t_1 + c\eta = t_2 \bmod r.$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

For fixed $t_1, t_2, \Delta t, c, \eta$, we recall that the probability of this $\lambda\rho$-double-collision is $\mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c)$.

– A ρ'-DOUBLE-COLLISION ON f. This is the second case of the delayed collision scenario. Here, t_1 and t_2 are the lengths of the two tails of the ρ, Δt is the distance between the two collision points, c is the cycle length of the ρ, and η is the number of cycles necessary around the ρ. Again, the trail from x_1 circles around the ρ, while the trail from x_2 waits for the λ-collision on f^r to happen. Thus, $t_1, t_2, \Delta t, c, \eta$ have the constraints

$$\left\lceil \frac{t_1 + c\eta}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2 + \Delta t}{r} \right\rceil \leq q_2, t_1 + c\eta = t_2 + \Delta t \bmod r.$$

Our strategy for bounding $\mathsf{cp}^r_\lambda[q_1, q_2]$ will be similar to the one we used for bounding $\mathsf{cp}^r_\rho[q]$: to take the bounds on $\mathsf{cp}_\lambda(t_1, t_2)$ for fixed t_1, t_2, $\mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c)$ for fixed $t_1, t_2, \Delta t, c$ and $\mathsf{cp}_{\rho'}(t_1, t_2, \Delta t, c)$ for fixed $t_1, t_2, \Delta t, c$ obtained in Sect. 3, and then use the union bound over all possible values these parameters can take.

APPLYING THE UNION BOUND TO $\mathsf{cp}^r_\lambda[q_1, q_2]$. Let $\mathcal{S}_1$ be the set of (t_1, t_2) values that satisfy the constraints

$$\left\lceil \frac{t_1}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2}{r} \right\rceil \leq q_2, t_1 = t_2 \bmod r,$$

and let

$$\mathsf{p}_1 := \sum_{\mathcal{S}_1} \mathsf{cp}_\lambda(t_1, t_2).$$

Let $\mathcal{S}_2$ be the set of $(t_1, t_2, \Delta t, c, \eta)$ values that satisfy the constraints

$$\left\lceil \frac{t_1 + \Delta t + c\eta}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2 + \Delta t}{r} \right\rceil \leq q_2, t_1 + c\eta = t_2 \bmod r,$$

and let

$$\mathsf{p}_2 := \sum_{\mathcal{S}_2} \mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c).$$

Let $\mathcal{S}_3$ be the set of $(t_1, t_2, \Delta t, c, \eta)$ values that satisfy the constraints

$$\left\lceil \frac{t_1 + c\eta}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2 + \Delta t}{r} \right\rceil \leq q_2, t_1 + c\eta = t_2 + \Delta t \bmod r,$$

and let

$$\mathsf{p}_3 := \sum_{\mathcal{S}_3} \mathsf{cp}_{\rho'}(t_1, t_2, \Delta t, c).$$

In addition, for the case where the trails reverse roles, we define $\mathcal{S}_4$ as the set of $(t_1, t_2, \Delta t, c, \eta)$ values that satisfy the constraints

$$\left\lceil \frac{t_1 + \Delta t}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2 + \Delta t + c\eta}{r} \right\rceil \leq q_2, t_1 = t_2 + c\eta \bmod r,$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

and

$$\mathsf{p}_4 := \sum_{\mathcal{S}_4} \mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c).$$

Similarly, we define $\mathcal{S}_5$ as the set of $(t_1, t_2, \Delta t, c, \eta)$ values that satisfy the constraints

$$\left\lceil \frac{t_1 + \Delta t}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2 + c\eta}{r} \right\rceil \leq q_2, t_1 + \Delta t = t_2 + c\eta \bmod r,$$

and

$$\mathsf{p}_5 := \sum_{\mathcal{S}_5} \mathsf{cp}_{\rho'}(t_1, t_2, \Delta t, c).$$

We state here the following bounds on $\mathsf{p}_1, \mathsf{p}_2, \mathsf{p}_3$, the proof of which we defer to Sect. 6:

Lemma 7. *Under the assumption that $N \log r > 90$,*

$$\begin{aligned}
p_1 &\leq \frac{q_1 q_2 r}{N}, \\
p_2 &\leq 8 \cdot (\log r)^2 \cdot \left(\frac{q_1 q_2 r}{N}\right)^2 + 24 \cdot (\log r)^3 \cdot \left(\frac{q_1 q_2 r}{N}\right), \\
p_3 &\leq 8 \cdot (\log r)^2 \cdot \left(\frac{q_1 q_2 r}{N}\right)^2 + 24 \cdot (\log r)^3 \cdot \left(\frac{q_1 q_2 r}{N}\right).
\end{aligned}$$

Final Bound for $\mathsf{cp}^r_\lambda[q_1, q_2]$. We observe that the bounds for p_2 and p_3 in Lemma 7 are symmetric over q_1 and q_2. Thus, we have

$$\begin{aligned}
\mathsf{p}_4 &\leq 8 \cdot (\log r)^2 \cdot \left(\frac{q_1 q_2 r}{N}\right)^2 + 24 \cdot (\log r)^3 \cdot \left(\frac{q_1 q_2 r}{N}\right), \\
\mathsf{p}_5 &\leq 8 \cdot (\log r)^2 \cdot \left(\frac{q_1 q_2 r}{N}\right)^2 + 24 \cdot (\log r)^3 \cdot \left(\frac{q_1 q_2 r}{N}\right).
\end{aligned}$$

Using the union bound, we get

$$\mathsf{cp}^r_\lambda[q_1, q_2] \leq \mathsf{p}_1 + \mathsf{p}_2 + \mathsf{p}_3 + \mathsf{p}_4 + \mathsf{p}_5.$$

This gives us the required bound, which we state next in the form of a lemma.

Lemma 8. *When $N \log r > 90$,*

$$\mathsf{cp}^r_\lambda[q_1, q_2] \leq 32 \cdot \left(\frac{q_1 q_2 r \log r}{N}\right)^2 + 97 \cdot (\log r)^2 \cdot \left(\frac{q_1 q_2 r \log r}{N}\right).$$

Proof. As $r \geq 2$, we can relax the bound of p_1 as

$$\mathsf{p}_1 \leq \frac{q_1 q_2 r}{N} \leq \frac{q_1 q_2 r}{N} \cdot (\log r)^3.$$

The rest follows from Lemma 7. □

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

4.3 A More General Collision Attack

Previously, we looked at two main approaches for a collision attack: the single-trail attack and the two-trail attack, and we bounded their success probabilities. Now, we will bound the success probability of a more general collision attack. More specifically, we consider collision attack subject to the restriction that is given in the statement of Theorem 1 in Sect. 1: every query is either chosen from a set of size m (with $m \leq q$) of predetermined starting points, or is the response of a previous query. First, let us introduce the notion of a transcript.

Transcript. Let us consider any adversary $\mathcal{A}$ that interacts with an oracle $\mathcal{O}$. This interaction can be represented as a transcript, that is, as a list of queries made and answers returned. Let the transcript tr be defined as the q-tuple of input-output pairs $\mathsf{tr} = ((x_1, y_1), (x_2, y_2), \ldots, (x_q, y_q))$. Without loss of generality, we do not consider adversaries here that repeat the same query, i.e., all q queries are distinct.

Sources and Trails. For $j, j' \in [q], j \neq j'$, we say that $x_{j'}$ is a *predecessor* of x_j if

$$f(x_{j'}) = x_j.$$

We call x_j a *source* if it does not have a predecessor. If there exists a non-empty subset of the queries for which every query has a predecessor that is in the same subset, and no query has a predecessor outside the set, we call this subset a *permutation cycle*. Note that a permutation cycle forms a rho-shape with a tail of length zero. For a permutation cycle, we define the query x_j of the permutation cycle with the smallest index j to be a *source*.

Suppose that there are m sources along the q queries, which we call $z_1, \ldots, z_m$. Then we can see the attack as an m-trail attack, with the m trails starting from $z_1, \ldots, z_m$ and of lengths $q_1, \ldots, q_m$ respectively. Thus, each point that is not a source must be on one of these m trails.

If the collision attack is successful, then for some $i, i' \in [q]$ with $i \neq i'$, we have

$$f(x_i) = f(x_{i'}).$$

In that case, one of the following must hold:

- x_i and $x_{i'}$ are on the same trail, say the one from z_p – in this case, a successful q_p-query single-trail attack starting from z_p has occurred;
- x_i and $x_{i'}$ are on different trails, say the ones from z_p and $z_{p'}$ respectively – in this case, a successful $(q_p, q_{p'})$-query two-trail attack starting from $(z_p, z_{p'})$ has occurred.

A Word on the Choice of $q_1, \ldots, q_m$. We note here that since we are allowing the trails to collide and merge with each other, the trail lengths $q_1, \ldots, q_m$ are not necessarily unique, since the queries on the merged trail can be counted on either trail, or both. We can get around this by choosing to count each merged trail as part of any one of the pre-merging trails, while the other is thought to stop at the point of collision. This way, we ensure that $\sum_{j=1}^{m} q_j = q$.

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

To bound the success probability of this more general collision attack, we can use the previously obtained bounds on the success probabilities of single-trail attacks and two-trail attacks along with the union bound. With notation as above we recall the following bounds:

– Single-Trail Attack. For a q-query single-trail attack, Lemma 6 gives us the bound

$$\mathsf{cp}^r_\rho[q] \leq 2 \cdot \left(\frac{q^2\sqrt{r}}{N}\right) + 2 \cdot \sqrt{\frac{q^2 r \log r}{N}}.$$

– Two-Trail Attack. For a (q_1, q_2)-query two-trail attack, Lemma 8 gives us the bound

$$\mathsf{cp}^r_\lambda[q_1, q_2] \leq 32 \cdot \left(\frac{q_1 q_2 r \log r}{N}\right)^2 + 97 \cdot (\log r)^2 \cdot \left(\frac{q_1 q_2 r \log r}{N}\right).$$

Let $\mathsf{cp}^r[q](\mathcal{A})$ denote the probability that the collision adversary $\mathcal{A}$ making q queries finds a collision on f^r. For $q_1, \ldots, q_m$, with

$$\sum_{i=1}^{m} q_i = q,$$

and let $\mathsf{cp}^r[q](q_1, \ldots, q_m)$ denote the probability that a collision attack with m trails of lengths $q_1, \ldots, q_m$ finds a collision on f^r. Thus,

$$\mathsf{cp}^r[q](\mathcal{A}) \leq \max_{\sum q_i = q} \mathsf{cp}^r[q](q_1, \ldots, q_m).$$

By the union bound, we have

$$\mathsf{cp}^r[q](q_1, \ldots, q_m) \leq \sum_{i=1}^{m} \mathsf{cp}^r_\rho[q_i] + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathsf{cp}^r_\lambda[q_i, q_j].$$

We bound the two terms separately.

$$\begin{aligned}
\sum_{i=1}^{m} \mathsf{cp}^r_\rho[q_i] &= \sum_{i=1}^{m} \left[2 \cdot \left(\frac{q_i^2\sqrt{r}}{N}\right) + 2 \cdot \sqrt{\frac{q_i^2 r \log r}{N}}\right] \\
&= 2 \cdot \left(\frac{\sqrt{r}}{N}\right) \cdot \sum_{i=1}^{m} q_i^2 + 2 \cdot \sqrt{\frac{r \log r}{N}} \cdot \sum_{i=1}^{m} q_i \\
&\leq 2 \cdot \left(\frac{\sqrt{r}}{N}\right) \cdot q^2 + 2 \cdot \sqrt{\frac{r \log r}{N}} \cdot q \\
&= 2 \cdot \left(\frac{q^2\sqrt{r}}{N}\right) + 2 \cdot \sqrt{\frac{q^2 r \log r}{N}};
\end{aligned}$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

$$\begin{aligned}
\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} \mathsf{cp}^r_\lambda[q_i,q_j] &= \sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\left[32\cdot\left(\frac{q_iq_jr\log r}{N}\right)^2\right.\\
&\qquad\left.+\,97\cdot(\log r)^2\cdot\left(\frac{q_iq_jr\log r}{N}\right)\right]\\
&= 32\cdot\left(\frac{r\log r}{N}\right)^2\cdot\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}q_i^2q_j^2\\
&\qquad+\,97\cdot(\log r)^2\cdot\left(\frac{r\log r}{N}\right)\cdot\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}q_iq_j\\
&\le 16\cdot\left(\frac{r\log r}{N}\right)^2\cdot q^4+49\cdot(\log r)^2\cdot\left(\frac{r\log r}{N}\right)\cdot q^2\\
&= 16\cdot\left(\frac{q^2r\log r}{N}\right)^2+49\cdot(\log r)^2\cdot\left(\frac{q^2r\log r}{N}\right).
\end{aligned}$$

Since these bounds are free of $q_1,\ldots,q_m$, this proves Theorem 1 of the paper.

5 Bounding the Advantage of Distinguishing f and f^r

5.1 Security Game

The Setup. An oracle $\mathcal{O}$ imitating a function g takes q queries $\{x_i \mid i\in[q]\}$ and returns

$$\{y_i = g(x_i) \mid i\in[q]\}.$$

The q-tuple of input-output pairs of the oracle is called the transcript, denoted as

$$\mathsf{tr} = ((x_1,y_1),(x_2,y_2),\ldots,(x_q,y_q)).$$

Both the real oracle $\mathcal{O}_{\text{REAL}}$ and the ideal oracle $\mathcal{O}_{\text{IDEAL}}$ will initially select a uniformly random function f. Then, $\mathcal{O}_{\text{REAL}}$ goes on to imitate f^r, while $\mathcal{O}_{\text{IDEAL}}$ imitates f itself. For any adversary $\mathcal{A}$, we want to bound its advantage, defined as

$$\mathbf{Adv}_{f,f^r}(q) = \left|\Pr\left[\mathcal{A}^{\mathcal{O}_{\text{IDEAL}}}(q)\to 1\right]-\Pr\left[\mathcal{A}^{\mathcal{O}_{\text{REAL}}}(q)\to 1\right]\right|.$$

As in the collision attack of Sect. 4.3, we can view the transcript tr as m trails of lengths $q_1,\ldots,q_m$ with sources $z_1,\ldots,z_m$, possibly with collisions, such that no query is counted in more than one trail, and hence

$$\sum_{j=1}^{m} q_j = q.$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

For $i \in [m]$, we shall use the notation

$$z_{i,1} := \mathcal{O}(z_i),$$
$$z_{i,j} := \mathcal{O}(z_{i,j-1}), 2 \leq j \leq q_i.$$

Good and Bad Transcripts. We partition the set of attainable transcripts into a set $\mathcal{T}_{\mathsf{good}}$ of good transcripts, and a set $\mathcal{T}_{\mathsf{bad}}$ of bad transcripts. We say $\mathsf{tr} \in \mathcal{T}_{\mathsf{bad}}$ if either of the following holds:

- For some $i \in [m]$,
$$z_{i,q_i} = z_i,$$
that is, the i-th trail forms a permutation cycle. Note that, by our construction of the trails, $z_{i_1,j}$ cannot equal z_{i_2} unless $i_1 = i_2$.
- For some $i_1, i_2 \in [m], j_1 \in [q_{i_1}], j_2 \in [q_{i_2}]$ with $(i_1, j_1) \neq (i_2, j_2)$, we have
$$z_{i_1,j_1} = z_{i_2,j_2},$$
that is, there is a ρ-collision on one of the trails ($i_1 = i_2$), or there is a λ-collision on two of the trails ($i_1 \neq i_2$).

5.2 Applying the H-Coefficient Technique

Let us denote the probability distribution of the transcripts in the real world by $\Pr_{\mathcal{O}_{\text{REAL}}}$, and in the ideal world by $\Pr_{\mathcal{O}_{\text{IDEAL}}}$. Our proof will use Patarin's H-coefficient technique [16].

Lemma 9 (H-Coefficient Technique). *Let $\mathcal{A}$ be an adversary, and let $\mathcal{T} = \mathcal{T}_{good} \cup \mathcal{T}_{bad}$ be a partition of the set of attainable transcripts. Let ε_1 be such that for all $\mathsf{tr} \in \mathcal{T}_{good}$:*

$$\frac{\Pr_{\mathcal{O}_{\text{REAL}}}[\mathsf{tr}]}{\Pr_{\mathcal{O}_{\text{IDEAL}}}[\mathsf{tr}]} \geq 1 - \varepsilon_1.$$

Furthermore, let $\varepsilon_2 = \Pr_{\mathcal{O}_{\text{IDEAL}}}[\mathsf{tr} \in \mathcal{T}_{bad}]$. Then $\mathbf{Adv}_{f,f^r}(q) \leq \varepsilon_1 + \varepsilon_2$.

Proof. For a proof and a detailed explanation of this technique, see Chen and Steinberger [9]. □

Probability of Bad Transcripts in Ideal Model. We can easily bound the probability that a transcript tr from the ideal oracle $\mathcal{O}_{\text{IDEAL}}$ is in $\mathcal{T}_{\mathsf{bad}}$. Suppose all of the q responses lie outside $\{z_i \mid i \in [m]\}$, and there is no collision between any of the responses. When this happens, tr cannot be in $\mathcal{T}_{\mathsf{bad}}$. The probability of this is at least $1 - \dfrac{2q^2}{N}$: two responses collide with probability at

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

most $\frac{q^2}{N}$; and a response collides with a z_i with probability at most $\frac{q^2}{N}$, since there are m different values of z_i, and $m \leq q$. Thus,

$$\varepsilon_2 := \Pr_{\mathcal{O}_{\text{IDEAL}}}[\mathsf{tr} \in \mathcal{T}_{\mathsf{bad}}] \leq \frac{2q^2}{N}.$$

Probability of Good Transcripts. We now focus only on transcripts in $\mathcal{T}_{\mathsf{good}}$. Let us consider a good and attainable transcript $\mathsf{tr} \in \mathcal{T}_{\mathsf{good}}$. For the ideal oracle, as the number of distinct inputs is q, we have

$$\Pr_{\mathcal{O}_{\text{IDEAL}}}[\mathsf{tr}] = \frac{1}{N^{\underline{q}}}.$$

Now we bound $\Pr_{\mathcal{O}_{\text{REAL}}}[\mathsf{tr}]$ for $\mathsf{tr} \in \mathcal{T}_{\mathsf{good}}$. Consider a $(q_1, \ldots, q_m)$-query m-trail collision attack on f^r, with sources $z_1, \ldots, z_m$ respectively. Theorem 1 tells us that this attack fails with probability at least $1 - \phi(q, r)$, where

$$\phi(q,r) := 2\left(\frac{q^2\sqrt{r}}{N}\right) + 2\sqrt{\frac{q^2 r \log r}{N}} + 16\left(\frac{q^2 r \log r}{N}\right)^2 + 49(\log r)^2\left(\frac{q^2 r \log r}{N}\right).$$

We now observe that when this attack fails, the attack transcript is either isomorphic as a graph to tr, or contains a permutation cycle.[7] A permutation cycle occurs when queries of f^r collide with a source z_i, which has probability at most $\frac{q^2 r}{N}$, since there are m different values of z_i and $m \leq q$. Thus, the attack transcript is isomorphic to tr with probability at least

$$1 - \phi(q,r) - \frac{q^2 r}{N}.$$

Now the graph of this attack transcript has $q + m$ nodes, all distinct. Of these, the m sources are already fixed. The rest can take values in $N^{\underline{q}}$ ways. Now all of these $N^{\underline{q}}$ graphs are equally likely to occur in the scenario described above, i.e., when the m-trail attack fails and does not contain a permutation cycle. One of the equally likely $N^{\underline{q}}$ graphs is the graph of tr. Thus,

$$\Pr_{\mathcal{O}_{\text{REAL}}}[\mathsf{tr}] \geq \left(1 - \phi(q,r) - \frac{q^2 r}{N}\right) \cdot \frac{1}{N^{\underline{q}}}.$$

Applying the H-Coefficient Technique. Let $R(\mathsf{tr})$ be the ratio of the probabilities of $\mathsf{tr} \in \mathcal{T}_{\mathsf{good}}$ under $\mathcal{O}_{\text{REAL}}$ and $\mathcal{O}_{\text{IDEAL}}$ respectively. Then we have shown above that

$$R(\mathsf{tr}) \geq \left(1 - \phi(q,r) - \frac{q^2 r}{N}\right) \cdot \frac{1}{\beta(q)}.$$

[7] Note that the graph isomorphism follows from a simple relabeling of inputs and outputs, starting with the sources of every trail. This is possible because excluding collisions and permutation cycles means that no two inputs will have the same output, and outputs never correspond to a source.

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security, ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

From Lemma 1, we have

$$\beta(q) \leq 1.$$

Thus,

$$R(\mathsf{tr}) \geq 1 - \varepsilon_1$$

where

$$\varepsilon_1 := \phi(q, r) + \frac{q^2 r}{N}.$$

Hence, by the H-coefficient technique of Lemma 9, we have

$$\mathbf{Adv}_{f,f^r}(q) \leq \varepsilon_1 + \varepsilon_2.$$

This proves Theorem 2 of the paper.

6 Proof of Lemma 7

Recalling the Setup. In Sect. 4 we defined three sets $\mathcal{S}_1$, $\mathcal{S}_2$, and $\mathcal{S}_3$. $\mathcal{S}_1$ is the set of (t_1, t_2) values that satisfy the constraints

$$\left\lceil \frac{t_1}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2}{r} \right\rceil \leq q_2, t_1 = t_2 \bmod r;$$

$\mathcal{S}_2$ is the set of $(t_1, t_2, \Delta t, c, \eta)$ values that satisfy the constraints

$$\left\lceil \frac{t_1 + \Delta t + c\eta}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2 + \Delta t}{r} \right\rceil \leq q_2, t_1 + c\eta = t_2 \bmod r,$$

$\mathcal{S}_3$ is the set of $(t_1, t_2, \Delta t, c, \eta)$ values that satisfy the constraints

$$\left\lceil \frac{t_1 + c\eta}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2 + \Delta t}{r} \right\rceil \leq q_2, t_1 + c\eta = t_2 + \Delta t \bmod r.$$

We further defined the following:

$$\mathsf{p}_1 = \sum_{\mathcal{S}_1} \mathsf{cp}_{\lambda}(t_1, t_2);$$
$$\mathsf{p}_2 = \sum_{\mathcal{S}_2} \mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c);$$
$$\mathsf{p}_3 = \sum_{\mathcal{S}_3} \mathsf{cp}_{\rho'}(t_1, t_2, \Delta t, c).$$

Lemma 7 claimed the following bounds for $\mathsf{p}_1, \mathsf{p}_2$ and p_3 (as long as $N \log r > 90$):

$$\mathsf{p}_1 \leq \frac{q_1 q_2 r}{N},$$
$$\mathsf{p}_2 \leq 6 \cdot (\log r)^2 \cdot \left(\frac{q_1 q_2 r}{N}\right)^2 + 18 \cdot (\log r)^3 \cdot \left(\frac{q_1 q_2 r}{N}\right),$$
$$\mathsf{p}_3 \leq 6 \cdot (\log r)^2 \cdot \left(\frac{q_1 q_2 r}{N}\right)^2 + 18 \cdot (\log r)^3 \cdot \left(\frac{q_1 q_2 r}{N}\right).$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

In this section, we establish these bounds.

BOUNDING p_1. For this we need to bound $\#\mathcal{S}_1$. This case is very simple. We observe the $t_1 \leq q_1 r$, so there are at most $q_1 r$ choices for t_1. Once t_1 is fixed, given the constraints $t_1 = t_2 \bmod r$ and $t_2 \leq q_2 r$, there are at most q_2 choices for t_2. Thus, we have

$$\#\mathcal{S}_1 \leq q_1 q_2 r,$$

which, using (5), gives the bound

$$\mathsf{p}_1 = \sum_{\mathcal{S}_1} \mathsf{cp}_\lambda(t_1, t_2) \leq \#\mathcal{S}_1 \cdot \frac{1}{N} \leq \frac{q_1 q_2 r}{N}.$$

TOWARDS BOUNDING p_2: COUNTING OVER t_1, t_2 AND Δt. This is the most involved part of the calculations. For simplicity of notation we define the function

$$\zeta(\alpha) := (\sqrt{2\alpha N} + 3)^2 = 2\alpha N + 6\sqrt{2\alpha N} + 9.$$

Recall that $\mathcal{S}_2$ is the set of all $(t_1, t_2, \Delta t, c, \eta)$ satisfying

$$\left\lceil \frac{t_1 + \Delta t + c\eta}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2 + \Delta t}{r} \right\rceil \leq q_2, t_1 + c\eta = t_2 \bmod r.$$

We begin by fixing a choice of c and η. We want to bound the number of choices for $(t_1, t_2, \Delta t)$. For this we relax the constraints a little. Let $\mathcal{S}_2' = \mathcal{S}_2'(c, \eta)$ be the set of values for $(t_1, t_2, \Delta t)$ satisfying

$$t_1 \leq q_1 r, \Delta t \leq q_2 r, t_2 \leq q_2 r, t_1 + c\eta = t_2 \bmod r.$$

Now we fix a real number $\alpha > 0$, and split $\mathcal{S}_2'$ into two disjoint sets:

$$\mathcal{S}_2'^{+}[\alpha] := \left\{ (t_1, t_2, \Delta t) \in \mathcal{S}_2' \mid \max(t_1, \Delta t) \geq \sqrt{2\alpha N} + 3 \right\},$$
$$\mathcal{S}_2'^{-}[\alpha] := \left\{ (t_1, t_2, \Delta t) \in \mathcal{S}_2' \mid \max(t_1, \Delta t) < \sqrt{2\alpha N} + 3 \right\}.$$

For $\mathcal{S}_2'^{+}[\alpha]$, there are at most $q_1 r$ choices for t_1 and at most $q_2 r$ choices for Δt, and for each of these choices, we have at most q_2 choices for t_2. Thus,

$$\#\mathcal{S}_2'^{+}[\alpha] \leq q_1 q_2^2 r^2.$$

For $\mathcal{S}_2'^{-}[\alpha]$, there are at most $\sqrt{2\alpha N} + 3$ choices for t_1 and at most $\sqrt{2\alpha N} + 3$ choices for Δt, and for each of these choices, since choosing t_1 also fixes $t_2 \bmod r$, we have at most q_2 choices for t_2. Thus,

$$\#\mathcal{S}_2'^{-}[\alpha] \leq (\sqrt{2\alpha N} + 3)^2 \cdot q_2 = \zeta(\alpha) \cdot q_2.$$

When $(t_1, t_2, \Delta t) \in \mathcal{S}_2'^{+}[\alpha]$,

$$t_1 + t_2 + \Delta t + c\eta \geq \sqrt{2\alpha N} + 3,$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

so that according to (6):

$$\mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c) \leq \frac{e^{-\alpha}}{N^2}.$$

When $(t_1, t_2, \Delta t) \in \mathcal{S}_2'^{-}[\alpha]$, (7) gives us

$$\mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c) \leq \frac{1}{N^2}.$$

Let

$$\begin{aligned}
\mathsf{p}_2(c, \eta) &:= \sum_{\mathcal{S}_2'} \mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c) \\
&= \sum_{\mathcal{S}_2'^{+}[\alpha]} \mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c) + \sum_{\mathcal{S}_2'^{-}[\alpha]} \mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c) \\
&\leq q_1 q_2^2 r^2 \cdot \frac{e^{-\alpha}}{N^2} + \zeta(\alpha) \cdot q_2 \cdot \frac{1}{N^2} \\
&= \frac{q_2}{N^2} \cdot \left[q_1 q_2 r^2 \cdot e^{-\alpha} + \zeta(\alpha) \right].
\end{aligned}$$

Towards Bounding p_2: Counting over c and η. We next bound the number of choices for (c, η) that satisfy the constraints. Again, we relax the constraints a little. Let $\mathcal{T}$ be the set of (c, η) values such that

$$c\eta \leq q_1 r.$$

Next we fix $d = \mathsf{gcd}(c, r)$. Let $\mathcal{T}[d]$ denote the set

$$\{(c, \eta) \in \mathcal{T} \mid \mathsf{gcd}(c, r) = d\}.$$

c now takes values over multiples of d. We split the counting into two parts:

- When $c \leq q_1 d$, we recall that η is defined as the smallest solution to $t_1 + c\eta = t_2 \bmod r$. From elementary number theory, we have $\eta \leq \frac{r}{d}$. Thus, there are q_1 choices of c and for each there are $\frac{r}{d}$ choices for η, so in all there are $\frac{q_1 r}{d}$ such choices for η and c.
- When $c > q_1 d$, we use the bounds $c \leq q_1 r$ and $\eta \leq \frac{q_1 r}{c}$. Let $z = \frac{c}{d}$. Thus, as c runs over all multiples of d from $(q_1 + 1) \cdot d$ to $q_1 r$, z takes all integer values from $q_1 + 1$ to $\frac{q_1 r}{d}$. Thus, the number of choices for η and c with $c > q_1 d$ is

$$\sum_{z=q_1+1}^{\frac{q_1 r}{d}} \frac{q_1 r}{zd} = \frac{q_1 r}{d} \cdot \sum_{z=q_1+1}^{\frac{q_1 r}{d}} \frac{1}{z} \leq \frac{q_1 r}{d} \cdot \log\left(\frac{r}{d}\right),$$

  the last step following from Lemma 2.

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

Putting these two together, we get

$$\#\mathcal{T}[d] \leq \frac{q_1 r}{d} \cdot \left(1 + \log\left(\frac{r}{d}\right)\right).$$

Now, d can take values over all factors of r, so we have

$$\begin{aligned}\#\mathcal{T} &= \sum_{d|r} \#\mathcal{T}[d] \leq \sum_{d|r} \frac{q_1 r}{d} \cdot \left(1 + \log\left(\frac{r}{d}\right)\right) \\ &\leq \sum_{d|r} \frac{q_1 r}{d} \cdot (1 + \log r) \leq q_1 \cdot (1 + \log r) \sum_{d|r} \frac{r}{d} \\ &\leq q_1 \cdot (1 + \log r) \cdot \sigma(r),\end{aligned}$$

the last step coming from Lemma 4.

Finally, we observe that whenever $(t_1, t_2, \Delta t, c, \eta) \in \mathcal{S}_2$, we have $(t_1, t_2, \Delta t) \in \mathcal{S}_2'(c, \eta)$, and $(c, \eta) \in \mathcal{T}$. Hence,

$$\mathsf{p}_2 = \sum_{\mathcal{S}_2} \mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c) \leq \sum_{\mathcal{T}} \sum_{\mathcal{S}_2'} \mathsf{cp}_{\lambda\rho}(t_1, t_2, \Delta t, c) = \sum_{\mathcal{T}} \mathsf{p}_2(c, \eta).$$

This gives us the bound

$$\mathsf{p}_2 \leq \frac{q_1 q_2}{N^2} \cdot (1 + \log r) \cdot \sigma(r) \cdot \left[q_1 q_2 r^2 \cdot e^{-\alpha} + \zeta(\alpha)\right]. \tag{12}$$

Bounding p_3. Recall that $\mathcal{S}_3$ is the set of all $(t_1, t_2, \Delta t, c, \eta)$ satisfying

$$\left\lceil \frac{t_1 + c\eta}{r} \right\rceil \leq q_1, \left\lceil \frac{t_2 + \Delta t}{r} \right\rceil \leq q_2, t_1 + c\eta = t_2 + \Delta t \bmod r.$$

The set $\mathcal{S}_3$ is almost identical to the set $\mathcal{S}_2$. However, the counting arguments are identical to those for p_2, as the relaxation of the constraints is valid for p_2 as well as p_3. Combined with (8), we have

$$\mathsf{p}_3 = \sum_{\mathcal{S}_3} \mathsf{cp}_{\rho'}(t_1, t_2, \Delta t, c) = \sum_{\mathcal{S}_3} \mathsf{cp}_{\lambda\rho}(t_1, t_2, 0, c) \leq \sum_{\mathcal{T}} \sum_{\mathcal{S}_2'} \mathsf{cp}_{\lambda\rho}(t_1, t_2, 0, c).$$

Thus, we have

$$\mathsf{p}_3 \leq \frac{q_1 q_2}{N^2} \cdot (1 + \log r) \cdot \sigma(r) \cdot \left[q_1 q_2 r^2 \cdot e^{-\alpha} + \zeta(\alpha)\right].$$

Simplifying the Bounds. Now we make a series of generous relaxations to get a simple easy-to-see bound for p_2 and p_3. Under the assumption that $\sqrt{2\alpha N}+3 \leq \sqrt{3\alpha N}$, we have $\zeta(\alpha) \leq 3\alpha N$. The assumption can be written as

$$(\sqrt{3} - \sqrt{2}).\sqrt{\alpha N} \geq 3.$$

In other words,

$$\alpha N \geq 9(\sqrt{3} + \sqrt{2})^2 = 9(5 + 2\sqrt{6}).$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security, ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

Now, $2\sqrt{6} < 5$, so a sufficient condition to ensure this is $\alpha N \geq 90$. We now put $\alpha = \log r$, and observe in passing that the ensuing assumption that $N \log r \geq 90$ is quite reasonable. For this choice of α, we have

$$\zeta(\alpha) \leq 3N \log r, \tag{13}$$

and

$$e^{-\alpha} = \frac{1}{r}. \tag{14}$$

Since $(5/3) \cdot \log r \geq 1$ for $r \geq 2$, we have

$$1 + \log r < \frac{5}{3} \log r + \log r = \frac{8}{3} \log r. \tag{15}$$

Finally, to bound $\sigma(r)$, we use Lemma 5, which gives us

$$\sigma(r) < 3r \log r. \tag{16}$$

Plugging (13)–(16) into (12), we have

$$\begin{aligned} \mathsf{p}_2 &\leq \frac{q_1 q_2}{N^2} \cdot 3r \log r \cdot \frac{8}{3} \log r \cdot (q_1 q_2 r^2 \cdot \frac{1}{r} + 3N \log r) \\ &= 8 \cdot (\log r)^2 \cdot \left(\frac{q_1 q_2 r}{N}\right)^2 + 24 \cdot (\log r)^3 \cdot \left(\frac{q_1 q_2 r}{N}\right). \end{aligned}$$

Similarly,

$$\mathsf{p}_3 \leq 8 \cdot (\log r)^2 \cdot \left(\frac{q_1 q_2 r}{N}\right)^2 + 24 \cdot (\log r)^3 \cdot \left(\frac{q_1 q_2 r}{N}\right).$$

This completes the proof of Lemma 7.

7 Conclusion and Future Work

We studied the iterated random function problem, and proved the first bound in this setting that is tight up to a factor of $(\log r)^3$. In previous work, the iterated random function problem was seen as a special case of CBC-MAC based on a random function f. We obtained our bound by analysing the probability of a common class of collision attacks, and applying Patarin's H-coefficient technique to bound the advantage of distinguishing f^r from f. Trying to improve the $(\log r)^3$ factor in the security bound is an interesting topic for future work.

References

1. Bellare, M., Kilian, J., Rogaway, P.: The Security of Cipher Block Chaining. In: CRYPTO 1994. LNCS, vol. 839, pp. 341–358 (1994)
2. Bellare, M., Kilian, J., Rogaway, P.: The Security of the Cipher Block Chaining Message Authentication Code. J. Comp. Syst. Sci. 61(3), 362–399 (2000)

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security, ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

3. Bellare, M., Pietrzak, K., Rogaway, P.: Improved Security Analyses for CBC MACs. In: CRYPTO 2005. LNCS, vol. 3621, pp. 527–545 (2005)
4. Bellare, M., Ristenpart, T., Tessaro, S.: Multi-instance Security and Its Application to Password-Based Cryptography. In: CRYPTO 2012. vol. 7417, pp. 312–329. Springer (2012)
5. Bellare, M., Rogaway, P.: The Security of Triple Encryption and a Framework for Code-Based Game-Playing Proofs. In: EUROCRYPT 2006. LNCS, vol. 4004, pp. 409–426 (2006)
6. Berke, R.: On the security of iterated MACs. Ph.D. thesis, ETH Zürich (2003)
7. Bernstein, D.J.: A short proof of the unpredictability of cipher block chaining (January 2005), `http://cr.yp.to/antiforgery/easycbc-20050109.pdf`
8. Bossi, S., Visconti, A.: What Users Should Know About Full Disk Encryption Based on LUKS. In: CANS 2015. LNCS, vol. 9476, pp. 225–237 (2015)
9. Chen, S., Steinberger, J.P.: Tight Security Bounds for Key-Alternating Ciphers. In: EUROCRYPT 2014. LNCS, vol. 8441, pp. 327–350 (2014)
10. Dodis, Y., Gennaro, R., Håstad, J., Krawczyk, H., Rabin, T.: Randomness Extraction and Key Derivation Using the CBC, Cascade and HMAC Modes. In: CRYPTO 2004. LNCS, vol. 3152, pp. 494–510 (2004)
11. Dodis, Y., Ristenpart, T., Steinberger, J.P., Tessaro, S.: To Hash or Not to Hash Again? (In)Differentiability Results for H^2 and HMAC. In: CRYPTO 2012. LNCS, vol. 7417, pp. 348–366 (2012)
12. Ferguson, N., Schneier, B.: Practical Cryptography. Wiley (2003)
13. Gaži, P., Pietrzak, K., Rybár, M.: The Exact PRF-Security of NMAC and HMAC. In: CRYPTO 2014. LNCS, vol. 8616, pp. 113–130 (2014)
14. Minaud, B., Seurin, Y.: The Iterated Random Permutation Problem with Applications to Cascade Encryption. In: CRYPTO 2015. LNCS, vol. 9215, pp. 351–367 (2015)
15. Nandi, M.: A Simple and Unified Method of Proving Indistinguishability. In: INDOCRYPT 2006. LNCS, vol. 4329, pp. 317–334 (2006)
16. Patarin, J.: The "Coefficients H" Technique. In: SAC 2008. LNCS, vol. 5381, pp. 328–345 (2008)
17. Preneel, B., van Oorschot, P.C.: MDx-MAC and Building Fast MACs from Hash Functions. In: CRYPTO 1995. LNCS, vol. 963, pp. 1–14 (1995)
18. Robin, G.: Grandes valeurs de la fonction somme des diviseurs et hypothèse de Riemann. J. Math. Pures Appl. 63, 187–213 (1984)
19. Sönmez Turan, M., Barker, E., Burr, W., Chen, L.: Recommendation for Key Derivation Using Pseudorandom Functions (Revised). NIST Special Publication 800-132, National Institute of Standards and Technology (NIST) (December 2010)
20. Wagner, D., Goldberg, I.: Proofs of Security for the Unix Password Hashing Algorithm. In: ASIACRYPT 2000. LNCS, vol. 1976, pp. 560–572 (2000)
21. Wuille, P.: Bitcoin Network Graphs (2017), `http://bitcoin.sipa.be/`
22. Yao, F.F., Yin, Y.L.: Design and Analysis of Password-Based Key Derivation Functions. In: CT-RSA 2005. LNCS, vol. 3376, pp. 245–261 (2005)
23. Yao, F.F., Yin, Y.L.: Design and Analysis of Password-Based Key Derivation Functions. IEEE Transactions on Information Theory 51(9), 3292–3297 (2005)

A Proofs of Lemmas from Sect. 2

Lemma 1. *Let $\alpha > 0$ be a real number. Then, for $i \geq \sqrt{2\alpha N} + 1$, we have*

$$\beta(i) \leq e^{-\alpha}.$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security, ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

Proof. Recall (1):

$$N^{\underline{i}} := \prod_{j=0}^{i-1}(N-j),$$

so that from (2):

$$\beta(i) := \frac{N^{\underline{i}}}{N^i} = \prod_{j=0}^{i-1}\left(1-\frac{j}{N}\right).$$

As $1-x \leq e^{-x}$ for any $x \in \mathbb{R}$,

$$\left(1-\frac{j}{N}\right) \leq e^{-\frac{j}{N}},$$

and we get

$$\beta(i) \leq \prod_{j=0}^{i-1} e^{-\frac{j}{N}} = e^{-\sum_{j=0}^{i-1}\frac{j}{N}}.$$

For $i \geq \sqrt{2\alpha N}+1$, we have

$$(i-1)^2 \geq 2\alpha N,$$

and we get

$$-\sum_{j=0}^{i-1} j = -\frac{i(i-1)}{2} \leq -\frac{(i-1)^2}{2} \leq -\alpha N,$$

so that

$$\beta(i) \leq e^{-\frac{\alpha N}{N}} \leq e^{-\alpha}.$$

□

Lemma 2. *For any two positive integers a and b with $b \geq a$,*

$$\sum_{i=a+1}^{b} \frac{1}{i} \leq \log\left(\frac{b}{a}\right)$$

Proof. For any positive integer i, we know that

$$e^{-\frac{1}{i}} \geq 1-\frac{1}{i} = \frac{i-1}{i},$$

so

$$e^{\frac{1}{i}} \leq \frac{i}{i-1}.$$

Thus,

$$e^{\left(\sum_{i=a+1}^{b}\frac{1}{i}\right)} = \prod_{i=a+1}^{b} e^{\frac{1}{i}} \leq \prod_{i=a+1}^{b}\frac{i}{i-1} = \frac{b}{a}.$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

Taking logarithms, we have

$$\sum_{i=a+1}^{b} \frac{1}{i} \leq \log\left(\frac{b}{a}\right).$$

□

Lemma 3. *For any positive integer b,*

$$\mathsf{d}(b) < 2\sqrt{b}.$$

Proof. Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x. For any divisor a of b such that $a > \sqrt{b}$, we have a unique $k < \sqrt{b}$ such that $ak = b$. Since k can have at most $\left\lceil \sqrt{b} \right\rceil - 1$ values, b can have at most $\left\lceil \sqrt{b} \right\rceil - 1 < \sqrt{b}$ divisors greater than $\sqrt{b}$. A similar reasoning shows that the number of divisors of b not greater than $\sqrt{b}$ is at most $\sqrt{b}$. Thus

$$\mathsf{d}(b) < 2\sqrt{b}.$$

□

Lemma 4. *For any positive integer b,*

$$\sum_{a|b} \frac{b}{a} = \sigma(b).$$

Proof. For a divisor a of b, suppose $ak = b$. Then k is also positive, so $k|b$. Moreover, k is distinct for each distinct a. As there are as many distinct values of a as there are divisors of b, k also runs precisely over the set of divisors of b. Hence

$$\sum_{a|b} \frac{b}{a} = \sum_{a|b} k = \sigma(b).$$

□

Lemma 5. *For any positive integer $b \geq 2$,*

$$\sigma(b) < 3b \log b.$$

Proof. Robin [18] showed that for all $b \geq 3$,

$$\sigma(b) < 1.7811 \cdot b \log\log b + \frac{0.6483 \cdot b}{\log\log b},$$

Since

$$(\log\log 10)^2 > 0.6483,$$

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

for $b \geq 10$ we have

$$\frac{0.6483 \cdot b}{\log \log b} < b \log \log b.$$

Or, as $\log \log b < \log b$ for $b > 1$, we have

$$\sigma(b) < 3b \log b$$

for $b \geq 10$. We can evaluate this inequality for smaller integer values of b, and find that the inequality is satisfied for any positive integer $b \geq 2$. □

Bhaumik, Ritam; Datta, Nilanjan; Dutta, Avijit; Mouha, Nicky; Nandi, Mrudil.
"The Iterated Random Function Problem."
Paper presented at The 23rd Annual International Conference on the Theory and Application of Cryptology and Information Security,
ASIACRYPT 2017, Hong Kong, China. December 3, 2017 - December 7, 2017.

A Method for Improving High-Insertion-Loss Measurements with a Vector Network Analyzer*

Jeffrey A. Jargon and Dylan F. Williams

National Institute of Standards and Technology, 325 Broadway, M/S 672.03, Boulder, CO 80305 USA
Email: jeffrey.jargon@nist.gov, Tel: +1.303.497.4961

***Abstract* — We present a method for improving high-insertion-loss measurements with a calibrated vector network analyzer (VNA) requiring only two additional pieces of hardware. By utilizing an amplifier and an attenuator, and measuring wave-parameters rather than scattering-parameters, we are able to increase the dynamic range of our measurements while decreasing uncertainties due to the noise floor of the VNA. We compare the results of our technique to standard methods for insertion-loss values up to 110 dB.**

***Index Terms* — attenuator, calibration, high insertion loss, measurement, uncertainty, vector network analyzer.**

I. Introduction

The need to accurately characterize high values of insertion loss is critical in applications such as line-of-sight channel measurements and near- to mid-range antenna measurements. In our case, we are currently participating in a cooperative effort between the National Institute of Standards and Technology (NIST) and the National Telecommunications and Information Administration (NTIA) to quantify the agreement among four different types of channel-sounder systems for both conducted and free-field environments at 3.5 GHz. The VNA is serving as a reference to which the other three systems are being compared, since we can use it to provide traceable, calibrated measurements with uncertainties characterized by the NIST Microwave Uncertainty Framework [1].

In a free-field environment, we will encounter channels with insertion losses varying from 50 dB to 100 dB, depending on the separation distance between transmit and receive antennas. Standard methods for measuring the upper end of these values are inadequate, as the noise floor of the VNA limits our ability to accurately characterize such channels.

In this paper, we describe a straightforward method to extend the dynamic range of our VNA measurements by measuring wave-parameters, as opposed to scattering-parameters (*S*-parameters), and utilizing an amplifier and an attenuator. In the following sections, we describe our measurement setup, and compare results from this technique to standard methods.

II. Measurement Setup

Figure 1 shows a simplified schematic diagram of a four-sampler VNA. The source is switched between ports 1 and 2 so all four *S*-parameters can be calculated from the measured incident and reflected signals. Normally, a calibration is first performed to correct for non-idealities in the VNA, and then calibrated *S*-parameters of devices under test (DUTs) can be measured. The setup, shown in Figure 1, works well for moderate insertion losses, but as these values increase, the measurements become noisier due to the inadequate dynamic range of the VNA. In an attempt to overcome this problem, we developed a method that makes use of an amplifier between the source and the couplers on port 1. With the increased power incident on port 1, we need to place a compensating attenuator between the coupler and the receiver responsible for measuring the incident power to prevent damage to the receiver. Figure 2 illustrates these modifications, where a 20 dB amplifier and a 20 dB attenuator have been inserted. For measurements with this setup, we cannot risk damage to the receiver responsible for measuring the transmitted signal on port 2 when power is applied to port 1 and a DUT is connected between the two ports. Thus, for this case, we restrict ourselves to measuring DUTs with at least 20 dB of attenuation.

Prior to making measurements with the modified setup, we perform a calibration with the amplifier removed, but with the attenuator left in. With the amplifier removed, we can ensure the calibration is performed in the linear regime of the VNA. The attenuator must be left in, however, for the calibration to remain valid. Additionally, we are required to measure wave-parameters rather than *S*-parameters, and our calibrations must be performed with an eight-term model rather than a twelve-term model [2]. The reason is that the switch terms, which correct for differences in the reflection coefficients of the terminating resistor switched between ports 1 and 2, will change with the amplifier placed in the system. Wave-parameters automatically compensate for this phenomena since incident and reflected signals are directly measured [3].

The wave-parameter calibration begins by transforming the uncalibrated wave-parameter measurements of each standard into uncalibrated *S*-parameters. Then, we apply a calibration designed to work on switch-term-corrected data. If the user wishes to determine calibrated wave-parameters, two additional terms are required: the magnitudes at each frequency, which can be determined with a calibrated power meter, and the phase relationships among frequencies, which can be determined with a characterized comb generator. We should note that for this particular application, neither magnitude nor phase calibrations are required since we are ultimately taking ratios of the measured wave-parameters and calculating *S*-parameters.

Jargon, Jeffrey; Williams, Dylan.
"A Method for Improving High-Insertion-Loss Measurements with a Vector Network Analyzer."
Paper presented at 89th ARFTG Microwave Measurement Symposium, Honolulu, HI, United States. June 9, 2017 - June 9, 2017.

TABLE I
COMPARING HIGH INSERTION-LOSS MEASUREMENTS OF FOUR DIFFERENT CALIBRATIONS

Attenuator Setting (dB)	$\|S_{21}\|$ Mean ± Std. Dev. (dB)			
	S-Parameters (-17 dBm)	S-Parameters (0 dBm)	Wave-Parameters (0 dBm)	Wave Parameters (0 dBm + Amp)
60	-60.69 ± 0.12	-60.71 ± 0.06	-60.71 ± 0.06	-60.70 ± 0.06
70	-70.72 ± 0.34	-70.74 ± 0.08	-70.73 ± 0.08	-70.72 ± 0.06
80	-80.79 ± 1.08	-80.95 ± 0.16	-80.71 ± 0.17	-80.68 ± 0.05
90	-90.81 ± 3.47	-90.87 ± 0.49	-90.64 ± 0.50	-90.63 ± 0.08
100	-97.27 ± 5.44	-100.70 ± 1.51	-100.58 ± 1.55	-100.52 ± 0.15
110	-98.66 ± 5.56	-110.59 ± 4.75	-110.26 ± 4.37	-110.52 ± 0.43

III. MEASUREMENT COMPARISON

We performed four separate calibrations, all of which made use of an Open-Short-Load-Thru (OSLT) calibration kit with Type-N coaxial connectors. Physical models of the calibration standards were developed and validated with a multiline Thru-Reflect-Line (TRL) calibration within the NIST Microwave Uncertainty Framework [4].

The first three calibrations were made with the standard VNA setup, as illustrated in Figure 1. In the first one, we measured *S*-parameters at an input power of -17 dBm, the default power setting on the VNA. The second calibration was also measured with *S*-parameters, but at an increased input power of 0 dBm. The third calibration was performed with wave-parameters also at 0 dBm, so we could verify that measuring wave-parameters and *S*-parameters provides comparable results. And the fourth calibration made use of the modified setup of Figure 2, where the amplifier was removed during calibration, and re-inserted for the actual DUT measurements. Our DUT was a variable attenuator with a range of 0 to 110 dB with steps of 10 dB. All of the measurements were performed on a frequency grid between 2.7-3.7 GHz (the bandwidth of our amplifier) with a spacing of 1 MHz (1001 points) and an IF bandwidth of 100 Hz with no averaging.

Table 1 lists the mean values and standard deviations of the magnitudes of S_{21} calculated over the measured frequencies at attenuator settings of 60-110 dB for the four calibrations.

Although all of the calibrations provided comparable mean values up to 90 dB, the standard deviations increased much more drastically with increased attenuator settings for the case where *S*-parameters were measured at an input power of -17 dBm. At this input power, the 100 dB and 110 dB measurements were limited by the VNA's noise floor, and thus the values were incorrect. At the highest attenuator settings, both the *S*-parameter and wave-parameter measurements at an input power of 0 dBm were still able to provide reasonable mean values, but their respective standard deviations became increasingly large. In contrast, the wave-parameter measurements with the amplifier and attenuator resulted in standard deviations approximately ten times lower than with the standard setup at these highest settings.

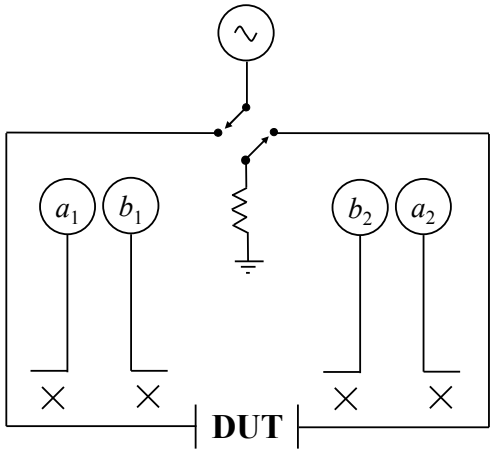


Fig. 1. Simplified schematic diagram of a four-sampler VNA.

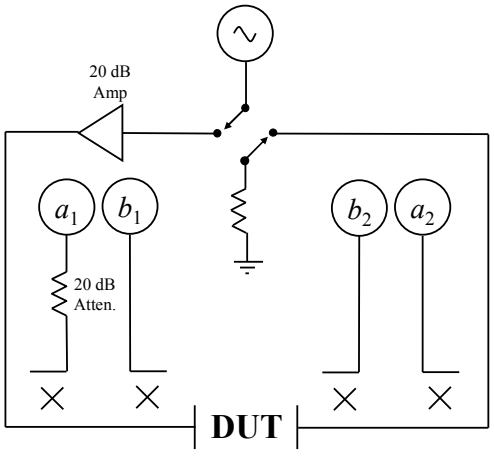


Fig. 2. Simplified schematic diagram of a four-sampler VNA with a 20 dB amplifier and 20 dB attenuator inserted for improved high-insertion-loss measurements.

Jargon, Jeffrey; Williams, Dylan.
"A Method for Improving High-Insertion-Loss Measurements with a Vector Network Analyzer."
Paper presented at 89th ARFTG Microwave Measurement Symposium, Honolulu, HI, United States. June 9, 2017 - June 9, 2017.

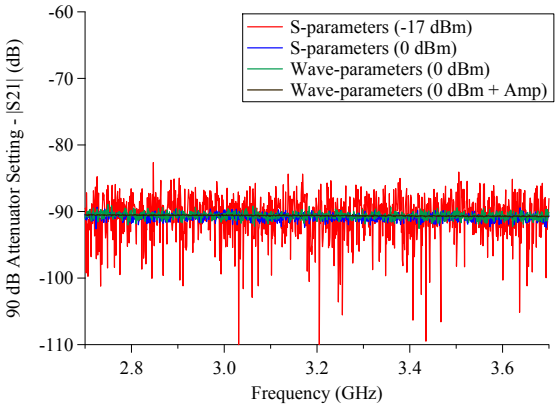


Fig. 3. Comparing measurements of $|S_{21}|$ at the 90 dB attenuator setting for four different calibrations.

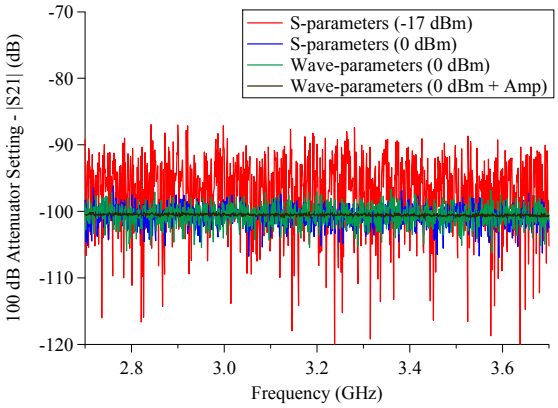


Fig. 4. Comparing measurements of $|S_{21}|$ at the 100 dB attenuator setting for four different calibrations.

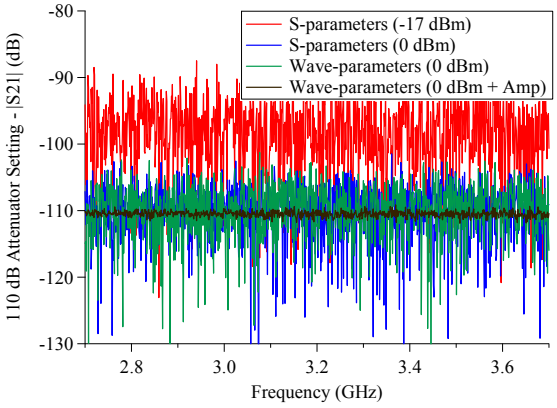


Fig. 5. Comparing measurements of $|S_{21}|$ at the 110 dB attenuator setting for four different calibrations.

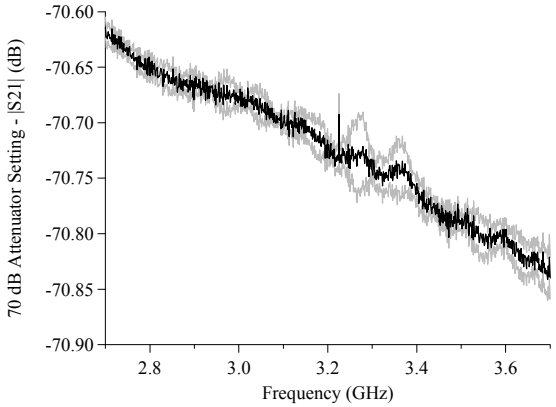


Fig 6. Nominal calibrated measurements (black curve) and 95% confidence intervals (grey curves) of $|S_{21}|$ for the 70 dB setting by use of the wave-parameter measurements at an input power of 0 dBm with the amplifier and attenuator.

Figures 3-5 illustrate the measurements of $|S_{21}|$ at the 90, 100, and 110 dB settings for the four different calibrations. Figure 6 plots the measurements and 95 % confidence intervals of $|S_{21}|$ for the variable attenuator at the 70 dB setting (a typical value we will likely encounter during free-field measurements) for the wave-parameter measurements at an input power of 0 dBm with the amplifier and attenuator.

IV. Conclusions

By utilizing only two additional pieces of hardware and measuring wave-parameters rather than *S*-parameters, we were able to increase the dynamic range of our VNA while decreasing uncertainties due to the noise floor. This modified setup decreased the standard deviations of $|S_{21}|$ by an order of magnitude for attenuator settings over 90 dB. Although we were limited to a maximum insertion loss of 110 dB due to our attenuator, we believe this method can provide adequate measurements for values up to approximately 130 dB.

For even higher values of insertion loss, encountered in applications such as non-line-sight channel measurements, we would most likely require additional hardware, including external couplers and a different calibration technique, such as Short-Open-Load-Attenuator. Additionally, we plan on examining the effects of performing high-insertion-loss measurements utilizing calibrations that include isolation terms. These topics remain the subject of future studies.

Jargon, Jeffrey; Williams, Dylan.
"A Method for Improving High-Insertion-Loss Measurements with a Vector Network Analyzer."
Paper presented at 89th ARFTG Microwave Measurement Symposium, Honolulu, HI, United States. June 9, 2017 - June 9, 2017.

ACKNOWLEDGEMENT

*This work was supported by the U.S. government, and is not subject to U.S. copyright.

The authors thank Gustavo Avolio, Paul Hale, Jeanne Quimby, and Damir Senic for their helpful discussions, and Robert Johnk and Chriss Hammerschmidt for the loan of their variable attenuator.

REFERENCES

[1] D. F. Williams, NIST Microwave Uncertainty Framework, Beta Version, www.nist.gov/services-resources/software/wafer-calibration-software, 2017.

[2] R. B. Marks, "Formulations of the Basic Vector Network Analyzer Error Model Including Switch Terms," 50th ARFTG Microwave Measurement Conference, Portland, OR, Dec. 1997.

[3] D. Senic, K. A. Remley, D. F. Williams, D. C. Ribeiro, C. M. Wang, and C. L. Holloway, "Radiated Power Based on Wave Parameters at Millimeter-wave Frequencies for Integrated Wireless Devices," 88th ARFTG Microwave Measurement Conference, Austin, TX, Dec. 2016.

[4] J. A. Jargon, D. F. Williams, and P. D. Hale, "Developing Models for Type-N Coaxial VNA Calibration Kits within the NIST Microwave Uncertainty Framework," 87th ARFTG Microwave Measurement Conference, San Francisco, CA, May 2016.

Jargon, Jeffrey; Williams, Dylan.
"A Method for Improving High-Insertion-Loss Measurements with a Vector Network Analyzer."
Paper presented at 89th ARFTG Microwave Measurement Symposium, Honolulu, HI, United States. June 9, 2017 - June 9, 2017.

Eurographics Workshop on 3D Object Retrieval (2017), pp. 1–11
I. Pratikakis, F. Dupont, and M. Ovsjanikov (Editors)

Point-Cloud Shape Retrieval of Non-Rigid Toys[†]

F. A. Limberger[‡1], R. C. Wilson[‡1], M. Aono[6], N. Audebert[9], A. Boulch[7], B. Bustos[7], A. Giachetti[5], A. Godil[13], B. Le Saux[8], B. Li[11], Y. Lu[12], H.-D. Nguyen[2,4], V.-T. Nguyen[2,3], V.-K. Pham[2], I. Sipiran[8], A. Tatsuma[6], M.-T. Tran[2], S. Velasco-Forero[10]

[1]Univeristy of York, United Kingdom, [2]University of Science, VNU-HCM, Vietnam, [3]University of Information Technology, VNU-HCM, Vietnam, [4]John von Neumann Institute, VNU-HCM, Vietnam, [5]University of Verona, Italy, [6]Toyohashi University of Technology, Japan, [7]Dept. Computer Science, University of Chile, Chile, [8]Dept. Engineering, Pontifical Catholic University of Peru, Peru, [9]ONERA, The French Aerospace Lab, France, [10]Centre de Morphologie Mathématique, MINES Paristech, France, [11]University of Southern Mississippi, USA, [12]Texas State University, USA, [13]National Institute of Standards and Technology, USA

Abstract

In this paper, we present the results of the SHREC'17 Track: Point-Cloud Shape Retrieval of Non-Rigid Toys. The aim of this track is to create a fair benchmark to evaluate the performance of methods on the non-rigid point-cloud shape retrieval problem. The database used in this task contains 100 3D point-cloud models which are classified into 10 different categories. All point clouds were generated by scanning each one of the models in their final poses using a 3D scanner. The retrieval performance is evaluated using seven commonly-used statistics (PR-plot, NN, FT, ST, E-measure, DCG, mAP). In total, there are 8 groups and 31 submissions taking part in *this contest. The evaluation results shown by this work suggest that researchers are* progressing *towards shape descriptors which can capture the main characteristics of 3D models, however, more tests still need to be made, since this is the first time we compare non-rigid signatures for point-cloud shape retrieval.*

Categories and Subject Descriptors (according to ACM CCS): H.3.3 [Computer Graphics]: Information Systems—Information Search and Retrieval

1. Introduction

With the rapid development of virtual reality (VR) and augmented reality (AR), especially in gaming, 3D data has become part of our everyday lives. Since the creation of 3D models is essential to these applications, we have been experiencing a large growth in the number of 3D models available on the Internet in the past few years. The problem now has been organizing and retrieving these models from databases. Researchers from all over the world are trying to create shape descriptors in a way to organize this huge amount of models, making use of many mathematical tools to create discriminative and efficient signatures to describe 3D shapes. The importance of shape retrieval is evidenced by the 11 years of the Shape Retrieval Contest (SHREC).

There are two distinct areas which concern shape retrieval: The first, non-rigid shape retrieval, which deals with the problem of articulations of the same shape [LGT*10, LGB*11, LZC*15], and second, comprehensive shape retrieval [BBC*10, LLL*14, SYS*16], which deals with any type of deformation, for example, scaling, stretching and even differences in topology. While comprehensive shape retrieval is more general, non-rigid shape retrieval is as important when it is necessary to carefully classify similar objects that are in distinct classes [PSR*16].

Three-dimensional point clouds are the immediate result of scans of 3D objects. Although there are efficient methods to create meshes from point clouds, sometimes this task can be complex, particularly when point-cloud data present missing parts or noisy surfaces, for example, fur or hair. In this paper, we are interested in the non-rigid shape retrieval task, therefore we propose to create a non-rigid point-cloud shape retrieval benchmark (PRoNTo: Point-Cloud Shape Retrieval of Non-Rigid Toys), which was produced given the necessity of testing non-rigid shape signatures computed directly from unorganized point clouds, *i.e.*, without any connectivity information. This is the first benchmark ever created to test, specifically, the performance of non-rigid point-cloud models.

This benchmark is important given the need to compare 3D non-rigid shapes based directly on a rough 3D scan of the object, which is a more difficult task than comparing signatures computed from well-formed 3D meshes. 3D scanners may introduce some sampling problems to the scanned models, given the difficulty of reaching all parts of the object by the scan head and given that some materials have specular properties and these can generate outliers.

Although some methods available in the literature use point sets to create their shape signatures, we have not seen these methods being used directly to address the non-rigid point-cloud shape-

DISCLAIMER: Any commercial product or company name in this paper is given for informational purposes only. Their use does not imply recommendation.

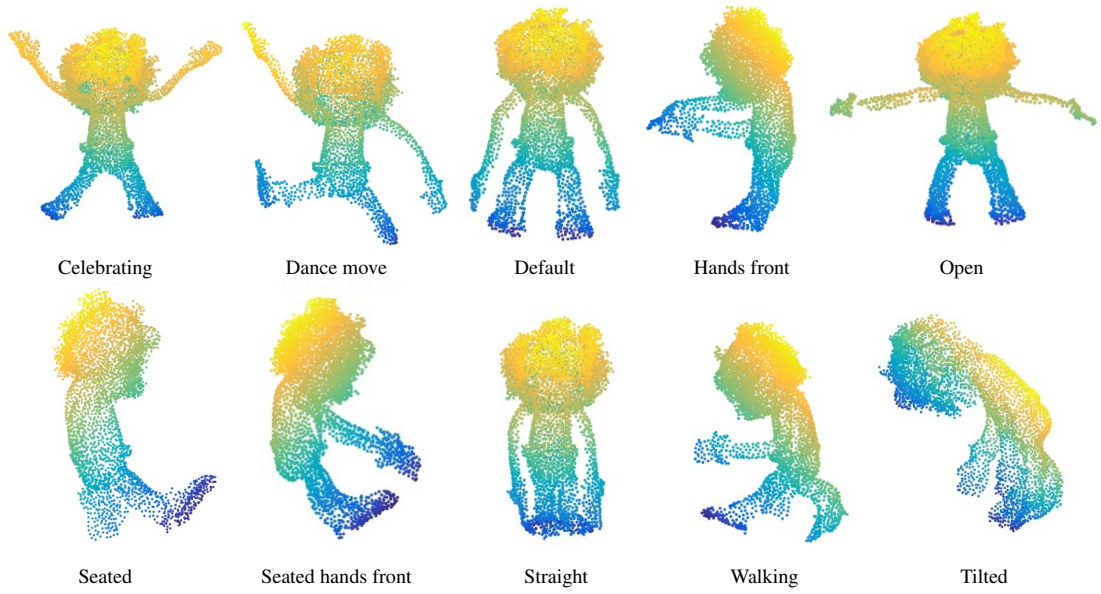


Figure 1: Different poses captured of the objects, showing model *Monster* as an example. Point clouds were coloured by Y and Z coordinates.

retrieval problem since there are no specific point-cloud datasets available for this purpose. Instead, these methods normally use mesh vertices, which sometimes can be a very bad idea, unless vertices are very well distributed along the surface of the shape.

2. Dataset

Our dataset consists of 100 models that are derived from 10 different real objects. Each real object was scanned in 10 distinct poses by articulating them around their joints. The different poses and each one of the objects used to create this database can be seen in Figures 1 and 2, respectively. After scanning all the poses we manually removed the supports used to scan the objects using MeshLab.

Objects were scanned using the Head & Face Color 3D Scanner of Cyberware. This scanner makes a 360 degrees scan around the object estimating x, y, and z coordinates of a vertical patch. The scanning process captures an array of digitized points and also the respective RGB colors although they are not used in this contest. The file format for the objects was chosen as the Object File Format (.off), which, in this case, contains only vertex information. We also resample models using the Poisson-Disk Sampling algorithm [CCS12] since the scan generates an arbitrary number of samples. This way, we control the sampling rate so that every model has approximately 4K points. Finally, we perform an arbitrary rotation of the model so that it is not always in the same orientation. The point clouds acquired by our scans suffer from common scanning problems like holes and missing parts resulted from self-occlusions of the shapes, and also from noise given that some toy's materials have specular reflection properties.

Figure 2: Different toys used to create the PRoNTo dataset.

3. Evaluation

The evaluation rules follow standard measures used in SHREC tracks in the past. We asked participants to submit up to 6 dissimilarity matrices. These matrices could be a result of different algorithms or different parameter settings, at the choice of the participant. A dissimilarity matrix is the result of a shape retrieval problem which gives the difference between every model in the database. It has the size $N \times N$, where N is the number of models of the dataset and the position (i, j) in the matrix gives the differ-

ence between models i and j. No class information is provided with the data, and supervised methods are not allowed in the track.

In total, seven standard quantitative evaluation measures were computed over the dissimilarity matrices submitted by the participants to test the retrieval accuracy of the algorithms: Precision-and-Recall (PR) curve, mean Average Precision (mAP), E-Measure (E), Discounted Cumulative Gain (DCG), Nearest Neighbor (NN), First-Tier (FT) and Second-Tier (ST).

4. Participants

During this contest, we had 8 groups taking part in the SHREC'17 PRoNTo contest and we received in total 31 dissimilarity-matrix submissions, as detailed below:

1. *MFLO-FV-IWKS*, *MFLO-SV-IWKS*, *PCDL-FV-IWKS*, *PCDL-SV-IWKS*, *GL-FV-IWKS* and *GL-SV-IWKS* submitted by Frederico A. Limberger and Richard C. Wilson.
2. *BoW-RoPS-1*, *BoW-RoPS-2*, *BoW-RoPS-DMF-3*, *BoW-RoPS-DMF-4*, *BoW-RoPS-DMF-5* and *BoW-RoPS-DMF-6* submitted by Minh-Triet Tran, Viet-Khoi Pham, Hai-Dang Nguyen and Vinh-Tiep Nguyen. Other team members: Thuyen V. Phan, Bao Truong, Quang-Thang Tran, Tu V. Ninh, Tu-Khiem Le, Dat-Thanh N. Tran, Ngoc-Minh Bui, Trong-Le Do, Minh N. Do and Anh-Duc Duong.
3. *POHAPT* and *BPHAPT* submitted by Andrea Giachetti.
4. *CDSPF* submitted by Atsushi Tatsuma and Masaki Aono.
5. *SQFD(HKS)*, *SQFD(WKS)*, *SQFD(SIHKS)*, *SQFD(WKS-SIHKS)* and *SQFD(HKS-WKS-SIHKS)* submitted by Benjamin Bustos and Ivan Sipiran.
6. *SnapNet* submitted by Bertrand Le Saux, Nicolas Audebert and Alexandre Boulch.
7. *AlphaVol1*, *AlphaVol2*, *AlphaVol3* and *AlphaVol4* submitted by Santiago Velasco-Forero.
8. *m3DSH-1*, *m3DSH-2*, *m3DSH-3*, *m3DSH-4*, *m3DSH-5* and *m3DSH-6* submitted by Bo Li, Yijuan Lu and Afzal Godil.

5. Methods

In the next sections, we detail all the participant methods that have successfully competed in the PRoNTo dataset contest. Experimental settings of each method are displayed at the end of each section.

5.1. Spectral Descriptors for Point Clouds, by Frederico Limberger and Richard Wilson

The key idea of this method is to test spectral descriptors computed directly from point clouds using different formulations for the computation of the Laplace-Beltrami operator (LBO). We test three different methods for computing the LBO: the Mesh-Free Laplace operator (MFLO), the Point-Cloud Laplace (PCDLaplace) [BSW09] and the Graph Laplacian (GL).

Our framework is as follows. We first compute the eigendecomposition of the different LBO methods. Then we compute local descriptors. We encode these local features using state-of-the-art encoding schemes (FV and SV). Furthermore, we compute the differences between shape signatures using Efficient Manifold Ranking. We now detail each part of our framework.

Laplace-Beltrami operator: The LBO is a linear operator defined as the divergence of the gradient, taking functions into functions over the 2D manifold $\mathcal{M}$

$$\Delta_{\mathcal{M}} f = -\bigtriangledown \cdot \bigtriangledown_{\mathcal{M}} f \tag{1}$$

given that f is a twice-differentiable real-valued function. Although we compute the LBO using three different methods, all these use the same parameters that are equivalent in each approach. The eigendecomposition of the LBO results in their eigenvalues and eigenfunctions, which are commonly known as the shape spectrum and these are further used to compute a local descriptor.

Local Descriptor: After computing the shape spectrum, we compute the Improved Wave Kernel Signature (IWKS) [LW15] which is a local spectral descriptor based on the Schrodinger equation and it is governed by the wave function $\psi(x,t)$.

$$i\Delta_{\mathcal{M}}\psi(x,t) = \frac{\partial \psi}{\partial t}(x,t), \tag{2}$$

The IWKS is an improved version of the WKS [ASC11]. It has a different weighting filter of the shape spectrum which captures, at the same time, the major structure of the object and its fine details, therefore being more informative than the WKS.

Encoding: For computing the encoding of local descriptors into global signatures for shape retrieval, we use state-of-the-art encodings: Fisher Vector (FV) [PD07] and Super Vector (SV) [ZYZH10]. These methods are based on the differences between descriptors and probabilistic distribution functions, which we approximate by Gaussian Mixture models. More details about these encodings can be found in [LW15].

Distances between signatures: Distances between signatures are computed using Efficient Manifold Ranking (EMR) [XBC*11], which accelerates the classic Manifold Ranking [ZWG*04]. EMR has similar evaluation performance to MR, however, it has much lower computation times when used in large databases.

Experimental settings: We compute the first 100 eigenvalues and eigenfunctions of the respective LBO using 15 nearest neighbours to compute the proximity graph for all methods. In PCDLaplace, we use the following additional parameters: $htype = ddr; hs = 2; rho = 2$. For the computation of the local descriptor, we use $iwksvar = 5$. For computing the Gaussian dictionary, we use the first 29 models of the database to create GMMs with 38 components for each signature frequency. For computing EMR we use 99 landmarks and we use k-means as the landmark selection method. The number of landmarks is usually chosen as a slightly smaller number than the number of models in total. For one model in the database, it takes approximately 15 seconds to compute the MFLO-IWKS, 8 seconds to compute the GL-IWKS and 11 seconds to compute the PCDL-IWKS. To compute the entire dissimilarity matrix it takes approximately 43 seconds with the FV and 56 seconds with the SV. All experiments were carried out in Matlab on a PC CPU i7-3770 3.4GHz, 8GB RAM.

Godil, Afzal.
"Point-Cloud Shape Retrieval of Non-Rigid Toys."
Paper presented at Eurographics 2017 Workshop on 3D Object Retrieval, April 23-24, 2017, Lyon, France, Lyon, France. April 23, 2017 - April 24, 2017.

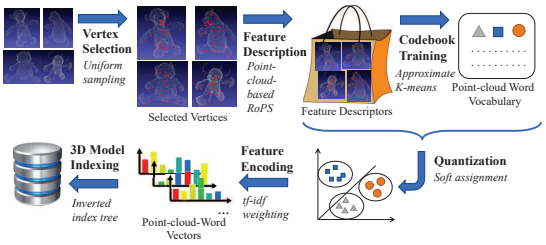


Figure 3: Bag-of-Words framework for 3D object retrieval

5.2. Bag-of-Words Framework for 3D Object Retrieval, by Minh-Triet Tran, Viet-Khoi Pham, Hai-Dang Nguyen and Vinh-Tiep Nguyen

We develop our framework for 3D object retrieval based on Bag-of-Words scheme for visual object retrieval [SZ03]. BoW method originated from text retrieval domain and is shown to be successfully applied on large-scale image [NNT*15] and 3D object retrieval [PSA*16]. Figure 3 illustrates the main components of our framework.

Preprocessing: Each point cloud is normalized into a unit cube and densified to reduce significant difference in the density between different parts.

Feature detector: For each model, we uniformly take random samples of $5\% \leq p_{Sampling} \leq 50\%$.

Feature descriptor: We describe the characteristics of the point cloud in a sphere with supporting radius r surrounding a selected vertex. We propose our point-cloud-based descriptor inspired by the idea of RoPS [GSB*13] to calculate the descriptor directly from a point cloud (without reconstructing faces). We first estimate the eigenvectors of the point cloud within a supporting radius r of a selected vertex, then transform the point cloud to achieve rotation invariant for the descriptor, and finally calculate the descriptor. We consider the supporting radius r from 0.01 to 0.1.

Codebook: All features extracted from the models are used to build a codebook with size relatively equal to 10% of the total number of features in the corpus, using Approximate K-Means.

Quantization: To reduce quantization error, we use soft-assignment [PCI*08] with 3 nearest neighbors.

Distance measure metric: instead of using a symmetric distance, we use L1, asymmetric distance measurement [ZJS13], to evaluate the dissimilarity of each pair of objects.

Our first two runs (1 and 2) are results of our BoW framework using random sampling with $p_{Sampling} = 45\%$ and codebook size of 18000. The radius of point-cloud based RoPS for run 1 and 2 are $r = 0.04$ and 0.05, respectively.

Each main component of our BoW framework is deployed on a different server. Codebook training module using Python 2.7 is deployed on Ubuntu 14.04 with 2.4 GHz Intel Xeon CPU E5-2620 v3, 64 GB RAM. It takes 30 minutes to create a codebook with 18,000 visual words from 180,000 features. 3D feature extraction and description module, written in C++, runs on Ubuntu 14.04, 2GHz Intel Xeon CPU E5-2620, 1GB RAM.

The retrieval process in Matlab R2012b with feature quantization and calculating the dissimilarity matrix is performed on Windows Server 2008 R2, 2.2GHz Intel Xeon CPU E5-2660, 12 GB RAM. The average time to calculate features of a model is 1-2 seconds and it takes on average 0.02 seconds to compare an object against all 100 objects.

5.2.1. Distance Matrix Fusion

With each setting for our BoW framework, we get a different retrieval model. We propose a simple method to linearly combine k distance matrices $\mathbf{D}_1, \mathbf{D}_2, ..., \mathbf{D}_k$ with different coefficients into a new distance matrix:

$$\mathbf{D}_{Fusion} = w_1\mathbf{D}_1 + w_2\mathbf{D}_2 + ...w_k\mathbf{D}_k. \quad (3)$$

Our objective is to take advantage of different retrieval models obtained from our framework with the expectation to increase the performance of the retrieval process.

We limit the number of seeds k of 2 or 3, and the value of a coefficient w_i is from 0 to 1, step 0.2. Our last four runs are combinations of Run1 and Run2 with different values of w_1 and w_2: **Run3**: $w_1 = 0.8$ and $w_2 = 0.6$; **Run4**: $w_1 = 1.0$ and $w_2 = 0.8$; **Run5**: $w_1 = 0.6$ and $w_2 = 0.2$; and **Run6**: $w_1 = 1.0$ and $w_2 = 0.4$. Experimental results show that the fusion runs can even yield better performance in retrieval than the two original seeds.

5.3. Simple meshing and Histogram of Area Projection Transform, by Andrea Giachetti

The method is based on simple automatic point cloud meshing followed by the estimation of the Histogram of Area Projection Transform descriptor [GL12]. Shapes are represented through a set of N voxelized maps encoding the area projected along the inner normal direction at sampled distances $R_i, i = 1..N$ in a spherical neighborhood of radius σ around each voxel center location $\vec{x}$. Values at different radii are weighted in order to have a scale-invariant behavior. Histograms of MAPT computed inside the objects are quantized in 12 bins and evaluated at 12 equally spaced radii values ranging from 3 to 39mm., with σ always taken as half the radius. Histograms computed at the different radii considered are concatenated creating an unique descriptor. Dissimilarity matrices are generated by measuring the histogram distances with the Jeffrey divergence.

This descriptor is robust against pose variation and inaccuracy due to holes, especially if histograms are estimated inside the shape only. For this reason we applied the HAPT estimation using the code publicly available at the web site www.andreagiachetti.it on closed meshes estimated on the original point clouds with two different procedures implemented as simple Meshlab [CCC*08] scripts.

We submitted matrices corresponding to each of these procedures.

Poisson reconstruction: In the first run, we just applied Poisson reconstruction [KBH]. Points' normals have been estimated on a 12 neighbors range, and the octree depth has been set equal to 9.

Ball Pivoting and Poisson: In the second run we first smoothed the point set using Moving Least squares, then we applied the ball pivoting method [BMR*99] to extract an open mesh. The mesh has been refined with triangle splitting and normals have been recomputed on the basis of the meshing. From mesh and normals obtained, a closed watertight mesh has been finally obtained with the Poisson reconstruction method.

Note that meshing would be not mandatory for the application of the method, as a point cloud implementation of the descriptor would be quite simple. However, we believe that Poisson meshing provides in general a better estimate of the inner part of the mesh and is effective in reconstructing missing parts in a reasonable way.

Meshlab scripts run in less than one second per model and HAPT estimation takes 10 seconds on average. Histogram distance estimation time, performed in Matlab, is negligible in comparison.

5.4. Covariance Descriptor with Statistics of Point Features, by Atsushi Tatsuma and Masaki Aono

For non-rigid human 3D model retrieval, we previously proposed the local feature extraction method [PSR*16] that calculates the histogram, mean, and covariance of geometric point features. In this track, we further calculate the skewness and kurtosis of geometric point features to obtain more discriminative local feature. 3D point-cloud object finally is represented with the covariance of the local features consisting of the histogram, mean, covariance, skewness, and kurtosis of geometric point features. We call our approach the Covariance Descriptor with Statistics of Point Features (CDSPF).

The overview of our approach is illustrated in Figure 4. We first calculate 4D point geometric feature $\mathbf{f} = [f_1, f_2, f_3, f_4]$ proposed in [WHH03]. The geometric feature is computed for every pair of points $\mathbf{p}_a$ and $\mathbf{p}_b$ in the point's k-neighborhood:

$$f_1 = \tan^{-1}(\mathbf{w}\cdot\mathbf{n}_b/\mathbf{u}\cdot\mathbf{n}_a), \quad (4)$$
$$f_2 = \mathbf{v}\cdot\mathbf{n}_b, \quad (5)$$
$$f_3 = \mathbf{u}\cdot(\mathbf{p}_b-\mathbf{p}_a/d), \quad (6)$$
$$f_4 = d, \quad (7)$$

where the normal vectors of $\mathbf{p}_a$ and $\mathbf{p}_b$ are $\mathbf{n}_a$ and $\mathbf{n}_b$, $\mathbf{u} = \mathbf{n}_a$, $\mathbf{v} = (\mathbf{p}_b-\mathbf{p}_a)\times\mathbf{u}/||(\mathbf{p}_b-\mathbf{p}_a)\times\mathbf{u}||$, $\mathbf{w} = \mathbf{u}\times\mathbf{v}$, and $d = ||\mathbf{p}_b-\mathbf{p}_a||$.

Next, we collect the point features in a 16-bin histogram $\mathbf{h}$. The index of histogram bin h is defined by the following formula:

$$h = \sum_{i=1}^{4} 2^{i-1} s(t, f_i), \quad (8)$$

where $s(t,f)$ is a threshold function defined as 0 if $f < t$ and 1 otherwise. The threshold value used for f_1, f_2 and f_3 is 0, while the threshold value for f_4 is the average value of f_4 in the k-neighborhood.

Furthermore, we calculate the mean, covariance matrix, skewness, and kurtosis of the point features. Let $\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N$ be the point features of size N. The mean μ, covariance matrix C, skewness s, and kurtosis k are calculated as follows [Mar70]:

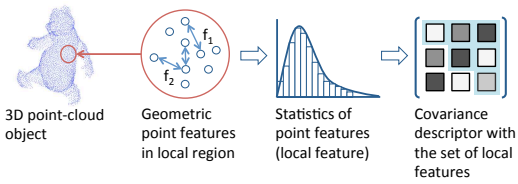


Figure 4: Overview of CDSPF extraction process.

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\mathbf{f}_i, \quad (9)$$
$$C = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{f}_i-\mu)(\mathbf{f}_i-\mu)^\top, \quad (10)$$
$$s = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\{(\mathbf{f}_i-\mu)^\top C^{-1}(\mathbf{f}_i-\mu)\}^3, \quad (11)$$
$$k = \frac{1}{N}\sum_{i=1}^{N}\{(\mathbf{f}_i-\mu)^\top C^{-1}(\mathbf{f}_i-\mu)\}^2. \quad (12)$$

Since the covariance matrix C lies on the Riemannian manifold of symmetric positive semi definite matrices, we map the covariance matrix onto a point in the Euclidean space by using Pennec et al.'s method [PFA06].

We finally obtain the local feature by concatenating the histogram, mean, covariance, skewness, and kurtosis of the point features. The local feature is normalized with the signed square rooting and ℓ_2 normalization [JC12]. To compare 3D point-cloud objects, we integrated the set of local features into a feature vector with the covariance descriptor approach [TPM06].

Since 3D point-cloud objects in the dataset do not have normal vector information, we used the Point Cloud Library [RC11] for estimating normal vector of each point. Moreover, we set the size of the neighborhood k to 30. We employ the Euclidean distance for the dissimilarity between two feature vectors.

The method was implemented in C++. Experiments were carried out under Debian Linux 8.7 on a CPU 3.4GHz Intel Core i7-6800K and 128GB DDR4 memory. The average time to calculate the shape descriptor for a 3D model is about 1.46 seconds and it takes approximately 10 seconds to compute the dissimilarity matrix.

5.5. Signature Quadratic Form Distance on Spectral Descriptors, by Ivan Sipiran and Benjamin Bustos

Our method combines the flexibility of the Signature Quadratic Form Distance (SQFD) [BUS09] with the robustness of intrinsic spectral descriptors. On the one hand, the SQFD distance has proven to be effective in multimedia domains where objects are represented as a collection of local descriptors [SLBS16]. On the other hand, intrinsic descriptors are useful to keep robustness to non-rigid transformations. Our proposal consists of representing the input 3D point cloud as a set of local descriptors which will be compared through the use of the SQFD distance.

Godil, Afzal.
"Point-Cloud Shape Retrieval of Non-Rigid Toys."
Paper presented at Eurographics 2017 Workshop on 3D Object Retrieval, April 23-24, 2017, Lyon, France, Lyon, France. April 23, 2017 - April 24, 2017.

Let P be a 3D point cloud. The first step of our method is to compute a set of local descriptors on P. The spectral descriptors depends on the computation of the Laplace-Beltrami operator on the point cloud. So a pre-processing step is needed to guarantee a proper computation of this operator. The pre-processing is performed as follows

- **Normal computation.** We compute a normal for each point in the point cloud. For a given point, we get the 20 nearest neighbors and compute the less dominant direction of the neighborhood.
- **Poisson reconstruction.** We reconstruct the surface for the point cloud using the screened Poisson reconstruction method [KH13]. We set the octree depth to eight and the depth for the Laplacian solver to six. The output is a Manifold triangle mesh that preserves the structure of the original point cloud.

Let M be the obtained mesh. We compute a local descriptor for each vertex in the mesh. We denote the set of local descriptors of the mesh M as F_M. The challenge now is how to compare two objects through their collections of local descriptors. The approach to use the SQFD distance establishes that we need to compute a more compact representation called a signature. Let, suppose the existence of a local clustering on F_M that groups similar local descriptors such that the number of clusters is n and $F_M = C_1 \cup C_2 \cup \ldots \cup C_n$. The signature is defined as $S^M = \{(c_i^M, w_i^M), i = 1 \ldots n\}$, where $c_i^M = \frac{\sum_{x \in C_i} x}{|C_i|}$ and $w_i^M = \frac{|C_i|}{|F_M|}$. Each element in the signature contains the average descriptor in the cluster (c_i^M) and a weight (w_i^M) to quantify how representative is the cluster in the collection of local descriptors.

Note that the local clustering is a key ingredient of the computation of the signatures. Here we briefly give some details about the clustering. We use an adaptive clustering method that searches groups of descriptors using two distance thresholds. The method uses an intra-cluster threshold λ that sets the maximum distance between descriptors in the same cluster. Also, the method uses an inter-cluster threshold β that sets the minimum distance between centroids of different clusters. In addition, the clustering method only preserves clusters with a number of descriptors greater than a parameter N_m. More details can be found in [SLBS16].

Given two objects M and N, and their respective signatures S^M and S^N, the Signature Quadratic Form Distance is defined as

$$SQFD(S^M, S^N) = \sqrt{(w^M | - w^N) \cdot A_{sim} \cdot (w^M | - w^N)^T} \qquad (13)$$

where $(w^A|w^B)$ denotes the concatenation of two weight vectors. The matrix A_{sim} is a block similarity matrix that stores the correlation coefficients between clusters. To transform a distance between cluster centroids to a correlation coefficient, we need to apply a similarity function. We use the Gaussian similarity function

$$sim(c_i, c_j) = \exp(-\alpha d^2(c_i, c_j)). \qquad (14)$$

Note that to compute the transformation, we need to choose the value of parameter α and the ground distance for descriptors. In all our experiments, we use $\alpha = 0.9$ and L^2 as ground distance. More details about the computation of signatures and the SQFD distance can be found in [SLBS16].

Experimental Settings: We provide five runs using different configurations. Here, we describe the parameters used in each run

- **SQFD(WKS).** We use the normalized Wave Kernel Signature [ASC11] as local descriptor. The parameters for local clustering are $\lambda = 0.2$, $\beta = 0.4$, $N_m = 30$.
- **SQFD(HKS).** We use the normalized Heat Kernel Signature [SOG09] as local descriptor. The parameters for local clustering are $\lambda = 0.1$, $\beta = 0.2$, $N_m = 20$.
- **SQFD(SIHKS).** We use the Scale-invariant Heat Kernel Signature [BK10] as local descriptor. The parameters for local clustering are $\lambda = 0.1$, $\beta = 0.2$, $N_m = 20$.
- **SQFD(WKS-SIHKS).** We use a distance function as a combination of distances. For every pair of objects, we compute the sum of the distances obtained with SQFD(WKS) and SQFD(SIHKS).
- **SQFD(HKS-WKS-SIHKS).** We use the combination of three distances. We use the weighted sum of distances SQFD(HKS), SQFD(WKS) and SQFD(SIHKS). The weights are 0.15, 0.15 and 0.7, respectively.

We implemented our method in Matlab under Windows 10 on a PC CPU i7 3.6 GHz, 12GB RAM. The average time to compute the shape signature for a 3D model is about 5 seconds and it takes approximately 0.3 seconds to compare a pair of signatures.

5.6. SnapNet for Dissimilarity Computation, by Alexandre Boulch, Bertrand Le Saux and Nicolas Audebert

The objective of this approach is to learn a classifier, in an unsupervised way, that will produce similar outputs for the same shapes with different poses. As we do not know the ground truth, i.e. the model used to generate the pose, we will train the classifier as if each pose was a different class. It is a 100-class problem.

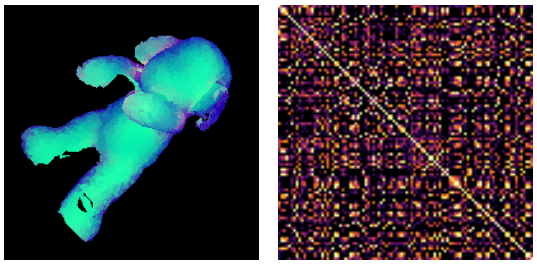
Training dataset: The training dataset is generated by taking snapshots around the 3D model [Gra14]. In order to create visually consistent snapshots, we mesh the point cloud using [MRB09]. The snapshots are 3-channel images. The first channel encodes the distance to the camera (i.e. depth map), the second is the normal orientation to the camera direction and the third channel is an estimation of the local noise in the point cloud (ratio of eigenvalues of the neighborhood covariance matrix). An example of such a snapshot is presented in Figure 5a.

CNN training: The CNN we train is a VGG16 [SZ14] with a last fully connected layer with a 100 outputs. We initialize the weights with the model trained on the ILSVRC12 contest. We then fine tune the network using a step learning rate policy.

Distance computation: The classifier is then applied to images and produces images classification vectors v_{im}. For each model we compute a prediction vector V_M based on the images:

$$V_M = \frac{\sum_{im \in M} v_{im}}{\| \sum_{im \in M} im \|^2} \qquad (15)$$

The distance matrix X contains the pairwise L^2 distances between



(a) Example of snapshot generated from shape point cloud

(b) Dissimilarity matrix

Figure 5: Snapshot example and dissimilarity matrix.

the V_M. Each line is then normalized using a soft max:

$$X_{i,j} = \frac{\exp(X_{i,j})}{\sum_j \exp(X_{i,j})} \tag{16}$$

Matrix X is not symmetrical. We finally define the symmetrical distance matrix as D is such that $D = X^T X$. The values of D are clipped according to the 5^{th} and 50^{th} percentiles and then re-scaled in $[0,1]$. The resulting matrix is presented on figure 5b.

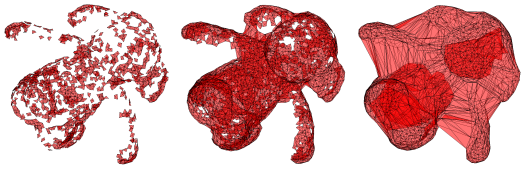
The method was implemented in Python and C++, using the deep learning framework *Pytorch*. We ran the experiments on Linux, CPU Intel Xeon(R) E5-1620 3.50GHz. The training part was operated on a NVidia Titan X Mawell GPU and the test part (predictions) on a NVidia GTX 1070. Generating the snapshots took around 10 seconds per model. The training took around 8 hours. The prediction vectors were generated in 2 seconds per model and the dissimilarity matrix is computed in less than 10s.

5.7. Alpha-shapes volume curve descriptor, by Santiago Velasco-Forero

Given $\mathcal{S}$ be the set of finite set of points in $\mathbb{R}^3$, we have computed a set of *three-dimensional alpha-shapes* of radius r, proposed by Edelsbrunner [EM94], and denoted by $\alpha_r(\mathcal{S})$. Rather than finding an optimal fixed value, we focus on a range of values for the scale parameter r, and our descriptor computes the volume of each $\alpha_r(\mathcal{S})$. Thus, the similarity of two shapes is then computed by the distance of their alpha-shapes volume curve in Euclidean norm. An example of different $\alpha_r(\mathcal{S})$ by varying r is illustrated in Figure 6. The values of parameter (r) used in the different submission are:

- AlphaVol 1: $r \in [0.02, 0.045, 0.07]$
- AlphaVol 2: $r \in [0.02, 0.045, 0.07, 0.095]$
- AlphaVol 3: $r \in [0.02, 0.045, 0.07, 0.095, 0.12]$
- AlphaVol 4: $r \in [0.02, 0.045, 0.07, 0.095, 0.12, 0.145]$

We have implemented our method in Matlab and carried out experiments under Mac on a PC CPU Intel Core i7 2.8 GHz, 16 GB RAM 1600MHz DDR3, and a NVIDIA GeForce GT 750M 2048 MB. The average computation times for the shape descriptors are as follows: AlphaVol1: 92 ms, AlphaVol2: 108 ms, AlphaVol3: 118 ms and AlphaVol4: 137 ms and it takes approximately 1.48 seconds to compute the dissimilarity matrix in the four cases.



(a) $\texttt{vol}(\alpha_{r=0.005}(\mathcal{S})) = 0.0001$

(b) $\texttt{vol}(\alpha_{r=0.01}(\mathcal{S})) = 0.0003$

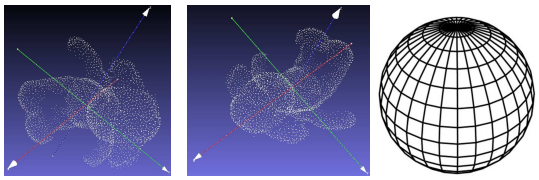
(c) $\texttt{vol}(\alpha_{r=0.035}(\mathcal{S})) = 0.0025$

Figure 6: Example of representation space by α-shapes

5.8. Modified 3D Shape Histogram for non-rigid 3D toy model retrieval (m3DSH), by Bo Li, Yijuan Lu and Afzal Godil

The 100 non-rigid point cloud toy models contain only 3D points to represent ten different poses for each of the ten toys. We can first reconstruct a 3D surface for each 3D point cloud such as to extract our previously developed 3D surface-based non-rigid shape descriptors. However, considering retrieval efficiency, the raw point cloud data is directly used for 3D shape descriptor extraction for shape comparison. For simplicity, we chose 3D Shape Histogram (3DSH) [AKKS99]. The original 3DSH descriptor uniformly partitions the surrounding space of a 3D shape into a set of shells, sectors or spiderweb bins and counts the percentage of the surface sampling points falling in each bin to form a histogram as the 3DSH descriptor. Rather than like the original 3DSH descriptor which divides the space uniformly, to increase its descriptiveness we developed a modified variation of 3DSH descriptor, that is m3DSH, by dividing the 3D space occupied by a 3D shape in a non-uniform way.

Figure 7 illustrates the overview of the feature extraction process: Principal Component Analysis (PCA) [Jol02]-based 3D model normalization, and extraction of a modified 3D Shape Histogram descriptor m3DSH. The details of our algorithm are described as follows.



(a) Original model (b) PCA normalization (c) m3DSH descriptor

Figure 7: Modified 3D Shape Histogram (m3DSH) feature extraction process.

1) PCA-based 3D shape normalization: PCA-based 3D shape normalization: We utilize PCA [Jol02] for 3D model normalization (scaling, translation and rotation). After this normalization, each 3D point cloud is scaled to be enclosed in the same bounding sphere with a radius of 1, centered at the origin, and rotated to have as-close-as-possible consistent orientations for different poses of the

Godil, Afzal.
"Point-Cloud Shape Retrieval of Non-Rigid Toys."
Paper presented at Eurographics 2017 Workshop on 3D Object Retrieval, April 23-24, 2017, Lyon, France, Lyon, France. April 23, 2017 - April 24, 2017.

same toy object. These are important for the following m3DSH descriptor extraction.

2) Modified 3D Shape Histogram descriptor m3DSH extraction: The original 3DSH descriptor only has two degrees of freedom (DOF), which are numbers of sectors and number of shells. As we know, a 3D space has three DOFs according to its spherical coordinate representation (ρ, ϕ, θ). The reason is that 3DSH uniformly divides ϕ and θ into the same number of bins, which forms a certain number of sectors. Here, in order to improve its flexibility and descriptiveness, we individually divide ϕ and θ into a number of vertical bins (V) and a number of horizontal bins (H), since the two dimensions do not have the same importance. We denote the number of radius bins for ρ as R. In the experiments, we tested two different combinations of V, H and R: V=5, H=8, R=6; and V=12, H=12, R=6.

3) Quadratic form shape descriptor distance computation and ranking: Similar as [AKKS99], we adopt the quadratic form distance to measure the distance between the extracted histogram features of the 3D models. It has a parameter σ to control the similarity degree of the resulting distance to Euclidean distance. In our experiments, we tested three σ [AKKS99] values: σ=1, σ=5, σ=10. Finally, we rank 3D models according to the computed shape descriptor distances in an ascending order.

We implemented our method in Java and carried out experiments under Windows 7 on a personal laptop with a 2.70 GHz Intel Core i7 CPU, 16GB memory. The average time to calculate the shape descriptor for a 3D model is about 0.03 seconds and it takes approximately 0.14 seconds to compute the dissimilarity matrix.

6. Results

In this section, we compare the results of all participant's runs. In total, we had 8 groups participating and we received 31 dissimilarity matrices. The retrieval scores computed from these matrices represent the overall retrieval performance of each method, *i.e.*, how well they perform on retrieving all models from the same class when querying every model in the database. The quantitative statistics used to measure the performance of methods are: NN, FT, ST, E-measure, DCG, mAP and the Precision-and-Recall plot. For the meaning of each measure, we refer the reader to [SMKF04].

Table 1 shows the method performances of all 31 runs. It is worth pointing out that some methods perform quite well on this database. By analysing particularly DCG, which is a very good and stable measure for evaluating shape retrieval methods [LZC*15], we can see that three methods have DCG greater than 0.900 (BoW-RoPS-DMF-3, BPHAPT and MFLO-FV-IWKS). Surprisingly, Tran's methods have DCG values greater than 0.990. The method clearly outperforms all other methods in the contest as evidenced by the Precision-and-Recall plot in Figure 8. BoW-RoPS can definitely capture the differences between classes and it seems robust to most of the non-rigid deformations presented in this database. Curiously, Tran's method uses asymmetric distance computation between descriptors, which leads to distances between models i and j being different from the distances between models j and i. This is clearly evidenced by their dissimilarity matrices.

Table 1: Six standard quantitative evaluation measures of all 31 runs computed for the PRoNTo dataset.

Participant	Method	NN	FT	ST	E	DCG	mAP
Boulch	SnapNet	0.8800	0.6633	0.8011	0.3985	0.8663	0.771
Giachetti	POHAPT	0.9400	0.8300	0.9144	0.4156	0.9419	0.900
	BPHAPT	0.9800	0.9111	0.9544	0.4273	0.9743	0.953
Li	m3DSH-1	0.4000	0.1656	0.2778	0.1824	0.4802	0.297
	m3DSH-2	0.4400	0.1867	0.2856	0.1932	0.4997	0.313
	m3DSH-3	0.4400	0.1767	0.2878	0.1917	0.5039	0.314
	m3DSH-4	0.4000	0.1511	0.2511	0.1712	0.4659	0.286
	m3DSH-5	0.4200	0.1722	0.2767	0.1815	0.4930	0.304
	m3DSH-6	0.4100	0.1700	0.2678	0.1712	0.4848	0.300
Limberger	GL-FV-IWKS	0.8200	0.5756	0.7244	0.3595	0.8046	0.702
	GL-SV-IWKS	0.7000	0.5267	0.6678	0.3327	0.7562	0.651
	MFLO-FV-IWKS	0.8900	0.7911	0.8589	0.4024	0.9038	0.858
	MFLO-SV-IWKS	0.9000	0.7100	0.7933	0.3702	0.8765	0.800
	PCDL-FV-IWKS	0.8200	0.6656	0.7978	0.3976	0.8447	0.764
	PCDL-SV-IWKS	0.8900	0.6656	0.7911	0.3732	0.8613	0.781
Sipiran	SQFD(HKS)	0.2900	0.2244	0.3322	0.2176	0.5226	0.344
	SQFD(WKS)	0.5400	0.3111	0.4467	0.2507	0.6032	0.427
	SQFD(SIHKS)	0.2900	0.2533	0.4133	0.2590	0.5441	0.377
	SQFD(WKS-SIHKS)	0.5000	0.3100	0.4500	0.2634	0.6000	0.425
	SQFD(HKS-WKS-SIHKS)	0.3900	0.2844	0.4389	0.2624	0.5722	0.403
Tatsuma	CDSPF	0.9200	0.6744	0.8156	0.4005	0.8851	0.794
Tran	BoW-RoPS-1	1.0000	0.9744	0.9967	0.4390	0.9979	0.995
	BoW-RoPS-2	1.0000	0.9778	0.9933	0.4385	0.9973	0.993
	BoW-RoPS-DMF-3	1.0000	0.9778	0.9978	0.4390	0.9979	0.995
	BoW-RoPS-DMF-4	1.0000	0.9778	0.9978	0.4390	0.9979	0.995
	BoW-RoPS-DMF-5	1.0000	0.9733	0.9978	0.4390	0.9979	0.995
	BoW-RoPS-DMF-6	1.0000	0.9733	0.9978	0.4390	0.9979	0.995
Velasco	AlphaVol1	0.7900	0.5878	0.7578	0.3980	0.8145	0.707
	AlphaVol2	0.7800	0.5122	0.6844	0.3751	0.7673	0.643
	AlphaVol3	0.7700	0.4567	0.6467	0.3629	0.7364	0.600
	AlphaVol4	0.7000	0.4356	0.6111	0.3454	0.7148	0.571

Considering all groups that have participated in this con-test, half of them (4) compute local features (MFLO-FV-IWKS, SQFD(WKS), CDSPF and BoW-RoPS-DMF-3) and the other half (4) compute global features (BPHAPT, SnapNet, m3DSH-3 and AlphaVol1). Our first guess was that local features would be more popular to represent non-rigid shapes, as evidenced by [LZC*15]. Our guess was based on the fact that ideally local features should be more similar than global features because same-class shapes were captured originally from the same 3D object, and locally they should be more similar than globally. For example, while a shape can be in a totally different pose, locally only joint regions are deformed. However, we also need to consider local noise in the formula, which does not affect global methods in the same level.

Tran's method is in the first place and uses local features. Clearly, in the second the place is Giachetti's method, which is based on global features from 3D meshes created from the point clouds. In total, 3 groups use meshing procedures before computing the descriptors (BPHAPT, SQFD(WKS) and SnapNet). Interestingly, two methods use quadratic form distance to compute dissimilarities between descriptors, one from a global descriptor (m3DSH-3) and other from a local descriptor SQFD(WKS).

Even though no training set was available in this track, Boulch's method uses a Convolutional Neural Network by employing an unsupervised learning architecture where every model is considered belonging to a different class. On the other hand, more methods also adopt unsupervised learning algorithms to create dictionaries using the Bag of Words encoding paradigm (BoW-RoPS-DMF-3

Godil, Afzal.
"Point-Cloud Shape Retrieval of Non-Rigid Toys."
Paper presented at Eurographics 2017 Workshop on 3D Object Retrieval, April 23-24, 2017, Lyon, France, Lyon, France. April 23, 2017 - April 24, 2017.

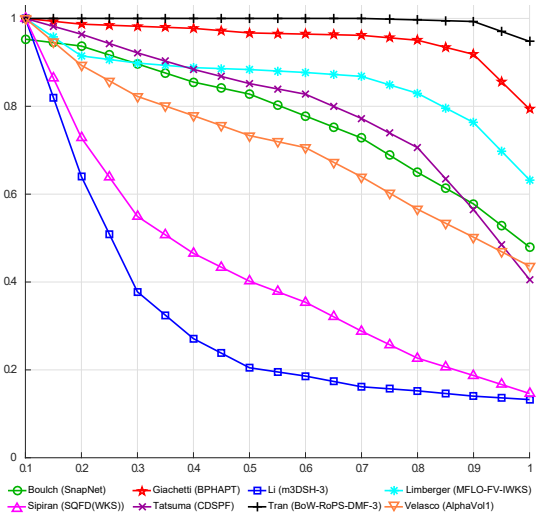


Figure 8: Precision-and-recall curves of the best runs of each group evaluated for the PRoNTo dataset.

and MFLO-FV-IWKS) being these ranked first and third on this contest, respectively, and showing that the BoW model is a good way of representing local features. Furthermore, two other methods use histogram encoding (vector quantization) to create a unique descriptor for each point cloud (BPHAPT and CDSPF).

We also observed a couple of new other ideas applied to PRoNTo dataset. For instance, Velasco uses alpha-shapes to represent point clouds; by varying the alpha-shape radius he compares models given their alpha-shape volume curve. Limberger's method uses a new formulation to compute the Laplace-Beltrami operator of point clouds, which leads to better results than the standard Graph Laplacian. Tatsuma computes additional statistics of point features in addition to the geometric feature proposed by [WHH03]. Two groups use matrix-fusion methods with different weights to improve the performance of their methods (Tran and Sipiran), however, these methods did not show a substantial improvement from the performance of the original descriptors.

For more information about this track, please refer to the official website [LW17] where the database, the corresponding evaluation code and classification file are available for academic use.

7. Conclusion

In this paper, we have created a non-rigid point cloud dataset which is derived from real toy objects. In the beginning, we discussed the importance of this data to future researchers. Then, we explained the dataset characteristics and we showed how the evaluation was carried out. Afterwards, we introduced each one of the 8 groups and their methods which competed on this track. In the end, we presented quantitative measures of the 31 runs submitted by the participants and analysed their results.

The interest in non-rigid shape retrieval is overwhelming and evident by the previous SHREC tracks. This track was not different. It has attracted a large number of participants (8 groups and 31 runs) given that it is the first time that a non-rigid point-cloud dataset is used in the SHREC contest. We believe that the organization of this track is just a beginning and it will encourage other researchers to further investigate this important research topic.

Several research directions in point-cloud shape retrieval can be pursued from this work and are listed as follows: (1) Create a larger dataset which contains more types of objects (not only hu-man shaped toys) to better evaluate shape signatures. (2) Create more discriminative local or global signatures for 3D point clouds.(3) Employ state-of-the-art Deep Learning techniques which do not depend on large training datasets.

ACKNOWLEDGMENTS

Frederico A. Limberger was supported by CAPES Brazil (Grant No.: 11892-13-7). Atsushi Tatsuma and Masaki Aono were supported by Kayamori Foundation of Informational Science Advancement, Toukai Foundation for Technology, and JSPS KAKENHI (Grant No.: 26280038 and 15K15992).

References

[AKKS99] ANKERST M., KASTENMÜLLER G., KRIEGEL H., SEIDL T.: 3D shape histograms for similarity search and classification in spatial databases. In *Advances in Spatial Databases, 6th International Symposium, Hong Kong, China, Proceedings* (1999), pp. 207–226. 7, 8

[ASC11] AUBRY M., SCHLICKEWEI U., CREMERS D.: The wave kernel signature: A quantum mechanical approach to shape analysis. In *Computer Vision Workshops (ICCV Workshops), IEEE International Conference on* (Nov 2011), pp. 1626–1633. 3, 6

[BBC*10] BRONSTEIN A. M., BRONSTEIN M. M., CASTELLANI U., FALCIDIENO B., FUSIELLO A., GODIL A.: SHREC 2010 : robust large-scale shape retrieval benchmark. *3DOR 2010* (2010). 1

[BK10] BRONSTEIN M., KOKKINOS I.: Scale-invariant heat kernel signatures for non-rigid shape recognition. In *Computer Vision and Pattern Recognition, 2010 IEEE Conference on* (June 2010), pp. 1704–1711. 6

[BMR*99] BERNARDINI F., MITTLEMAN J., RUSHMEIER H., SILVA C., TAUBIN G.: The ball-pivoting algorithm for surface reconstruction. *IEEE transactions on visualization and computer graphics 5*, 4 (1999), 349–359. 5

[BSW09] BELKIN M., SUN J., WANG Y.: Constructing laplace operator from point clouds in Rd. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms* (Philadelphia, PA, USA, 2009), SODA '09, Society for Industrial and Applied Mathematics, pp. 1031–1040. 3

[BUS09] BEECKS C., UYSAL M. S., SEIDL T.: Signature quadratic form distances for content-based similarity. In *Proc. ACM Int. Conf. on Multimedia* (New York, USA, 2009), MM '09, ACM, pp. 697–700. 5

[CCC*08] CIGNONI P., CALLIERI M., CORSINI M., DELLEPIANE M., GANOVELLI F., RANZUGLIA G.: Meshlab: an open-source mesh processing tool. In *Eurographics Italian Chapter Conference* (2008), vol. 2008, pp. 129–136. 4

[CCS12] CORSINI M., CIGNONI P., SCOPIGNO R.: Efficient and flexible sampling with blue noise properties of triangular meshes. *IEEE Transactions on Visualization and Computer Graphics 18*, 6 (June 2012), 914–924. 2

[EM94] EDELSBRUNNER H., MÜCKE E. P.: Three-dimensional alpha shapes. *ACM Trans. Graph. 13*, 1 (Jan. 1994), 43–72. 7

[GL12] GIACHETTI A., LOVATO C.: Radial symmetry detection and shape characterization with the multiscale area projection transform. In *Computer Graphics Forum* (2012), vol. 31, Wiley Online Library, pp. 1669–1678. 4

[Gra14] GRAHAM B.: Spatially-sparse convolutional neural networks. *CoRR abs/1409.6070* (2014). 6

[GSB*13] GUO Y., SOHEL F. A., BENNAMOUN M., LU M., WAN J.: Rotational projection statistics for 3D local surface description and object recognition. *International Journal of Computer Vision 105*, 1 (2013), 63–86. 4

[JC12] JÉGOU H., CHUM O.: Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening. In *Proc. of the 12th European Conf. on Computer Vision* (2012), vol. 2, pp. 774–787. 5

[Jol02] JOLLIFFE I.: *Principal Component Analysis, 2nd edn.* Springer, Heidelberg, 2002. 7

[KBH] KAZHDAN M., BOLITHO M., HOPPE H.: Poisson surface reconstruction. 2006. In *Symposium on Geometry Processing, ACM/Eurographics*, pp. 61–70. 4

[KH13] KAZHDAN M., HOPPE H.: Screened poisson surface reconstruction. *ACM Trans. Graph. 32*, 3 (July 2013), 29:1–29:13. 6

[LGB*11] LIAN Z., GODIL A., BUSTOS B., DAOUDI M., HERMANS J., KAWAMURA S., KURITA Y., LAVOUÃL' G., NGUYEN H., OHBUCHI R., OHKITA Y., OHISHI Y., PORIKLI F., REUTER M., SIPIRAN I., SMEETS D., SUETENS P., TABIA H., VANDERMEULEN D.: SHREC'11 track: shape retrieval on non-rigid 3D watertight meshes. In *3DOR 2011* (2011), Eurographics Assoc., pp. 79–88. 1

[LGT*10] LIAN Z., GODIL A., T. F., FURUYA T., HERMANS J., OHBUCHI R., SHU C., SMEETS D., SUETENS P., VANDERMEULEN D., WUHRER S.: SHREC'10 Track: Non-rigid 3D Shape Retrieval. In *Eurographics Workshop on 3D Object Retrieval* (2010), Daoudi M., Schreck T., (Eds.), The Eurographics Assoc. 1

[LLL*14] LI B., LU Y., LI C., GODIL A., SCHRECK T., AONO M., CHEN Q., CHOWDHURY N. K., FANG B., FURUYA T., JOHAN H., KOSAKA R., KOYANAGI H., OHBUCHI R., TATSUMA A.: Large scale comprehensive 3D shape retrieval. In *3D Object Retrival Workshop* (2014), pp. 131–140. 1

[LW15] LIMBERGER F. A., WILSON R. C.: Feature encoding of spectral signatures for 3D non-rigid shape retrieval. In *Proceedings of the British Machine Vision Conference* (2015), BMVA Press, pp. 56.1–56.13. 3

[LW17] LIMBERGER F. A., WILSON R. C.: SHREC'17 - Point-Cloud Shape Retrieval of Non-Rigid Toys, 2017. URL: https://www.cs.york.ac.uk/cvpr/pronto/. 9

[LZC*15] LIAN Z., ZHANG J., CHOI S., ELNAGHY H., EL-SANA J., FURUYA T., GIACHETTI A., GULER R. A., LAI L., LI C., LI H., LIMBERGER F. A., MARTIN R., NAKANISHI R. U., NETO A. P., NONATO L. G., OHBUCHI R., PEVZNER K., PICKUP D., ROSIN P., SHARF A., SUN L., SUN X., TARI S., UNAL G., WILSON R. C.: Non-rigid 3d shape retrieval. In *Proceedings of the 2015 Eurographics Workshop on 3D Object Retrieval* (Aire-la-Ville, Switzerland, Switzerland, 2015), 3DOR, Eurographics Association, pp. 107–120. 1, 8

[Mar70] MARDIA K. V.: Measures of multivariate skewness and kurtosis with applications. *Biometrika 57*, 3 (1970), 519–530. 5

[MRB09] MARTON Z. C., RUSU R. B., BEETZ M.: On Fast Surface Reconstruction Methods for Large and Noisy Datasets. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (Kobe, Japan, May 12-17 2009). 6

[NNT*15] NGUYEN V., NGO T. D., TRAN M., LE D., DUONG D. A.: A combination of spatial pyramid and inverted index for large-scale image retrieval. *International Journal of Multimedia Data Engineering and Management 6*, 2 (2015), 37–51. 4

[PCI*08] PHILBIN J., CHUM O., ISARD M., SIVIC J., ZISSERMAN A.: Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008)* (2008). 4

[PD07] PERRONNIN F., DANCE C.: Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition. IEEE Conference on* (June 2007), pp. 1–8. 3

[PFA06] PENNEC X., FILLARD P., AYACHE N.: A riemannian framework for tensor computing. *International Journal of Computer Vision 66*, 1 (2006), 41–66. 5

[PSA*16] PRATIKAKIS I., SAVELONAS M. A., ARNAOUTOGLOU F., IOANNAKIS G., KOUTSOUDIS A., THEOHARIS T., TRAN M.-T., NGUYEN V.-T., PHAM V.-K., NGUYEN H.-D., LE H.-A., TRAN B.-H., TO H.-Q., TRUONG M.-B., PHAN T. V., NGUYEN M.-D., THAN T.-A., MAC C.-K.-N., DO M. N., DUONG A.-D., FURUYA T., OHBUCHI R., AONO M., TASHIRO S., PICKUP D., SUN X., ROSIN P. L., , MARTIN R. R.: Partial shape queries for 3D object retrieval. In *3DOR: Eurographics Workshop on 3D Object Retrieval* (2016). 4

[PSR*16] PICKUP D., SUN X., ROSIN P. L., MARTIN R. R., CHENG Z., LIAN Z., ET AL.: Shape retrieval of non-rigid 3d human models. *International Journal of Computer Vision 120*, 2 (2016), 169–193. 1, 5

[RC11] RUSU R. B., COUSINS S.: 3D is here: Point Cloud Library (PCL). In *Proc. of the IEEE International Conference on Robotics and Automation* (2011), ICRA'11, pp. 1–4. 5

[SLBS16] SIPIRAN I., LOKOC J., BUSTOS B., SKOPAL T.: Scalable 3d shape retrieval using local features and the signature quadratic form distance. *The Visual Computer* (2016), 1–15. 5, 6

[SMKF04] SHILANE P., MIN P., KAZHDAN M., FUNKHOUSER T.: The princeton shape benchmark. In *Proceedings of the Shape Modeling International 2004* (Washington, DC, USA, 2004), SMI '04, IEEE Computer Society, pp. 167–178. 8

[SOG09] SUN J., OVSJANIKOV M., GUIBAS L.: A concise and provably informative multi-scale signature based on heat diffusion. In *Proceedings of the Symposium on Geometry Processing* (2009), SGP '09, Eurographics Association, pp. 1383–1392. 6

[SYS*16] SAVVA M., YU F., SU H., AONO, ET AL.: Large-Scale 3D Shape Retrieval from ShapeNet Core55. In *Eurographics Workshop on 3D Object Retrieval* (2016), Ferreira A., Giachetti A., Giorgi D., (Eds.), The Eurographics Association. 1

[SZ03] SIVIC J., ZISSERMAN A.: Video google: A text retrieval approach to object matching in videos. In *9th IEEE International Conference on Computer Vision (ICCV 2003)* (2003), pp. 1470–1477. 4

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556* (2014). 6

[TPM06] TUZEL O., PORIKLI F., MEER P.: Region covariance: A fast descriptor for detection and classification. In *Proc. of the 9th European Conf. on Computer Vision* (2006), ECCV'06, pp. 589–600. 5

[WHH03] WAHL E., HILLENBRAND U., HIRZINGER G.: Surflet-pair-relation histograms: A statistical 3D-shape representation for rapid classification. In *Proceedings of the 4th International Conference on 3D Digital Imaging and Modeling* (2003), 3DIM '03, pp. 474–482. 5, 9

[XBC*11] XU B., BU J., CHEN C., CAI D., HE X., LIU W., LUO J.: Efficient manifold ranking for image retrieval. In *Proceedings of the 34th Annual International ACM SIGIR Conference (SIGIR'11)* (2011). 3

[ZJS13] ZHU C., JEGOU H., SATOH S.: Query-adaptive asymmetrical dissimilarities for visual object retrieval. In *IEEE International Conference on Computer Vision, ICCV 2013* (2013), pp. 1705–1712. 4

[ZWG*04] ZHOU D., WESTON J., GRETTON A., BOUSQUET O., SCHÖLKOPF B.: Ranking on data manifolds. In *Advances in Neural Information Processing Systems 16*, Thrun S., Saul L. K., Schölkopf B., (Eds.). MIT Press, 2004, pp. 169–176. 3

[ZYZH10] ZHOU X., YU K., ZHANG T., HUANG T. S.: Image classification using super-vector coding of local image descriptors. In *Proceedings of the 11th European Conference on Computer Vision: Part V* (Berlin, 2010), ECCV'10, Springer-Verlag, pp. 141–154. 3

Positively Charged Ag-dendron Conjugates: Stability Enhanced AgNPs for Biomedical Applications

Tae Joon Cho*, Jingyu Liu, and Vincent A. Hackley

Material Measurement Laboratory, National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg, MD 20899-1070
*email: taejoon.cho@nist.gov

ABSTRACT

We developed positively charged silver nanoparticles ([Ag-Ds]$^+$), nominally 20 nm in diameter, using dendron chemistry combined with reduction of silver nitrate in the presence of sodium borohydride. The conjugate was developed within the context of potential biomedical applications. Rational design was applied to yield a dendron capping agent that enhances surface anchoring, hydrophilicity, and cationic surface charge. The colloidal stability and physico-chemical properties of the conjugates were evaluated under physiologically relevant conditions using dynamic light scattering, zeta potential, UV-vis absorbance, and transmission electron microscopy. Properties evaluated include size, size distribution, shape and uniformity, positive surface charge, and surface plasmon response. Colloidal stability was extensively investigated with respect to shelf-life over 6 months, temperature variation, pH, and interaction with proteins in cell culture media. Overall, the investigation confirmed the successful development of a stable positive complex with remarkable stability in biologically relevant test media containing proteins and electrolytes, and with a shelf-life exceeding 6 months. The excellent aqueous stability for this conjugate enhances its potential use as a test material for investigating interactions between positively charged NPs and bio-entities, as an antibacterial agent or a vehicle for drug delivery.

Keywords: silver nanoparticles, dendrons, positively charged, characterization, stability

1 INTRODUCTION

Among the many engineered nanomaterials, silver nanoparticles (AgNPs) are most widely utilized in commercial products [1 - 4] due to their (1) antimicrobial activities (*silver*) and (2) unique physico-chemical properties (*nano-scale structure*). In the past decades, preparation methods for AgNPs have been developed [2, 3] using different approaches to acquire successful (ideal) nanoparticles exhibiting properties such as good dispersion, uniformity, and stability (free from agglomeration or aggregation, etc.) for their biological applications. Herein, we describe the development of dendron-stabilized silver nanoparticles by modification of gold-dendron conjugate synthesis used in our previous study [5] to yield enhanced colloidal stability. Moreover, by introducing a trimethyl ammonium end group on the dendron structure, the resulting silver-dendron conjugates (hereafter abbreviated [Ag-Ds]$^+$) possess positive charges on the surface. In addition to general antimicrobial activity of AgNPs, the cationic AgNPs could potentially provide additional biological activities such as cellular uptake [6] or transfection efficiency [7] (like positively charged gold nanoparticles) that are induced by electrostatic interaction with negatively charged cell surfaces, however, there are only a few reported materials [8, 9] for this purpose. The [Ag-Ds]$^+$ developed in this report were characterized by dynamic light scattering (DLS), transmission electron microscopy (TEM), UV-Vis absorbance, and inductively coupled plasma mass spectrometry (ICP-MS) for investigation of size distribution, shape uniformity, surface plasmon resonance (SPR) behavior, and for quantifying nanoparticles mass, respectively. Also, very importantly, we systematically examined the colloidal stability of the [Ag-Ds]$^+$, including long-term shelf life, in physiological media (e.g., phosphate buffered saline (PBS), and Dulbecco's modified Eagle's medium (DMEM)), as a function of pH and temperature. Additionally, we evaluated ion release rates for Ag^+ ions from the [Ag-Ds]$^+$, and other conjugates including citrate- and polyvinylpyrrolidone (PVP) coated AgNPs by ICP-MS, to determine potential relative antibacterial activity of the [Ag-Ds]$^+$ by comparison to known AgNPs.

2 EXPERIMENTS

2.1 Materials and Instruments[1]

AgNPs (nominal 20 nm) were purchased from Ted Pella, Inc. (Redding, CA). Reference Material (RM) 8017 (PVP coated AgNPs nominally 75 nm) was obtained from NIST. [11] Other specific reagents used in this study are identified in the reference [5]. All chemicals were used without further purification. A quadrupole ICP-MS (X

[1] The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology.

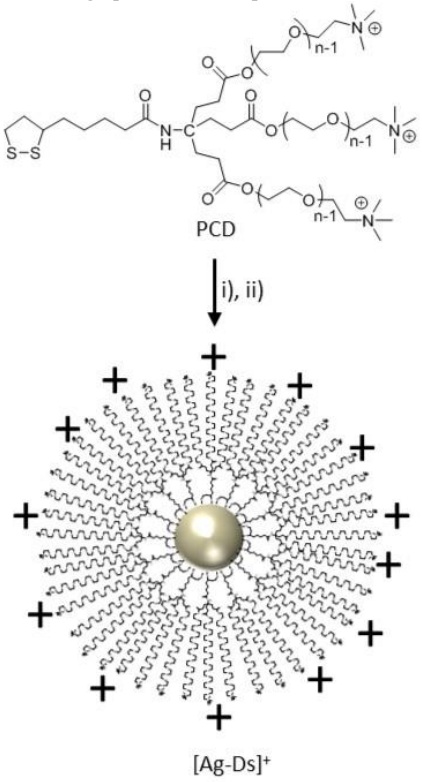
TechConnect Briefs 2017, TechConnect.org, ISBN 978-0-9988782-0-1

Cho, Tae Joon; Hackley, Vincent; Liu, Jingyu.
"Positively Charged Ag-dendron Conjugates: Stability Enhanced AgNPs for Biomedical Applications."
Paper presented at TechConnect World Innovation Conference, Washington, DC, United States. May 15, 2017 - May 17, 2017.

series[II], ThermoFisher Scientific, Waltham, MA, USE) with a PFA-ST nebulizer (Elemental Scientific, Omaha, NE, USA) and an impact bead spray chamber was used for single particle analysis. The instrument was tuned daily for maximum ^{115}In sensitivity and minimum $^{156}CeO/^{140}Ce$ oxide level. The sample uptake rate was measured every day in triplicate by weighing a vial containing DI water before and after 5 min of aspiration, and was relatively constant at (0.18 to 0.19) mL/min. Details regarding other instruments (DLS, UV-Vis, and TEM) and methodology are also provided in reference [5]. The uncertainty of size and zeta potential represent the mean and one standard deviation of at least three measurements under repeatability conditions.

2.2 Preparation of $[Ag\text{-}Ds]^+$

To an aqueous solution of $AgNO_3$ (10 mL, 2.5 mmol/L, 99.9 %, Aldrich), 1 mL of aqueous positively charged dendron (PCD, 2.5 mmol/L) and 1 mL of freshly prepared $NaBH_4$ (50 mmol/L in H_2O) were added sequentially at room temperature. The color of the reaction mixture changed from pale yellow to brown immediately after the addition was completed. After stirring for 2 h, the crude colloidal silver solution (reddish brown) was dialyzed against DI water (MWCO = 10kD, cellulose ester membrane) for 2 d and passed thru a 0.1 µm filter to remove any traceable large particles or impurities such as dust.



Scheme 1. Synthesis of positively charged dendron stabilized AgNP ($[Au\text{-}Ds]^+$); i) AgNOs, ii) $NaBH_4$, r.t., 2 h

3 RESULTS AND DISCUSSIONS

3.1 Synthesis of $[Ag\text{-}Ds]^+$

As shown in scheme 1, the designed 1→3 branched cationic dendron (PCD) [5] is composed of a thioctic acid moiety, polyethylene glycol (PEG, $M_r \approx 600$) chains, and quaternary ammonium terminal groups to provide reactivity, hydrophilicity/aqueous stability and pH-independent cationic sites. The $[Ag\text{-}Ds]^+$ were prepared from $AgNO_3$ with PCD in the presence of sodium borohydride ($NaBH_4$) as a reducing agent (Scheme 1) to yield a translucent solution with a reddish brown color. This product was purified by dialysis against deionized water.

3.2 Characterization of $[Ag\text{-}Ds]^+$

The physico-chemical properties of $[Ag\text{-}Ds]^+$ were determined by a combination of complementary and orthogonal measurement techniques including DLS, UV-Vis, TEM, and ICP-MS. The concentration of silver mass in initially purified $[Ag\text{-}Ds]^+$ was determined to be (335 ± 1.94) µg/mL by ICP-MS. The z-average diameter of the $[Ag\text{-}Ds]^+$, obtained by DLS, was (19.1 ± 0.1) nm with a monomodal size distribution (polydispersity index = 0.17, Figure 1a) and the calculated hydrodynamic size distribution gave no indication of significant aggregation between the particles. The measured zeta potential was (+24.0 ± 0.4) mV at pH 8.1, which confirmed a positively charged corona surrounding the silver core. The uncertainty of z-average diameter and zeta potential represent the mean and one standard deviation of at least three measurements under repeatability conditions. UV-Vis measurements reveal an SPR band near 408 nm (Figure 1b), a slightly red-shifted value compared to citrate-stabilized AgNPs (at 390 nm) in this size range.

TEM yields a mean diameter of (6.8 ± 2.1, figure 1c) nm for the silver core of $[Ag\text{-}Ds]^+$, and indicates a spherical uniformity in particle shape. As expected through previous work, [5] the TEM diameter of the $[Ag\text{-}Ds]^+$ is smaller than the diameter obtained by DLS due to the presence of the dendron corona that contributes to the hydrodynamic envelope of the particles but is transparent to TEM.

Cho, Tae Joon; Hackley, Vincent; Liu, Jingyu.
"Positively Charged Ag-dendron Conjugates: Stability Enhanced AgNPs for Biomedical Applications."
Paper presented at TechConnect World Innovation Conference, Washington, DC, United States. May 15, 2017 - May 17, 2017.

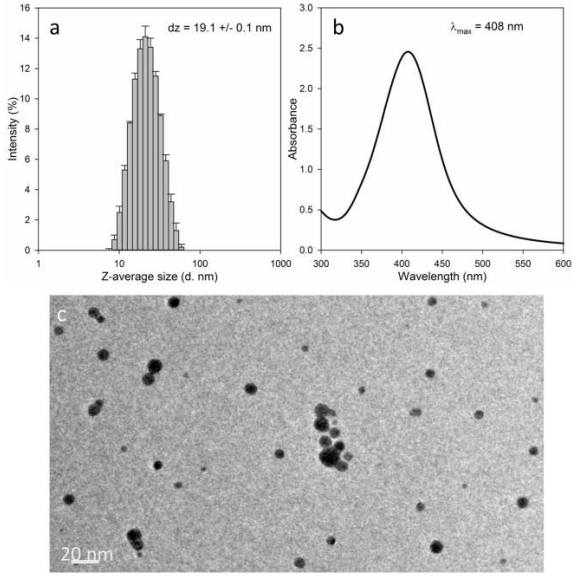


Figure 1. Characterization of $[Ag\text{-}Ds]^+$: (a) DLS size distribution. Error bars represent standard deviation, (b) SPR band by UV-Vis absorbance for initial purified sample; dilution factor is 10 for UV-vis measurements, and (c) TEM image; scale bar is 20 nm.

3.3 Stability study of $[Ag\text{-}Ds]^+$

Colloidal stability is an important issue for any commercial application of AgNPs. We evaluated the stability of the $[Ag\text{-}Ds]^+$ over a range of relevant conditions utilizing previously established protocols. [10] Native $[Ag\text{-}Ds]^+$ aged for 6 months under ambient laboratory conditions yielded a size distribution and SPR band (Figure 2a, b) that were almost identical to the freshly prepared and purified product. These results suggest that there is no significant change in the physico-chemical properties with respect to at least a 6-month shelf life under ambient conditions.

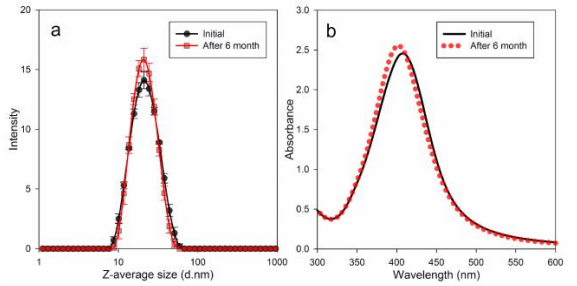


Figure 2. Shelf-life test of $[Ag\text{-}Ds]^+$; (a) DLS size distributions for the initial product (black line) and after 6 months (red line). Error bars represent standard deviation, (b) UV-Vis absorbance showing SPR band for the initial product (black line) and after 6 months (red line).

For biological application, in particular, stability in physiological media is critical. Based on UV-vis, citrate-stabilized AgNPs showed immediate instability in PBS (Figure 3a), otherwise $[Ag\text{-}Ds]^+$ exhibited excellent stability in PBS over a 48 h period relevant to cell exposure assays (figure 3b). We attribute this stability to the hydrophilicity of the inserted PEG chains and dendritic steric repulsions that substantially reduce the charge screening effect. In addition, $[Ag\text{-}Ds]^+$ were tested in DMEM, which is another common biological test medium for cell assays. The results for DMEM (Figure 3c) show the SPR band intensity reduced by about 50 % over 48 h, and the reduction in the SPR absorbance can be attributed to removal of material, perhaps by agglomeration followed by rapid sedimentation. A similar observation was reported in a previous study [5] on cationic Au-dendron conjugates. Furthermore, we evaluated the colloidal behavior of $[Ag\text{-}Ds]^+$ with protein in physiological medium, such as 10 % bovine serum albumin (BSA) in DMEM, and it appeared that similar result was observed (Figure 3d) compared to that in BSA free DMEM, over the same time period. This indicates that there is no significant interaction between $[Ag\text{-}Ds]^+$ and BSA to induce any distinct behaviors of $[Ag\text{-}Ds]^+$ in protein abundant medium.

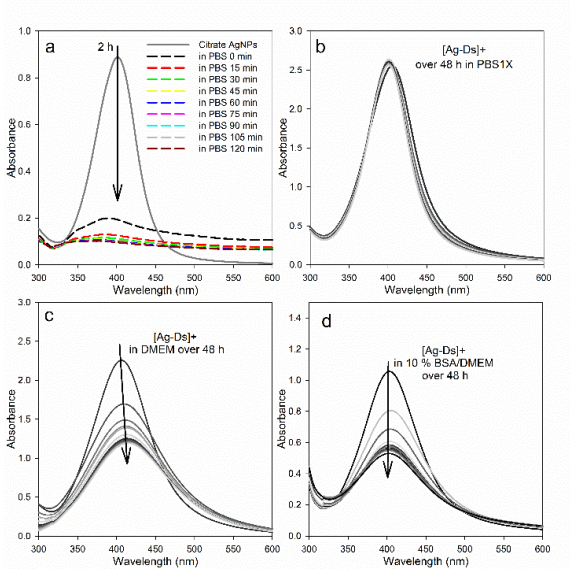


Figure 3. Stability of $[Ag\text{-}Ds]^+$ in biological test media over time, as monitored by UV-Vis: (a) citrate AgNPs, (b) PBS, (c) DMEM, and (d) 10 % BSA in DMEM.

Stability over a wide range of pH values was also investigated. From mildly acidic to basic conditions (pH 3 ~ 10), $[Ag\text{-}Ds]^+$ showed remarkable stability. Under harsher pH conditions, they exhibit gradually decreasing absorbance (45 % and 20 % reductions for 50 mmol/L HCl or 50 mmol/L NaOH, respectively; data omitted) over 12 h. Overall, the resistance against acid destabilization is greatly

TechConnect Briefs 2017, TechConnect.org, ISBN 978-0-9988782-0-1

Cho, Tae Joon; Hackley, Vincent; Liu, Jingyu.
"Positively Charged Ag-dendron Conjugates: Stability Enhanced AgNPs for Biomedical Applications."
Paper presented at TechConnect World Innovation Conference, Washington, DC, United States. May 15, 2017 - May 17, 2017.

improved relative to citrate AgNPs. [12]

Thermal stability of $[Ag\text{-}Ds]^+$ was evaluated by UV-Vis over the range from (20 to 60) °C, which covers the relevant range for most biological assays. Samples were incubated for 30 min at each temperature before measurements were conducted. The constancy of the SPR band (from UV-Vis spectra, data omitted) confirm that the $[Ag\text{-}Ds]^+$ are stable with respect to temperature variations over the tested range.

3.4 Release of Ag^+ ions from AgNPs

The antimicrobial activity of AgNPs is closely related to the oxidative dissolution process that releases bioactive Ag^+ ions. [13] Briefly, the total silver mass in $[Ag\text{-}Ds]^+$ was determined with ICP-MS after sample digestion using 70 % HNO_3. To examine the Ag^+ release process, $[Ag\text{-}Ds]^+$, citrate AgNPs, and PVP AgNPs were diluted in 5 mmol/L of acetate buffer (pH 4) to Ag mass concentration of 5 µg/mL, which were then incubated at room temperature in the dark. Aliquots of AgNP suspensions were taken at desired time points and were subjected to centrifugal ultrafiltration (Amicon-0.5 filter, 3 kDa), followed by quantification of the dissolved fraction in the filtrates by ICP-MS.

The ion release behavior of $[Ag\text{-}Ds]^+$ was compared with that of citrate and PVP coated AgNPs at pH 4. As shown in Figure 4, a continuous increase of dissolved Ag species was observed regardless of surface functionalization. The $[Ag\text{-}Ds]^+$ exhibit a release profile comparable to citrate AgNPs, suggesting possible use of this novel positively charged AgNP for antimicrobial applications.

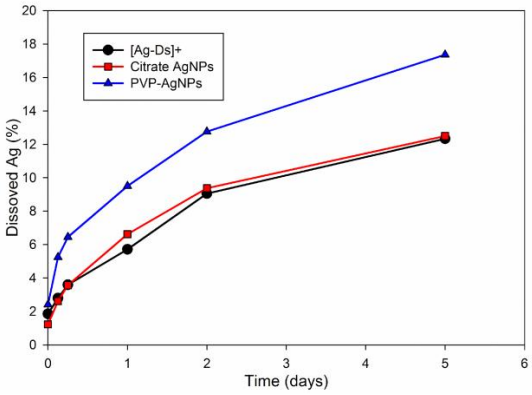


Figure 4. Time resolved Ag^+ release from AgNPs of different functionality. The release experiment was conducted in 5 mmol/L and pH 4 acetate buffer at AgNP concentration of 5 µg/mL.

4 CONCLUSION

The positively charged silver nanoparticles, $[Ag\text{-}Ds]^+$ were developed as a candidate for a nanoscale test material for biological application such as cellular assays, antimicrobial agent and/or a drug carrier in nanomedicines. In summary, we recently created a novel 1→3 directional first generation cationic dendron (PCD) and successfully developed its silver conjugate. The critical physico-chemical properties including size and size distribution, optical property, shapes, and surface charge of the $[Ag\text{-}Ds]^+$ were fully characterized by DLS, UV-Vis, TEM, and zeta potential measurements. Very importantly, especially for its biological application, the colloidal behaviors of the $[Ag\text{-}Ds]^+$ were investigated under physiologically relevant conditions, and exhibited remarkably enhanced stability relative to the control citrate AgNPs and therefore satisfactory for consideration in applications requiring long shelf-life, good dispersion in physiological media, wide range of pHs and temperatures. The ion release behavior of $[Ag\text{-}Ds]^+$ was evaluated by ICP-MS, and showed a similar release profile as citrate AgNPs suggesting possible use as a antibacterial agent.

REFERENCES

[1] S. Chernousova and M. Epple, Angew. Chem. Int. Ed. 52, 1636, 2013.

[2] K. Chaloupka, Y. Malam and A. M. Seifalian, Trends in Biotechnol. 28, 580, 2010.

[3] C. A. Dos Santos, M. M. Seckler, A. P. Ingle, I. Gupta, S. Galdiero, M. Galdiero, A. Gade, M. Rai, J. Pharm. Sci. 103, 1931, 2014.

[4] L. Rizzello, P. P. Pompa, Chem. Soc. Rev. 43, 1501, 2014.

[5] T. J. Cho, R. I. MacCuspie, J. Gigault, J. M. Gorham, J. T. Elliott, and V. A. Hackley, Langmuir, 30, 3883, 2014.

[6] E. C. Cho, J. W. Xie, P. A. Wurm, Y. N. Xia, Nano Lett. 9, 1080, 2009.

[7] T. Niidome, K. Nakashima, H. Takahashi, Y. Niidome, Chem. Commun. 1978, 2004.

[8] Z. M. Sui, X. Chen, L. Y. Wang, L. M. Xu, W. C. Zhuang, Y. C. Chai, C. J. Yang, Physica E, 33, 308, 2006.

[9] H. J. Lee, S. G. Lee, E. J. Oh, H. Y. Chung, S. I. Han, E. J. Kim, S. Y. Seo, H. D. Ghim, J. H. Yeum, J. H. Choi, Colloids and Surface B; Biointerfaces, 88, 505, 2011.

[10] available at https://www-s.nist.gov/srmors/view_detail.cfm?srm=8017.

[11] T. J. Cho, R. A. Zangmeister, R. I. MacCuspie, A. K. Patri, and V. A. Hackley, Chem. Mater. 23, 2665, 2011.

[12] R. I. MacCuspie, J. Nanopart. Res. 13, 2893, 2011.

[13] C. Levard, E. M. Hotze, G. V. Lowry, G. E. Brown, Jr. Environ. Sci. Technol. 46, 6900, 2012.

Cho, Tae Joon; Hackley, Vincent; Liu, Jingyu.
"Positively Charged Ag-dendron Conjugates: Stability Enhanced AgNPs for Biomedical Applications."
Paper presented at TechConnect World Innovation Conference, Washington, DC, United States. May 15, 2017 - May 17, 2017.

Channel Modeling and Performance of Zigbee Radios in an Industrial Environment

Mehrdad Damsaz, Derek Guo, Jeff Peil, Wayne Stark
Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109

Nader Moayeri, Richard Candell
National Institute of Standards and Technology
Gaithersburg, MD 20899

***Abstract*—In this paper we describe measurements of wireless propagation characteristics to develop path loss models in industrial environments. The models for path loss we develop are two-slope models in which the path loss is a piecewise linear relation with the log distance. That is, the path loss is a inverse power law with two regions, two exponents and a break point, that are optimized to find the best fit to the measured data. Second, the multipath power delay profile is determined. We use a reference measurement and the CLEAN algorithm for processing the measurements in order to determine an estimate for the impulse response of the channel. From this the delay spread of the channel can be determined. Finally we discuss the performance of Zigbee receivers. We compare the performance of different receiver structures for the O-QPSK type of modulation used as one Zigbee physical layer.**

I. Introduction

The pervasive application of wireless communications is well known. One such application is to an indoor factory environment. This environment creates challenges for reliable communications. One potential communication system that could be used to provide wireless communication in this environment is a Zigbee radio. Zigbee radios use of 2 MHz of bandwidth in the 2.4 GHz ISM band. To understand the performance of a Zigbee radio (or any other radio) in this environment the first task is to understand the propagation effects of such an environment. The National Institute for Standards and Technology (NIST) has carried out a measurement campaign at various factory environments, including their own machine shop. We have used these measurements to develop channel models that are suitable for evaluating the performance of wireless communication systems with bandwidth up to about 20 MHz such as a Zigbee radio or a WiFi (802.11) based system. This paper presents the results of processing the measurements to obtain channel models. From the measurements we determine the propagation loss as a function of distance, the shadowing level, the rms delay spread of the channel. Finally, we present results on the performance of a Zigbee radio when used on the channels considered.

The rest of the paper is organized as follows. In Section II we describe the measurements and the methodology to determine the impulse response for a particular transmitter, receiver location. In Section III the path loss models are described. In section IV the methodology to generate the impulse response from the measurements are discussed. The performance of the Zigbee physical layer is discussed in Section V followed by conclusions.

II. Measurements

The channel measurement or sounding campaign was carried out by NIST at various factory or factory-like environments. One location was the NIST machine shop in Gaithersburg, MD. Another location was an automotive assembly plant. NIST researchers from Boulder, CO transmitted a sounding signal and measured the response. The NIST researchers used a cart containing a mobile receiver that moved along a set path defined and measured the received signal from a transmitter located in the shop. The transmitted signal was a 40 MHz wideband signal using a pseudo-noise signal (m-sequence) that was mixed to a carrier frequency (2.4 GHz and 5 GHz). The receiver mixes the received signal to baseband and then samples the signal at an 80MHz rate. The transmitted signal was generated from 8188 (=4x2047) samples from an pseudo-noise sequence (m-sequence) generator. The receiver then sampled the received signal after mixing down to baseband with an IQ demodulator.

Figure 1 shows the layout of the machine shop in Gaithersburg where one set of the measurements were

Candell, Richard; Damsaz, Mehrdad; Guo, Derek; Moayeri, Nader; Peil, Jeff; Stark, Wayne.
"Channel Modeling and Performance of Zigbee Radios in an Industrial Environment."
Paper presented at 13th IEEE International Workshop on Factory Communication Systems, Trondheim, Norway. May 31, 2017 - June 2, 2017.

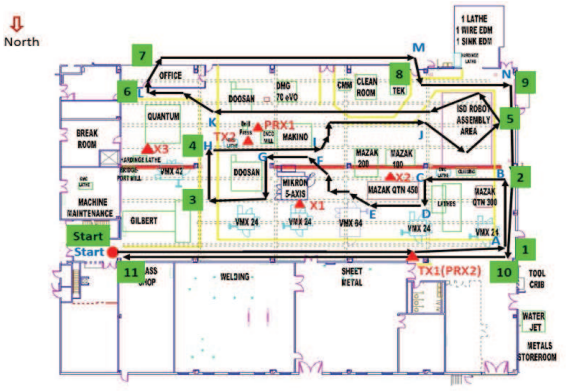


Fig. 1. Layout of Room

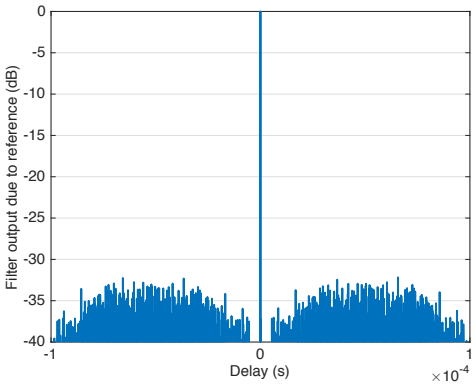


Fig. 3. Output of matched filter for reference system

made. There are a number of industrial machines in the room. The receiver was moved from the "start" location through the room and ended up back at the start (shown as location 11 on map). Various check points with known locations (e.g. locations "Start", 1,...,11) were identified with particular acquisitions of received responses. In between these known locations for certain acquisitions the location was determined by assuming that the receiver moved at a constant speed. By knowing the coordinates of the different check points and the associated measurements, the location of the receiver for other measurements could be determined. In each run 10,500 measurements were taken. Various antenna configurations (e.g. polarizations) and two different transmitter heights were used for different runs.

The basic setup of the channel sounding is illustrated in Figure 2. The transmitter and receiver have clocks that were initially synchronized. While this would allow accurate determination of the delay, it was not essential in the measurements channel models we developed. The transmitted signal was a m-sequence of length 2047 sampled four times per chip and then up-converted to a carrier frequency.

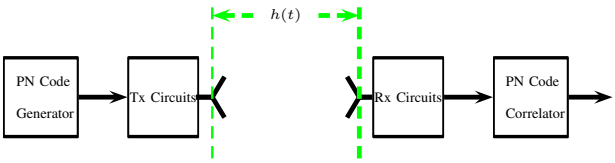


Fig. 2. Single-input, single-output channel sounding system

The PN code is an m-sequence of length 2047 using shift register feedback connection. The signal is generated by first mapping the m-sequence values, 0 and 1, to +1 and -1 respectively and then repeating each chip four times at a sample rate of 80 M samples/second. The duration of the signal is $T = 8188/(80 \times 10^6) = 102.35\mu s$. Corresponding to each transmission there is a recording of the received signal after mixing down to baseband. The recorded signal is a complex signal corresponding to an IQ demodulator.

In order that the equipment not influence the estimation of the channel characteristic, a measurement was made with only an attenuator inserted between the transmitter and receiver (without the antennas). This reference measurement provides a baseline for determining the effect of the antennas and the channel but not the measuring equipment.

To determine the equipment and channel characteristics (e.g. impulse response) we process the received signal with a filter that is matched to the transmitted signal from the m-sequence generator at the transmitter. The magnitude of the normalized output of the filter matched to the m-sequence is shown in Figure 3 where the normalization is such that the peak output value is 1 (0dB). Figure 3 shows the output due only to the equipment without any channel but with an attenuator between the transmitter and receiver. The sidelobes of the response are roughly 35dB lower than the main lobe (at zero delay). In order to accurately estimate the channel we will "remove" the effect of the sidelobes of the reference signal using a CLEAN-type algorithm.

III. Path Loss Models

There are several parts of our effort to characterize the channel. The first part is to determine the average

Candell, Richard; Damsaz, Mehrdad; Guo, Derek; Moayeri, Nader; Peil, Jeff; Stark, Wayne.
"Channel Modeling and Performance of Zigbee Radios in an Industrial Environment."
Paper presented at 13th IEEE International Workshop on Factory Communication Systems, Trondheim, Norway. May 31, 2017 - June 2, 2017.

received power as a function of distance and to generate an appropriate model. In this part of our characterization it is only the received power that is of importance, as opposed to the actual channel impulse response, which we will calculate later. To determine the path loss we measured the power in the received signal and then compared that to the power in the reference signal (the signal received when the antennas were replaced by an attenuator). By taking into account the attenuation used without the channel and the power of the reference signal we can determine the path loss of the channel (including the antennas) at each distance. The average received power as the receiver moved through various places is shown in Figure 4 for one particular run with one particular type of antenna polarization. The number of measurements for a particular polarization and frequency and transmitter location was 10,500. Each of these is called an acquisition. The average received power as a function of acquisition number and the distance as a function of acquisition number is shown in Figure 4.

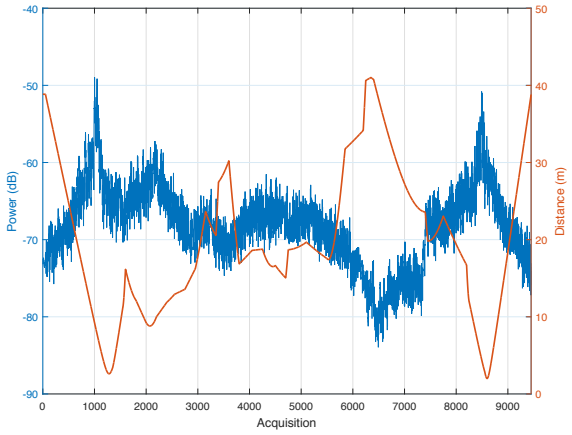


Fig. 4. Received Power vs. Acquisition and Distance vs. Acquisition, Cross Polarization, 2.4 GHz, Transmitter Location 1.

Clearly there are various power levels received at a given distance. This is the shadowing of the channel, typically modeled as a lognormal random variable with a certain variance. The path loss models the average received power as a function of distance. We will discuss the shadowing (that adds a variance to the average received power) later. The model for the average received power as a function of distance is typically an inverse power law where the power received is inversely proportional to the distance raised to a power: $P_r = k/d^{\alpha}$ for $d > d_0$. Various estimates for the parameters (k, α, and d_0) have been developed for the (average) path loss, $PL(d)$, expressed in dB as a function of distance. As a baseline the free space path loss for $f = 2450$ MHz (assuming isotropic antennas) is [1]

$$PL(d) = 40.28 + 20\log_{10}(d).$$

This is a single slope relationship between the distance and path loss since when the path loss in dB is plotted versus distance on a log scale it results in a straight line with a single slope . The generic single slope model for path loss (in dB) is

$$PL(d) = PL(d_0) + 10\alpha\log_{10}(d/d_0), d > d_0. \qquad (1)$$

Wloczysiak [2] has proposed the following received power model for indoor applications, although exactly what type of indoor environment is not specified (industrial versus residential versus office).

$$PL(d) = 50.3 + 40\log_{10}(d).$$

This model has a slope of 40dB decrease in power per decade of distance, or a received power exponent of 4. In this case $PL(d_0) = 50.3$ and $\alpha = 4$ and d_0 is larger than roughly 10m. Li et al. [3] have proposed models for the received power in a residential environment. The model has additional attenuation for going through walls and for going through floors. These models have a range of slopes between 1.13 and 1.61 for different houses with an overall proposed model with $\alpha = 1.37$.

Monti [4] has proposed a path loss model based on measurements in an office-like environment:

$$PL(d) = 54.5 + 16.4\log_{10}(d)$$

which has a path loss exponent of $\alpha = 1.6$. Jansen et. al. [5] also proposed models for indoor radio channels in an office/laboratory like environment. A range between 1.86 and 4.46 is given for the path loss exponent. Larger path loss exponents are given for non line-of-sight environments than line-of-sight. Tanghe et. al. [6] proposed path loss models in an industrial-like environment (e.g. manual or automated production line and warehouse). Their path loss models are of the form given in (1) where both α and $PL(d_0)$ are chosen to provide the best fit. They call this the non-fixed intercept model compared to the fixed intercept models described earlier. In the non-fixed intercept models the value of $PL(d_0)$ is chosen to minimize the mean square error of the fit along with the path loss exponent α. The models in [6] have exponents between 1.52 and 2.16 depending on line-of-sight power models for indoor radio channels and $PL(d_0)$ between 67.43 and 80.48dB. For 2.4 GHz

frequencies they have the following parameters for the best match choice for $PL(d_0)$. Here we distinguish between different environments between the transmitter and receiver: line of sight (LOS), non line of sight (NLOS) and combined.

Conditions	$PL(d_0)$	α
LOS	67.43	1.72
NLOS (light clutter)	72.71	1.52
NLOS (heavy clutter)	80.48	1.69
All	71.84	2.16

Finally another paper [7] for industrial applications uses the single slope model with a fixed intercept to obtain a path loss exponent between 1.86 and 2.7.

In our model we use a two slope model in which at close in distances the path loss has one slope and at a larger distance the path loss has a second slope. The transition between the two different slopes is optimized to obtain the best overall least squares fit. Our model then is a piecewise linear in that over some initial range of distances there is one value for the slope, α, and then at larger distances there is a second value for α. The path loss in dB then has the form

$$PL(d) = \begin{cases} k_1 + \alpha_1 10\log_{10}(d), & d < \beta \\ k_2 + \alpha_2 10\log_{10}(d), & d > \beta \end{cases}$$

with the boundary condition that the path loss is continuous where the slope changes. This model has the advantage that the slope is not influenced by measurements very close to the transmitter. At such distances the path loss is relatively unimportant because the received power will be relatively high (except perhaps to determine amount of interference generated). The model is also simple enough to be used without undue complications. Other approaches, like a second order regression could also be used but would seem to be more complicated. Our model at sufficiently large distances is just an inverse power law model with essentially the minimum distance of applicability determined. The channel model constants k_1, k_2, α_1, α_2 and β are to be determined from the measurements. Some of the results for this model are shown below based on the Gaithersburg measurements. In our measurements we have some minimum distance (about 2 meters) and some maximum distance (about 40 meters). We plot the generated model as a solid line between these two limits and a dashed line at smaller distances than the minimum and larger distances than the maximum. Figure 5 shows the attenuation for a 2.4 GHz system with horizontal polarized antennas. Figure 6 shows the attenuation for a 2.4 GHz system with vertical polarized antennas. Figure 7 shows the attenuation for a 2.4 GHz system with cross polarized antennas. The data for vertical polarized antennas mostly follows a single slope model but for a few distances the attenuation shows an increase in the attenuation. The two slope model finds the best break point between the two slopes and the best slopes such that continuity is maintained. By separating the two regions and finding the optimal α_1, α_2, and β we can find accurate models for the path loss at distances where the received power level is important. The Matlab Shape Language Modeling toolbox was used to find the best parameters for these models. In Figure 8 we compare the models for different polarizations. As can be seen over a range of distances between 10 and about 30m the path loss exponent is very similar for the different polarizations. Also plotted is the best two-slope piecewise linear model for the aggregate of all polarizations. Here the slope for large distances is about 1.96.

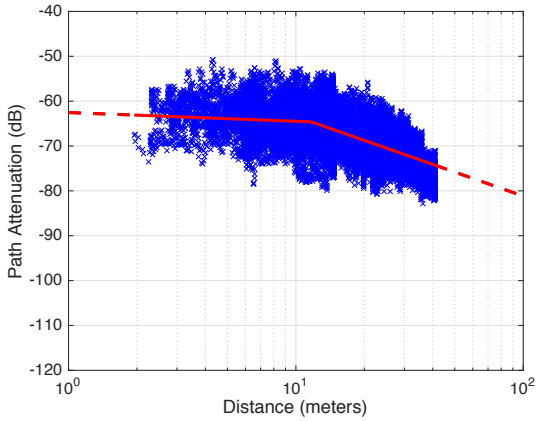


Fig. 5. Attenuation vs. distance, horizontal polarization, 2.4 GHz.

Additional measurements were made at 5 GHz. Figures 9, 10, and 11 show the attenuation for horizontal polarization, vertical polarization and cross polarization at 5 GHz.

The parameters of the model were found based on finding the smallest mean squared error between the model and the measurements. The parameters of the overall model are shown in the Table I. Note that the mean square error of the measurements is also the variance corresponding to a log-normal distribution of the received power. For the above received power versus distance, the inverse power law in the high distance region started at about 12 meters with an exponent of

Candell, Richard; Damsaz, Mehrdad; Guo, Derek; Moayeri, Nader; Peil, Jeff; Stark, Wayne.
"Channel Modeling and Performance of Zigbee Radios in an Industrial Environment."
Paper presented at 13th IEEE International Workshop on Factory Communication Systems, Trondheim, Norway. May 31, 2017 - June 2, 2017.

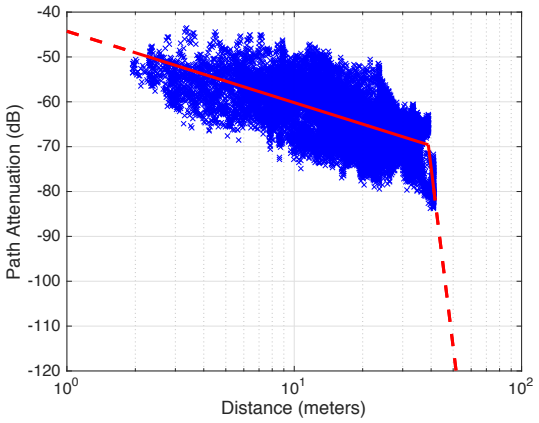


Fig. 6. Attenuation vs. distance, vertical polarization, 2.4 GHz.

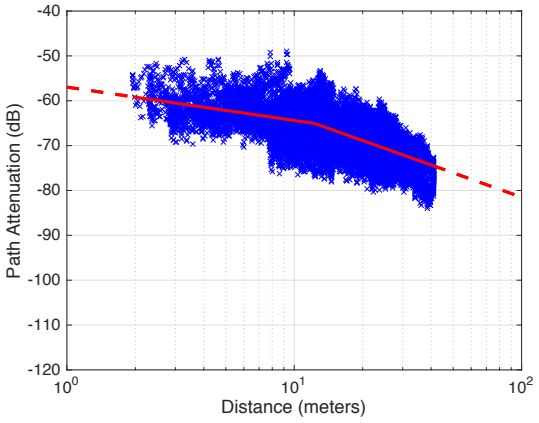


Fig. 7. Attenuation vs. distance, cross polarization, 2.4 GHz.

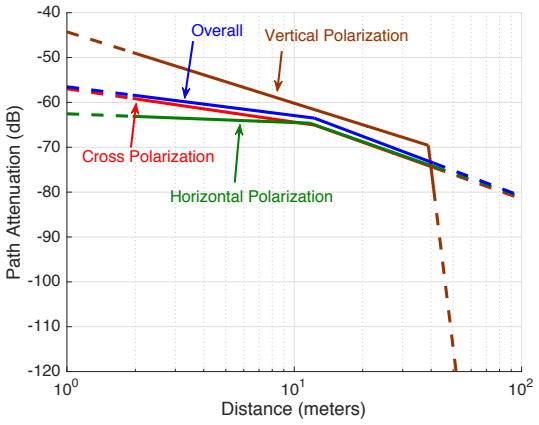


Fig. 8. Models of attenuation vs. distance, 2.4 GHz.

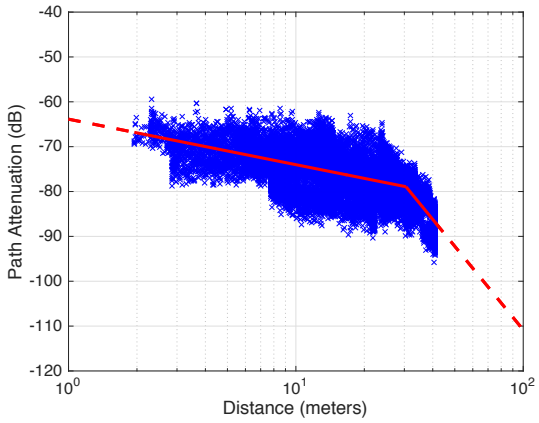


Fig. 9. Attenuation vs. distance, horizontal polarization, 5 GHz.

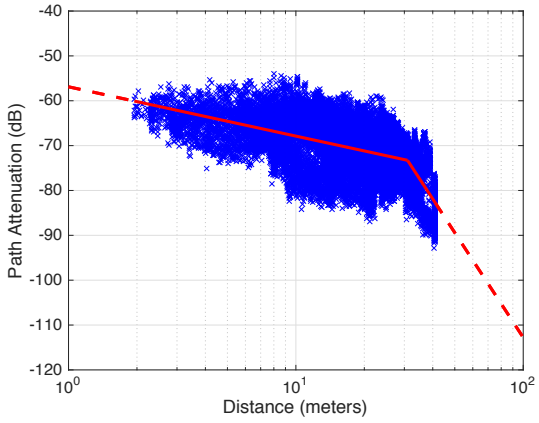


Fig. 10. Attenuation vs. distance, vertical polarization, 5 GHz.

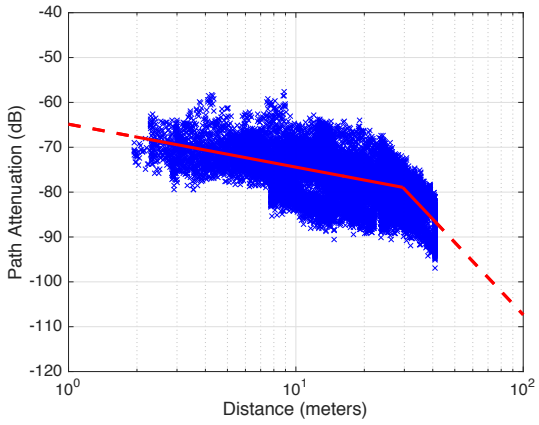


Fig. 11. Attenuation vs. distance, cross polarization, 5 GHz.

Candell, Richard; Damsaz, Mehrdad; Guo, Derek; Moayeri, Nader; Peil, Jeff; Stark, Wayne.
"Channel Modeling and Performance of Zigbee Radios in an Industrial Environment."
Paper presented at 13th IEEE International Workshop on Factory Communication Systems, Trondheim, Norway. May 31, 2017 - June 2, 2017.

1.91 and a path loss of about 65 dB at a distance of 12 meters.

TABLE I
PARAMETERS FOR CROSS POLARIZATION, 2.4 GHZ, T1

Parameter	Value
α_1	0.64
α_2	1.91
β	12.25
k_1	56.52
k_2	42.72
σ^2	12.60

In Figure 12 we compare the received power for these models versus distance. All other models have a simple linear representation of received power in dB versus log distance, we have a piecewise linear model with two different slopes. Overall our model has a power loss exponent close to that of free space at large distances but has a smaller exponent at small distances compared to the other models. The other models mainly were for office spaces as opposed to an industrial setting.

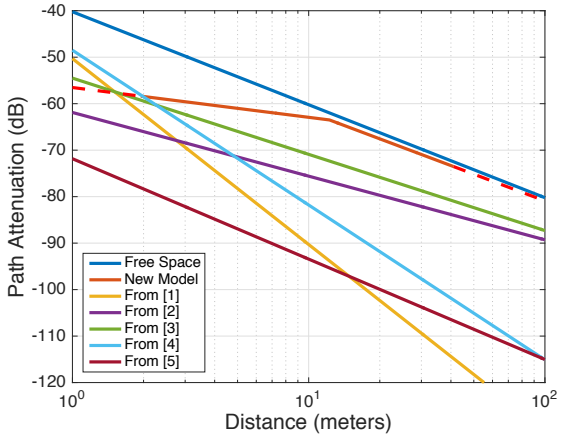


Fig. 12. Various models for overall received power (2.4 GHz) as a function of distance

Shadowing is another factor in determining the performance of a communication system. Shadowing is generally modeled as a log normal random variable. That is, the received power, expressed in dB, is a Gaussian random variable. The mean of the random variable is a function of distance as determined by the path loss model. The variance of the Gaussian random variable measures the effect due to shadowing. Our estimation of the path loss model, by finding the best piecewise linear attenuation model to minimize the mean squared error also results in a mean squared error that is the variance of the Gaussian random variable that models the shadowing. Our results indicate a shadowing variance between 7dB and 14dB. The path loss model used for all polarizations corresponded to a shadowing parameter of about 12dB. This tends to be somewhat larger than other models. So while the average path loss seems to be smaller than other models, the variance tends to be larger.

IV. IMPULSE RESPONSE

The measurement procedure described above allows us to estimate the channel impulse response or the power delay profile of the multipath channel. Consider the output of the reference system after performing a matched filtering and the corresponding output of the measurement system. The output of the reference system where the antennas have been replaced by an attenuator is given by

$$\begin{aligned} z_r(t) &= y_r(t) * h_{MF}(t) \\ &= [s_T(t) * h_T(t) * h_R(t) * h_{MF}(t)]A. \end{aligned}$$

where $s_T(t)$ is the signal generated by the m-sequence generator, $h_T(t)$ and $h_R(t)$ are the impulse responses for the transmitter and receiver circuitry, $h_{MF}(t)$ is the impulse response of the matched filter and A is the attenuation value. The corresponding output for the measurement system is

$$\begin{aligned} z_m(t) &= y_m(t) * h_{MF}(t) \\ &= [s_T(t) * h_T(t) * h(t) * h_R(t) * h_{MF}(t)] \\ &= [s_T(t) * h_T(t) * h_R(t) * h_{MF}(t)] * h(t) \\ &= \frac{z_r(t) * h(t)}{A}. \end{aligned}$$

Thus the output for the measurement system is the reference response with an additional filtering due to the channel but without the factor due to the attenuator. The output of the matched filter for the reference systems (as seen in Figure 3), $z_r(t)$ is nearly an ideal impulse function but with sidelobes about 35dB lower than the main lobe. As a result, the channel $h(t)$ could be estimated simply as $\hat{h}(t) \approx z_m(t)A$. Here we want to account for the sidelobes of the reference signal to more accurately estimate the impulse response of the channel. Often the multipath aspect of a channel is modeled as a series of impulses of the form

$$h(t) = \sum_i \beta_i \delta(t - \tau_i)$$

Candell, Richard; Damsaz, Mehrdad; Guo, Derek; Moayeri, Nader; Peil, Jeff; Stark, Wayne.
"Channel Modeling and Performance of Zigbee Radios in an Industrial Environment."
Paper presented at 13th IEEE International Workshop on Factory Communication Systems, Trondheim, Norway. May 31, 2017 - June 2, 2017.

where β_i is a complex path gain at delay τ_i. For this channel model the result of the measurement would be

$$\begin{aligned} z_m(t) &= \frac{1}{A} z_r(t) * h(t) \\ &= \frac{1}{A} z_r(t) * \sum_i \beta_i \delta(t - \tau_i) \\ &= \frac{1}{A} \sum_i \beta_i z_r(t - \tau_i). \end{aligned}$$

Our goal is to determine the values for β_i and τ_i. Since the function $z_r(t)$ is known, the approximation used above is that $h(t)$ is just a normalized version of $z_m(t)$. However, we can also calculate the β_i by using the known value of $z_r(t)$. In particular we can determine the largest value of β_i by looking at the largest value of the measurement and the associated delay and associating that output value with of β_i. With that determined we can subtract off the effect of the largest β_i, namely $\beta_i z_r(t - \tau_i)$ and continue the process to find the second largest value of β_i and the associated delay. This is generally known as the CLEAN algorithm [8]. In Figure 13 we show in blue the the result of applying the CLEAN algorithm to estimate the impulse response. The error, shown in red, is the left over signal after 250 iterations of the CLEAN algorithm where in each iteration the largest magnitude signal is accounted for by a particular delay and coefficient of the channel. Note that a particular delay could correspond to several iterations that have the largest magnitude residual error signal. In this case it is the (complex) sum of these coefficients that determine the final coefficient at that delay.

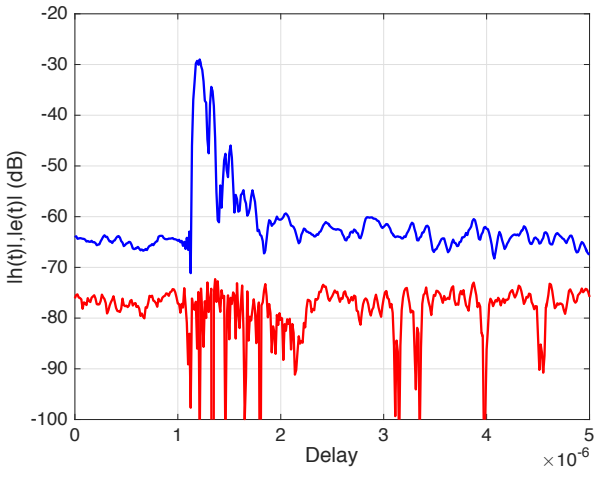


Fig. 13. Result of processing measurements with CLEAN algorithm

This approach of estimating the impulse response was applied to 10,500 different acquisitions for a particular run. The magnitude of the impulse responses is shown in Figure 14.

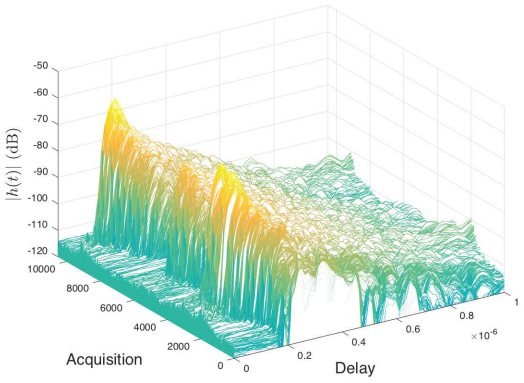


Fig. 14. Impulse response for various acquisitions

With an estimate of the impulse response of the channel various other channel parameters can be calculated. Often the rms delay spread is used to characterize a channel. A large rms delay spread can degrade the performance of certain systems. The rms delay spread can be calculated as follows. Let $h(t)$ be the impulse response of the channel. First we define the power delay profile as

$$P(\tau) = \frac{|h(\tau)|^2}{\int |h(t)|^2 dt}.$$

The power delay profile is a probability density function since it integrates to 1 and is non-negative. The absolute received power level is normalized out in determining the power delay profile. The mean excess delay spread is calculated as

$$\bar{\tau} = \int t P(t) dt.$$

The mean square delay spread is

$$\sigma_\tau^2 = \int (t - \bar{\tau})^2 P(t) dt.$$

Then the rms delay spread is σ_τ. From the set of impulse responses we can determine the power delay profile for each acquisition and the corresponding rms delay spread. The Gaithersburg machine room indicated rms delay spreads in the range of 90 ns to as much as 400ns. This would indicate a coherence bandwidth of more than 2.5 MHz. Thus, for this delay spread, there should

Candell, Richard; Damsaz, Mehrdad; Guo, Derek; Moayeri, Nader; Peil, Jeff; Stark, Wayne.
"Channel Modeling and Performance of Zigbee Radios in an Industrial Environment."
Paper presented at 13th IEEE International Workshop on Factory Communication Systems, Trondheim, Norway. May 31, 2017 - June 2, 2017.

not be significant intersymbol (or interchip) interference in a Zigbee system. However, the measurements in an automotive assembly building indicated delays spreads in the range of 1-5 μs. This corresponds to a coherence bandwidth as small as 200 kHz and the interchip interference would play a role in determining the performance.

V. Zigbee Performance

In this section we consider the error probability for a Zigbee communication system. As is known, the 802.15.4 standard specifies how signals are to be transmitted but not how signals are to be received. While there are multiple physical layers defined in the standard, our focus in on signals in the 2.4 GHz band. These signals are designed to transmit data at a maximum rate of 250 kbps but could be smaller. One physical layer defined in the standard is called Offset QPSK. This is a modulation technique that maps groups of four information bits into complex sequences of length 16 chips and then uses offset QPSK with half sine pulse shaping. This is essentially MSK at the chip level. This modulation technique can be demodulated in various ways. A coherent receiver with soft decision demodulation will be the most complex receiver but have the best performance. A noncoherent receiver that does coherent integration over a chip sequence but does not require a coherent phase reference will have worse performance. A receiver that makes a hard decision on each chip using noncoherent demodulation and then finds which of the 16 chips sequences is closest in Hamming distance would have even worse performance.

First we consider a comparison of a purely orthogonal signal set with a perfectly coherent receiver and evaluate the symbol error probability. Note that, in a typical Zigbee application the packet error probability will be of the most interest rather than the bit error probability or the symbol error probability. However, to understand the effects of different modulation techniques and demodulation techniques we evaluate the symbol error probability of a four bits symbol. Figure 15 compares two different modulation techniques and two different receivers. One modulation is an orthogonal signal set. The second modulation is the Zigbee signal set. One receiver is a coherent receiver that requires ideal synchronization and perfect phase estimation. The second receiver is a noncoherent receiver. This noncoherent receiver assumes a constant phase offset for the duration of the time for transmission of the signals (e.g. 16 times the length of a chip). For Zigbee this would be about 16 μs. As can be seen from the figure, the Zigbee signal set with coherent demodulation requires about 0.6dB more signal-to-noise ratio (E_b/N_0 (dB)) than an orthogonal signal set at a symbol error probability of 10^{-5}. A noncoherent receiver with an ideal orthogonal signal set at a symbol error rate of 10^{-5} has the same required signal-to-noise ratio (E_b/N_0) as coherent demodulation of the Zigbee signals. However, at higher error rates the coherent receiver for Zigbee signals performs better than the noncoherent receiver for orthogonal signal. The Zigbee signals with noncoherent reception has worse performance at a symbol error rate of 10^{-5} by a little more than 0.6dB than orthogonal signals with noncoherent reception. Note that a symbol error rate of 10^{-5} might correspond to a packet error rate in the range of 10^{-3} with packets on the order of 100 symbols (50 bytes). A receiver making hard decisions on each chip would be expected to be about 2dB worse performance than the receivers shown here.

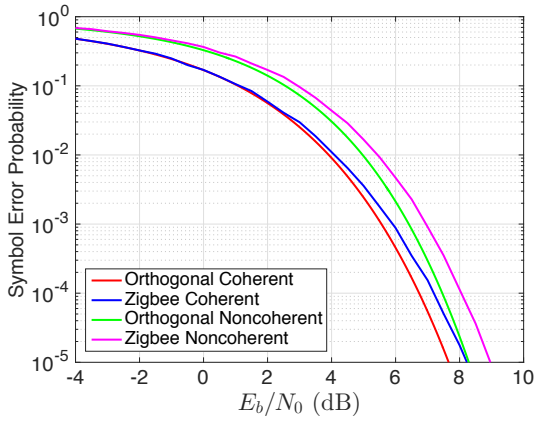


Fig. 15. Symbol error probability, orthogonal vs. Zigbee, coherent vs. noncoherent reception.

The packet error probability for transmission of information depends first on being able to detect the presence of a transmission and then being able to synchronize to the transmitted signal (timing). After that demodulation of each symbol in a packet is required for the packet to be correct. For Zigbee there is no error control coding technique that could correct symbol errors. The only notion of coding is in the construction of the signal set. In Zigbee a pair of symbols determine a byte of information. A packet in Zigbee can have at most 127 data bytes but could have as few as 9 (ignoring the preamble bytes).

In Figure 16 the packet error probability of an IEEE 802.15.4 system with a coherent receiver and a noncoherent receiver for a packet of length 127 bytes is shown for an additive white Gaussian noise (AWGN)

Candell, Richard; Damsaz, Mehrdad; Guo, Derek; Moayeri, Nader; Peil, Jeff; Stark, Wayne.
"Channel Modeling and Performance of Zigbee Radios in an Industrial Environment."
Paper presented at 13th IEEE International Workshop on Factory Communication Systems, Trondheim, Norway. May 31, 2017 - June 2, 2017.

channel. As can be seen the coherent receiver is less than 2dB better than the noncoherent receiver. One reason for such a small gap is that the modulation used in Zigbee is a version of 16-ary orthogonal modulation. As is known, orthogonal modulation has asymptotically (for large number of signals) the same performance for coherent reception and noncoherent reception. Here the signal-to-noise ratio (E_b/N_0) is the average *received* energy per information bit to noise power spectral density.

In indoor and outdoor applications, radio systems need to have a good performance which means a reasonable amount of information loss. As with any other radio system, in order to evaluate Zigbee performance, we started with simulating Zigbee in an additive white Gaussian noise (AWGN) channel as well as Ricean fading channels. In IEEE 802.15.4, at the beginning of PPDU of each packet, there is a 4 bytes-long preamble which consists of 32 zero bits for all packets. We are using these 32 zero bits to find the start of each packet using a matched filter which is matched to each symbol (4 zeros) of the preamble.

The transmitted signal is passed through a complex AWGN channel and the output of the channel is fed into the Zigbee receiver. As the first block of any radio-system receiver, a synchronization block is designed to find the start of each packet. Since there is a fixed pattern in preamble part of each packet, the receiver uses a matched-filter to locate the separating flag between any two consecutive packets. After finding start of packet, it is possible to pass preamble and demodulate the length of payload-byte of PPDU. Knowing the packet start and the packet length, then the next step is to demodulate the payload which carries the information bits. The demodulation is 16-orthogonal demodulation and is used to detect payload of each packet. The magnitudes only of the 16 demodulator outputs are used to make a decision about the data for the non-coherent receiver. To do coherent demodulation, the real part of outputs of inner products are considered and the maximum is selected. Both coherent and non-coherent receiver have been simulated and their performances in terms of Packet Error Rate (PER) are compared above in Figure 16.

Indoor channel environments are not always well modeled by an AWGN channel. Multipath propagation and obstacle reflections can have a significant impact on system performance. In order to model the multi-path propagation, which is a serious factor in indoor-communication applications, a Rician fading channel has been simulated. Rician fading is a stochastic model for the radio propagation when the signal arrives at the receiver by several different paths. Rician fading can nicely model the environment specially, when one path, which is usually line of sight path, is much stronger than others. This appears to be the case for some of the indoor industrial channel for Zigbee since the bandwidth is relatively small (2 MHz) compared to the bandwidth for WiFi (20 MHz). Our simulation models the amplitude gain using a Rician distribution. Rayleigh fading is used to model the multipath propagation when there is no line of sight. A Rician model with different ratio of direct line-of-sight power versus diffuse power, known as the K factor has been used in our simulation. The packet error rate (PER) for a packet of length 9 and 127 bytes with a coherent receiver is shown in Figure 17 and Figure 18. As expected, the PER converges to AWGN packet error rate curve as the K factor gets large. Notice that there is only a slight performance degradation of the larger packet size relative to the small packet size in these figures, especially for Rayleigh fading. This is due to the fact that the fading is assumed to be a constant for the duration of a packet. If the fade is a bad fade (i.e. destructive interference) then the error probability will be large for symbols and for the packet as a whole. While a good fade (i.e. constructive interference) will result in correct symbols and the packet as a whole. So the packet error probability is dominated by the probability of a good fade versus a bad fade.

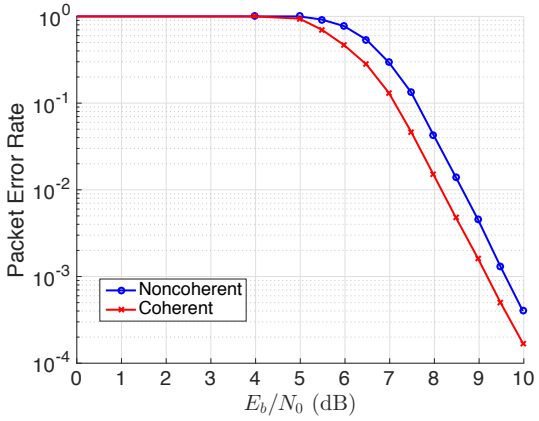


Fig. 16. Packet error probability, block length 127 bytes: coherent vs. noncoherent, AWGN

VI. Conclusions

In this paper we have used measurements to obtain models for indoor industrial environment channels. Our models are piecewise linear relations between the received power (in dB) and the log of the distance. Perhaps the most useful part of the propagation model

Candell, Richard; Damsaz, Mehrdad; Guo, Derek; Moayeri, Nader; Peil, Jeff; Stark, Wayne.
"Channel Modeling and Performance of Zigbee Radios in an Industrial Environment."
Paper presented at 13th IEEE International Workshop on Factory Communication Systems, Trondheim, Norway. May 31, 2017 - June 2, 2017.

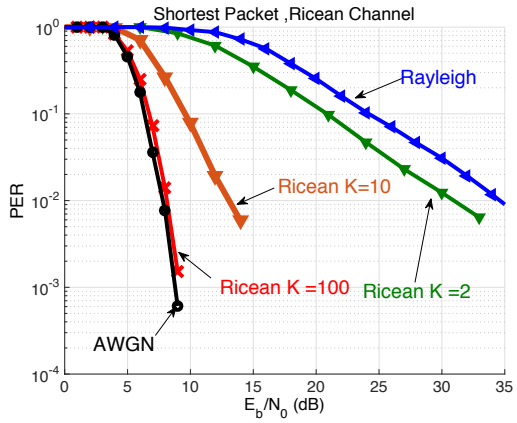


Fig. 17. Packet error probability, block length 9 bytes, Rician, Rayleigh channel, coherent reception

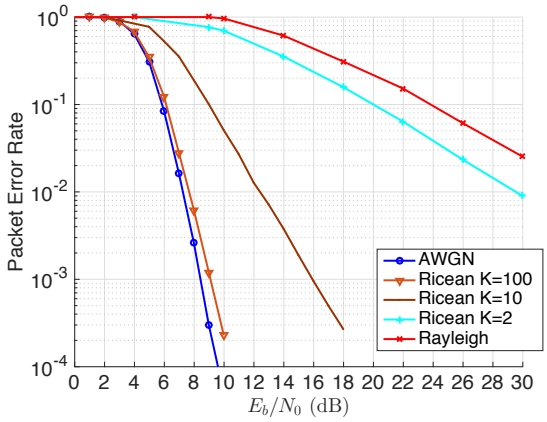


Fig. 18. Packet error probability, block length 127 bytes: coherent Rician, Rayleigh, coherent reception

occurs after the breakpoint in the piecewise linear model where the power received becomes small. The received power at short distances is larger than other models while at higher distances the received power is less than most other models. At distances smaller than the break point in the piecewise linear model the received power is going to be quite large and the exact value of the received power is probably not important as the system will have more than adequate power to decode a packet correctly. We have used the CLEAN algorithm to determine the multipath channel characteristics. The multipath delay spread is generally less that 0.5 μs and is comparable to the inverse bandwidth of a Zigbee system. That is, most of the multipath components will be within a single chip duration of a Zigbee signal. We have used the measurements to evaluate the shadowing parameter for this environment and our results show a log normal shadowing of between 7 and 12 dB. A Zigbee radio system with different receivers has been simulated and the performance in different channel environments has been determined. While there is a small difference between coherent and noncoherent receivers (e.g. about 2dB), there is a large gap between AWGN performance and Rayleigh faded performance. This is to be expected since the Zigbee signals do not employ error-correcting codes or wide enough bandwidth so that the fading is mitigated.

Acknowledgement: This research was supported by the National Institute of Standards and Technology (NIST) under grant 70NANB14H316. Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

REFERENCES

[1] D. B. Green and A. Obaidat, "An accurate line of sight propagation performance model for ad-hoc 802.11 wireless LAN (WLAN) devices," *IEEE International Conference on Communications*, vol. 5, pp. 3424–3428, 2002.

[2] S. Wloczysiak, "Extending 2.4 GHz ZigBee short range radio performance with Skyworks front-end modules," *Microwave Journal*, pp. 1–10, August 2009.

[3] H. Li, L. Zhao, M. J. Darr, and P. Ling, "Modeling wireless signal transmission performance path loss for ZigBee communication protocol in residential houses," *2009 ASABE Annual International Meeting*, 2009.

[4] C. Monti, A. Saitto, and D. Valletta, "Indoor radio channel models for ieee 802.15. 4 technology," *Proceedings of EURASIP Workshop on RFID*, 2008.

[5] G. J. Janssen, P. Stigter, R. Prasad, et al., "Wideband indoor channel measurements and BER analysis of frequency selective multipath channels at 2.4, 4.75, and 11.5 GHz," *Communications, IEEE Transactions on*, vol. 44, no. 10, pp. 1272–1288, 1996.

[6] E. Tanghe, W. Joseph, L. Verloock, L. Martens, H. Capoen, K. V. Herwegen, and W. Vantomme, "The industrial indoor channel: large-scale and temporal fading at 900, 2400, and 5200 MHz," *IEEE Transactions on Wireless Communications,*, vol. 7, no. 7, pp. 2740–2751, 2008.

[7] J. Ferrer-Coll, P. Ängskog, J. Chilo, and P. Stenumgaard, "Characterisation of highly absorbent and highly reflective radio wave propagation environments in industrial applications," *IET Communications*, vol. 6, no. 15, pp. 2404–2412, 2012.

[8] Q. Spencer, M. Rice, B. Jeffs, and M. Jensen, "A statistical model for angle of arrival in indoor multipath propagation," *IEEE 47th Vehicular Technology Conference*, vol. 3, pp. 1415–1419 vol.3, May 1997.

Adaptive Synchronization Reference Selection for Out-Of-Coverage Proximity Services

Samantha Gamboa, Fernando J. Cintrón, David Griffith, Richard Rouil
National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
Email: {samantha.gamboa, fernando.cintron, david.griffith, richard.rouil}@nist.gov

***Abstract*—The introduction of Proximity services (ProSe) in Long Term Evolution Advanced (LTE-A) allows User Equipments (UEs) to communicate directly without routing the data through the LTE access network. This is a major step towards supporting mission-critical communication for first responders who need the ability to communicate ubiquitously. To properly receive data, the UEs must be synchronized. Thus, reducing the synchronization delays is important to avoid service disruption. When operating outside of the network coverage, UEs cannot rely on the synchronization information provided by the base station. In such cases, a distributed protocol is required to announce and detect the synchronization information within devices in proximity. In this paper, we present an adaptive algorithm that reduces the out-of-coverage synchronization delays while meeting the requirements specified in the LTE-A standard. The algorithm takes into account the UE traffic and synchronization conditions to achieve these goals. We evaluate the algorithm performance using our ns-3 ProSe implementation and show that fast convergence time to a synchronized state can be achieved using the proposed algorithm while satisfying the standard performance constraints.**

I. Introduction

Proximity Services (ProSe) is a Long Term Evolution Advanced (LTE-A) function that was introduced in release 12. ProSe allows LTE-A User Equipments (UEs) to perform device-to-device (D2D) communication [1]. The UEs use a direct link called sidelink to transfer information between them without the need of using traditional links through a base station (downlink/uplink). ProSe is defined to work in-coverage with or without network assistance, and out-of-coverage in an autonomous way. The ability to work out-of-coverage is crucial for public safety mission-critical use cases, as it allows first responders to communicate regardless of the location of the incident or the network status [2].

UEs need to be synchronized to be able to decode the information transmitted over the sidelink. Thus, the UEs need to follow the same Synchronization Reference (SyncRef), which indicates the common timing, frequency, and system configuration to use. In-coverage UEs follow the SyncRef indicated by the network, while out-of-coverage UEs follow the SyncRef of other UEs.

Achieving fast convergence to a synchronized state within a group of out-of-coverage UEs is challenging, as the synchronization protocol is a distributed process. In this paper, we assess how the frequency of triggering the SyncRef selection algorithm impacts the convergence time. Intuitively, the more often the UEs execute the SyncRef selection algorithm, the faster the convergence is. However, there are several limiting factors.

First, the sidelink is half-duplex since it uses the same frequency for transmission and reception. The operations needed for the SyncRef selection algorithm (i.e., detection and signal strength measurement of SyncRefs in proximity) require the UE to be in reception mode. Data transmissions scheduled during these periods are preempted, as synchronization operations have priority over other ProSe functions [3]. As a result, the ProSe standard defines a maximum transmission drop rate due to SyncRef selection of 2 %, which limits the algorithm triggering frequency [4].

Second, the receiver circuitry is active during SyncRef selection. Thus, the synchronization process consumes power even when no data is transmitted. This is an important factor to consider in the implementation of the synchronization process, especially for first responder UEs used during mission-critical tasks.

The simplest synchronization scheme is to execute the SyncRef selection algorithm periodically, as is done in the LTE downlink [5]. However, we have shown that a periodic SyncRef selection scheme can lead to problems when two SyncRefs are transmitting simultaneously and continuously over time [6], although we did not address the period selection procedure nor did we consider traffic patterns other than the case of saturated UEs that transmit continuously.

In this paper, we analyze the period selection procedure considering the above mentioned constraints, which are traffic and scenario dependent. We consider on-off traffic patterns with different activity factors, and we show that a synchronization period chosen for a given activity factor may not satisfy the constraints if the traffic varies. Moreover, a period chosen for a worst-case scenario could lead to infrequent synchronization and large convergence times, which is undesirable for public safety mission-critical scenarios. To address these

Cintron, Fernando; Gamboa Quintiliani, Samantha; Griffith, David; Rouil, Richard.
"Adaptive synchronization reference selection for out-of-coverage Proximity Services (ProSe)."
Paper presented at 28th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Montreal, QC, Canada.
October 8, 2017 - October 13, 2017.

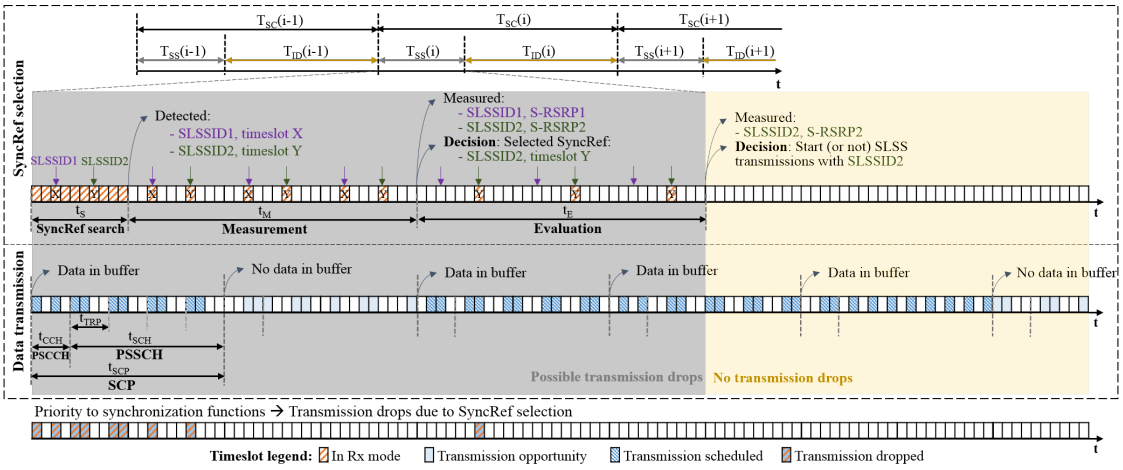


Fig. 1. System model scheme. Timeline of a UE performing SyncRef selection and sidelink communication data transmissions concurrently.

limitations, we propose an algorithm that triggers the SyncRef selection dynamically, based on the local traffic condition and configuration of each UE.

Given the novelty of ProSe, the literature related to out-of-coverage ProSe synchronization protocol is scarce. Most of the existing studies focused either on SyncRef detection procedures [7], or on the decision process required to select the adequate SyncRef after detection of multiple ones (see [6] and references therein). These studies led to the standard design and procedures explained in this paper. To the best of our knowledge, our work is the first one to focus on the SyncRef selection triggering function, which is one of the topics left to implementation by the LTE-A standardization body.

The rest of the paper is organized as follows. In Section II, we characterize the system model and in Section III, we define the problem in study. In Section IV, we describe the proposed algorithm and detail the results of the performance evaluation in Section V. Finally, we conclude the paper in Section VI.

II. System model

ProSe UEs partition time in blocks of 1 ms, called subframes (SFs). Thus, we will use the terms timeslot, SF, and ms interchangeably in this paper. We assume that during a given SF, the UE can be either in reception (Rx) mode or in transmission (Tx) mode. We assume UEs switch between modes instantaneously. We consider the evaluation period of N SFs in which the UE is performing sidelink communication and synchronization functions concurrently. We assume the UE is out-of-coverage. The notation used throughout the paper is listed in Table I.

A. Sidelink synchronization

This function comprises two concurrent processes [8]:

1) The transmission of synchronization information: where the UE advertises its synchronization information by transmitting several signals and a message [9]. From now on, we will refer to that set of elements as the Sidelink Synchronization Signal (SLSS). An SLSS has a duration of one SF and is transmitted periodically every 40 ms. The SLSS encodes an ID (SLSSID) which identifies the synchronization information being transmitted. UEs transmitting SLSSs are called SyncRef UEs and the conditions for becoming a SyncRef are dependent on the synchronization status of the UE.

2) The selection of synchronization reference: where the UE acquires the synchronization information transmitted by nearby SyncRefs and selects and synchronizes to the most suitable SyncRef. We model this process as a chain of three sub-processes. First, the UE performs a SyncRef search in which it is continuously in Rx mode during t_S SFs . Second, the UE measures the Sidelink Reference Signal Received Power (S-RSRP) of the SLSSs transmitted by the n_{SR} detected SyncRefs. The UE takes l samples of each SyncRef within a given period of time (t_M). As the SLSSs are transmitted with fixed periodicity, the UE only needs to be in Rx mode for the known corresponding SFs, as depicted in Figure 1. The UE uses the information contained in the SLSS and the S-RSRP measurements to select the most suitable SyncRef and synchronize to it. Third, the UE evaluates the selected SyncRef if any, measuring its S-RSRP for a given period of time (t_E). This information supports the decision process used to determine if the UE itself needs to become a SyncRef.

We denote the duration of the ith SyncRef selection process as $T_{SS}(i)$, which is variable, as each sub-process is conditionally executed depending on the result of the previous sub-process. An idle period of length $T_{ID}(i)$ follows the ith SyncRef selection process, in which no

Cintron, Fernando; Gamboa Quintiliani, Samantha; Griffith, David; Rouil, Richard.
"Adaptive synchronization reference selection for out-of-coverage Proximity Services (ProSe)."
Paper presented at 28th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Montreal, QC, Canada.
October 8, 2017 - October 13, 2017.

TABLE I
LIST OF SYMBOLS

Symbol	Definition
t_{S}	Duration of the SyncRef search sub-process
t_{M}	Duration of the S-RSRP measurement sub-process
l	Number of S-RSRP measurement samples
t_{E}	Duration of the SyncRef evaluation sub-process
$T_{\mathrm{SS}}(i)$	Duration of the SyncRef selection process i
$T_{\mathrm{ID}}(i)$	Duration of the idle period after the SyncRef selection process i
$T_{\mathrm{SC}}(i)$	Duration of the synchronization cycle i
$n_{\mathrm{SR}}(i)$	Number of detected SyncRefs during the SyncRef selection process i
$n_{\mathrm{RX}}(i)$	Number of SFs in Rx mode during the synchronization cycle i
Δ_{RX}^{X}	Ratio of time the UE spent in Rx mode during a given period of duration X SFs
t_{SCP}	Duration of the SCP
t_{CCH}	Duration of the PSCCH
t_{SCH}	Duration of the PSSCH
t_{TRP}	Duration on the TRP
n_{TRP}	Number of repetitions of the TRP within the PSSCH
k_{TRP}	Number of SFs available for transmission in each TRP
$n_{\mathrm{TO}}(i)$	Number of SFs with transmission opportunities during the synchronization cycle i
$n_{\mathrm{TS}}(i)$	Number of SFs with scheduled transmissions during the synchronization cycle i
$\beta_{\mathrm{TX}}(i)$	Ratio of transmission opportunities with scheduled transmissions during the synchronization cycle i
$n_{\mathrm{DR}}(i)$	Number of SFs with dropped transmission during the synchronization cycle i
$\beta_{\mathrm{DR}}(i)$	Ratio of SFs with dropped transmissions during the synchronization cycle i
Δ_{DR}^{X}	Transmission drop rate for a given period of duration X SFs
γ	Maximum ratio of time the UE is allowed to be in Rx mode due to the synchronization function
δ	Maximum allowed transmission drop rate
T	Length of the synchronization cycles when using periodic SyncRef selection triggering algorithm
T_{RX}	Feasible T satisfying the constraint in the time in Rx mode
T_{DR}	Feasible T satisfying the transmission drop rate constraint
E_i	Duration of the estimation period for the adaptive SyncRef selection triggering algorithm
C_i	Duration of the calculation period for the adaptive SyncRef selection triggering algorithm
$n_{\mathrm{TS}}^{\mathrm{ID}}(i)$	Number of SFs with scheduled transmissions during the idle period of the synchronization cycle i
$n_{\mathrm{TS}}^{\mathrm{SS}}(i)$	Number of SFs with scheduled transmissions during the SyncRef selection i

synchronization operations requiring the UE to be in Rx mode are performed. The ith synchronization cycle, of duration $T_{\mathrm{SC}}(i)$, is the period comprising the SyncRef selection i and the corresponding idle period.

During the ith synchronization cycle, the UE spends $n_{\mathrm{RX}}(i)$ SFs in Rx mode. We assume one SF per S-RSRP measurement sample for the measurement and evaluation sub-processes. Thus, $n_{\mathrm{RX}}(i)$ is given by:

$$n_{\mathrm{RX}}(i) = t_{\mathrm{S}} + n_{\mathrm{SR}}(i) \times l + l. \tag{1}$$

For a given evaluation period of N SFs in which m synchronization cycles occurred, the fraction of time the UE spent in Rx mode (Δ_{RX}^{N}) is

$$\Delta_{\mathrm{RX}}^{N} = \frac{1}{N} \sum_{i=1}^{m} n_{\mathrm{RX}}(i). \tag{2}$$

The LTE-A standard defines some parameters and constraints related to the SyncRef selection process [4]. First, the UE should be able to identify newly detectable SyncRef UEs within 20 s, thus, the SyncRef selection process should be executed at least once every 20 s. Second, it provides the values for the measurement and evaluation period length, i.e., t_{M} = 400 ms and t_{E} = 800 ms. Third, the UE should be able to measure up to six (6) SyncRef in each SyncRef selection process. Last, the UE can drop a maximum of 2 % of its sidelink communication transmissions at the physical layer for the purpose of SyncRef UE selection within a 20 s period.

B. Sidelink communication

The sidelink communication function is performed over periodically repeating Sidelink Communication Periods (SCPs) of duration t_{SCP} [10]. Each SCP is composed of two channels: the Physical Sidelink Control Channel (PSCCH) of duration t_{CCH} and the Physical Sidelink Shared Channel (PSSCH) of duration t_{SCH}. When the UE has data to transmit, it uses the PSCCH to send the Sidelink Control Information (SCI) message, and the PSSCH to send the data. The SCI contains the information needed by receiving UEs to decode the data in the PSSCH if it is intended for them: the data destination, the modulation and coding scheme (MCS), and the PSSCH resource assignment in time and frequency, among other parameters.

The SCI message is sent twice in the PSCCH and the resource assignment is done randomly, i.e., the used SFs and Resource Blocks (RBs) vary from one SCP to another. The data is allocated in the PSSCH following a Time Resource Pattern (TRP), which is a SF indication bitmap of fixed length t_{TRP}, and that is repeated n_{TRP} times during the PSSCH, where $n_{\mathrm{TRP}} = \lfloor \frac{t_{\mathrm{SCH}}}{t_{\mathrm{TRP}}} \rfloor$. Each TRP has k_{TRP} SFs available for transmission, and the set of TRPs to be used is defined in [10]. For each SCP, the UE randomly selects the TRP to use as well as the RBs within each SF of the TRP [11].

We denote as $n_{\mathrm{TO}}(i)$ the number of SFs with transmission opportunities the UE has during the synchronization cycle i. The parameter $n_{\mathrm{TO}}(i)$ is calculated using Eq. (3), given that two transmissions are needed for the PSCCH and that the k_{TRP} transmissions are repeated n_{TRP} times in the PSSCH.

$$n_{\mathrm{TO}}(i) = \frac{T_{\mathrm{SC}}(i)}{t_{\mathrm{SCP}}} \left(2 + n_{\mathrm{TRP}}\, k_{\mathrm{TRP}}\right). \tag{3}$$

We assume the UE will use all the transmission opportunities within a SCP if it has data to transmit. However, if the UE does not have data to transmit at the beginning of a given SCP, the transmission opportunities within that SCP are not used, and these timeslots are free to be used to perform other operations. We denote as $n_{\mathrm{TS}}(i)$ the number of timeslots in which the UE

Cintron, Fernando; Gamboa Quintiliani, Samantha; Griffith, David; Rouil, Richard.
"Adaptive synchronization reference selection for out-of-coverage Proximity Services (ProSe)."
Paper presented at 28th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Montreal, QC, Canada.
October 8, 2017 - October 13, 2017.

scheduled a transmission during the synchronization cycle i. Thus, $n_{\text{TS}}(i)$ is given by Eq. (4), where $\beta_{\text{TX}}(i)$ is the ratio of transmission opportunities the UE intended to use to perform transmissions within the synchronization cycle i. The parameter $\beta_{\text{TX}}(i)$ characterizes the traffic pattern of the transmitting UE, as it is an indication of the ratio of time the UE has data to transmit.

$$n_{\text{TS}}(i) = \beta_{\text{TX}}(i)\, n_{\text{TO}}(i). \tag{4}$$

C. Transmission drops due to SyncRef selection

We denote as $n_{\text{DR}}(i)$ the number of timeslots in which the UE drops a transmission during the ith synchronization cycle. This implies that transmissions were scheduled for these SFs, but the UE was in Rx mode performing operations associated to the SyncRef selection. Nevertheless, the UE does not drop transmissions every time it is in Rx mode for synchronization purposes, as can be seen in Figure 1. The specific SFs in which the UE is in Rx mode during a SyncRef selection process depend on external factors to the UE, e.g., number of SyncRefs in proximity and their timing. The specific SFs where the UE schedules its communication transmissions depend on random processes, e.g., selection of TRP and PSCCH timeslots. We define $\beta_{\text{DR}}(i)$ to be the fraction of SFs during the ith SyncRef selection in which the UE was in Rx mode and a transmission drop occurred. Thus,

$$n_{\text{DR}}(i) = \beta_{\text{DR}}(i)\, n_{\text{RX}}(i). \tag{5}$$

Furthermore, the transmission drop rate for the evaluation period of N SFs (Δ^N_{DR}) in which m synchronization cycles occurred is

$$\Delta^N_{\text{DR}} = \frac{\sum_{i=1}^{m} n_{\text{DR}}(i)}{\sum_{i=1}^{m} n_{\text{TS}}(i)}. \tag{6}$$

D. System constraints

The length of the synchronization cycles occurring during the evaluation period of N SFs should be chosen to guarantee Eq. (7) and Eq. (8) for that period. The parameter γ is the maximum fraction of time the UE is allowed to be in Rx mode due to the synchronization function, and δ is the maximum allowed transmission drop rate.

$$\Delta^N_{\text{RX}} \leq \gamma. \tag{7}$$

$$\Delta^N_{\text{DR}} \leq \delta. \tag{8}$$

For the rest of the paper we will use δ = 0.02 and $N = 20\,000$ ms (20 s) in order to align with the LTE-A standard requirements mentioned in Section II-A2. The fraction of time spent in Rx mode, γ, is an indicator of the maximum power that the UE can allot to the SyncRef selection process, since the UE must expend power to actively listen for SyncRef signals and to do the computations to determine which SyncRef to follow.

TABLE II
EVALUATION PARAMETERS

	Param.	Value		Param.	Value
Eval. period	N (ms)	20000	**Sidelink comm.**	t_{SCP} (ms)	40
SyncRef selection	t_{S} (ms)	40		t_{CCH} (ms)	8
	t_{M} (ms)	400		t_{SCH} (ms)	32
	t_{E} (ms)	800		t_{TRP} (ms)	8
	l	4		k_{TRP}	2

At present, the standard does not give a value for γ, so we consider several values in this paper.

III. Problem Formulation

When using a periodic algorithm to trigger the SyncRef selection process, all the synchronization cycles have the same length T, namely, $T_{\text{SC}}(1) = T_{\text{SC}}(2) = ... = T_{\text{SC}}(m) = T$. The selection of T is critical in order to satisfy Eq. (7) and Eq. (8) because it determines the number of SyncRef selection processes to be executed within an evaluation period. This selection is challenging as the values of Δ^N_{RX} and Δ^N_{DR} are dependent on multiple variable factors. However, it is possible to estimate a set of possible values for T that satisfy the constraints in worst case conditions.

For a fixed period T, the number of synchronization cycles within an evaluation period of N SFs is given by $m = \frac{N}{T}$. By grouping Equations (1) – (8), and assuming that all the synchronization cycles in the evaluation period are identical and that the UE detects the maximum allowed number of SyncRefs each time ($n_{\text{SR}} = 6$), we find expressions for:

- The values of T satisfying Eq. (7):

$$T_{\text{RX}} \geq \frac{(t_{\text{S}} + 6\,l + l)}{\gamma}. \tag{9}$$

- The values of T satisfying Eq. (8):

$$T_{\text{DR}} \geq \frac{\beta_{\text{DR}}\,(t_{\text{S}} + 6\,l + l)\,t_{\text{SCP}}}{\beta_{\text{TX}}\,(2 + n_{\text{TRP}}\,k_{\text{TRP}})\,\delta}. \tag{10}$$

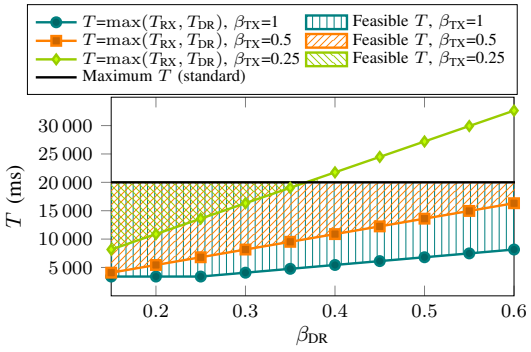


Fig. 2. Feasible values of the synchronization cycle length for the periodic SyncRef selection triggering, considering different traffic intensities (β_{TX}) and ratio of transmission drops (β_{DR}). Constraints: δ = 0.02 and γ = 0.02.

Cintron, Fernando; Gamboa Quintiliani, Samantha; Griffith, David; Rouil, Richard.
"Adaptive synchronization reference selection for out-of-coverage Proximity Services (ProSe)."
Paper presented at 28th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Montreal, QC, Canada.
October 8, 2017 - October 13, 2017.

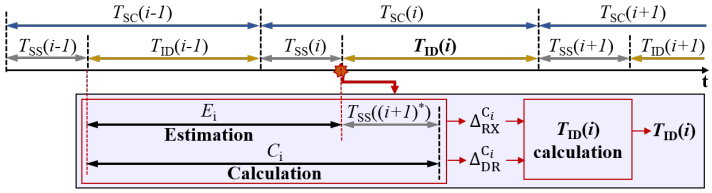


Fig. 3. Proposed adaptive SyncRef selection triggering algorithm.

Thus, one should choose T to satisfy Eq. (11), which considers both constraints and $T < 20$ s (Section II-A2).

$$\max(T_{RX}, T_{DR}) \leq T \leq 20\,000 \text{ ms}. \tag{11}$$

Figure 2 shows feasible values for the period T for different values of activity factors β_{TX} and transmission drops β_{DR} using the parameters in Table II. We display only the possible values for β_{DR} given that configuration. We see that the set of feasible values for T decreases with β_{TX}, but most importantly, we can see that there is no feasible T for $\beta_{TX} = 0.25$ when $\beta_{DR} > 0.37$. This implies that a single fixed value of T is not able to guarantee that the SyncRef selection algorithm will satisfy the required performance constraints for all situations.

To achieve fast convergence in the distributed out-of-coverage synchronization algorithm, T should be as small as possible to increase the frequency of SyncRef selection. Choosing T following a worst case scenario for a given value of β_{TX} is inefficient for larger values of β_{TX}, as T is unnecessarily large. For example, in the worst case ($\beta_{DR} = 0.6$), the minimum feasible value is $T = 16\,320$ ms for $\beta_{TX} = 0.5$, which is double the needed T for $\beta_{TX} = 1$ ($T = 8\,160$ ms). Moreover, the values of n_{SR}, β_{TX}, and β_{DR} vary from one SyncRef selection to another, making a worst case T selection even more inefficient. To overcome those limitations, we propose an algorithm that makes the value of T variable, adjusting it depending on the conditions experienced by the UE.

IV. Proposed Proactive Algorithm

We developed an adaptive SyncRef selection triggering algorithm, whose objective is to reduce the convergence time by triggering the selection process as soon as possible. This is done while respecting the constraints of time allowed in receiving state (Eq. (7)) and maximum packet drop rate (Eq. (8)) for every evaluation period. Thus, the length of the synchronization cycles (T_{SC}) varies depending on the UE conditions. A schematic representation of the proposed algorithm is presented in Figure 3 and described in this section.

The proposed algorithm estimates the suitable duration for the synchronization cycle in progress ($T_{SC}(i)$). To do so, it calculates the value of $T_{ID}(i)$ at the end of the current SyncRef selection process as can be seen in Figure 3.

The UE uses two sets of information:

- The data collected for the proactive *estimation period* of duration $E_i = T_{ID}(i-1) + T_{SS}(i)$, which comprises the idle period of the previous synchronization cycle, and the current SyncRef selection. The data includes the amount of time spent in Rx mode and the number of transmission drops.
- A prediction of the information for the next SyncRef selection, denoted as $(i+1)^*$, with a duration $T_{SS}((i+1)^*)$. We consider two predictions: a strict (S) prediction that assumes the worst case scenario for the next SyncRef selection (e.g., $n_{SR} = 6$, selection of a SyncRef and $\beta_{DR} = 1$); and a historical (H) prediction that assumes the next process will be similar to the one that occurred during the estimation period.

Using this information, the UE calculates the fraction of time in Rx mode and the transmission drop rate for the proactive *calculation period* of duration $C_i = E_i + T_{SS}((i+1)^*)$, namely $\Delta_{RX}^{C_i}$ and $\Delta_{DR}^{C_i}$, using Eq. (12) and Eq. (13) respectively.

$$\Delta_{RX}^{C_i} = \frac{n_{RX}(i) + n_{RX}((i+1)^*)}{C_i}. \tag{12}$$

$$\Delta_{DR}^{C_i} = \frac{n_{DR}(i) + n_{DR}((i+1)^*)}{n_{TS}^{ID}(i-1) + n_{TS}^{SS}(i) + n_{TS}^{SS}((i+1)^*)}. \tag{13}$$

The denominator of Eq. (13) is composed of the scheduled transmissions during the calculation period, i.e., during the previous idle period ($n_{TS}^{ID}(i-1)$) and SyncRef selection ($n_{TS}^{SS}(i)$), and the prediction for the next SynRef selection ($n_{TS}^{SS}((i+1)^*)$).

Next, the UE estimates $T_{ID}(i)$ relative to the previous idle period using Eq. (14). Thus, $T_{ID}(i)$ increases compared to $T_{ID}(i-1)$ if any of the constraints are not met during the calculation period, and is reduced otherwise.

$$T_{ID}(i) = \max\left(\frac{\Delta_{RX}^{C_i}}{\gamma} T_{ID}(i-1), \frac{\Delta_{DR}^{C_i}}{\delta} T_{ID}(i-1)\right). \tag{14}$$

Finally, the UE sets a timer $T_{ID}(i)$ and performs the $(i+1)$st SyncRef selection upon its expiration.

Cintron, Fernando; Gamboa Quintiliani, Samantha; Griffith, David; Rouil, Richard.
"Adaptive synchronization reference selection for out-of-coverage Proximity Services (ProSe)."
Paper presented at 28th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Montreal, QC, Canada.
October 8, 2017 - October 13, 2017.

V. Evaluation

In previous work, we extended the LTE module of the ns-3 simulation platform [12] to consider ProSe functionalities [13]. The following evaluation was performed using this implementation.

A. Configuration

We considered 24 out-of-coverage UEs in proximity in a broadcast scenario. Half of the UEs transmit data to the group, and the other half is silent but receives the transmitted data. Each UE was configured with a different random SLSSID, frame, and subframe number at the beginning of each simulation, creating an initially non-synchronized environment. The convergence time is defined as the time required for all UEs to acquire the same synchronization parameters. We used the parameters in Table II for the configuration of the sidelink communication and SyncRef selection.

The transmitter UEs followed an on-off traffic pattern with On and Off periods exponentially distributed with mean μ_{ON} and μ_{OFF} respectively. We set $\mu_{\text{ON}} = 2.5\,\text{s}$, and we varied μ_{OFF} to evaluate scenarios with different TAF[1]. The periodic algorithms were worst-case configured ($\beta_{\text{DR}} = 0.6$ and $\beta_{\text{TX}} = \text{TAF}$). The adaptive algorithm was configured with $\delta = 0.02$. Both types of prediction, strict (S) and historical (H), were tested for the proposed proactive (Pro) algorithm. The simulation time was 1000 s and each configuration was simulated 150 times using different random seeds.

For all evaluations, we monitored the fraction of time in Rx mode ($\Delta^{20\text{s}}_{\text{RX}}(i,t)$) and the transmission drop rate ($\Delta^{20\text{s}}_{\text{DR}}(i,t)$) for each UE i and for each monitoring period of 20 s in the simulation. To this purpose, we use a monitoring sliding window of 20 s length, advancing every 1 ms.

B. Results

The results presented in this section correspond to the scenario with TAF = 50 %, unless otherwise stated. Similar trends were observed for the other scenarios, and figures are omitted due to lack of space.

We observe that the adaptive algorithm is able to reduce the convergence time of the scenarios while satisfying the system constraints in most of the monitoring periods, as shown in Table III and Figure 4. Table III shows the percentage of monitoring period instances in the whole evaluation that do not satisfy the transmission drop rate constraint. These cases are observed because a monitoring window containing part of a SyncRef selection process with transmission drops, and sliding from a period with successful transmissions to a period without any transmission, will exhibit an increased transmission drop rate for that monitoring period. This is confirmed by the increase of such cases when the transmitters have larger off periods, i.e., with smaller TAF. However, we see that all the values in Table III are below 1 %, which represents very good performance considering the granularity of the monitoring.

[1]The Traffic Activity Factor (TAF) denotes the average fraction of time the UE is transmitting, and it is calculated as $\text{TAF} = \frac{\mu_{\text{ON}}}{\mu_{\text{ON}}+\mu_{\text{OFF}}}$.

TABLE III
Percentage of monitoring periods where $\Delta^{20\text{s}}_{\text{DR}}(i,t) > \delta$, considering all transmitters in all simulations

			Scenario	
			TAF = 75 %	TAF = 50 %
Algorithm	Fix		0	0.0033
	γ = 0.02	Pro S	0	0.0019
		Pro H	0.0004	0.0390
	γ = 0.12	Pro S	0	0.0025
		Pro H	0.0063	0.4573
	γ = 0.24	Pro S	0	0.0016
		Pro H	0.0068	0.4536
	γ = 1.00	Pro S	0	0.0010
		Pro H	0.0550	0.4538

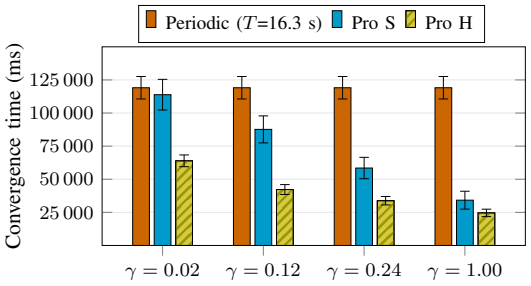


Fig. 4. Convergence time. Average with 95% confidence interval are shown.

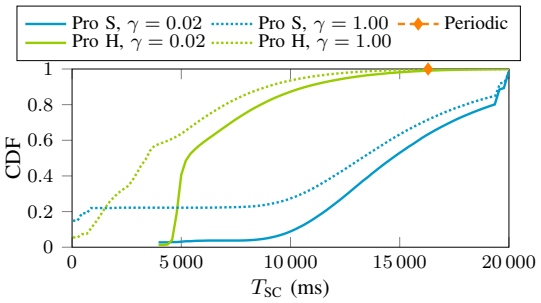


Fig. 5. Cumulative distribution function of the length of the synchronization cycles (T_{SC}) for the transmitter UEs.

For all the scenarios and configurations considered, $\Delta^{20\text{s}}_{\text{RX}}(i,t) < \gamma$ for all UEs. The length of the synchronization cycles is mostly influenced by the constraint in the transmission drop rate for the transmitters, which is stricter. For the receivers, there is no large variation in the time spent in Rx mode between consecutive synchronization cycles, which indicates that the proactive algorithm can easily adapt to respect the constraint.

A larger γ allows us to perform the SyncRef selection more often in periods of low transmission drop rate or

Cintron, Fernando; Gamboa Quintiliani, Samantha; Griffith, David; Rouil, Richard.
"Adaptive synchronization reference selection for out-of-coverage Proximity Services (ProSe)."
Paper presented at 28th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Montreal, QC, Canada.
October 8, 2017 - October 13, 2017.

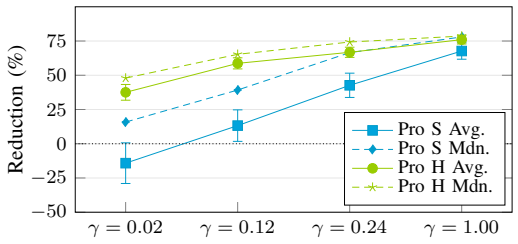


Fig. 6. Percentage of reduction in the convergence time regarding the periodic algorithm. Average with 95% confidence interval and median are shown.

when transmitters are in off periods. This is reflected in Figure 5, where we observe larger proportion of smaller synchronization cycles with $\gamma = 1.00$ than with $\gamma = 0.02$. Thus, the transmitters synchronize faster and the receivers react faster to these SyncRef changes when the algorithms are configured with larger γ, which reduces the convergence time as can be seen in Figure 4.

As expected, the strict prediction handles sudden increases in the transmission drop rate better than its historical counterpart. This is reflected by fewer cases in which $\Delta_{\text{DR}}^{20s}(i,t) > \delta$ for the strict algorithm (Table III). However, the strict criterion provides less flexibility and the reductions in convergence time are smaller than when using historical prediction. The performance gap decreases when γ increases, as larger γ increases performance by acting in the off or low transmission drop periods, balancing the transmission drops in the predictions as shown in Figure 4.

Although the adaptive algorithm reduces the average convergence time, it did not produce an improvement in all cases. Figure 6 depicts the percentage of convergence time reduction of the proactive algorithm compared to the periodic algorithm. It is calculated per simulation, i.e., it compares the algorithms under the same initial conditions. From Figure 6, we observe that the average value is below the median value for all of the cases, which reflects the influence of few cases with small reductions or even increases (negative reductions) in the convergence time. A good example can be seen for the algorithm with strict prediction and $\gamma = 0.02$. However, the central tendency is better represented by the median in this case, and, since it is positive, it highlights the overall performance gain achieved by the adaptive algorithm.

A configuration with $\gamma = 1.00$ allows UEs to be in Rx mode all the time, yet this was never the case in our evaluation. UEs were always able to detect SyncRefs and were in Rx mode during only a fraction of the measurement and evaluation periods for performing the S-RSRP sampling. The maximum value for the fraction of time in Rx mode attained by the UEs in our evaluation was 14.7 %. This value will be greater if the UEs cannot detect any SyncRef and are constantly performing the SyncRef search. Thus, adapting the value of γ depending on the UE conditions is crucial to control power consumption, e.g., $\gamma = 1.00$ when detecting multiple SyncRefs to ensure fast convergence, and reducing its value when convergence is achieved or when no SyncRef is detected in order to preserve battery charge.

VI. Conclusion

In this paper, we studied the synchronization protocol used by out-of-coverage ProSe-enabled UEs, such as those that will be used by first responders. We showed that choosing a fixed period for triggering the synchronization reference selection can be challenging and inefficient if constraints in the transmission drop rate and the time in reception mode are considered. We proposed an algorithm that reduces the synchronization delays by adapting the time to trigger the process depending on the local conditions of each UE. We evaluated the efficiency of the algorithms using system level simulations and we pointed out the trade-offs that should be considered to achieve a given level of performance and satisfy the LTE-A standard requirements.

References

[1] 3GPP, "Technical Specification Group Services and System Aspects; Proximity-based services (ProSe); Stage 2 v.12.7.0," 3rd Generation Partnership Project (3GPP), TS 23.303, 2015.

[2] G. Fodor, S. Parkvall, S. Sorrentino *et al.*, "Device-to-device communications for national security and public safety," *IEEE Access*, vol. 2, pp. 1510–1520, 2014.

[3] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation v.12.8.0," 3rd Generation Partnership Project (3GPP), TS 36.211, 2016.

[4] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA); Requirements for support of radio resource management v.12.12.0," 3rd Generation Partnership Project (3GPP), TS 36.133, 2014.

[5] J. Salo and J. Reunanen, "Interlayer Mobility Optimization," in *LTE Small Cell Optimization: 3GPP Evolution to Release 13*. John Wiley & Sons, Ltd., 2016.

[6] S. Gamboa, F. J. Cintrón, D. Griffith *et al.*, "Impact of timing on the Proximity Services (ProSe) synchronization function," in *2017 IEEE 14th Consumer Communications & Networking Conference (CCNC)*, 2017.

[7] S. L. Chao, H. Y. Lee, C. C. Chou *et al.*, "Bio-Inspired Proximity Discovery and Synchronization for D2D Communications," *IEEE Communications Letters*, vol. 17, no. 12, 2013.

[8] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification v.12.8.0," 3rd Generation Partnership Project (3GPP), TS 36.331, 2016.

[9] M. Cannon, "On the Design of D2D Synchronization in 3GPP," in *IEEE ICC 2015 - Workshop on Device-to-Device Communication for Cellular and Wireless Networks*, 2015.

[10] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures v.12.11.0," 3rd Generation Partnership Project (3GPP), TS 36.213, 2016.

[11] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification v.12.9.0," 3rd Generation Partnership Project (3GPP), TS 36.321, 2016.

[12] NS-3 Consortium, "ns-3 Network Simulator," Available at: https://www.nsnam.org/.

[13] R. Rouil, F. J. Cintrón, A. Ben Mosbah *et al.*, "Implementation and Validation of an LTE D2D Model for ns-3," in *2017 Workshop on ns-3 (WNS3)*, 2017.

Cintron, Fernando; Gamboa Quintiliani, Samantha; Griffith, David; Rouil, Richard.
"Adaptive synchronization reference selection for out-of-coverage Proximity Services (ProSe)."
Paper presented at 28th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Montreal, QC, Canada.
October 8, 2017 - October 13, 2017.

Unsupervised Clustering for Millimeter-Wave Channel Propagation Modeling

Jian Wang[1], Camillo Gentile[1], Jelena Senic[2], Ruoyu Sun[2], Peter B. Papazian[2], Chiehping Lai[1]

[1]Wireless Networks Division, National Institute of Standards and Technology, Gaithersburg, MD, USA
[2]RF Technology Division, National Institute of Standards and Technology, Boulder, CO, USA
e-mail: {jian.wang, camillo.gentile, jelena.senic, ruoyu.sun, peter.papazian, chiehping.lai}@nist.gov

***Abstract*—To date, we have designed and assembled millimeter-wave channel sounders at 60 GHz and 83 GHz. They can estimate the angle-of-departure and angle-of-arrival of channel multipath components as well as their delay and Doppler frequency shift. In addition, due to the fast acquisition time and because the receiver is mounted on a mobile robot, the systems can collect measurements for hundreds of different transmitter-receiver configurations in just minutes. It follows that channel-model reduction, including the multipath-component clustering process, must be reliable, consistent, and unsupervised. In this paper, we describe a simple clustering process tailored to the properties of millimeter-wave channels that fully exploits the multi-dimensionality of the extracted multipath components and requires only a few tunable parameters. Through extensive experimentation, we have verified that the process is robust and delivers consistent results across five different environments and across both frequency bands investigated. Illustrative examples are provided.**

Index Terms— 5G; double-directional channel; mmWave.

I. Introduction

The inherent drawback of millimeter-wave (mmWave) bands compared to sub-6-GHz bands is greater path loss due to free-space[1], oxygen-absorption, and/or penetration losses. To compensate the link budget, extremely high-gain – in turn extremely narrow-beam – antennas will be employed. Given their limited beamwidth, the antennas must be electronically steered in azimuth and elevation along the angle-of-departure (AoD) and angle-of-arrival (AoA) of any viable propagation paths between the respective transmitter (TX) and receiver (RX) to provide omnidirectional fields of view. Hence from a channel-modeling perspective, it is fundamental to understand how many paths are available in an environment and their distribution in both delay and 3D double-directional angle (azimuth AoD, elevation AoD, azimuth AoA, elevation AoA).

Diffraction at mmWave frequencies has been demonstrated to be significantly weaker than at sub-6-GHz [1]. The absence of diffraction renders the channel "sparse," meaning that only a few dominant multipath components (MPCs) will be detected. The strongest will be the direct path (in line-of-sight (LOS) conditions) while the others will originate from specular reflections off ambient objects. Each specular reflection gives rise to scattering of the incident wave into a dominant specular multipath component and weaker diffuse components surrounding in the multi-dimensional delay-angle space; altogether they form a *cluster*. The dominant paths can be exploited to send multiple data streams between the TX and RX. In other words, the number of clusters will determine the maximum number of independent streams that can be sent in one polarization[2].

Whether the clusters are distributed randomly in the delay-angle space (e.g. WINNER II [3], COST 2100 [4] 5GCMSIG [5]) or follow a deterministic map-based distribution (e.g. METIS2020 [6], MiWEBA [7], mmMAGIC [8]), a fundamental process in reducing channel models is clustering the MPCs extracted from measurements. Besides their distribution in space, also important are the delay- and angle-dispersion characteristics of the clusters, i.e. their shape. When clusters overlap, separating them is a challenge. There are two reasons why the challenge is diminished in mmWave systems compared to legacy systems: 1. the channel is sparse, hence there are less components a priori; 2. 5G systems will fully exploit the multi-dimensionality of the channel – for proper radio and network design, the path loss of the extracted MPCs should be polarization-dependent (VV, VH, HV, HH) and indexed according to delay, 3D double-directional angle, and Doppler frequency shift – hence cluster overlap in all six dimensions is less likely.

To date, we have designed and assembled mmWave channel sounders at 60 GHz [9] and at 83 GHz [10]; the latter features a three-dimensional switched array at the RX for AoA discrimination in both azimuth and elevation while the former features an array at the TX as well for AoD discrimination. Besides high delay resolution (up to 0.5 ns), the systems are capable of estimating Doppler shift. In addition, due to the fast acquisition time and because the RX is mounted on a mobile robot, the systems can collect channel measurements for hundreds of different TX-RX configurations in just minutes. It follows that model reduction, including the MPC clustering process, must be reliable, consistent, and unsupervised.

In this paper, we describe a simple clustering process tailored to the properties of both mmWave channels and 5G systems that accepts only a few tunable parameters. Through extensive measurements, we verified the process to be robust and deliver consistent results across five different environments (lecture room, lobby, hallway, basement, and data center) in LOS and non-LOS conditions and across both frequency bands. The remainder of this paper is organized as follows. In Section II, we present our clustering process. In Section III, we show some illustrative examples. Finally, we draw some important conclusions in Section IV.

[1]Assuming a frequency-dependent aperture size, not a fixed physical antenna size.

[2]Dually polarized streams are provisioned for mmWave systems [2].

Gentile, Camillo; Papazian, Peter; Senic, Jelena; Sun, Roy; Wang, Jian.
"Unsupervised Clustering for Millimeter-Wave Channel Propagation Modeling."
Paper presented at 2017 IEEE Vehicular Technology Conference - Fall, Toronto, ON, Canada. September 24, 2017 - September 27, 2017.

II. Clustering Process

In this section, we describe the process through which the multipath components extracted from a channel measurement are clustered. Before the process begins, we check whether the TX-RX configuration lies in LOS conditions; if so, the direct path – the strongest of all MPCs – is easily detected and removed[3]. As there is no scattering associated with it, nor are there diffuse components surrounding it. As such, the clustering is only applied to the remaining components.

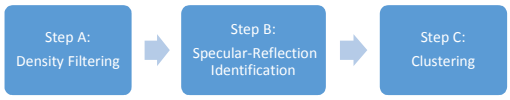
The components to be clustered are indexed through i. Each is characterized by path gain[4] PG_i (in dB) in the space $\boldsymbol{x}_i = (\tau_i, \theta_i^{A,AoD}, \theta_i^{E,AoD}, \theta_i^{A,AoA}, \theta_i^{E,AoA}, \nu_i)$, where τ denotes delay, θ the azimuth (A) or elevation (E) AoD or AoA, and ν the Doppler frequency shift. Fig. 1(a) displays an example of MPCs extracted from a measurement in a data center with our 60 GHz system. The floorplan of the environment and the TX-RX configuration are shown in Fig. 1(b). Because each dimension j will have a different range, the space is normalized as

$$\hat{x}_i^j = \frac{x_i^j - \min_i x_i^j}{\max_i x_i^j - \min_i x_i^j}$$

so that every $\hat{x}_i^j$ falls within the range [0,1]. The clustering is then carried out in the normalized space.

Often clusters from a single acquisition contained too few components, prohibiting sufficient statistical characterization of the shape. As such, for each TX-RX configuration, in reality MPCs from 8 acquisitions were aggregated while the RX was moving on the mobile robot; since acquisitions were at least a wavelength apart, the scattering between them is known to be independent. Fig. 2 shows a cluster identified in the data center indexed in the delay dimension only. Each symbol represents a different acquisition. Note that the specular components (blue) are significantly stronger than the diffuse components (green).

The remainder of this section is partitioned into three subsections, each describing a sequential step of the clustering process:



A. *Density Filtering*

The first step of the clustering process is to filter out any spurious paths, typically originating from weak diffraction or small ambient objects. To this end, we apply the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm [11]. In the algorithm, each MPC is scanned to determine whether it is surrounded by at least N_m paths (including itself) within some radius ε, using the Euclidean distance metric in the space. If so,

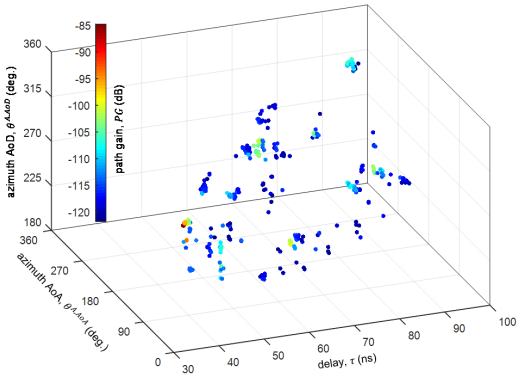


Fig. 1(a). Multipath components extracted from a measurement in a data center with our 60 GHz system. Each component is indexed in delay, azimuth AoD, and azimuth AoA (elevation AoD, elevation AoA, and Doppler shift are omitted) and is color-coded against the path-gain legend.

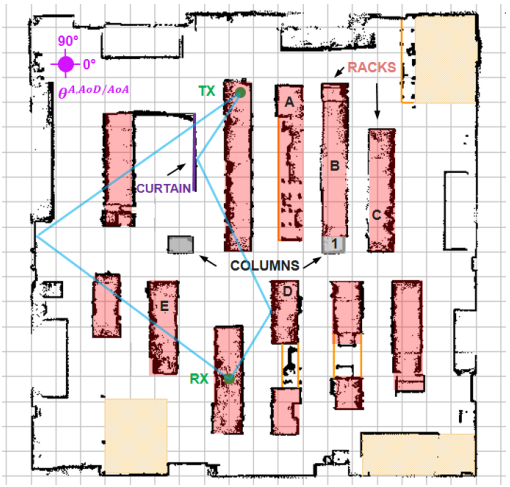


Fig. 1(b). Floorplan of a data center and example configuration with the TX and RX on top of racks. Reflectors that generated the detected multipath components in the channel measurement are labeled. Shown in cyan are example first- and second-order reflections from ambient objects.

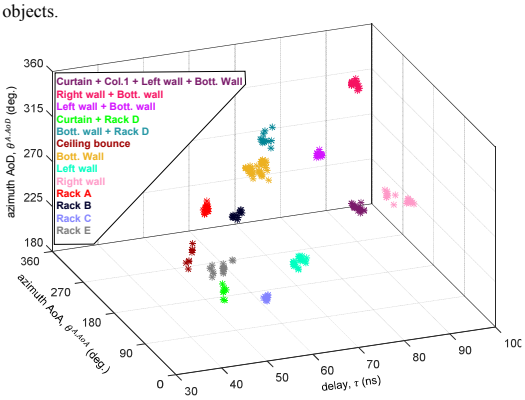


Fig. 1(c). Clustering of multipath components in Fig. 1(a). Each cluster is displayed as a different color and is labeled in reference to the reflector(s) in Fig. 1(b) that generated it.

[3]In non-LOS conditions, the direct path is assumed to go undetected as penetration loss at mmWave frequencies is very high.

[4]In our measurements, we only had VV polarization; but if multiple polarizations are available, they could be exploited as separate dimensions for enhanced discrimination.

Gentile, Camillo; Papazian, Peter; Senic, Jelena; Sun, Roy; Wang, Jian.
"Unsupervised Clustering for Millimeter-Wave Channel Propagation Modeling."
Paper presented at 2017 IEEE Vehicular Technology Conference - Fall, Toronto, ON, Canada. September 24, 2017 - September 27, 2017.

the MPC is designated as a *core point*. Then all neighbors (within the same radius ε) of a core point (including the core point itself) are designated as *reachable* and so kept; otherwise they are deemed outliers and consequently discarded. This step yields multiple *regions*, each a collection of mutually reachable MPCs.

In our experiments, we set $\varepsilon = 0.04$ fixed and N_m to vary between 3 and 8 as

$$N_m = \left\lfloor \frac{8}{3} \cdot 3^{\widehat{PG}_i} \right\rfloor,$$

$$\widehat{PG}_i = \frac{PG_i - \min_i PG_i}{\max_i PG_i - \min_i PG_i}.$$

Hence the condition on the number of neighbors is relaxed with lower path gain. While these parameters can be tuned through visual inspection on a sample set of measurements, we found them to be robust and as such were maintained constant across the five environments and two center frequencies tested.

B. Specular-Reflection Identification

A defining feature of clusters at mmWave frequencies is that each one contains a single specular component marked by a peak in path gain. It may occur that multiple specular components fall within a region delineated in Step A. This will occur especially if the dimensionality of the MPC space is reduced due to the limitations of the measurement system. What happens is that the MPCs are projected into a lower dimensional space, increasing chances of cluster overlap. In order to resolve separate clusters, we identify peaks within a region under the premise that each peak is linked to a specular component.

As can be observed from the example in Fig. 2, local peaks will arise due to random fluctuation of the path gain in delay. Although the delay dimension only is shown, it holds true for the other dimensions as well. As a means to isolate the global peaks in each region, we filter the path gain in the multi-dimensional space to smooth out the local peaks. First we interpolate the path gain between the MPCs through the LOWESS (Locally Weighted Scatterplot Smoothing) algorithm with robust bisquare weights [12]. This algorithm is used in regression analysis to create a continuous surface while attenuating the effect of outliers. Then we apply a Gaussian filter to average over any residual local peaks. The standard deviation of the filter is set to $\sigma = \Delta x^{\max}/12$, where $\Delta x^{\max}$ is the maximum length of the region over all dimensions. Finally, we isolate the global peaks by comparing them to their neighboring values in the smoothed space.

C. Clustering

The clustering step is initialized by pinning the clusterheads to the coordinates of the global peaks identified over all regions in Step B; hence the number of clusters corresponds to the number of global peaks. We then apply the K-Power Means Clustering algorithm [13]. In one iteration, individual MPCs are assigned to the clusterhead for which the Euclidean distance is minimum. Once the MPCs have been all assigned, the clusterhead is recomputed as the weighted centroid of the MPCs assigned to the cluster; the path gain is the weight. The iterations continue until convergence.

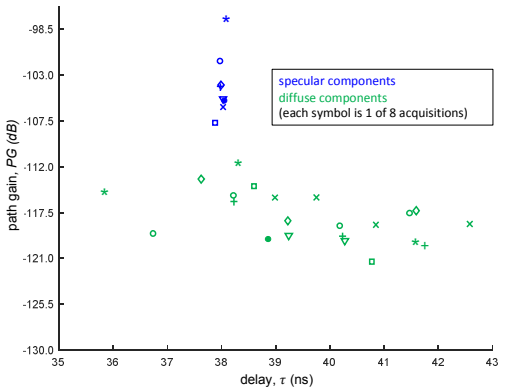


Fig. 2. Multipath components of an example cluster aggregated over eight small-scale acquisitions (each one shown with a different symbol).

III. Experimental Results

Fig. 1(c) shows the clustering results for the example measurement in the data center. Each cluster is marked with a different color. The results were validated through an exercise parallel to the measurements: the specular reflections were raytraced for the TX-RX configuration given the floorplan of the environment. Then the delay, AoD, and AoA of the raytraced paths were compared with the same properties of the specular reflections identified from the measurements and paired accordingly. The purpose was to discern the number and type of reflectors on each path from the paired raytraced path (for which this information was furnished). In Fig. 1(c), the reflectors on each of the 13 paths (one for each cluster) were labeled in reference to Fig. 1(b). Fig. 1(b) also displays paired raytraced paths for two of the 13 (the other 11 were omitted to avoid clutter). Good agreement between the locations of the reflectors in the floorplan and their delay-angle properties was witnessed, validating the mmWave cluster model.

To further substantiate the effectiveness of our clustering process, we illustrate another example with our 83 GHz system in a lecture room. Fig. 3(a) shows the MPCs extracted from the measurement for the TX-RX configuration in Fig. 3(b). Note that since there is no TX array for this system – only a single element – we could not estimate AoD. Finally, Fig. 3(c) shows the clustering results: 10 clusters are shown and each cluster is traced back to the main reflectors in Fig. 3(b).

Gentile, Camillo; Papazian, Peter; Senic, Jelena; Sun, Roy; Wang, Jian.
"Unsupervised Clustering for Millimeter-Wave Channel Propagation Modeling."
Paper presented at 2017 IEEE Vehicular Technology Conference - Fall, Toronto, ON, Canada. September 24, 2017 - September 27, 2017.

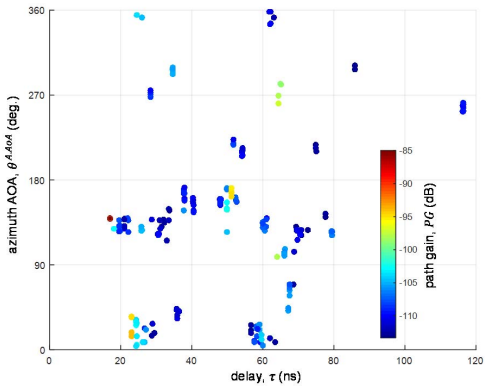


Fig. 3(a). Multipath components extracted from a measurement in a lecture room with our 83 GHz channel sounder. Each component is indexed in delay and azimuth AoA (elevation AoA and Doppler shift are omitted) and is color-coded against the path-gain legend.

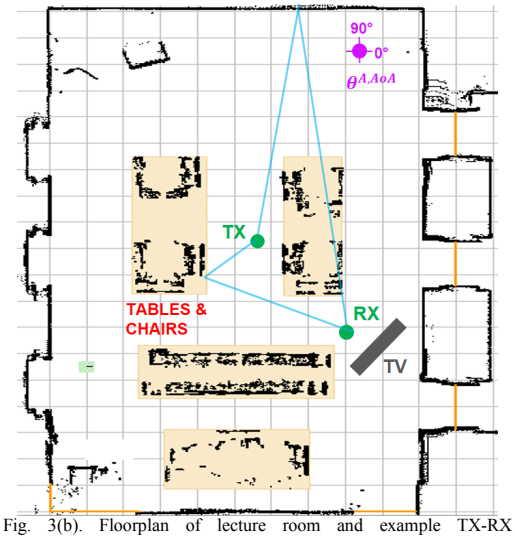


Fig. 3(b). Floorplan of lecture room and example TX-RX configuration. Reflectors that generated the detected multipath components in the channel measurement are labeled. Reflections off the Top wall and the Tables & Chairs are shown in cyan.

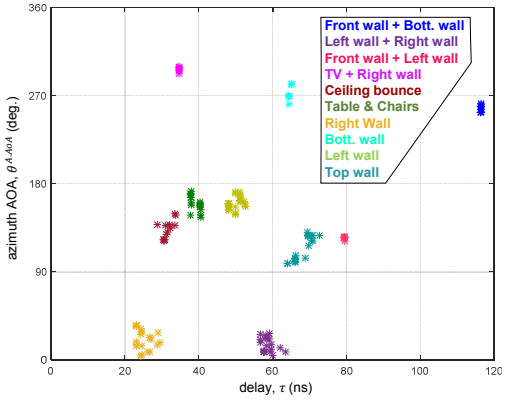


Fig. 3(c). Clustering of multipath components in Fig. 3(a). Each cluster is displayed as a different color and is labeled in reference to the reflector(s) in Fig. 3(b) that generated it.

IV. Conclusions

In this paper, we describe a clustering process for multipath components extracted from measurements with millimeter-wave channel sounders at 60 GHz and 83 GHz. The process contains few tunable parameters which are verified to be robust against measurements taken in five different environments and across the two frequency bands. While a total of six MPC dimensions are viable for clustering, the key findings of our analysis are:

1. The azimuth dimension is much more distinct than elevation because, given typical TX, RX, and ceiling heights, ground and ceiling bounces will impinge at shallow elevation angles with little variability amongst them. Elevation angle is most useful to discriminate ground from ceiling bounces at short distances (less than 5 m or so).
2. AoD and AoA are both critically important because two paths will often arrive (depart) at similar angles but with different departure (arrival) angles. They are otherwise inseparable.
3. Doppler frequency shift maps directly to angle-of-arrival (angle-of-departure) when the RX (TX) is mobile [14] and so does not provide additional information for clustering.

References

[1] J. Senic, C. Gentile, P.B. Papazian, et al., "Analysis of E-band Path Loss and Prpagation Mechanisms in the Indoor Environment," *IEE Trans. on Antennas and Propagation*, June 2017.

[2] http://www.ieee802.org/11/Reports/tgay_update.htm

[3] P. Kyosti, J. Meinila, L. Hentila, et al., "WINNER II Channel Models," Technical Report EC FP6, Sept. 2007.

[4] L. Liu, C. Oestges, J. Poutanen, et al., "The COST 2100 MIMO Channel Model," *IEEE Wireless Communications*, vol. 19, no. 6, pp. 92-99, 2012.

[5] 5G Special Interest Group homepage: http://www.5gworkshops.com/5GCM.html

[6] METIS II Project homepage: https://metis-ii.5g-ppp.eu/

[7] A. Maltsev, A. Pudeyev, I. Karls, et al., "Quasi-deterministic Approach to mmWave Channel Modeling in a Non-stationary Environment," *IEEE GLOBECOM*, Dec. 2014.

[8] mm MAGIC Project homepage: https://5g-mmmagic.eu/

[9] R. Sun, P.B. Papazian, J. Senic, et al., "Design and Calibration of a Double-directional 60 GHz Channel Sounder for Multipath Component Tracking," *IEEE EuCAP*, March 2017.

[10] P. B. Papazian, C. Gentile, K.A. Remley, et al., "A Radio Channel Sounder for Mobile Millimeter-Wave Communications: System Implementation and Measurement Assessment," *IEEE Trans. on Microwave Theory and Techniques*, vol. 64, no. 9, pp. 2924-2932, Aug. 2016.

[11] M. Ester, H. P. Kriegel, J. Sander, et al., "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *ACM KDD*, Aug. 1996.

[12] W. S. Cleveland and S. J. Devlin, "Locally Weighted Regression: An Approach to Regression Analysi by Local Fitting," Journal of the American Statistical Association, vol. 83, pp. 596-610, 1988.

[13] N. Czink, P. Cera, J. Salo, et al., "A Framework for Automatic Clustering of Parametric MIMO Channel Data Including Path Powers," *IEEE Vehicular Tech. Conf., Fall*, Sept. 2006.

[14] J. Wang, C. Gentile, P.B. Papazian, et al., "Quasi-Deterministic Model for Doppler Spread in Millimeter-wave Commmunication Systems," *IEEE Antennas and Wireless Propagation Letters*, May 2017.

Gentile, Camillo; Papazian, Peter; Senic, Jelena; Sun, Roy; Wang, Jian.
"Unsupervised Clustering for Millimeter-Wave Channel Propagation Modeling."
Paper presented at 2017 IEEE Vehicular Technology Conference - Fall, Toronto, ON, Canada. September 24, 2017 - September 27, 2017.

2017 IEEE 13th International Conference on eScience

Towards a Hybrid Human-Computer Scientific Information Extraction Pipeline

Roselyne B. Tchoua*, Kyle Chard†, Debra J. Audus‡, Logan T. Ward†,
Joshua Lequieu§, Juan J. de Pablo§ and Ian T. Foster*†¶
*Department of Computer Science, University of Chicago, Chicago, IL, USA
Email: roselyne@uchicago.edu
†The Computation Institute, University of Chicago and Argonne, Chicago, IL, USA
‡Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD, USA
§Institute for Molecular Engineering, University of Chicago, Chicago, IL, USA
¶Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA

***Abstract*—The emerging field of materials informatics has the potential to greatly reduce time-to-market and development costs for new materials. The success of such efforts hinges on access to large, high-quality databases of material properties. However, many such data are only to be found encoded in text within esoteric scientific articles, a situation that makes automated extraction difficult and manual extraction time-consuming and error-prone. To address this challenge, we present a hybrid Information Extraction (IE) pipeline to improve the machine-human partnership with respect to extraction quality and person-hours, through a combination of rule-based, machine learning, and crowdsourcing approaches. Our goal is to leverage computer and human strengths to alleviate the burden on human curators by automating initial extraction tasks before prioritizing and assigning specialized curation tasks to humans with different levels of training: using non-experts for straightforward tasks such as validation of higher accuracy results (e.g., completing partial facts) and domain experts for low-certainty results (e.g., reviewing specialized compound labels). To validate our approaches, we focus on the task of extracting the glass transition temperature of polymers from published articles. Applying our approaches to 6 090 articles, we have so far extracted 259 refined data values. We project that this number will grow considerably as we tune our methods and process more articles, to exceed that found in standard, expert-curated polymer data handbooks while also being easier to keep up-to-date. The freely available data can be found on our Polymer Properties Predictor and Database website at http://pppdb.uchicago.edu.**

***Index Terms*—Information Extraction, Crowdsourcing, Machine Learning, Polymers, Glass transition**

I. Introduction

Materials informatics [1–3], often referred to as the fourth paradigm of materials discovery [4, 5], combines large datasets and computational models to identify candidates for new materials, with the goal of reducing both time-to-market and development costs. As such methods rely on access to large, machine-readable databases, the traditional text-based physical handbooks will not suffice. However, there are few examples of these scientific digital databases and constructing new ones is a monumental task requiring years of expert labor, as the data that populate these databases must often be extracted manually from free-text publications. One excellent example of a digital database is PolyInfo [6], which contains the records for over 200 000 properties of polymers extracted from more than 12 000 articles—a process that required years of expert curation effort. Achieving databases as large and useful as PolyInfo for different material properties at a rate commensurate with the time-to-market goals of modern materials engineering is a daunting task. It might appear that automated methods of extraction could solve this problem; however, despite considerable progress in natural language processing (NLP) and machine learning [7–13], fully automated extraction is not yet possible due to the complexity by which such properties are encoded in publications. Instead, human effort is needed to develop rules, define training sets, and validate results [14–16].

In response, we propose a hybrid Information Extraction (IE) pipeline that combines automation and crowdsourcing in ways that leverage the complementary strengths of computational modules and humans. This pipeline first extracts candidate properties automatically and subsequently assigns various curation tasks to humans with the goal of maximizing throughput and accuracy while minimizing the burden on human curators. We applied a preliminary version of this concept to extract 263 values for the Flory-Huggins interaction parameter, a measure of miscibility between two entities—typically a polymer and either another polymer or a solvent [17]. In that case, we automatically browsed and searched a relevant journal in polymer science for this property to identify candidate articles and trained student reviewers to extract data: an effective but still relatively costly approach [18]. Here, we extend that work, increasing the automation to develop an integrated IE pipeline that combines a general-purpose NLP toolkit to parse text and perform preliminary recognition; specialized domain-specific models to identify entities and relationships; a ranking system to prioritize crowdsourced tasks; and a crowdsourcing framework to review candidate relationships. We apply this system to extract the *glass transition temperature* (T_g) of polymers. This important property in the design of new polymeric materials quantifies the temperature at which polymers transition from a glassy state into a rubbery state. Values for this parameter are often found in the text of scientific articles.

We used this new IE pipeline to process 6 090 articles

Audus, Debra; Chard, Kyle; Foster, Ian; Joshua, Lequieu; Tchoua, Roselyne; Ward, Logan; de Pablo, Juan.
"Towards a Hybrid Human-Computer Scientific Information Extraction Pipeline."
Paper presented at 2017 IEEE 13th International Conference on e-Science, Auckland, New Zealand. October 24, 2017 - October 27, 2017.

published over the last decade in *Macromolecules*, a prominent journal in polymer science. In the first pipeline step, an NLP-based extraction process identified 1 442 **T_g candidates** in these articles—text fragments with characteristics suggestive of a T_g value, but often with various irregularities. Subsequent automated and crowdsourcing curation steps then processed these candidates, in some cases confirming and/or completing a polymer–T_g value and in others establishing that no such value is in fact present. Curating the output of the NLP extraction required only a half-hour of expert time and a combined six hours of untrained crowds. To date, we have extracted 259 T_g values from a subset of our articles and expect this number to increase dramatically as we improve our pipeline and apply it to new data. In comparison, the recent edition of the expert-curated *Physical Properties of Polymer Handbook* [19], last published in 2007, contains only ≈600 T_g values. The most recent and machine-accessible output of our IE pipeline is freely available at http://pppdb.uchicago.edu and https://materialsdatafacility.org [20].

The primary contributions of this paper are: (1) the design of a hybrid extraction pipeline that combines computer and human strengths; (2) demonstration that this method can accurately extract properties from publications; and (3) design and evaluation of extraction and curation tools, including a rule-based parser for T_g, a polymer identification module for distinguishing polymers from other chemical compounds, a polymer proximity search module for recovering polymer names from related text, crowdsourcing modules for identifying unrecognized polymers and flagging anomalous polymer names, and a prioritization model to guide curation effort.

The rest of this paper is as follows. Section II reviews related work in the information extraction of scientific facts. Section III motivates the problem by introducing T_g and discusses the challenges associated with automated extraction. Section IV describes the design and implementation of our IE pipeline. Section V evaluates the accuracy of the various stages in our pipeline. We discuss future work in Section VI before concluding in Section VII.

II. Related Work

IE methods have been applied in various scientific domains. The medical community has long been interested in the automated extraction and aggregation of data from medical text. Medical Language Extraction and Encoding System (MedLEE) [21, 22], cTAKES [23], and medKAT [24] are NLP tools specialized for the medical domain. These tools are designed to extract clinical information from text documents and to translate entities and terms to controlled ontologies and vocabularies. Much research in this domain has focused on the complexity of clinical text, for example there are significant challenges identifying negation, family relationships, temporality, and uncertainty. The general purpose nature of these tools also allows more sophisticated and specialized applications to be developed. For example, MedLEE has been adapted to build biomolecular and genotype-phenotype networks (GENIES [25] and BioMedLEE [26], respectively). These tools tend to be specialized and rely heavily on the development of ontologies, a tedious and time consuming process. Similarly, several NLP tools have recently been developed to mine data from patents and scientific literature in chemistry and materials science [11, 12, 27].

With the recent advances in machine learning and statistical inference approaches, scientific applications are turning their attention to deep learning tools such as DeepDive [13]. PaleoDeepDive [28], built upon DeepDive, automatically extracts paleontological data from text, tables, and figures in scientific publications. GeoDeepDive [29] performs similar tasks in the geosciences. For good performance in such applications, IE software often relies on and extends large databases: for example, PaleoDeepDive builds on PaleoDB [30] and GeoDeepDive builds on Macrostrat [31]. However, many fields, including materials science, do not yet have access to large and structured sets of texts that deep learning systems can use to learn scientific facts and relationships. The IE pipeline is an intermediary, but essential, step towards accumulating such structured data.

Because of the challenges in fully automated IE systems (e.g., dependence on ontologies and/or large training datasets) but also for validation purposes, humans are often involved in the extraction of scientific facts as domain experts. There is also recent interest in using crowdsourcing or "human computation" to solve problems that computers cannot handle correctly or cost-efficiently. Previous work has leveraged crowdsourcing to support extraction of data from tables within PDF documents [16] and also to ensure human quality control (i.e., expert curation) [15] while extracting empirical observations from literature. CrowdDB [32] uses human input to answer queries that neither database systems nor search engines can adequately answer due to the nature of the queries (e.g., discovering new data not included in a database). In our work, we aim to identify such cases—where humans are better suited for a task—and use the complementary strengths of humans and computers to populate a database of scientific facts. Wallace et al. [33] also pursue this goal, using a hybrid machine learning and crowdsourcing approach to identify published randomized controlled trials (RCTs) [33]. They use machine learning classifiers to recognize citations that are deemed highly unlikely to describe RCTs, deferring to crowdsourcing otherwise.

III. Motivation

We first provide a brief description of polymers and T_g and review both the state-of-the-art in NLP and the challenges associated with its application to the T_g problem.

A. Glass Transition Temperature

Polymers are molecules formed by covalently bonding small molecules, referred to as monomers, together. As the resulting polymer molecules generally have large molecular masses, potentially exceeding three orders of magnitude greater than water, they are sometimes referred to as macromolecules. Due in part to their large molecular masses, often in the form of

Audus, Debra; Chard, Kyle; Foster, Ian; Joshua, Lequieu; Tchoua, Roselyne; Ward, Logan; de Pablo, Juan.
"Towards a Hybrid Human-Computer Scientific Information Extraction Pipeline."
Paper presented at 2017 IEEE 13th International Conference on e-Science, Auckland, New Zealand. October 24, 2017 - October 27, 2017.

long chains, polymers have a variety of useful properties. For example, the long chains of *poly(ethylene terephthalate)* become entangled, making them harder to pull apart; this results in strong but lightweight water bottles. In addition to being strong, many synthetic polymers are also extremely cheap as they can be synthesized from petroleum-based feedstocks. The combination of low cost and useful properties has resulted in polymers becoming a ubiquitous part of life.

In the design of new polymeric materials, the temperature relative to the T_g can have a profound effect on the properties of the polymeric material. T_g is defined as the temperature at which a polymer transitions from a solid, amorphous, glassy state to a rubbery state as the temperature is increased. Physically, when polymers are in the glassy state, the molecules are trapped and cannot move past each other due to a lack of thermal energy, while when they are in the rubbery state, the molecules are mobile. As the properties for the two states are drastically different, the glass transition plays a key role in both choosing a polymer for a given application and in the processing of the polymeric material. For example, plexiglass (*poly(methyl methacrylate)*), used as a lightweight substitute for glass, has a high T_g of roughly 110 °C, while neoprene (*polychloroprene*), used for laptop sleeves, has a low T_g of roughly -50 °C [34]. Exact, as opposed to rough, values of T_g require additional contextual information such as the molecular mass. We plan to capture such information in future work. However, as extracting contextual information is significantly more challenging than the already difficult task of extracting polymer–T_g pairs from literature, we focus on the polymer–T_g pairs first.

B. Natural Language Processing

Rule-based methods are commonly used for simple information extraction tasks. Such methods are straightforward to understand and allow developers to trace and fix errors; they are suitable for simple, well-defined problems (e.g., extracting spouses by identifying the subject and object in sentences containing the word *married*). However, they require tedious effort to construct and modify, as many rules are typically required to extract the same information expressed in various forms. In contrast, statistical and machine learning techniques are trainable, adaptable, and require little manual labor; however, they are opaque and require training data. Researchers often combine the two methods to increase the completeness and accuracy of extracted information [35]. Still, challenges remain, including the lack of the annotated corpora need to train machine-learning models. The lack of corpora is particularly common in fields such as bioinformatics [36] and our own, polymer science. Other challenges, not limited to specific scientific domains, include automatically deciphering subtleties in the English language, in general, and language particular to the domain itself. In polymer synthesis papers, for example, authors sometimes omit the name of the polymer, instead referencing or describing the underlying chemistry. In these cases, the polymer name is not readily apparent, and may require an expert polymer scientist to extract that information.

IV. Design and Implementation

The desired output of our pipeline is a set of polymer–T_g pairs, which can then be used to construct a machine-accessible database of values. Thus, the task can be seen as a two-part process consisting of recognizing polymer names and temperatures and establishing a relationship (t is a T_g of p) between pairs of entities. In order to reduce the burden on curators, we combine complementary human and machine strengths throughout our pipeline. We base our pipeline on a leading materials NLP toolkit, ChemDataExtractor [12], and develop automated and crowdsourcing modules to extract and curate polymer–T_g pairs. We focus here on extracted text excerpts containing a single T_g value. While multiple T_g values may be reported for a single polymer (e.g., prepared with different processing methods), we focus on pairs of polymers mapped to a single T_g for this work. In this section, we first describe the pipeline at a high level and then present the NLP toolkit, our various extraction and curation models, and methods used to prioritize human review.

A. Our Pipeline

Figure 1 illustrates our current pipeline with its six main stages. In stage 1, an extended version of a general-purpose materials NLP toolkit called ChemDataExtractor is used to extract a set of **T_g candidates** from text; in stage 2, compound names identified by the NLP Module are processed to create a polymer dictionary. As we describe below, the candidates identified in stage 1 can be in various forms: compound–T_g pairs; solitary T_gs, with no associated compound; and label–T_g pairs, in which the T_g is associated with a label rather than a compound. Each form requires further processing, which is performed in stage 3 via two automated curation modules and one crowdsourcing module. The results of those three modules are combined as the **proposed polymer–T_g pairs**. Stage 4 engages crowds in flagging erroneous results, stage 5 prioritizes final validation and curation of the proposed pairs, and stage 6 applies final expert review.

Designing a system to make use of crowds requires tailoring tasks to the expertise of the participants. For example, it is significantly easier for a nonexpert to mark a polymer–T_g pair as correct or incorrect than to extract the pair from a paragraph of text. Thus, we focus our crowdsourcing modules on simple micro-curation tasks. We have developed crowdsourcing modules to address two curation tasks: resolving labels that refer to polymer names and flagging anomalous polymer names.

The output of our pipeline is a set of **confirmed polymer–T_g pairs**, each associating a polymer name (with acronyms and/or synonyms) with a single T_g. These pairs are represented in a JSON format that can be easily processed and loaded into a database. Listing 1 shows an example record.

B. Natural Language Processing Module

The first phase of our pipeline requires the identification and extraction of structured representations of information embedded within text. There has been a wealth of research into creating specialized systems for extracting materials [11, 27]

Audus, Debra; Chard, Kyle; Foster, Ian; Joshua, Lequieu; Tchoua, Roselyne; Ward, Logan; de Pablo, Juan.
"Towards a Hybrid Human-Computer Scientific Information Extraction Pipeline."
Paper presented at 2017 IEEE 13th International Conference on e-Science, Auckland, New Zealand. October 24, 2017 - October 27, 2017.

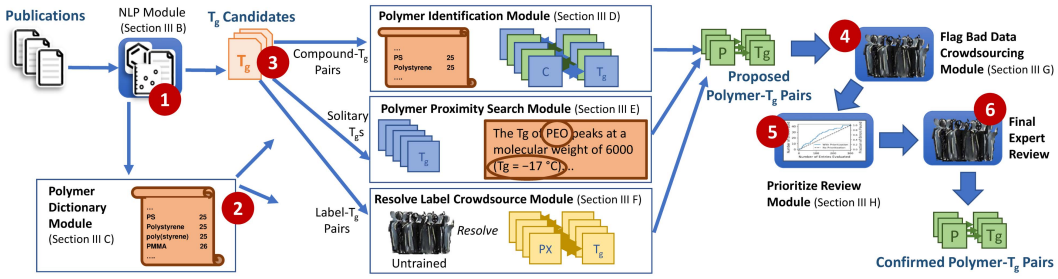


Fig. 1: The six-stage hybrid IE pipeline, showing (1) the NLP Module, which identifies T_g candidates; (2) the Polymer Dictionary Module, which identifies polymer names in NLP output; (3) the three automated extraction and crowdsourcing modules used to process different forms of candidates; (4) the Flag Bad Data Crowdsource Module, in which crowds flag anomalous results, (5) the Prioritize Review Module, which ranks extracted polymer–T_g pairs to prioritize expert validation, and (6) the Final Expert Review.

```
{
      "names": [
            "PBMA",
            "poly(butyl methacrylate)"
      ],
      "glass_transitions": [
            {
              "units": "°C",
              "value": "20"
            }
      ]
}
```

Listing 1: This polymer–T_g record indicates that the polymer *poly(butyl methacrylate)*, also known as *PBMA*, has a T_g of 20 °C.

and other domain-specific [14, 36–38] content from text. Thus, we choose to extend an existing NLP toolkit, ChemDataExtractor [12], to extract T_g values from documents.

1) ChemDataExtractor: ChemDataExtractor is a best-of-breed system for materials extraction, as evidenced by its performance in the relevant chemical compound and drug name recognition (CHEMDNER) community challenge [39]. It implements an extensible end-to-end text-mining pipeline that can process common publication formats including Portable Document Format (PDF), HyperText Markup Language (HTML), and eXtensible Markup Language (XML); it also supports extraction from headings, paragraphs, and captions, and produces machine-readable structured output data that can be used for subsequent processing. ChemDataExtractor automatically extracts chemical named entities and their associated properties, measurements, and relationships from scientific documents. It uses a combination of machine learning (linear-chain conditional random field) models, dictionary-based approaches, and regular expressions for entity recognition. It also detects and associates acronyms and synonyms with polymer names. Entity properties are extracted using a rule-based approach customized for specific properties. Extractors are provided for properties such as melting point and spectrum types, but not T_g.

2) Extending ChemDataExtractor for Glass Transition Temperatures: Our T_g extraction module incorporates specialized knowledge about the forms in which T_g values are expressed in scientific articles. Adapting the format of ChemDataExtractor's melting point extractor, our module contains rules that detect a prefix for a temperature (e.g., "a glass transition temperature of") and then detect and extract the associated temperature (e.g., "20 °C"). ChemDataExtractor then links these values with the associated compound(s). Of course, T_gs are expressed in many formats and therefore our rules must include variations of such statement structures. For instance, we include rules that match various quantifiers, such as "a glass transition temperature **range** of." Similarly, our rules capture approximate values, where temperatures are preceded by terms such as **ca.** or **around**. Further, our rules support variations of glass transition temperature including T_g, glass transition temp. and more. In total, we defined two dozen rules to address different variations and representations of glass transition temperature. Our T_g extractor has since been integrated into ChemDataExtractor.

The output of our extended ChemDataExtractor is a set of JSON records, each containing one or more T_g values and, optionally, an associated chemical compound name plus any automatically-detected acronyms and synonyms.

C. Polymer Dictionary Module

The materials literature includes references to a wide range of compounds beyond just polymers. The original ChemDataExtractor does not distinguish polymers from non-polymers and thus we face the challenge of correctly identifying which chemical name entities in a paper correspond to polymers. Unfortunately, no complete dictionary for polymer names exists and the standardized International Union of Pure and Applied Chemistry (IUPAC) naming conventions [40] often result in lengthy and, hence, rarely used names. Thus polymers are expressed using a combination of common names, IUPAC names, and trade names. The polymer identification problem is further complicated by the fact that values are often

Audus, Debra; Chard, Kyle; Foster, Ian; Joshua, Lequieu; Tchoua, Roselyne; Ward, Logan; de Pablo, Juan.
"Towards a Hybrid Human-Computer Scientific Information Extraction Pipeline."
Paper presented at 2017 IEEE 13th International Conference on e-Science, Auckland, New Zealand. October 24, 2017 - October 27, 2017.

reported for *copolymers*, in which two or more monomers are used during synthesis.

The Polymer Dictionary Module implements heuristics for identifying those compound names extracted by stage 1 that likely correspond to polymers, and collects the resulting names in a **polymer dictionary**. These heuristics include rules related to text-based names (e.g., prefixes of "P" and "poly") as well as rules prescribed by the IUPAC guide [41] for forming polymer names. The latter is valuable for identifying copolymers. For example, names containing the substring "*-alt-*" indicate copolymers comprising two species of monomeric units in alternating sequence.

This module also handles synonyms and acronyms, a common occurrence in polymer science. For example, we may find the polymer *Polystyrene* represented in the same or different articles by the synonym *poly(styrene)* or the acronym *PS*. ChemDataExtractor includes mechanisms for identifying and grouping synonyms and acronyms. We record these groups in our polymer dictionary. We also include both singular and plural representations, for example *polystyrene* and *polystyrenes*. To avoid confusion with acronyms, we only consider plurals for names longer than four characters. Thus, for example, *PSS*, the acronym for *poly(styrene sulfonate)*, is not identified as the plural form of *PS*. We exclude copolymers from our dictionary as these are easily recognizable via our implemented IUPAC polymer heuristics.

To bootstrap the polymer dictionary, we ran our polymer identification heuristics over all 6 090 full-text HTML publications from *Macromolecules* and thereby populated the dictionary with 12 814 polymer names and acronyms in 9 178 different detected groups.

D. Polymer Identification Module

For cases where compound-T_g pairs were idenified using the NLP Module, the Polymer Idenification Module determines which of those compounds are polymers and which are not. To do so, the module simply labels any compound present in the polymer dictionary produced by the Polymer Dictionary Module as a polymer and all other entires as non-polymers.

E. Polymer Proximity Search Module

One significant type of error for text extraction are T_g values that are not associated with a polymer name. To correct these errors, we have developed a proximity-based approach for determining whether the polymer name is mentioned nearby where the temperature was found (for example, in the previous sentence or paragraph). For each sentence in the document, we determine whether it contains a T_g value, and if so, return the closest polymer name (using the polymer dictionary) within the sentence, if any such name is to be found. If no polymer is found, we extend the search to the preceding sentence, as illustrated in Figure 2.

This process increases the number of polymer–T_g pairs discovered; however, it may decrease the accuracy of the extracted pairs. We discuss validation in Section IV-G.

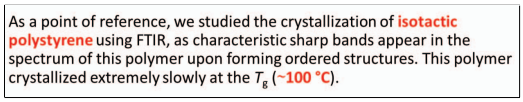


Fig. 2: The NLP Module yields a solitary T_g record in this example text [42], as the corresponding compound is mentioned in the previous sentence. The Polymer Proximity Search Module disambiguates the reference and proposes *isotactic polystyrene* as a (correct) candidate match for the T_g.

F. Resolve Label Crowdsource Module

This first crowdsource module addresses errors where the text extraction matched T_g values to labels (e.g., *Polymer A*) rather than the actual polymer name. These labels frequently occur in the polymer literature to avoid repetition of complex polymer names, such as the following.

```
poly(1,2:3,4-di-O-isopropylidene-6-O-(2′—formyl-4′-
    vinylphenyl)-d-galactopyranose)
```

We created an interface that presents labels and the paper in which each appears, and asks the crowd to enter the polymer name for each label. As this task requires little knowledge of polymer science, we use an untrained crowd to resolve references. We provide these people with just a simple training guide (less than one page) to describe the task. In an attempt to quantify accuracy, we allow crowd members to specify their confidence (1–5) along with their input. Our goal is to use this confidence score to prioritize results for future review.

G. Flag Bad Data Crowdsource Module

The second crowdsource module presents users with a list of polymer–T_g pairs and asks them to flag whether the polymer names are incomplete or incorrect. The polymer names identified by the text extraction tool are sometimes not specific enough to identify the polymer being studied. As one example, the term "*hydroxyl copolyimides*" describes a family of polymers rather than a specific polymer, and therefore cannot be attributed a single T_g value. Given the complexity we use an expert crowd of polymer scientists to complete this task. Our flagging interface does not delete any data from our set, but rather records user "votes." We then use this information to prioritize further review.

H. Prioritize Review Module

Every stage of the pipeline uses a variety of methods to extract values with varying confidence. Thus, each proposed polymer–T_g pair has an associated probability of accuracy. For example, a pair extracted from a single sentence using our NLP rules and subsequently reviewed by an expert is likely to be accurate. In contrast, a pair in which the polymer name is a synonym, was found in the sentence preceding that containing the T_g value, and was not reviewed by a human, is less likely to be accurate. To formalize this concept, we explore methods for estimating confidence in a particular value and use this metric to prioritize (crowdsourced) curation tasks.

Audus, Debra; Chard, Kyle; Foster, Ian; Joshua, Lequieu; Tchoua, Roselyne; Ward, Logan; de Pablo, Juan.
"Towards a Hybrid Human-Computer Scientific Information Extraction Pipeline."
Paper presented at 2017 IEEE 13th International Conference on e-Science, Auckland, New Zealand. October 24, 2017 - October 27, 2017.

Our initial approach for the prioritization method relies on characteristics of polymer names and their associated T_g values. Hypothesizing that polymer names that appear more frequently in the database have a higher likelihood of being correct than infrequently used names, we assign a confidence to each polymer name based on its frequency of occurence. Further hypothesizing that outlier or extreme temperatures are more likely to indicate errors, we determine the minimum, mean, and maximum of all T_g values in our current database and use those values to identify outliers, to which we assign lower confidence values. These two scoring methods can be combined. For example, if two records appear equally infrequently, we prioritize for review the one with temperature farthest from the mean. Entries with confidence scores under a fixed threshold will then be funnelled to Stage 6 of the pipeline for expert review as shown in Figure 1.

V. Evaluation

We quantitatively evaluated our pipeline by comparing results against a gold standard, human-reviewed dataset. In this section, we describe our input dataset and then present our evaluation of each module in our pipeline.

A. Dataset

Our **input dataset** comprised of 6 090 publications in full-text HTML format. To obtain these publications, we automatically searched the journal *Macromolecules* using the keyword "T_g" over the ten-year period 2006–2016. We downloaded the full-text publications matching this query and sampled additional *Macromolecules* issues from the last decade to increase and diversify our corpus. This is the same dataset that we used to build our polymer dictionary, as described previously.

B. Natural Language Processing Module

Execution of the T_g–extended ChemDataExtractor NLP module described in Section IV-B identified 364 561 records, of which 1 330 were candidate T_g values from 927 distinct publications: 846 compound–T_g pairs, 456 solitary T_gs, and 28 label–T_g pairs. (Another 112 linked more than one compound and/or T_g value, a case that we leave for future work.) We stored these records in a database for convenient access to their features, which include the name of the associated compound, when present, and any synonyms for that compound.

C. Assembling a Gold Standard Dataset

We manually selected a subset of 50 papers for which the NLP module had identified one compound–T_g pair for which the compound contained the string "poly." We then had two polymer scientists each read 25 of these publications to identify all polymer–T_g pairs that they contain. The result is a gold standard dataset containing a total of 62 polymer–T_g pairs. We used this dataset for various evaluation steps.

To gain some initial experience with the use of this dataset, we also asked our experts to evaluate the accuracy of the 50 compound–T_g pairs identified in these papers by the NLP module. In evaluating precision, we assigned points to each extracted entry as follows: 1 point for **fully correct** entries, i.e., entries that were completely unambiguous and correct; 0.5 points to **partially correct** entries, in which information was missing (e.g., the module extracted *polyurethanes.11*, a correct but idiosyncratic name, which an expert clarified by adding *polyurethanes with various side chains*); and 0 points to other **incorrect** cases, such as those with an incomplete polymer name (e.g., the module extracted *hydroxyl copolyimides* instead of *APAF-ODA hydroxyl copolyimides*: the former describes a vast family of polymers and cannot be clarified without additional information).

The NLP module extracted 17 fully correct and 4 partially correct polymer–T_g pairs from the 50 articles, for a precision of 38 %. As our experts identified 62 T_g values in the 50 articles, the recall was 31 %. While the expert reviews, being aimed at assembling a gold standard, were particularly rigorous, these low values emphasize the difficulty of our task and the need for a hybrid solution. In most cases, errors were related to identification of the polymer name rather than the T_g value. In fact, for the subproblem of locating T_g values, our T_g extraction rule achieved 88 % precision (44 out of 50 cases) and 71 % recall (18 T_g values missed out of 62 total).

Precision: We attribute our low precision to three main reasons. A first is that the compound name was incorrectly or partially identified ≈50 % of the time. The low performance in polymer name recognition may be explained by the fact that the entity recognition component of ChemDataExtractor was trained on biomedical newspaper and biomedical training corpora, supplemented with unsupervised word cluster features derived from chemistry articles. The use of biomedical training data is due to the lack of appropriate annotated corpora for training machine learning models for polymer name recognition, a general problem in materials informatics. Moreover, our experts noted that some polymer names were difficult even for humans to extract, as they were not named but rather described in terms of their components: e.g., "*A cross-linked polymer with DABBF linkages was prepared by polyaddition of poly(propylene glycol) (PPG) (Mn = 2700), hexamethylene diisocyanate (HDI), dihydric DABBF, and triethanolamine (TEA) as a cross-linker in the presence of di-n-butyltin dilaurate (DBTDL, catalyst) in N,N-dimethylformamide (DMF) in a manner similar to that previously reported (Figure 1)*" [43].

A second difficulty, which arose in 8 % of the cases, was that one of our T_g extraction rules was loosely defined as simply "transition," to avoid tokenizing issues around the term "glass-transition." We expected that in the context of polymer science the most common transition temperature would be T_g. However, while this rule sometimes functioned as expected, it also matched sentences with "gel transition" and "phase transition" temperatures. We could redefine the loose transition rule, but while this would increase precision, it would also decrease recall. Initially, we view high recall as a preferable to high precision in our "big-data" approach, as we expect later pipeline stages to improve the precision.

A third difficulty, arising in 4 % of the cases, was that

Audus, Debra; Chard, Kyle; Foster, Ian; Joshua, Lequieu; Tchoua, Roselyne; Ward, Logan; de Pablo, Juan.
"Towards a Hybrid Human-Computer Scientific Information Extraction Pipeline."
Paper presented at 2017 IEEE 13th International Conference on e-Science, Auckland, New Zealand. October 24, 2017 - October 27, 2017.

complex sentence structure led to incorrect T_g values being extracted. For example, in sentences describing increases or decreases in temperature relative to a previously mentioned value, the software identified the difference as T_g: e.g., "*Comparing DSC results for dried composites (Figure 3b), a drop in Tg of 17 °C was observed for the clay composite, whereas the corresponding drop in Tg of the aerogel composite was only 3 °C*" [44]. One way to improve precision in such cases would be to analyze sentence complexity, as indicated by features such as number of words and the use of comparison terms such as "lower/greater" and "decrease/increase," and then defer to trained crowds for sentences above a certain threshold.

Recall: We view improving recall as an iterative process as we continue to find additional ways that T_g is expressed in the literature. During the evaluation of the NLP module, we inspected the results and added new rules to our T_g extractor to increase recall. For example, sometimes authors referred to the *"T_g value of"*; the extra *"value"* term was not included in the original parser. Another slightly more complex example consists of capturing a temperature expressed in the form *"T_g of **<polymer name>** is/was ..."*. This rule depends on correctly identifying the polymer name in the sentence, as some polymer names, which sometimes include dashes, spaces, and colons, will not always correspond to the regular expression class of words. We plan to use a larger dataset for evaluation, examine more cases of missed T_gs and identify new general rules that will further improve recall.

D. Polymer Identification Module

To test the polymer name classifier described in Section IV-D, we selected 100 papers: the 50 used in Section V-B plus 50 additional papers with compound–T_g records for which the compound names did not include "poly."

Using our full polymer name dictionary (prefixes and IUPAC guidelines as well as simple "poly" keyword search), we classified the compounds from the 100 papers. We achieved 91.8 % precision and 93.2 % recall. In other words, we correctly classified 91.8 % of the compounds as polymers and misclassified 6.8 % of the extracted compounds. An example of a false positive is identifying a class of polymers (e.g., *polyimides*) rather than an individual polymer. An example of a false negative is the copolymer *UPy-OPG-MAA*: as none of its three components existed in the polymer name dictionary, our heuristics could not identify it as a copolymer. The addition of polymer heuristics improved the performance of our polymer classification by correctly discovering additional polymers (16 % of the compounds initially classified as "non-polymers"), which were not detected by a simple string search of the names, hence potentially increasing the number of polymer–T_g pairs in the final output. They are particularly useful for detecting copolymers using IUPAC conventions (e.g., *PPDL-block-PLLA*) formed of previously seen polymer components (e.g., *PPDL* or *PLLA*). We have since composed a list of common polymer families that will further improve our classification results.

E. Polymer Proximity Search Module

Recall from Section IV-E that this module seeks to address the problem of T_g values that were extracted without a polymer name. To test this module, we first identified 115 records containing solitary T_g values. The module returned a polymer for 74 out of these 115 records (64.3 %). We executed the proximity search heuristic to consider the same and previous sentences and compared the identified polymer names to those identified by an expert. Our proximity search suggested correct polymer matches for 31 of the 63 records (49.2 %) in which the matching polymer was located within the same sentence. Its search of the preceding sentence identified correct polymer matches in 6 of the 11 records (54.5 %) in which the matching polymer was in that sentence. Together, searching both the T_g and preceding sentence led to the recovery of 37 polymers: 50.0 % of the original 74 solitary T_g mentions or 32.1 % of the test dataset, which includes false negatives. See Table I for a summary of the results.

TABLE I: Polymer proximity search module evaluation.

	True Positives	False Positives	Gold
Same sentence	31	32	63
Previous sentence	6	5	11
No candidate returned			41
Total	37	37	115

We note that success here requires correct identification of both the polymer and the temperature to be linked. Some compounds were only partially identified and the complete polymer–T_g pairs were not correctly recovered. Since the proximity search module uses the polymer database, improving polymer name recognition and the T_g parser will in turn increase proximity search performance. In some cases, proximity search introduced false positives for a different reason, as the compound closest to the temperature was used for comparison and was not associated with the extracted T_g for instance. Nevertheless, confirming or rejecting matches from this module is a less difficult task than extracting the polymer–T_g pairs.

F. Crowdsourcing Modules

Recall from Sections IV-F and IV-G that we have deployed two crowdsourcing modules: one to recover polymer names from author-defined labels and one to flag polymer–T_g pairs deemed to require further review.

In the first case, we presented three (non-expert) reviewers with polymer name labels and asked them to extract the polymer name from the full text. We also asked them to state their confidence (1-5). We identified 28 records for review (based on regular expression matching of the form "Polymer *[a-zA-Z0-9]*"). The three reviewers correctly identified 82.1 %, 78.6 %, and 35.7 % of those 28 records and reported an average of two hours of work. A simple consensus method across our three reviewers (selecting the answer from two or more reviewers in agreement) obtained 78.6 % accuracy when resolving these labels. Only in two cases did no reviewer

Audus, Debra; Chard, Kyle; Foster, Ian; Joshua, Lequieu; Tchoua, Roselyne; Ward, Logan; de Pablo, Juan.
"Towards a Hybrid Human-Computer Scientific Information Extraction Pipeline."
Paper presented at 2017 IEEE 13th International Conference on e-Science, Auckland, New Zealand. October 24, 2017 - October 27, 2017.

identify the correct label, seemingly indicating that this task was at an appropriate level of difficulty for our crowd. These results show that the use of untrained crowds can reduce the need for expert validation substantially. Table II summarizes reviewer performance and confidence scores. It shows the number of correct answers from reviewers with their reported confidence scores. Reviewer 2 correctly identified 21/21 labels with high confidence, 2/3 with medium confidence and 0/4 with low confidence.

TABLE II: Crowdsourcing for resolving polymer labels.

	Correct	Confidence (correct/total) High (1-2)	Med (3)	Low (4-5)	Time spent (hours)
Reviewer 1	23	23/28	0	0	3
Reviewer 2	23	21/21	2/3	0/4	2
Reviewer 3	10	0	0	10/28	1

In the second crowdsourcing task, we presented an expert polymer scientist with 302 compound–T_g pairs extracted by the NLP module for which the compound matched the string "poly." The reviewer took about 30 minutes to identify 43 (14 %) of these values as incomplete or incorrect, leaving the 259 confirmed polymer–T_g pairs noted in the abstract. Erroneous values included names that describe a class of polymers as opposed to a specific polymer (e.g., *polyolefin*) and unrecognized labels (e.g., *copolymer 10*), and additional descriptors (e.g., *macroporous poly(N-isopropylacrylamide) gel*). Overall, these results suggest that our extractor performs as expected in the majority of cases.

We will next apply this same process to the additional proposed polymer–T_g pairs produced by our system. In addition to improving our dictionary, we are currently compiling a list of common polymer family names and working on a list of common descriptors to ignore.

G. Prioritizing Review

We applied our scoring model (using polymer name frequency and T_g value distance from the median) to 302 compound–T_g pairs for which the compound name matched the string "poly." We compared the pairs prioritized by the scoring model against those flagged by experts in the previous crowdsourcing step. After ordering these pairs by confidence, we observed that 10 of the first 50 entries had been flagged as erroneous by our reviewers (see Figure 3), which is 40% more than would be expected if entries were randomly selected (≈7 errors). While not an extraordinary decrease in the number of reviews, it was achieved by a basic ranking scheme; we expect more sophisticated approaches to further reduce the human effort required to improve the quality of our database.

We plan to use a similar scheme to score entries in the polymer dictionary. The scheme will consider frequency, number of synonyms, and number of duplicate entries (same acronym for different polymers) to assign a confidence score to each entry. We anticipate that this approach will be able to detect additional unrecognized polymers: for example, *poly(2,4′-BFa)* where author-defined monomer *2,4′-BF-a* is specified elsewhere in the publication.

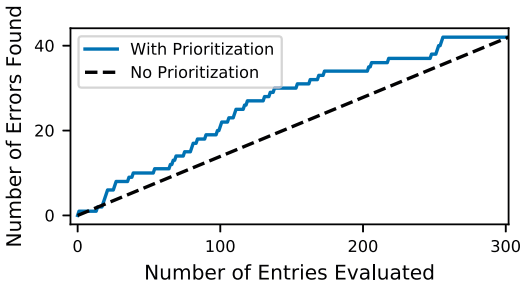


Fig. 3: Results of prioritizing crowdsourcing. The blue, solid line shows the number of errors found as a function of the number of expert reviews if the entries are evaluated following our prioritization scheme. The black, dotted line shows the number of errors found if entries are evaluated in a random order.

H. Summary of Results

Table III aggregates the results of our evaluation across the four types of T_g candidates that we have examined. The **Initial** column gives the number of each type extracted from our 6 090 articles, with poly–T_g here denoting compound–T_g pairs for which the compound name contains the string "poly" and nonpoly–T_g the remaining compound–T_g pairs. The **Yield** column indicates the number of T_g candidates of each type that are estimated to be correct, based on review. (For polymer–T_g and label–T_g, this is a full review; for compound–T_g and solitary T_g, the numbers are estimates based on expert review of a subset.) The **Pairs** column gives the number of polymer–T_g pairs that we expect from each method. Thus, we expect the final number of pairs extracted from our initial set of 6 090 articles to increase significantly—perhaps by 145 % to approximately 500—once we complete expert review.

TABLE III: Summary of module performance and expected number of polymer–T_g output from initial data.

Input Type	Initial	Module	Yield	Pairs
poly–T_g	302	Flag Bad Data	86.0 %	259
nonpoly–T_g	544	Polymer Identification	16.0 %	87
solitary T_g	456	Proximity Search	32.1 %	146
label–T_g	28	Resolve Labels	78.6 %	22
Totals	**1 330**			**514**

VI. Future Work

While our pipeline initially focuses on extracting polymer–T_g pairs, our approaches are equally applicable to other properties and forms of data.

Polymer properties, such as T_g, are often dependent on important contextual information, such as molecular mass and geometry (confined or bulk) as well as the experimental methods used to calculate values. We intend to develop methods to capture such information to provide context to

Audus, Debra; Chard, Kyle; Foster, Ian; Joshua, Lequieu; Tchoua, Roselyne; Ward, Logan; de Pablo, Juan.
"Towards a Hybrid Human-Computer Scientific Information Extraction Pipeline."
Paper presented at 2017 IEEE 13th International Conference on e-Science, Auckland, New Zealand. October 24, 2017 - October 27, 2017.

extracted values. As previously mentioned, we also plan to make improvements to the dictionary and evaluate its accuracy using experts.

Given the significant cost of manual curation, we are also investigating more advanced methods to prioritize where human effort should be used. Here we discuss two ideas relevant to this topic: validation of extracted data via machine learning models and experiment design.

A. Machine Learning Validation

The T_g of amorphous polymers is the most important and widely studied polymeric property because many other polymer properties, such as heat capacity and viscosity, are affected by this transition [45]. Many researchers have developed machine learning models for T_g [45–48] that we could retrain, using the entries in our database, to make predictions that would in turn validate extracted values. These T_g models provide rough estimates or reasonable ranges for the T_g values of various polymers, which would serve as physics-based validation of our extracted values and help prioritize curation.

B. Experiment Design

A major challenge with a hybrid pipeline is determining when to employ human expertise and, when human expertise is needed, what form of expertise to apply. While we work towards higher levels of accuracy from our automated modules, we do not expect the need for human input to disappear. We have explored several methods including expert review, untrained confidence scores, and a scoring mechanism for prioritization; however, none is without limitation. As the number of publications processed by our pipeline increases, this careful scrutiny of the data will become costly and eventually unworkable. We want to identify when and how to inject different types of human input into the pipeline efficiently. In other words, we want to increase accuracy while minimizing the quantity and cost of crowd input.

We plan to explore a more rigorous approach to automatic partitioning and assignment of extraction tasks by applying techniques from *optimal experiment design* [49–51] to maximize the accuracy of extracted data while minimizing the time and cost of human involvement. To this end, we expect to:

- Calculate the accuracy of values derived from a variety of automated and crowdsourcing modules.
- Assign values to datasets, for example in terms of their yield in polymer–T_g pairs and/or the rarity of those values, and then measure how dataset value changes with each automated and crowdsourced task.
- Assign levels of difficulty to tasks based on completeness and accuracy of the data to be processed and/or the information needed to complete the task, to help decide where to crowdsource various tasks.
- Assign costs to module usage so that we can compare, for instance, the costs of computational vs. crowdsourced modules; determine the cost of using crowds (e.g., person-hours); and quantify the differences in cost between a trained and untrained crowd.

We plan to investigate these topics as we develop the next generation of our pipeline.

VII. Conclusion

Despite significant progress in natural language processing and machine learning approaches to information extraction, there remains a gap between the current data extraction needs in fields such as materials science and the capabilities of state-of-the-art tools. We have described a hybrid human-machine IE pipeline that we have so far used to extract 259 glass transition temperature (T_g) values for polymers from 6 090 scientific articles, with an expectation of many more as we improve our methods and process more articles.

Our pipeline uses domain-specific automated and crowdsourcing extraction and curation modules to extract high-quality and accurate polymer–T_g pairs. The polymer classifier module achieved 91.8 % precision and 93.2 % recall. The polymer proximity search module correctly identified missing polymers for 50.0 % of those T_g values without polymers. We crowdsourced the recovery of unrecognized polymer names for an additional 22 polymer–T_g pairs and demonstrated that using untrained crowds for simple, well-defined domain-specific tasks can decrease the need for expert validation by about three fourth (78.6 % labels resolved by non-experts using concensus method). We have started the validation of automatically extracted data and presented a simple scoring scheme to prioritize the process. Our initial results show that even a simple method for assessing the quality of extracted data can effectively increase the impact of human curation.

While the size of our T_g database is not yet best-in-class, the hybrid pipeline presented in this work offers a sustainable and accelerated route to producing new materials property datasets. With only a few hours of effort from expert and non-expert curators, we were able to screen over 6 000 articles and produce a refined dataset of 259 polymer–T_g pairs from just 927 articles. Thus, our results demonstrate the considerable potential of combining automated and crowdsourcing modules to extract scientific facts from literature in an efficient and cost-effective manner. We continue to refine our automated extraction tools and develop yet more effective ways of prioritizing human curation for maximum benefit, and to use these tools to populate our open database. Our verified polymer–T_g pairs are available at both http://pppdb.uchicago.edu and https://materialsdatafacility.org.

Acknowledgments

We thank our crowd members for their help. This work was supported in part by NIST contract 60NANB15D077, the Center for Hierarchical Materials Design, and DOE contract DE-AC02-06CH11357. Official contribution of the National Institute of Standards and Technology; not subject to copyright in the United States.

References

[1] N. Nosengo, "Can artificial intelligence create the next wonder material?" *Nature*, vol. 533, no. 7601, pp. 22–25, may 2016. [Online]. Available: http://www.nature.com/doifinder/10.1038/533022a

Audus, Debra; Chard, Kyle; Foster, Ian; Joshua, Lequieu; Tchoua, Roselyne; Ward, Logan; de Pablo, Juan.
"Towards a Hybrid Human-Computer Scientific Information Extraction Pipeline."
Paper presented at 2017 IEEE 13th International Conference on e-Science, Auckland, New Zealand. October 24, 2017 - October 27, 2017.

[2] J. Hill, G. Mulholland *et al.*, "Materials science with large-scale data and informatics: Unlocking new opportunities," *MRS Bulletin*, vol. 41, no. 05, pp. 399–409, 2016. [Online]. Available: http://www.journals.cambridge.org/abstract_S0883769416000932

[3] J. J. de Pablo, B. Jones *et al.*, "The Materials Genome Initiative, the interplay of experiment, theory and computation," *Current Opinion in Solid State and Materials Science*, vol. 18, no. 2, pp. 99–117, 2014.

[4] K. M. Tolle, D. S. W. Tansley *et al.*, "The fourth paradigm: Data-intensive scientific discovery," *Proceedings of the IEEE*, vol. 99, no. 8, pp. 1334–1337, Aug 2011.

[5] A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science," *APL Materials*, vol. 4, no. 5, p. 053208, 2016.

[6] S. Otsuka, I. Kuwajima *et al.*, "PoLyInfo: Polymer database for polymeric materials design," in *International Conference on Emerging Intelligent Data and Web Technologies*. IEEE, 2011, pp. 22–29.

[7] S. Bird, E. Klein *et al.*, *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[8] S. Hellmann, J. Lehmann *et al.*, "Integrating NLP using linked data," in *International Semantic Web Conference*. Springer, 2013, pp. 98–113.

[9] C. D. Manning, M. Surdeanu *et al.*, "The Stanford CoreNLP natural language processing toolkit," in *ACL (System Demonstrations)*, 2014, pp. 55–60.

[10] N. A. Lewinski and B. T. McInnes, "Using natural language processing techniques to inform research on nanotechnology," *Beilstein Journal of Nanotechnology*, vol. 6, no. 1, pp. 1439–1449, 2015.

[11] L. Hawizy, D. M. Jessop *et al.*, "ChemicalTagger: A tool for semantic text-mining in chemistry," *Journal of Cheminformatics*, vol. 3, no. 1, p. 17, 2011.

[12] M. C. Swain and J. M. Cole, "ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature," *Journal of Chemical Information and Modeling*, vol. 56, no. 10, pp. 1894–1904, 2016.

[13] C. De Sa, A. Ratner *et al.*, "DeepDive: Declarative knowledge base construction," *ACM SIGMOD Record*, vol. 45, no. 1, pp. 60–67, 2016.

[14] A. Rzhetsky, I. Iossifov *et al.*, "GeneWays: A system for extracting, analyzing, visualizing, and integrating molecular pathway data," *Journal of Biomedical Informatics*, vol. 37, no. 1, pp. 43–53, 2004.

[15] C. Seifert, M. Granitzer *et al.*, "Crowdsourcing fact extraction from scientific literature," in *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data*. Springer, 2013, pp. 160–172.

[16] J. Takis, A. Islam *et al.*, "Crowdsourced semantic annotation of scientific publications and tabular data in PDF," in *11th International Conference on Semantic Systems*. ACM, 2015, pp. 1–8.

[17] R. B. Tchoua, J. Qin *et al.*, "Blending education and polymer science: Semiautomated creation of a thermodynamic property database," *Journal of Chemical Education*, vol. 93, no. 9, pp. 1561–1568, 2016.

[18] R. B. Tchoua, K. Chard *et al.*, "A hybrid human-computer approach to the extraction of scientific facts from the literature," *Procedia Computer Science*, vol. 80, pp. 386–397, 2016.

[19] H. B. Eitouni and N. P. Balsara, "Thermodynamics of polymer blends," in *Physical Properties of Polymers Handbook*. Springer, 2007, pp. 339–356.

[20] B. Blaiszik, K. Chard *et al.*, "The Materials Data Facility: Data services to advance materials science research," *JOM*, vol. 68, no. 8, pp. 2045–2052, 2016.

[21] C. Friedman, P. O. Alderson *et al.*, "A general natural-language text processor for clinical radiology," *Journal of the American Medical Informatics Association*, vol. 1, no. 2, pp. 161–174, 1994.

[22] C. Friedman, G. Hripcsak *et al.*, "Representing information in patient reports using natural language processing and the extensible markup language," *Journal of the American Medical Informatics Association*, vol. 6, no. 1, pp. 76–87, 1999.

[23] G. K. Savova, J. J. Masanz *et al.*, "Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications," *Journal of the American Medical Informatics Association*, vol. 17, no. 5, pp. 507–513, 2010.

[24] "medkat," http://ohnlp.sourceforge.net/MedKATp, accessed Sep, 2017.

[25] C. Friedman, P. Kra *et al.*, "GENIES: A natural-language processing system for the extraction of molecular pathways from journal articles," in *ISMB (supplement of bioinformatics)*, 2001, pp. 74–82.

[26] L. Chen and C. Friedman, "Extracting phenotypic information from the literature via natural language processing," *Studies in Health Technology and Informatics*, vol. 107, no. 2, pp. 758–762, 2004.

[27] D. M. Jessop, S. E. Adams *et al.*, "OSCAR4: A flexible architecture for chemical text-mining," *Journal of Cheminformatics*, vol. 3, no. 1, p. 41, 2011.

[28] S. E. Peters, C. Zhang *et al.*, "A machine reading system for assembling synthetic paleontological databases," *PLoS One*, vol. 9, no. 12, p. e113523, 2014.

[29] C. Zhang, V. Govindaraju *et al.*, "GeoDeepDive: Statistical inference using familiar data-processing languages," in *ACM SIGMOD International Conference on Management of Data*, pp. 993–996.

[30] "Paleodb," http://paleodb.org, accessed Sep, 2017.

[31] "Macrostrat," http://macrostrat.org, accessed Sep, 2017.

[32] M. J. Franklin, D. Kossmann *et al.*, "CrowdDB: Answering queries with crowdsourcing," in *ACM SIGMOD International Conference on Management of Data*, 2011, pp. 61–72.

[33] B. C. Wallace, A. Noel-Storr *et al.*, "Identifying reports of randomized controlled trials (RCTs) via a hybrid machine learning and crowdsourcing approach," *Journal of the American Medical Informatics Association*, 2017.

[34] J. Brandrup, E. H. Immergut *et al.*, Eds., *Polymer Handbook*, 4th ed. Wiley-Interscience, 1999.

[35] L. Chiticariu, Y. Li *et al.*, "Rule-based information extraction is dead! Long live rule-based information extraction systems!" in *Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 827–832.

[36] D. Zhou and Y. He, "Extracting interactions between proteins from the literature," *Journal of Biomedical Informatics*, vol. 41, no. 2, pp. 393–407, 2008.

[37] F. Rinaldi, G. Schneider *et al.*, "Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach," *Artificial Intelligence in Medicine*, vol. 39, no. 2, pp. 127–136, 2007.

[38] M. Krauthammer and G. Nenadic, "Term identification in the biomedical literature," *Journal of Biomedical Informatics*, vol. 37, no. 6, pp. 512–526, 2004.

[39] M. Krallinger, F. Leitner *et al.*, "CHEMDNER: The drugs and chemical names extraction challenge," *Journal of Cheminformatics*, vol. 7, no. 1, p. S1, 2015.

[40] R. G. Jones, E. S. Wilks *et al.*, Eds., *Compendium of Polymer Terminology and Nomenclature*. The Royal Society of Chemistry, 2009. [Online]. Available: http://dx.doi.org/10.1039/9781847559425

[41] R. C. Hiorns, R. J. Boucher *et al.*, "A brief guide to polymer nomenclature," *Polymer*, vol. 54, no. 1, pp. 3–4, 2013.

[42] J. Mattia and P. Painter, "A comparison of hydrogen bonding and order in a polyurethane and poly (urethane- urea) and their blends with poly (ethylene glycol)," *Macromolecules*, vol. 40, no. 5, pp. 1546–1554, 2007.

[43] K. Imato, A. Takahara *et al.*, "Self-healing of a cross-linked polymer with dynamic covalent linkages at mild temperature and evaluation at macroscopic and molecular levels," *Macromolecules*, vol. 48, no. 16, pp. 5632–5639, 2015.

[44] S. Bandi and D. A. Schiraldi, "Glass transition behavior of clay aerogel/poly (vinyl alcohol) composites," *Macromolecules*, vol. 39, no. 19, pp. 6537–6545, 2006.

[45] B. E. Mattioni and P. C. Jurs, "Prediction of glass transition temperatures from monomer and repeat unit structure using computational neural networks," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 2, pp. 232–240, 2002.

[46] A. DiBenedetto, "Prediction of the glass transition temperature of polymers: A model based on the principle of corresponding states," *Journal of Polymer Science Part B: Polymer Physics*, vol. 25, no. 9, pp. 1949–1969, 1987.

[47] T. Le, V. C. Epa *et al.*, "Quantitative structure-property relationship modeling of diverse materials properties," *Chemical Reviews*, vol. 112, no. 5, pp. 2889–2919, may 2012. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/cr200066h

[48] X. Yu, "Support vector machine-based QSPR for the prediction of glass transition temperatures of polymers," *Fibers and Polymers*, vol. 11, no. 5, pp. 757–766, 2010.

[49] V. V. Fedorov, *Theory of optimal experiments*. Elsevier, 1972.

[50] A. F. Emery and A. V. Nenarokomov, "Optimal experiment design," *Measurement Science and Technology*, vol. 9, no. 6, p. 864, 1998.

[51] V. Fedorov, "Optimal experimental design," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 5, pp. 581–589, 2010.

Audus, Debra; Chard, Kyle; Foster, Ian; Joshua, Lequieu; Tchoua, Roselyne; Ward, Logan; de Pablo, Juan.
"Towards a Hybrid Human-Computer Scientific Information Extraction Pipeline."
Paper presented at 2017 IEEE 13th International Conference on e-Science, Auckland, New Zealand. October 24, 2017 - October 27, 2017.

Cloud Security Automation Framework

Cihan Tunc[1,2], Salim Hariri[1], Mheni Merzouki[2], Charif Mahmoudi[2], Frederic J. de Vaulx[2], Jaafar Chbili[2], Robert Bohn[2], Abdella Battou[2]
[1] The University of Arizona, Tucson, Arizona, USA
[2] National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA
[1]{cihantunc, hariri}@email.arizona.edu
[2]{cihan.tunc, mheni.merzouki, charif.mahmoudi, fdevaulx, jaafar.chbili, robert.bohn, abdella.battou}@nist.gov

***Abstract*—Cloud services have gained tremendous attention as a utility paradigm and have been deployed extensively across a wide range of fields. However, Cloud security is not catching up to the fast adoption of its services and remains one of the biggest challenges for Cloud Service Providers (CSPs) and Cloud Service Consumers from the industry, government and academia. These institutions are increasingly faced with threats affecting the confidentiality, integrity and availability of the cloud resources such as DoS/DDoS attacks, ransomware attacks, and data breaches to name a few. In the current cloud systems, security requires manual translation of security requirements into controls. Such an approach can be for the most part labor intensive, tedious and error-prone leading to inevitable misconfigurations rendering the system at hand vulnerable to misuse be it malicious or unintentional. Therefore, it is of utmost importance to automate the configuration of the cloud systems per the client's security requirements steering clear from the caveats of the manual approach. Furthermore, cloud systems need to be continuously monitored for any misconfiguration, and therefore lack of the required security controls. In this paper, we present a methodology allowing for cloud security automation and demonstrate how a cloud environment can be automatically configured to implement the required NIST SP 800-53 security controls. Also, we show how the implementation of these controls in the cloud systems can be continuously monitored and validated.**

Keywords—cloud computing, cybersecurity, automation, security controls

I. INTRODUCTION

Cloud services have been one of the most important paradigms of today's IT world due to its salient features of on-demand, flexible, scalable, ubiquitous computing with minimal resource management effort for the end-users. The National Institute of Standards and Technology (NIST) defines the cloud computing as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction [1]. The sheer fact that cloud services offer an on-demand model and therefore promote efficient spending on IT departments is reason enough for the ongoing movement to deploy cloud systems in both government and non-government institutions. Thus, per International Data Corporation (IDC), by 2018, at least half of the IT expenses will be cloud-based with a reach of 60% of all IT infrastructures and 60-70% of all software services by 2020 [2]. And, Wikibon predicts that by 2022, Amazon AWS platform by itself will be approximately $43 billion revenue per year, providing 8.2% of all cloud business [3].

Even though cloud computing is considered a major IT movement, the cloud security remains a plaguing challenge, according to a RightScale report (25% of respondents cited cloud security, lack of resources/expertise, and managing cloud spending as the main challenges) [4]. The cumulative cost of cyberattacks to an organization averages $3.5 million annually [5], which gives an economic illustration of the importance of cloud security. Stakeholders from all fields steadily realize that Cloud services face numerous threats: data breaches, compromised credentials, account hijacking, permanent data loss and DoS/DDoS attacks just to name a few [6]. Add to that the lagging CSPs security default capabilities which do not meet the organization's security and privacy requirements [7]. From the customer's perspective, more institutions from the private sector are developing interest in the NIST Cyber Security Framework (CSF) and Risk Management Framework (RMF) to address and manage security risk, define requirements and security controls implementing them. However, to the best of our knowledge, one cannot find a well-defined practical approach translating the said security controls into actionable items running on the cloud environments. Therefore, in this paper, we present a methodology to automatically create a cloud computing environment implementing NIST SP800-53 security controls to satisfy cybersecurity requirements of the cloud systems at hand.

In summary, this paper aims at answering the following questions: (a) What are the security requirements from both the user side and the CSP side? (b) How can a user specify the security requirements? (c) How can we automate the deployment of cloud systems that meet user security requirements? and (d) How can we validate that the cloud systems meet the security needs and provide a comprehensive compliance report to support?

The rest of the paper is organized as follows. In Section II, we present the background information on the cloud security and standards domains. We present a cloud computing security taxonomy in Section III to be used for better understanding and categorizing each aspect of cloud security. The proposed methodology is presented in Section IV. Next, we present a proof-of-concept implementation in Section V. Finally, Section VI concludes the paper.

Battou, Abdella; Bohn, Robert; Chbili, Jaafar; Mahmoudi, Charif; Merzouki, Mheni; Tunc, Cihan; de Vaulx, Frederic; hariri, Salim.
"Cloud Security Automation Framework."
Paper presented at The IEEE Workshop on Automation of Cloud Configuration and Operations, Tucson, AZ, United States. September 18, 2017 - September 22, 2017.

II. Background

A. Cloud Computing and Security Appraoches

Cloud computing involves the delivery of services through different models such as Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and Infrastructure-as-a-Service (IaaS). While this paper mainly focuses on the IaaS model, it can be extended to other service models as well. For the IaaS, the users are offered virtual machines (VMs) residing through a hyper vision technology such as Xen or VMware among others. Due to the complex structure of the cloud services, cloud computing suffers from multiple security issues [8]. All cyber-attacks targeting physical machines in the physical domain exist on virtual machines running on the cloud environment, with added burden from the hypervisor. As the cloud offers a pool of resources that users share, the importance of the hypervisor security increases. The dependency of cloud computing on the virtualized environment raises more security issues, like hypervisor exploitations [9]. One instance of these attacks is the injection of malware in the publicly available virtual machine image, which subsequently affects the services of the cloud. Another major security issue for cloud computing systems is the insider attacks [10].

Many solutions have been proposed to solve security issues [11-14] in cloud computing. Some cloud security systems implemented a recovery-based intrusion tolerant algorithm that enhances the availability and resilience of cloud services or focused on hiding the data as a method to increase services' resilience to attacks [13]. Many of the security solutions developed for the cloud focus on efficiently protecting the cloud storage against diverse range of attacks including roll-back attacks [15]. Some of the proposed security solutions use innovative risk-based analysis for the security testing of the cloud environment. This risk-based analysis reduces the number of possible misuse cases of the cloud [14].

B. Autonomic Computing

By definition, cloud computing should have dynamic management and self-organization [1, 16, 17]. Nevertheless, most of the current cloud implementations are missing those two essential attributes. Understanding the functional requirements of cloud services is an essential key in defining the architecture of the system providing those services [18]. Applying software design approaches to the development of autonomic management systems provides a solid ground to maintain the service level agreement, protect against external attacks, prevent failures and enables recovery [19]. Proactively managing failures and providing self-healing is an important attribute of the cloud [20]. Self-configuration allows the cloud to adapt to changes by tuning the allocation of virtual resources and the applications parameters, thereby achieving self-optimization [21].

C. Cloud Standards

As cloud computing is being used ubiquitously, there are multiple cloud standards works in the literature such as the one from the Cloud Standards Customer Council (CSCC) aiming at cloud systems' successful adoption and solving the security and interoperability issues [22]. Hence, in terms of security, CSCC presented a reference for the organizations adopting the cloud computing and its security impacts [23]. Organization for the Advancement of Structured Information Standards (OASIS) provides open standards for the IT world in collaboration with the industry and academia [24]. Topology and Orchestration Specification for Cloud Applications (TOSCA) is one of OASIS standards intended to define services and applications, and their relationships, especially for cloud systems with a focus on automated management, portability and deployment [24]. For this purpose, TOSCA introduces a grammar for describing service templates, requirements, capabilities, and policies with the help of YAML. While TOSCA is a standard model, there are multiple implementations such as OpenTOSCA [25], Cloudify [26] in and OpenStack Heat [27]. However, current implementations of this standard do not address the cloud security challenges and only focus on the management and automation of the application deployment.

D. Discussion

In summary, the current efforts on cloud automation focus heavily on performance and predominantly overlook the vital security aspect. Also, to the best of our knowledge there have not been any studies with focus on automation of security mechanisms implementation in cloud systems. Therefore, this paper is the first attempt in this direction.

III. Cloud Computing Security Taxonomy

To fully understand the cloud computing security issues, we first developed a cloud security taxonomy based on NIST SP 800-53 [28] and Federal Risk and Authorization Management Program (FedRAMP) [29] security assessment framework. Next, we utilized the taxonomy to implement the required security controls and their management processes. Our security taxonomy has three sub-categories: Security components, privacy, and compliance, as shown in Figure 1.

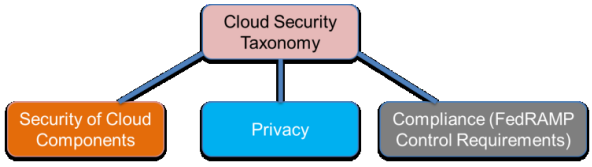


Figure 1. Cloud security taxonomy used for identifying the security controls.

Securing cloud systems involve securing the infrastructure, network, hosts, applications and data through a wide range of mechanisms such as Identity, Credentials and Access Management (ICAM), Data Segregation and Regulatory Compliance (Figure 2).

Please note that we only consider the security of the VMs in this work rather than the underlying physical hosts.

Figure 2. Cloud components security taxonomy.

One of the main goals of securing the cloud is ensuring users' data privacy. Figure 3 depicts a breakdown of data

Battou, Abdella; Bohn, Robert; Chbili, Jaafar; Mahmoudi, Charif; Merzouki, Mheni; Tunc, Cihan; de Vaulx, Frederic; hariri, Salim.
"Cloud Security Automation Framework."
Paper presented at The IEEE Workshop on Automation of Cloud Configuration and Operations, Tucson, AZ, United States. September 18, 2017 - September 22, 2017.

privacy concerns and governing principals which can be used to define the users' privacy requirements.

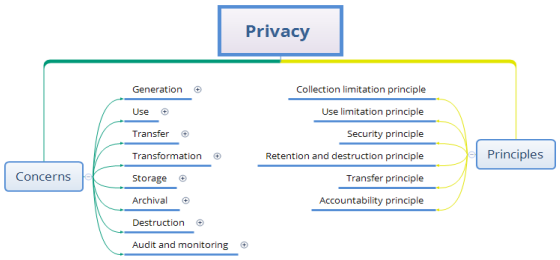


Figure 3. Privacy of the cloud can be described by the concerns and principles.

For these systems to be secured, cloud service customers, be it from the government or the private sector, need to make sure the cloud service providers satisfy a given set of security requirements. This is where FedRAMP comes into play as it assists government agencies in meeting the mandated FISMA requirements for cloud systems, and may be used by cloud service customers from the private sector as guidance when implementing their cloud security requirements. This paper does not intend to discuss the specifics or FedRAMP but rather use its security controls based on the NIST SP 800-53. The categorization (Low, Moderate, High) of the system at hand is done through FIPS PUB 199. Then the set of security controls corresponding to the baseline need to be implemented. The security controls can be grouped into three categories: Technical, Operational, and Management. In this paper, we do not address all the security controls but only the technical ones which need to be implemented on the VMs.

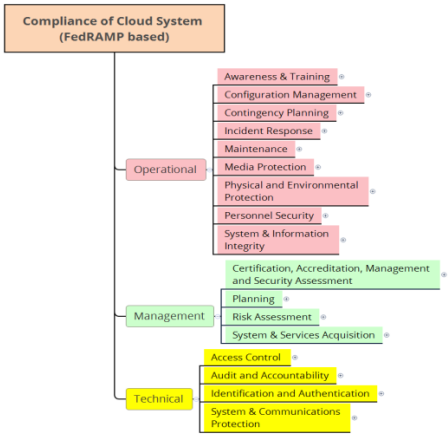


Figure 4. For the compliance, FedRAMP based taxonomy [29] can be used.

IV. Cloud Security Management Automation

In our cloud security management approach, we focus on the security automation for the Infrastructure-as-a-Service (IaaS) services offered by CSPs where users create, operate, and manage VMs with the requested virtual resources (e.g., number of cores, amount of memory, storage requirements, co-processors such as GPU) with full access in most cases using a dashboard since manual security management of the environment is improper.

The proposed cloud security management approach is shown in Figure 5. In our architecture, the user specifies the security requirements of the needed cloud environment and the VM definitions (e.g., the VM required resources) through an editor. Through a communication channel (e.g., Secure Shell (SSH), or Secure Sockets Layer (SSL)), the user's requirements are passed to the configuration engine. Next, the configuration engine interfaces with the CSP to create the requested cloud environment that satisfies the user security requirements. Once the system is set-up, the configuration engine interfaces with the CSP to continuously monitor the cloud environment and to notify the users if any abnormal configuration or behavior is detected. Below, we provide further details of each step in the form of a proof of concept.

As an example, a user wants to create a cloud environment for a web application that consists of a web server and a database (DB) server (Figure 6). In this example, the user wishes to specify the ports which will be used for communications among VMs, administrators, and the users of the web application. In addition to the required VM deployment information (such as the VM name, VM size), the user also needs to implement the selected security controls to satisfy the security requirements.

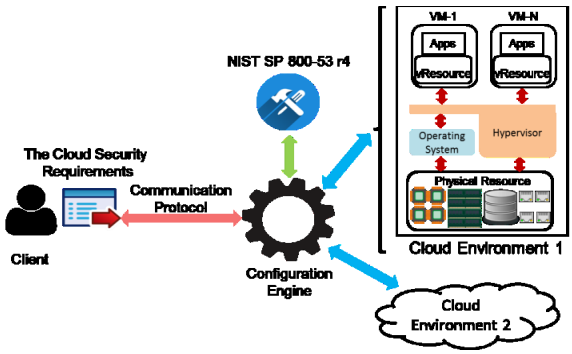


Figure 5. Cloud security management architecture.

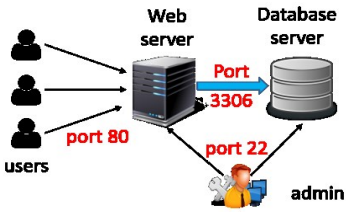


Figure 6. An example of a web application deployment.

One of the main questions that needs to be answered is how the users will specify their security requirements. NIST SP 800-53 revision 4 is a comprehensive catalog of security controls for all U.S. federal information systems except those related to national security. Therefore, we suggest using NIST SP 800-53 to implement the user requirements. Please note that not all the requirements can be met by the CSP due to lack of capabilities, which can create some gaps initially. In other scenarios, during the lifecycle of a cloud environment, the CSP capabilities may

Battou, Abdella; Bohn, Robert; Chbili, Jaafar; Mahmoudi, Charif; Merzouki, Mheni; Tunc, Cihan; de Vaulx, Frederic; hariri, Salim.
"Cloud Security Automation Framework."
Paper presented at The IEEE Workshop on Automation of Cloud Configuration and Operations, Tucson, AZ, United States. September 18, 2017 - September 22, 2017.

change (and cannot provide the previous capabilities anymore). In such cases, the user may seek other providers.

For the security control selection, we define the parameters based on the taxonomy we have presented in Section III. Like TOSCA standard, we believe that a user-friendly data serialization standard with an easy to read and edit syntax should be used; therefore, we chose to encode the requirements in YAML [24].

Figure 7 depicts the YAML based definition for the given example. The user encodes the VM deployment information as well as required security controls. The user defines the VM resource requirements (flavors) as ***m1.large*** and ***m1.medium*** (1 core, 2GB memory, and 20GB storage size), respectively (line 3 and 10). Then, based on the already implemented security controls by the CSP, the configuration engine gets the lists of the remaining security controls, script name and parameters, to be applied to satisfy the user's requirements. For the given example, the user selects "***AC-2*** Account Management" that requires monitoring the use of information system accounts, "***AC-7*** Unsuccessful Logon Attempts" that limits unsuccessful logon attempts, and finally "***SC-7*** Boundary Protection Control Enhancement (5) Boundary Protection | Deny By Default / Allow By Exception" that blocks all the network ports by default unless they are specified to be open. CSP offers a security control mechanism that checks the logged in accounts to verify if the account management is being fulfilled (line 6 and line 14). To control which users are valid, the script requires a list of users (comma separated) in this scenario. Hence, the user defines that the web server needs to have only ***user1*** and ***user2*** logged in to the system and any other user would be considered as a malicious user. Similarly, only ***user3*** is accepted for the DB server.

```
1.  VM:
2.    - name: web_server
3.      flavor: m1.large
4.      controls:
5.        - script: check_current_users
6.          parameters: "-l user1,user2"
7.        - script: check_login_attempts
8.          parameters: "-w 3 –c 5"
9.    - name: db_server
10.     flavor: m1.medium
11.     controls:
12.       - script: check_open_ports
13.         parameters: "-p 22,3306"
14.       - script: check_current_users
15.         parameters: "-l user3"
```

Figure 7. VM environment definition given to the configuration engine.

In addition, since DB server will be acting in the background and does not need to interact with the end-users, the admin can whitelist port 22 and 3306 (lines 12 and 13) so that the use of any other port will be flagged as an illegal action. Finally, for the security control ***AC-7*** to check unsuccessful logon attempts, the CSP requires a warning and a critical level (line 7). When the unsuccessful login attempts exceed the given number of warnings, the user is notified. When the number of attempts is beyond the critical level, the remote connection can be temporarily blocked as a security measure mitigating a brute force attack or a dictionary attacks carried on the VMs.

Please note that each security control is an individual script based on the user requirements and CSP capabilities that operate as an agent on VMs.

The configuration engine is responsible for creating the requested VMs (with the given resources requirements), applying the specified security controls, and monitoring the VMs continuously to see if there are any requirements which are not met and notify the users of the current state of the environment. In Figure 8, we present the configuration engine algorithm. In this approach, the configuration engine reads the configuration file given in Figure 8 (line 1) and creates the requested VMs with the given resource set and VM names (line 2). This results in a list of created VMs with their IPs, called VM_list. In order to upload the security controls implementation scripts (line 4) and configure the VMs (line 5), the configuration engine waits until the VMs are active and accessible. The security controls the CSP is offering are uploaded to the VMs based on the requirements and then the VMs are configured accordingly to operate. Next, the monitoring control center (the monitoring capabilities of the configuration engine) is configured so that the created VMs can be monitored and their states are logged to a database (line 6). And, during their lifecycles, the VMs are monitored continuously (line 7). If there is any state that does not meet the requirements (for example, if any user other than user1 and user2 exist on the web_server), the users are notified so that they act accordingly.

```
1.  configuration ← read_yaml
2.  VM_list = create _VMs(configuration["VM"])
3.  wait_VMs_active(VM_list)
4.  upload_security_scripts(VM_list,
    configuration["VM"])
5.  configure_VMs(VM_list, configuration["VM"])
6.  configure_monitoring_control_center(VM_list)
7.  while(true):
8.      states = monitor_VMs(VM_list,
    configuration["VM"])
9.      If(states not expected):
10.         notify_user(states)
```

Figure 8. Configuration engine algorithm.

Furthermore, during the lifecycle of the VMs, the user's security requirements may change. Thus, the configuration engine should also allow updating the security controls.

V. Proof of Concept Implementation

In this section, we present the proof of concept implementation environment, the tools used, and how they are configured and automated. For the implementation and testing, we have created an OpenStack based private cloud environment. OpenStack is an open-source cloud stack for building public/private clouds using multiple homogenous and heterogeneous systems and managing large pools of compute, storage, and networking resources through a dashboard or using

Battou, Abdella; Bohn, Robert; Chbili, Jaafar; Mahmoudi, Charif; Merzouki, Mheni; Tunc, Cihan; de Vaulx, Frederic; hariri, Salim.
"Cloud Security Automation Framework."
Paper presented at The IEEE Workshop on Automation of Cloud Configuration and Operations, Tucson, AZ, United States. September 18, 2017 - September 22, 2017.

APIs [31]. Figure 9 shows the topology of the testbed to experiment with and evaluate the proposed methodology. Our OpenStack environment consists of one separate controller node (which manages the cloud environment) and multiple compute nodes (enabling VM operations using a hypervisor). The controller node has the required OpenStack services such as Glance (image service) and Keystone (the identity manager) as well as OpenStack Python libraries for the automation (which is shown as management middleware). For the compute nodes, we have allocated three Dell XPS 8700 towers with i7 4770 processors and 12GB memory, running Ubuntu 16.04 Server for the hosts operating systems. For the virtualization, we have chosen Kernel-based Virtual Machine (KVM) hypervisor as it is supported by the Linux kernel. For the communication of the OpenStack services, an internal network switch is used.

For the continuous monitoring of the cloud environment, Nagios 3 [32] has been chosen as it is the go-to tool for remote system monitoring with flexible agent support. We have installed and configured Nagios on an individual VM and integrated with NDOUtils to provide database support which will be used to query the current and previous states of the VMs. As shown in the previous section example, multiple agents have been created based on the NIST SP 800-53 using Nagios NRPE. NRPE allows us to create custom scripts for the VMs' security controls that can be monitored by Nagios server.

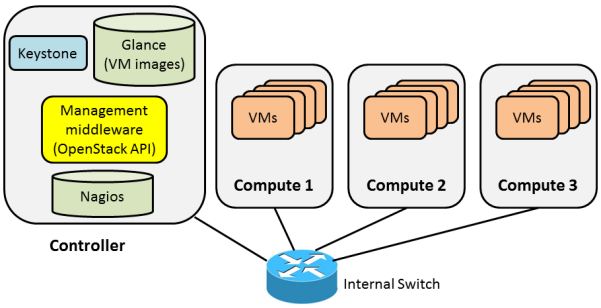


Figure 9. Proof of concept testbed architecture.

In Figure 10, we present a breakdown of the configuration time. The engine spends most of its configuration time, 33 seconds, creating the VMs and waiting for them to be accessible (i.e., active) while uploading the scripts and configuring the VMs (and Nagios server) take 2 seconds each. After the VMs are configured, it only takes 0.04 seconds to get the update from the Nagios DB. Not to mention that the agents used for the security controls implementations are lightweight in that their load on the system is negligible, and are checked periodically for a configurable amount of time, 5 seconds in this case; consequently, the agents do not introduce overhead which would impact the users' operations on the cloud environment.

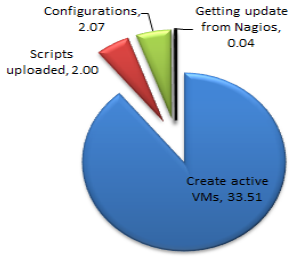


Figure 10. Time distribution of major configuration engine operations (in seconds).

Below, in Figure 11, we show an example of the user notification. In Figure 11(a), the services show that there are no unauthorized users, i.e. the logged in users are from the set of given allowed users. In Figure 11(b), the user is notified of ports that were not supposed to be open. Port 5666 is used by Nagios communication (i.e., NRPE service) and since it was not specified as an allowed port, the user is notified.

```
OK -- No unauthorized users
```

(a) Logged in users with no security violation

```
CRITICAL -- Unauthorized ports open.
Unauthorized ports: 5666
```

(b) Unauthorized open ports are notified

Figure 11. User notification example

VI. Conclusion

Even though the cloud computing systems are highly popular for personal usage, for organizations and government agencies, security of their cloud infrastructure is still a major concern. Current techniques for cloud security are manual and error prone which introduces additional vulnerabilities. Hence, it is critically important to develop a cloud security automation methodology that will be used to configure cloud systems that meet user security requirements. Thus, in this paper, we demonstrated a methodology that is based on the NIST SP 800-53 security controls and Nagios monitoring tool to implement the selected security controls using cloud service provider's capabilities. We have demonstrated a proof of concept implementation using OpenStack in a private cloud environment. As future work, we are planning on including multiple cloud systems as well as the ability to launch automated/semi-automated actions when there is a security control violation.

Acknowledgment

This work is partly supported by National Science Foundation (NSF) research project NSF CNS-1624668 and National Institute of Standards and Technology (NIST).

DISCLAIMER

Any mention of commercial products or organizations is for informational purposes only; it is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the products identified are necessarily the best available for the purpose.

The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

REFERENCES

[1] R. Mell, and T. Grance, "The NIST definition of cloud computing," NIST Special Publication 800-145, 2011.

[2] Report IDC FutureScape "Worldwide IT Industry, (2016), Leading Digital Transformation to Scale", New York

[3] "How Big Can AWS Get?", [Online] URL: http://wikibon.com/how-big-can-aws-get/, Accessed: June 2017

[4] "RightScale 2017 State of the Cloud Report" [Online] URL: http://www.rightscale.com/blog/cloud-industry-insights/cloud-computing-trends-2017-state-cloud-survey, Accessed: June 2017

[5] "Security Beyond the Traditional Perimeter Executive Summary", Ponemon Institute: July 2016, [Online] URL: http://cdn2.hubspot.net/hubfs/30658/Ponemon_External_Threat_2016__ExecSumm.pdf, Accessed: June 2017

[6] CSA Top Threats Working Group. "The Treacherous 12: Cloud Computing Top Threats in 2016." Cloud Security Alliance (CSA), Feb (2016).

[7] R. Chandramouli, S. Garfinkel, S. Nightingale, S. Rose, "Trustworthy Email," NIST Special Publication 800-177, Sep. 2016.

[8] B.P. Rimal, E. Choi, and I. Lumb, "A Taxonomy and Survey of Cloud Computing Systems," NCM 9, 2009, pp. 44-51.

[9] M. Schmidt, L. Baumgartner, P. Graubner, D. Bock and B. Freisleben, "Malware Detection and Kernel Rootkit Prevention in Cloud Computing Environments," 19th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, 2011.

[10] C. Modi, D. Patel, B. Borisaniya, A. Patel and M. Rajarajan, "A survey on security issues and solutions at different layers of Cloud computing," The Journal of Supercomputing, pp. 1-32, 2012

[11] M. Abbasy and B. Shanmugam, "Enabling Data Hiding for Resource Sharing in Cloud Computing Environments Based on DNA Sequences," IEEE World Congress, 2011.

[12] J. Feng, Y. Chen, D. Summerville, W. Ku, and Z. Su, "Enhancing cloud storage security against roll-back attacks with a new fair multi-party non-repudiation protocol," Consumer Communications and Networking Conference, 2011.

[13] Q. Nguyen and A. Sood, "Designing SCIT architecture pattern in a Cloud-based environment," IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops, 2011.

[14] P. Zech, "Risk-Based Security Testing in Cloud Computing Environments," IEEE Fourth International Conference on Software Testing, Verification and Validation, 2011.

[15] L. Kaufman, "Data security in the world of cloud computing," IEEE Security and Privacy Journal, vol. 7, no. 4, pp. 61-64, 2009

[16] F. Heylighen, and C. Gershenson, "The Meaning of Self-organization in Computing," IEEE Intelligent Systems, 18(4), 2003, pp. 72-75.

[17] M. Parashar, and Salim Hariri, "Autonomic computing: An overview," Unconventional Programming Paradigms, 2005, pp. 97-97.

[18] H. Takabi, J. BD Joshi, and G.-J. Ahn, "Security and privacy challenges in cloud computing environments," IEEE Security & Privacy 8, no. 6, 2010, pp. 24-31.

[19] C. Tunc, F. Fargo, Y. Al-Nashif, S. Hariri, and J. Hughes, "Autonomic Resilient Cloud Management (ARCM) Design and Evaluation," International Conference on Cloud and Autonomic Computing, London, 2014, pp. 44-49.

[20] P.K. Patra, H. Singh, and G. Singh, "Fault tolerance techniques and comparative implementation in cloud computing," International Journal of Computer Applications, 2013, 64(14).

[21] M. Agarwal, V. Bhat, H. Liu, V. Matossian, V. Putty, C. Schmidt, G. Zhang, L. Zhen, M. Parashar, B. Khargharia, and S. Hariri, "Automate: Enabling autonomic applications on the grid," IEEE Autonomic Computing Workshop, 2003, pp. 48-57.

[22] "The Cloud Standards Customer Council" URL: http://www.cloud-council.org

[23] "Security for Cloud Computing Ten Steps to Ensure Success Version 2.0," URL: http://www.cloud-council.org/deliverables/CSCC-Security-for-Cloud-Computing-10-Steps-to-Ensure-Success.pdf

[24] "OASIS Topology and Orchestration Specification for Cloud Applications (TOSCA)", URL: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=tosca

[25] T. Binz, U. Breitenbücher, F. Haupt, O. Kopp, F. Leymann, A. Nowak, and S. Wagner, "OpenTOSCA–a runtime for TOSCA-based cloud applications," Springer International Conference on Service-Oriented Computing, 2013, pp. 692-695

[26] L. Trammell, "TOSCA Cloud Orchestration for Beginners" URL: http://cloudify.co/2015/07/21/what-is-TOSCA-cloud-application-orchestration-tutorial-cloudify.html, Accessed: June 2017

[27] OpenStack Heat-Translator, [Online] URL: https://wiki.openstack.org/wiki/Heat-Translator

[28] Security and Privacy Controls for Federal Information Systems and Organizations, NIST Special Publication 800-53 Revision 4

[29] "FedRAMP Security Controls Baseline", Accessable: https://www.fedramp.gov/resources/documents-2016/

[30] CloudFlare DDoS protection, URL: www.cloudflare.com/lp/ddos-a/

[31] "OpenStack: Open source software for creating private and public clouds", URL: https://www.openstack.org

[32] "Nagios: The Industry Standard in IT Infrasturecture Monitoring", https://www.nagios.com

Battou, Abdella; Bohn, Robert; Chbili, Jaafar; Mahmoudi, Charif; Merzouki, Mheni; Tunc, Cihan; de Vaulx, Frederic; hariri, Salim.
"Cloud Security Automation Framework."
Paper presented at The IEEE Workshop on Automation of Cloud Configuration and Operations, Tucson, AZ, United States. September 18, 2017 - September 22, 2017.

Composition of CT Lung Density Reference Using Prompt Gamma Activation Analysis

H. Heather Chen-Mayer, Danyal Turkoglu, Rick Paul, Zachary Levine

National Institute of Standards and Technology, Gaithersburg, Maryland, USA

INTRODUCTION

The Computed Tomography (CT) density measures in Hounsfield Units (HU) have been used as a quantitative image biomarker for the diagnosis and monitoring of lung density changes due to emphysema, a feature of chronic obstructive pulmonary disease (COPD) [1]. To reduce variability in the density metrics specified by CT attenuation, measured in Hounsfield Units (HU), phantom studies in a variety of scanner models were conducted to provide assessments of the accuracy and precision of the density metrics across platforms solely due to machine calibration [2]. To provide a calibration reference for these phantoms, NIST has developed a suite of lung density reference foams, Standard Reference Material (SRM) 2088, which are certified for the absolute density [3-5]. The SRM is aimed at establishing the HU-electron density relationship and removing scanner dependence for the CT lung density measures. However, the composition of the polyurethane foam material (LAST-A-FOAM® FR-7100 series, General Plastics, USA) is not well known. We employed Prompt Gamma-ray Activation Analysis (PGAA) to determine the elemental composition by measuring the characteristic energies and intensities of prompt gamma rays emitted from H, N and C following neutron capture by nuclei [6].

DESCRIPTION OF THE ACTUAL WORK

PGAA was performed on SRM 2088 polyurethane foam blocks with five different nominal densities: 0.060 g/cm^3, 0.120 g/cm^3, 0.185 g/cm^3, 0.230 g/cm^3, and 0.325 g/cm^3. Each was cut into about 4 cm × 2 cm × 1 cm block, and mounted by Teflon® strings in an evacuated sample chamber for neutron beam irradiation and for gamma-ray spectroscopy with a high-purity germanium detector. The collection time varied from 1 h to 16 h. The relative atom fractions were determined by taking the ratio of the numbers of atoms, obtained from Eq. 1.

$$n_{x,\gamma_i} = \frac{A_{x,\gamma_i}}{\varepsilon_{\gamma_i}\ \sigma_{x,\gamma_i}\ \varphi_{th}} \qquad (1)$$

The quantities of interest in Eq. 1 are defined as:
n_{x,γ_i}: the number of atoms for element x (i.e., H, C, or N) for gamma ray γ_i.
i: peak index, in cases where there are multiple gamma-ray peaks for the element. (For H, $i = 1$.)
A_{x,γ_i}: count rate of the characteristic gamma-ray peak i from element x.
σ_{x,γ_i}: partial gamma-ray production cross section for gamma ray γ_i from element x.
ε_{γ_i}: detection efficiency for gamma ray γ_i.
φ_{th}: thermal-equivalent neutron flux in sample.

By taking the ratio the number of atoms determined by Eq. 1, the neutron flux characterization is bypassed for homogenous samples since it is the same for elements in the same sample. Thus, the resulting ratio is independent of sample characteristics (mass, composition, shape, etc.).

RESULTS

Gamma-ray spectra were acquired for the foam blocks with five different density. Fig.1 shows a comparison of the resulting gamma-ray spectra from the highest and lowest density samples. The peak analysis was performed using PeakEasy (Los Alamos National Laboratory). Energies of prompt gamma-ray peaks used for the analysis are:

- C: 1,261 keV; 3,684 keV; 4,945 keV
- H: 2,223 keV
- N: 5,269 keV; 5,298 keV; 5,533 keV; 10,829 keV
- Cl: 1,164 keV (not labeled)

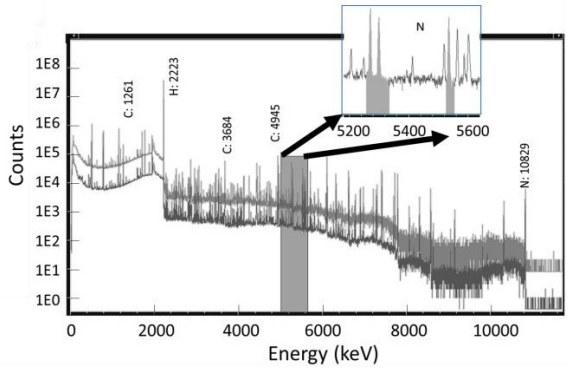


Fig. 1. Gamma-ray spectra acquired (normalized by detector live time) from the (top) highest density sample and from the (bottom) lowest density sample.

The number of atoms determined using Eq. 1 (with an arbitrary neutron flux value) were plotted (Fig. 2) for C and N versus H. The uncertainties due to counting statistics are small in comparison to the different values obtained from the different peaks for C and N, the average and the standard deviation (over i) of which are plotted in Fig. 2.

Chen-Mayer, Huaiyu; Levine, Zachary; Paul, Rick; Turkoglu, Danyal.
"Composition of CT Lung Density Reference Using Prompt Gamma Activation Analysis."
Paper presented at 2017 ANS Winter Meeting and Nuclear Technology Expo, Washington, DC, United States. October 29, 2017 - November 2, 2017.

The number of atoms increased proportionally to the density of the sample, as expected. The slope of each linear trendline represents Eq. 2 and can be used to determine the stoichiometry of the material because the samples are considered to have identical compositions but varying physical density.

Of the many possible "polyurethane" compositions, only two, or possibly three, have the H/C and N/C ratios that simultaneously fall within or close to the 95% confidence interval of the ratios determined by PGAA (Fig. 3). This information will be useful for CT number calibration using the SRM 2088 in a lung density phantom. Future directions include the investigation of sources of uncertainties and analysis of other minor elements in the matrix that have effects on x-ray attenuation.

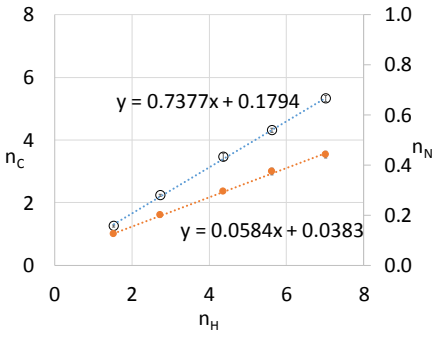


Fig. 2. Determination of atom ratios of C, H, and N based on Eq. 1. The slope and uncertainty are used to create the 95% confidence intervals of the most likely ratios in Fig. 3.

DISCLAIMER

Certain commercial products identified in this paper do not imply recommendation or endorsement by the authors or by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose. Contributions of the National Institute of Standards and Technology are not subject to copyright.

REFERENCES

1. D. A. Lynch, J. H. M. Austin, J. C. Hogg, et al. "CT-definable subtypes of chronic obstructive pulmonary disease", *Radiology* **277** 192–205 (2015)
2. H. H. Chen-Mayer, M. K. Fuld, B. Hoppel, P. F. Judy, J. P. Sieren, D. A. Lynch, A. Possolo, and S. B. Fain, "Standardizing CT lung density measure across scanner manufacturers", *Med. Phys.* **44**, 974-985 (2017).
3. Z. H. Levine, M. Li, A. P. Reeves, D. F. Yankelevitz, J. J. Chen, E. L. Siegel, A. Peskin, and D. N. Zeiger, A Low-Cost Density Reference Phantom for Computed Tomography, *Med. Phys.* **36**, 286–288 (2009).
4. www-.nist.gov/srmors/view_detail.cfm?srm=2088
5. Z. H. Levine, H. H. Chen-Mayer, A. L. Pintar, and D. S. Sawyer IV, "Standard Reference Materials for Medical CT", 490 OSA Proc. Conf. on Quantitative Medical Imaging, Bethesda, MD, 2013, *OSA Technical Digest (online)*, paper QW1G.3 (2013). doi:10.1364/QMI.2013.QW1G.3
6. R. Paul, R. Lindstrom, "Prompt Gamma-Ray Activation Analysis: Fundamentals and Applications", *J. Radioanal. Nucl. Chem.,* 243(1), 181 (2000).

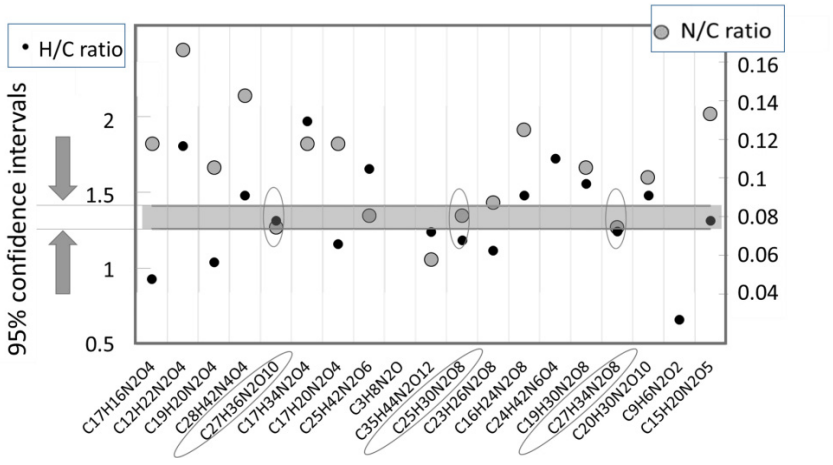


Fig. 3. Summary of the H/C and N/C ratios from the possible polyurethane compositions, with the ones closest to the ratios (circled) determined by PGAA.

Chen-Mayer, Huaiyu; Levine, Zachary; Paul, Rick; Turkoglu, Danyal.
"Composition of CT Lung Density Reference Using Prompt Gamma Activation Analysis."
Paper presented at 2017 ANS Winter Meeting and Nuclear Technology Expo, Washington, DC, United States. October 29, 2017 - November 2, 2017.

Advancing HL7 v2 to New Heights: A Platform for Developing Specifications, Test Plans, and Testing Tools

Robert Snelick

National Institute of Standards and Technology (NIST), Gaithersburg, USA

Abstract

Development of HL7 v2 data exchange interface specifications has long been problematic, plagued with ambiguous and inconsistent requirement specifications. This situation leads to potential misinterpretation by implementers, thus limiting the effectiveness of the specification and creating artificial and unnecessary barriers to interoperability. Likewise, the ability to test implementations effectively for conformance to the specifications is hindered. The current approach of specification development and test plan creation relies on word processing tools, meaning implementers and testers must read and interpret the information in these documents and then translate it into machine-computable requirements and test assertions. This approach is error prone—a better methodology is needed. We present a set of productivity tools in an integrated platform that allow users to define and constrain HL7 v2 specifications and to develop test plans that result in machine-computable artifacts. A testing infrastructure and framework subsequently uses these artifacts to create conformance testing tools automatically. We present and demonstrate the utility of a platform for developing specifications, writing test plans, and creating testing tools. The value proposition of this end-to-end methodology is explained for authors writing HL7 v2 specifications, for developers implementing interfaces, and for testers creating validation tools.

Keywords

Conformance; Healthcare Data Exchange Standards; Healthcare Information Systems; Interoperability; Specification Development Tools; Testing Tools

Correspondence to:

Prof. Dr. Robert Snelick
National Institute of Standards and Technology, 100 Bureau Drive
Stop 8970, Gaithersburg, MD 20899, USA.
E-mail: robert.snelick@nist.gov

EJBI 2017; 13(1):01-08
received: June 08, 2017
accepted: July 19, 2017
published: October 10, 2017

1 Introduction

For 30 years, HL7 (Health Level 7) Version 2 (v2) has been the predominant standard used for the exchange of healthcare administrative and clinical data. Healthcare information systems use the HL7 v2 protocol to develop standardized interfaces to connect to and exchange data with other systems. HL7 v2 covers a broad spectrum of domains including Patient Administration, Laboratory Orders and Results, and Public Health Reporting. The base HL7 v2 standard [1] is a framework that contains many message events, and for each event it provides an initial template (starting point) that is intended to be constrained for a specific use case. The application of constraints to a message event is referred to as profiling [2, 3]. For example, the VXU V04 (Unsolicited Vaccination Record Update) message event is a generic template for communicating information about a patient's immunization related events. The base message template is composed of mostly optional data elements. For a given use case, e.g., Send Unsolicited Immunization Update for the US Realm [4], the message template is "profiled". That is, elements can be constrained to be required, content can be bound to a set of pre-coordinated codes, and so on. The base message event (e.g., VXU V04) that has been constrained for a particular use (e.g., submitting immunization events) is referred to as a conformance profile[1]. An implementation guide is a collection of conformance profiles organized for a workflow (e.g., submitting, acknowledging, querying, and responding to/for immunization events). In this example, four conformance profiles exist, each with different message events; one for submitting an immunization event, for sending an acknowledgment, for querying for an immunization history, and for providing an immunization history. To date, HL7 v2 implementation guides have been created using word processing programs, which has resulted in ambiguous and inconsistent specification of requirements. This practice has hindered consistent interpretation among implementers, which has created an unnecessary barrier to interoperability.

[1]Also, referred to as a message profile.

Snelick, Robert.
"Advancing HL7 v2 to New Heights: A Platform for Developing Specifications, Test Plans, and Testing Tools."
Paper presented at 17th International HL7 Interoperability Conference(IHIC 2017), Athens, Greece. October 19, 2017 - October 24, 2017.

We present an end-to-end methodology and platform for developing specifications (implementation guides), writing test plans, and creating testing tools in the HL7 v2 technology space [5]. The platform includes three key foundational components:

- A tool to create implementation guides and conformance profiles
- A tool to create test plans, test cases, and associated test data
- A testing infrastructure and test framework to build testing tools

A key to the approach is that the "normal" process of creating implementation guides, test plans, and testing tools is "reversed". Instead of creating requirements using a natural language and subsequently interpreting the requirements to create test plans and test assertions, the requirements are captured with tools that internalize the requirements as computable artifacts.

Figure 1 illustrates a high-level overview of the methodology. Domain experts develop use cases, determine the message events that correspond to the interactions in the use cases, and then proceed to define the requirements. Using the methodology, they accomplish these tasks by entering this information into the Implementation Guide Authoring and Management Tool (IGAMT). During this process, the domain experts constrain the message events according to the requirements needed by the use case. Section 2 will elaborate more on this process and on the details of how the requirements are constrained. The output of IGAMT is a set of artifacts that are represented in Word, HTML, and XML formats. The complete implementation guide, including the narrative and messaging requirements, can be created in IGAMT and then exported in Word or HTML. Such formats are suitable for ballot at standards development organizations such as HL7 or IHE (Integrating the Healthcare Enterprises [6]). In May 2017, two HL7 v2 implementation guides that were generated by IGAMT were submitted for ballot. Each conformance profile can be exported as XML[2]. The XML format contains all the messaging requirements in a machine-computable representation, which is the most important aspect of IGAMT, since the XML conformance profiles have many uses including a computable definition of the message interface, message validation, test case and message generation, and source code generation.

The XML conformance profiles can be imported into the Test Case Authoring and Management Tool (TCAMT). TCAMT is used to create targeted test cases for interactions (profiles) defined in the implementation guide. The output is an additional set of constraints in an XML format. The entirety of the output generated from IGAMT and TCAMT is called a "resource bundle"[3].

[2]The XML format is defined by NIST and is publicly available but is not yet standardized. NIST intends to propose the format to HL7 for adoption. Additionally, there is no relationship between this format and other HL7 profiling formats such as the Templates Implementable Technology Specification (ITS) standard and FHIR.

[3]Is not related to a resource bundle in FHIR (Fast Healthcare Interoperability Resources).

The NIST platform includes a testing infrastructure of common utilities used for testing, such as a message validation engine, along with a testing framework that provides various testing tool components, such as a communication framework and a profile viewer. Testing Tool instances are then created using both the testing infrastructure and framework components as well as the resource bundle output generated from IGAMT and TCAMT.

The NIST platform allows end users to create conformance testing tools by means of a set of productivity tools. This streamlined approach can greatly reduce today's problems with conformance test tools. These problems include: tools often don't exist, they are expensive to build, they are difficult to update in a timely fashion, they are not adaptable for local refinements, and their time to market is lengthy. Additionally, the platform provides value through enforcing consistent and rigorous rules for requirements specifications.

The remainder of this paper explains the NIST platform in more detail in the context of how it can be applied in real-world use case settings. We first describe how IGAMT is used to define and constrain conformance profiles. One important aspect is the application of recently developed methods and best practices for requirements specification. Additionally, a brief overview of the validation process is given. Next, an explanation of how a set of targeted test cases are created in TCAMT is provided. In Section 4 we discuss a testing infrastructure and framework components. Next, an overview of the resulting test tools and how they are created is presented. Finally, there is a discussion on how the platform supports testing capabilities beyond the scope of the HL7 v2 interoperability specification. One goal of this paper is to inform the reader about the ease with which HL7 v2 implementation guides, test cases, and testing tools can be created using the NIST platform compared to the current laborious methods used today.

2 IGAMT

IGAMT [5] is a tool used to create HL7 v2.x implementation guides that contain one or more conformance profiles. The tool provides capabilities to create both narrative text (akin to a word processing program) and messaging requirements in a structured environment. Our focus in this paper is on the messaging requirements.

IGAMT contains a model of all the message events for every version of the HL7 v2 standard. Users begin by selecting the version of the HL7 v2 standard and the message events they want to include and refine in their implementation guide. For example, the message events VXU^V04, ACK, QBP^K11, and RSP^K11 are used to create eight conformance profiles in the immunization implementation guide [4]. Each message event is profiled (constrained) to satisfy the requirements of the use case. The QBP and RSP

Snelick, Robert.
"Advancing HL7 v2 to New Heights: A Platform for Developing Specifications, Test Plans, and Testing Tools."
Paper presented at 17th International HL7 Interoperability Conference(IHIC 2017), Athens, Greece. October 19, 2017 - October 24, 2017.

message types are used more than once to specify different uses.

Rules for building an abstract message definition are specified in the HL7 message framework, which is hierarchical in nature and consists of building blocks generically called elements [1]. These elements are segment groups, segments, fields, and data types (i.e., components and sub-components). The requirements for a message are defined by the message definition and the constraints placed on each data element. The constraint mechanisms are defined by the HL7 conformance constructs, which include usage, cardinality, value set, length, and data type. Additionally, explicit conformance statements are used to specify other requirements that can't be addressed by the conformance constructs. The process of placing additional constraints on a message definition is called profiling. The resulting constrained message definition is called a conformance profile (also referred to as a message profile). An example of a constraint is changing optional usage for a data element in the original base standard message definition to required usage in the conformance profile.

IGAMT provides, in a table format user interface, the mechanisms to constrain each data element at each level in the structure definition. The rows of the table list the data elements according to the structure definition being constrained (segments, fields, and data types). The columns list the conformance constructs that can be constrained for a data element, including the binding to a value set. Figure 2 shows a screen capture of the navigation and the segment profiling panels. On the left-hand side, the user can select the object to edit. The right-hand side displays the list of fields in the segment and the requirements that can be specified for the field.

One key philosophy of IGAMT is the capability of creating and reusing building block components. These lower level building blocks can be used to create higher level constructs efficiently. The building blocks include data type flavors, segment flavors, and profile components. A base data type can be constrained for a given use; the resulting data type is called a data type flavor (or data type specialization). A given base data type may have multiple data type flavors. These flavors can be saved in libraries and reused as needed. A similar process applies to creating segment flavors.

A profile component represents a subset of requirements that can be combined with other profiling building blocks. One such example is the definition of a profile for submitting immunizations. The Centers for Disease and Control and Prevention (CDC)

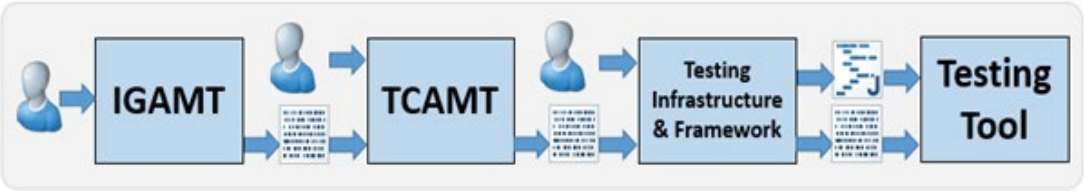


Figure 1: NIST HL7 v2 standards development and testing platform overview.

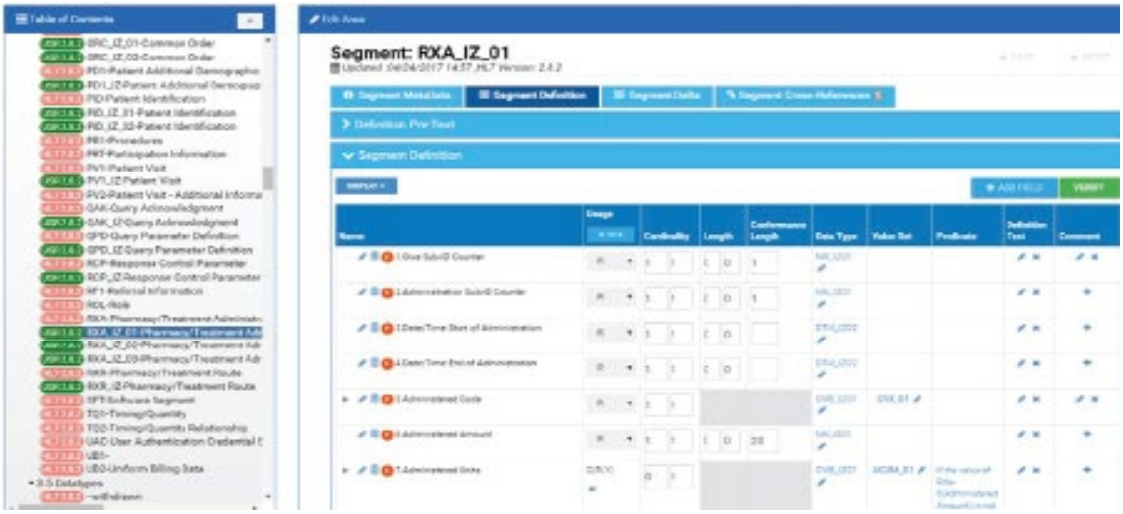


Figure 2: IGAMT screen capture: navigation and segment profiling view.

Snelick, Robert.
"Advancing HL7 v2 to New Heights: A Platform for Developing Specifications, Test Plans, and Testing Tools."
Paper presented at 17th International HL7 Interoperability Conference(IHIC 2017), Athens, Greece. October 19, 2017 - October 24, 2017.

creates a national level profile, however, individual states may have additional local requirements that can be documented in a profile component. Only the delta between the national and local requirements is documented in the profile component. Combining the national level profile and the state profile component yields a complete (composite) profile definition for a given state. Another example is for the case of sending laboratory results and reportable laboratory results to public health. The use cases are very similar. The reportable laboratory results have additional requirements; therefore, a profile should be created for sending laboratory results, followed by a profile component for reportable laboratory results. A composite profile for the public health use case can be created by combining the profile and the profile component. This design principle provides a powerful and effective approach for leveraging existing profiles and profile components [2].

A utility for creating and managing value sets is also provided. Specific value sets can be created and bound to data elements. For example, a base HL7 v2 table can be cloned and modified ("constrained") to create a value set for a specific use, thus enabling more granular value set bindings [2]. Instead of binding an entire HL7 v2 table to an element (typical practice), a value set containing only codes relevant to that element for a particular use is specified. Using this approach, multiple value sets are derived from a single HL7 v2 table, which provides clear requirements for implementers. Mechanisms for creating value set libraries are provided to promote reuse.

2.1 Improved Requirements Specification

In the effort to create conformance test tools for the Office of the National Coordinator (ONC) certification in support of the US Centers for Medicare and Medicaid Services (CMS) Meaningful Use (MU) program, it quickly became apparent that the HL7 v2 specifications named in the ONC rule were ambiguous, under-specified, and inconsistent. This made it difficult to create rigorous, comprehensive, and meaningful test tools and test cases to adequately validate vendor implementations for the ONC stated goal of enabling interoperability. If implementers can interpret and implement requirements in different ways, interoperability is impeded. To improve this situation, NIST worked closely with the specification authors and other stakeholders to gain clarity and subsequently co-published addendums and errata. This effort revealed deficiencies in the mechanisms for specification of requirements and approaches for creating implementation guides. As a remedy, new and improved methods for specifying requirements emerged along with a set of best practices. IGAMT incorporates these methods and encapsulates, automates, and simplifies how the requirements are specified. Table 1 provides a list of the most important methods, concepts, and best practices for improved specifications (beyond current practices).

2.2 IGAMT Message Model and Validation Process

IGAMT has an internal model of all HL7 v2 messages for each version of the standard (Figure 3). HL7 v2 publishes the standard in human readable text documents. Message definitions and accompanying structures are *codified* into a data base, which is available from HL7. IGAMT reads the data base and converts the message definitions into the IGAMT message model. The message model is the anchor on which all IGAMT functions and features are based. IGAMT reveals the model via a graphical user interface (GUI) where the user can constrain the message as needed. The user interface displays panels for the Message, Segment, Data Type, Value Set, Profile Components, Condition Predicates, and Conformance Statements. IGAMT exports the constrained message definition (a profile) as an XML profile instance. IGAMT ensures that the XML profile instance adheres to the rules of the Profile Schema. Validation is performed by validating a message instance against the constraints defined in the XML Profile. The validation engine interprets the requirements as documented in the XML Profile and makes assertions against the message instance accordingly. A Validation Report is generated. The validation process forms the basis of the conformance test tools.

3 TCAMT

TCAMT [5] is a tool used to create HL7 v2.x test plans that contain one or more (typically many) test cases. Key features in TCAMT include test plan creation (narrative and computable), IGAMT XML profile import, HL7 v2 message creation and import, constraint editing, constraint and messaging templates, and multiple export formats. A test case can consist of one or more test steps. A test step can be an HL7 v2.x interaction or a manual step such as visually inspecting the contents of an application's display screen. Each test case and test step can consist of a test description, pre- and post-conditions, objectives, evaluation criteria, and additional notes and comments. Test steps for an HL7 v2.x interaction contain an HL7 v2 message (with specific data) that aligns with the XML conformance profile created from IGAMT[4].

Targeted test cases are critical for assessing the capabilities of a system. TCAMT allows domain experts to create test cases (that include example messages) for certain scenarios and capabilities. Test cases provide context, which expands the scope of testing. Without context, a validation tool cannot test a message exhaustively to all requirements specified in the implementation guide. For example, elements with "required, but may be empty (RE)" usage, elements with "conditional usage (C)", or elements with cardinality greater than "1" cannot be assessed without targeted tests. A message

[4]Not necessarily conformant data; invalid data may be used in the testing process

Table 1: Methods, concepts, and best practices for improved specifications.

Concept	Issue	Feature/Improvement
Explicit Condition Predicates	Conditional usage is specified but lacks conditional statement or an explicit conditional statement	Explicit condition predicate with defined format, style, and pre-defined patterns
Condition Predicate True/ False Outcomes	Limited True/False outcomes for conditional usage (C and CE only)	Full range of true/false outcomes; for example, C(R/ RE) and C(RE/O)
Explicit Conformance Statements	Statements that hinted at being requirements are hidden in narrative sections of the specification	Explicit conformance statements with defined format, style, identification, and pre-defined patterns
Data Type Flavors	Conflated specializations of data type constraints, in-line constraints, un-managed data type flavors	Explicit data type flavor definitions, naming conventions, and style
Data Type Flavor Library	No notion of creating a library of data flavors for reuse by the community at-large	Master set of data type flavors and defined process for user defined flavors; promote consistency and reuse
Segment Flavors	Segments typically are defined to account for requirements for use in more than one message definition—resulting in conflation of requirements	Provide mechanisms to allow specific segment definition via segment flavors, profile components or explicit conformance statements
Profiling Multiple Occurrences	Capability to assign different data type flavors to multiple occurrences to a field element; defined in v2.8	Implemented in IGAMT and in XML profile instance; can vary by "type code", "order", and "one of"
Co-constraints	Missing, inconsistent, or lack of detailed specification of relationship among data element content; typically, in elements OBX-2, OBX-3, and OBX-5	Mechanism to define data element content relationships and dynamic data type flavor mapping for OBX-2 and OBX-5[1]
Value Set Specification	No explicit value set or code table specifications; often the base HL7 or HL7 User table is bound to an element (or elements) with no further constraints	Explicit value set definition creation and value set binding strength
Value Set Profiling	No formal methodology to constrain code systems for specific element binding and use	Explicit value set definition usage indicator for codes and attributes to indicate extensibility and stability
Profile Components	No constructs or methods to define profile building blocks of constraints for reuse	Profile components are introduced to defined a set of arbitrary requirements that when combined with a profile or other profile components create a complete profile (Composite Profile)
Delta Profiles	Complete specifications for closely related use cases	"delta" specifications can be created leveraging the concept of profile components
IG Template	No guidance on what implementation guides should contain	IGAMT incorporates several default templates and export options
Conformance Keywords	Non-existence and inconsistent definition and use of verbs to express requirements	Explicit definition and use of conformance keywords as part of the IG template; based on RFC 2119

[6] For example, based on different codes in OBX-3, different data type flavors of the same base data type can Tbe specified in OBX-2 that indicates the requirements in OBX-5. This enables precise requirements definition.

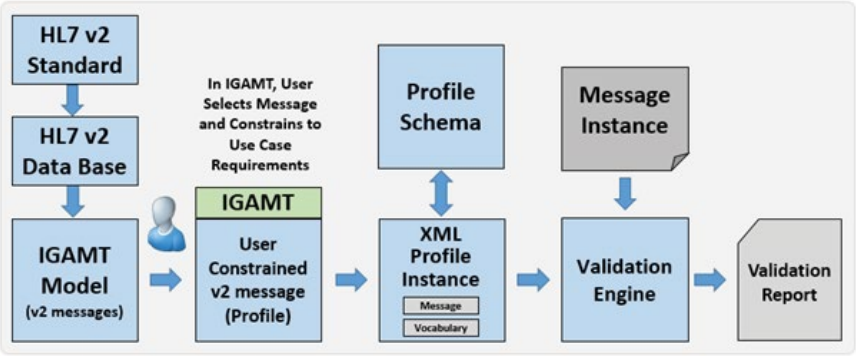


Figure 3: IGAMT message model and validation process.

that is validated against the requirements of a conformance profile without any provided context is called "context-free testing". A message that is validated against the requirements of a conformance profile and with a provided context is called "context-based testing" [2]. The test cases provide context, and TCAMT is a tool that allows users to create the test cases.

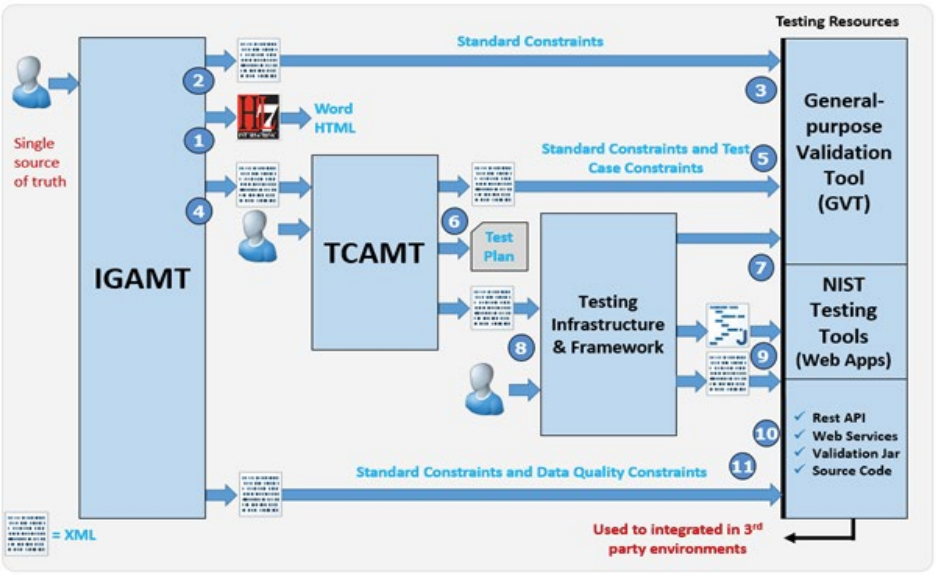


Figure 4: NIST HL7 v2 standards development and testing platform architecture.

A key design component in TCAMT is its use of the XML profiles created in IGAMT as a foundation. The message definition defined in the profile provides the foundation such that data associated with each message element of interest can be specified. TCAMT also allows the user to enter additional assertion indicators based on what they want to test. For example, for an element with a usage of "RE", the user can provide data that are expected to be entered into the sending system for the element and can select an assertion indicator. There are several assertion indicators that could be selected, for example, "presence". In this case, if the user provides test data and selects the indicator of "presence", a constraint is generated by TCAMT and is provided to the validation. For elements with "RE" usage, the element must be supported by the system-under test (SUT), but in a given message instance the element may not be populated. For this construct, the tester wants to ensure that the implementation has, in fact, included support for the element.

In a context-free environment, the absence of data in a message is not a conformance violation for elements with "RE" usage. However, in the example test case described above, data were provided and a presence constraint was specified. Now, when a message created for this test case is validated, the additional constraint triggers an assertion for the presence of data for this element. This method is one way to determine support for the element.

Via TCAMT, the user can create an unlimited number of test cases and test a broad spectrum of requirements. Other constraint indicators can be used to test for specific content or for the non-presence of an element. Additionally, test data can be provided to trigger conditional elements. In other instances, support for certain observations may need to be ascertained. In such cases, test data for specific observations (e.g., in an immunization forecast, the vaccine group, earliest date to give, and due date) can be provided, requiring the message instance to contain an OBX segment for each observation. The test case might be set up to expect certain LOINC[5] codes to ensure each observation (capability) is implemented by the system. TCAMT provides the mechanisms to conveniently and consistently create test cases. Output from TCAMT provides the additional constraints that are interpreted by the validation engine.

4 Testing Infrastructure and Framework

NIST has built an HL7 v2.x testing infrastructure and framework to aid in the process of creating conformance testing tools. The testing infrastructure provides a set of services utilized by the test tool framework to build specific instances of tools. A test tool can be built for a specific need or to be a general-purpose tool to handle multiple implementation guides and profiles. The latter tool is a web application where a user can upload implementation guides, conformance profiles, and test plans to "create" a test tool. The test tool is "built-on-the-fly" and can be generated as a by-product "for free" once the XML profile and associated artifacts have been created (in IGAMT and TCAMT). This process allows domain experts to "build" the test tool. Alternatively, the framework can be leveraged, customized, and installed locally. Using the framework, developers can choose to create customized, specific, or general-purpose

[5]Logical Observation Identifiers Names and Codes

web application conformance test tools, and they can access the validation via web services or incorporate validation via a JAR (Java Archive) file or source code. Regardless of the use, the platform can significantly improve the quality of implementation guides, assist in the creation and maintenance of test plans, expedite the stand-up of a validation tool, and, overall, reduce the cost and time of the entire process.

Figure 4 shows in more detail the end-to-end methodology and platform. A key design principle is that there is a single source of truth in the creation of implementation guides and test plans. Modifications are made in one place and are propagated to associated services, utilities, and tools. IGAMT is a tool used by domain expert authors to define requirements for interface specifications. Human readable (1) and machine computable (2) artifacts are exported. A context-free conformance test tool is automatically generated when the IGAMT XML profiles are loaded in the general-purpose validation tool (3). At this level, validation is based on the technical requirements defined in the profile. No context is associated when validating the message instance against the requirements defined in the profile. This type of validation is called context-free testing.

Point (4) shows the XML Profile as input into TCAMT. Test scenarios provide a context, that is, a real-world story with associated data. Additional constraints are generated from having context. The profile and context constraints are loaded into the general-purpose validation tool to create a context-based validation tool automatically (5). Point (6) indicates a human readable export of the Test Plan.

Point (7) indicates that the testing infrastructure and framework components are used as the basis for the general-purpose validation tool. The general-purpose validation tool is itself a tool that takes as input the resource bundle (XML Profile, TCAMT constraint file, etc.) to automatically generate a conformance test tool. Points (8) and (9) indicate the process by which developers can leverage the testing infrastructure and framework to create customized conformance test tools. Point (10) indicates that validation can be accessed via other methods that allow a user to integrate it into their local environments. The platform provides access to the tool validation via REST and web services. Additionally, the validation JAR and source code are available. Point (11) indicates that additional constraints can also be included that go beyond the scope of typical interface requirements. These can include data quality business rules, for example, ensuring that a vaccine dose reported is consistent in terms of the manufacturer, lot number, and date given. More on this topic is given in Section 6.

5 Conformance Test Tools

As shown, conformance testing tools are built using the testing infrastructure and framework, the IGAMT-produced conformance profiles, and the TCAMT-produced test plan. Testing tools are web-based applications that can support both context-free and context-based validation [5]. In addition to performing message validation, the tools provide a browse-able view of the requirements for each conformance profile. In the context-based mode, the test story, test data, and an example message are provided for each test step.

In the context-free mode, the user simply selects the conformance profile to validate against and then imports the message. The validation is performed automatically and a report is given. In the context-based mode, the user selects the test step and imports the message to validate. The test tool sets the validation to the conformance profile linked to the test step, performs the validation, and provides a report. In both modes, a tree structure of the message is shown on the left panel of the validation screen and can be used to inspect the content of individual data elements.

Test plans can be executed in non-transport mode and transport mode. Non-transport mode provides an interface to upload (cut/paste or load file) a message into the validation edit box. Transport mode allows an application to connect to the test tool to exchange messages interactively. The test tool can act as an initiator or responder as directed by the test plan. Various transport protocols are supported including MLLP and SOAP. Test Cases can also include manual test steps in addition to automated test steps that contain an HL7 v2 message exchange.

6 Requirements beyond the Interface Specification

The intent of HL7 v2 is specifically scoped to defined requirements for exchanging data between applications. The specifications typically do not impose any requirements on how the data are processed. Other specifications, in conjunction with the interface specification may specify such requirements (e.g., IHE integration profiles and functional requirements specifications). In real world settings, exchange partners need to account for more than just conformance to the exchange requirements. Data quality, business rules, and functional requirements are necessary to satisfy the desired outcome of the use case scenario. Mechanisms to define such requirements, and testing support that can verify that the complete workflow is implemented as intended, are beneficial.

The generic constraint generation utility in IGAMT can be used to create data quality constraints. Certain business rules can be applied to a message to determine if it meets the requirements necessary for incorporation by the receiver. A simple data quality rule for reporting an immunization record is that the date of administration must be after the date of birth. This constraint likely is never given in an HL7 v2 interface specification, however, data quality rules such as these are important at the local level.

Users can create these rules in IGAMT that provide additional validation (point (11) in Figure 4).

TCAMT can be used to create test cases to test functional requirements. For example, a scenario can be crafted in which three different immunization records for the same patient are created from different providers and sent to an immunization information system (IIS). A subsequent query to the IIS to return a complete immunization history can be performed. The response message can be examined to see if the consolidated record contains the expected combined immunization history. TCAMT provides the capability to create such a scenario and the additional content validation constraints. Testing for invalid (or negative) test case scenarios also can be created. The platform provides the capabilities for the tester to create unlimited test scenarios using convenient and powerful tooling.

7 Conclusion

We presented an end-to-end methodology and platform for developing standards, writing test plans, and creating testing tools in the HL7 v2 technology space. The platform includes three key foundational components: (1) a tool to create implementation guides and conformance profiles; (2) a tool to create test plans, test cases, and associated test data; and (3) a testing infrastructure and test framework to build testing tools. Requirements are captured in IGAMT and exported as conformance profiles. TCAMT is used to create a set of test cases based on the conformance profiles. A conformance test tool is created by combining the validation and associated artifacts with the testing infrastructure and framework.

Acknowledgements

I'd like to thank the NIST development and analysis team: H. Affo, W. Jung Yub, H. Tamri, I. Mellouli, A. El Ouakili, C. Rosin, N. Crouzier, M. Lefort, and S. Martinez; reviewers S. Taylor and F. Oemig, and finally C. Newman for providing user feedback on the tool suite.

References

[1] Health Level Seven International, Inc. Health Level 7 (HL7) Standard Version 2.7, ANSI/HL7, January, 2011, http://www.hl7.org.

[2] Oemig F, Snelick R. Healthcare Interoperability Standards Compliance Handbook. Basel: Springer International Publishing Switzerland; 2016. ISBN 978-3-319-44837-4.

[3] Snelick R, Oemig F. Principles for Profiling Healthcare Data Communication Standards. Software Engineering Research and Practice (SERP13), WORLDCOMP'13 July 22-25, 2013, Las Vegas, NV, 2013.

[4] Health Level Seven International, Inc. HL7 Version 2.5.1 Implementation Guide for Immunization Messaging; Release 1.5, October 1, 2014. http://www.cdc.gov/vaccines/programs/iis/technical-guidance/downloads/hl7guide-1-5-2014-11.pdf

[5] National Institute for Standards and Technology. NIST Resources and Tools in Support of HL7 v2 Standards. http://hl7v2tools.nist.gov/

[6] Integrating the Healthcare Enterprise (IHE). http://www.ihe.net

Development of A New Atom-Based SI Traceable Electric-Field Metrology Technique

Christopher L. Holloway, Matt T. Simons, and Joshua A. Gordon
RF Technology Division
National Institute of Standards and Technology (NIST), Boulder, CO 80305
holloway@boulder.nist.gov : 303-497-6184

***Abstract*—We are developing a fundamentally new atom-based approach for electric (E) field metrology. This technique has the capability of becoming a new international standard for E-field measurements and calibrations. Since this new approach is based on atomic transitions of alkali atoms (mainly caesium and rubidium atoms), the probe is self-calibrating and has a capability of performing measurements over a large bandwidth (from 10's MHz to the THz range). This new approach will lead to a self-calibrated, SI traceable, E-field measurement, and has the capability to perform measurements on a fine spatial resolution in both the far-field and near-field. We will report on the development of this new metrology approach, including the first fiber-coupled vapor-cell for E-field measurements, which allows for easier and more flexible measurements. We discuss key applications, including self-calibrated measurements, millimeter-wave and sub-THz measurements, field mapping, and sub-wavelength and near-field imaging. We show results for free-space measurements of E-fields, for measuring the E-field distribution along the surface of a circuit board, and for measuring the directivity pattern of a horn antenna.**

I. Introduction

One of the keys to developing new technologies is to have sound metrology tools and techniques. Whenever possible, we would like these metrology techniques to make absolute measurements of a physical quantity. Preferably, we would like to make measurements directly traceable to the International System of Units (SI). Measurements based on atoms provide such a direct SI traceability path and enable absolute measurements of physical quantities. Atom-based measurements have been used for several years; most notable are time (s), frequency (Hz), and length (m). We would like to extend these atom-based techniques to other physical quantities, including electric (E) fields.

We are developing a fundamentally new atom-based approach that will lead to a self-calibrated, SI traceable E-field measurement that has the capability to perform measurements on a fine spatial resolution in both the far-field and near-field [1]-[9]. This new approach is significantly different from currently used field measurement techniques in that it is based on the interaction of radio-frequency (RF) E-fields with Rydberg atoms (alkali atoms placed in a glass vapor-cell that are excited optically to Rydberg states). The Rydberg atoms act like an RF-to-optical transducer, converting an RF E-field strength to an optical-frequency response. In this new approach, we employ the phenomena of electromagnetically induced transparency (EIT) and Autler-Townes splitting [1]-[3], [10]-[13]. This splitting is easily measured and is directly proportional to the applied RF E-field amplitude and results in an absolute SI traceable measurement. The technique is very broadband allowing self-calibrated measurements over a large frequency band including 500 MHz to 500 GHz (and possibly up to 1 THz and down to 10's of megahertz). Various other benefits of the new approach are listed in [1] and [2].

Besides having a self-calibrating, SI-traceable probe, various other applications are possible, including millimeter-wave and sub-THz measurements, field mapping and sub-wavelength imaging, and far-field and near-field measurements. This technique is demonstrated by showing free-space measurements of E-fields, measurements of the E-field distribution along the surface of a circuit board, and measurements of the directivity pattern of a horn antenna. Some of the measurements presented here are performed with a new fiber-coupled probe. This fiber-coupled design allows for a moveable form-factor probe which makes measurements easier and more flexible for various applications.

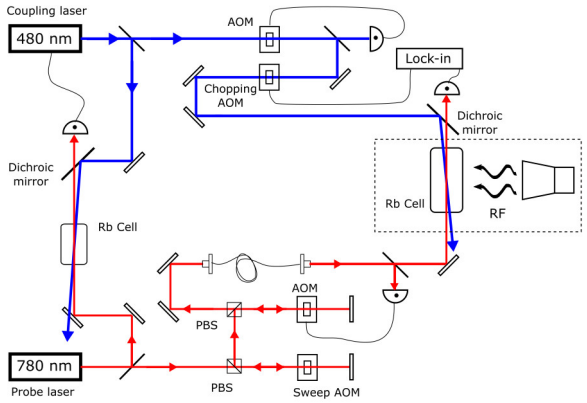
II. Description of Technique

The basic concept of this measurement approach uses a vapor of alkali atoms (placed in a glass cell, referred to as a "vapor" cell) as the active medium for the radio frequency (RF) E-field measurement. The basic concept is that by manipulating alkali atoms with both optical (laser) fields and RF fields, it is possible to cause a laser to transmit through a vapor cell where it would normally be absorbed by the atoms in the vapor cell. Rubidium (^{85}Rb) and cesium (^{133}Cs) are the two atomic species that are typically used in the approach.

A typical measurement setup is shown in Fig. 1. This measurement approach can be represented by the four-level atomic system shown in Fig. 2, see [2], [3], [14] for details. In effect, the "probe" laser is used to probe the response of the ground-state transition of the atoms (level 1 to level 2 in Fig. 2), and a second laser ("coupling" laser) is used to excite the atoms to a high energy Rydberg state (level 3 in Fig. 2). In the presence of the coupling laser, a destructive quantum interference occurs and the atoms become transparent to the resonant probe laser (this is the concept of EIT). A transparency window is opened for the probe laser light: probe light transmission is increased. The coupling laser wavelength

Gordon, Joshua; Holloway, Christopher; Simons, Matthew.
"Development of A New Atom-Based SI Traceable Electric-Field Metrology Technique."
Paper presented at Antenna Measurements Techniques Association: 39th Annual Symposium, Atlanta, GA, United States. October 15, 2017 - October 20, 2017.

(a) photos of setup



(b) block diagram

Fig. 1. Experimental setup for E-field measurements using EIT: a) photo of of the setup and (b) block diagram, polarizing beam splitter (PBS) and acousto-optic modulator (AOM).

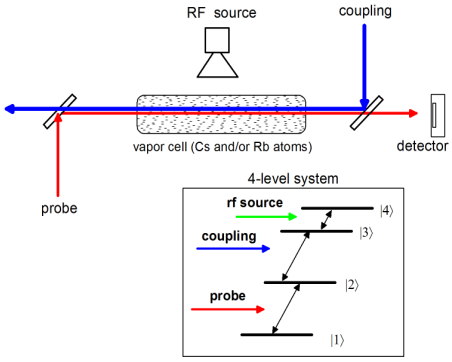


Fig. 2. Illustration of a four-level system, and the vapor cell setup for measuring EIT, with counter-propagating probe and coupling beams. The RF is applied transverse to the optical beam propagation in the vapor cell.

is chosen such that the atom is in a sufficiently high state (a Rydberg state) such that a RF field couples two Rydberg states (levels 3 and 4 in Fig. 2).

A detailed explanation both from an atomic physics viewpoint and experimental approach is given in [1] and [2]. Experimentally, the approach is explained as follows: If the probe laser is tuned to a ground state transition of alkali atoms in a vapor cell (levels 1 and 2 in Fig. 2), after propagation through the vapor cell the atoms will absorb the light and little power will be detected. The power measured on the detector when the laser is scanned across this wavelength is shown in the bottom curve in Fig. 3(a) ($\Delta p = \omega_o - \omega_p$; ω_o is the on-resonance angular frequency of the ground state transition and ω_p is the angular frequency of the probe laser.) This is the typical signal one obtains when performing atomic spectroscopy experiments (i.e., the classical Doppler profile). The minimum in the curve indicates the resonance frequency of the ground state transition of the alkali atom. When the coupling laser is allowed to propagate through the cell (the coupling laser is counter-propagating on top of the probe laser) an interference between the two atomic states occurs, hence allowing the probe laser to pass through the vapor cell with less absorption (an increase in the probe laser transmission). This is the concept of EIT, i.e., a medium that was normally absorbing becomes transparent with the presence of the coupling laser. This is shown in the top curve in Fig. 3(a) (note the wings of all three curves normally would lay on top of one another, but they are shifted here for ease of viewing). Notice at $\Delta_p = 0$, the power on the detector is larger than the Doppler background, i.e., the global inverted bell-shaped behavior. The wavelength of the coupling laser is chosen judiciously such that the atoms are excited to a very high energy, where an RF source is at a resonant frequency that causes an atomic transition to a nearby state (i.e., an RF atomic transition). When the RF source is turned on, the EIT signal splits into two (this splitting is called Autler-Townes (AT) splitting), see the middle curve in Fig. 3(a). The EIT signal and the splitting can be weak at times. To increase the EIT signal-to-noise, we modulate the coupling-laser amplitude with a 50/50 duty-cycle 30 kHz square wave and detect any resulting modulation of the probe transmission with a lock-in amplifier. This removes the Doppler background and isolates the EIT signal. Fig. 3(b) shows a typical EIT signal from the lock-in amplifier. The splitting of the EIT peak is indicated by Δf_m.

This splitting (Δf_m) of the probe laser spectrum is easily measured and is directly proportional to the applied RF E-field amplitude. Once this Δf_m is measured, the RF E-field strength is obtained by [2], [3], [4]:

$$|E| = 2\pi \frac{\hbar}{\wp} \frac{\lambda_p}{\lambda_c} \Delta f_m = 2\pi \frac{\hbar}{\wp} \Delta f_o \ , \qquad (1)$$

where $\hbar$ is Planck's constant, $\wp$ is the atomic dipole moment of the RF atomic transition (see [2]), $\Delta f_o = \frac{\lambda_p}{\lambda_c} \Delta f_m$, and λ_p and λ_c are the wavelengths of the probe and coupling laser, respectively. The λ_p/λ_c ratio is needed to account for the Doppler mismatch of the probe and coupling lasers [10], when the probe laser is scanned during the experiments. One can also scan the coupling laser and not the probe laser during the experiments. If the coupling laser is scanned, it is not required to correct for the Doppler mismatch, and λ_p/λ_c ratio is not needed, see [14] for details. In this case, $\Delta f_o = \Delta f_m$.

Gordon, Joshua; Holloway, Christopher; Simons, Matthew.
"Development of A New Atom-Based SI Traceable Electric-Field Metrology Technique."
Paper presented at Antenna Measurements Techniques Association: 39th Annual Symposium, Atlanta, GA, United States. October 15, 2017 - October 20, 2017.

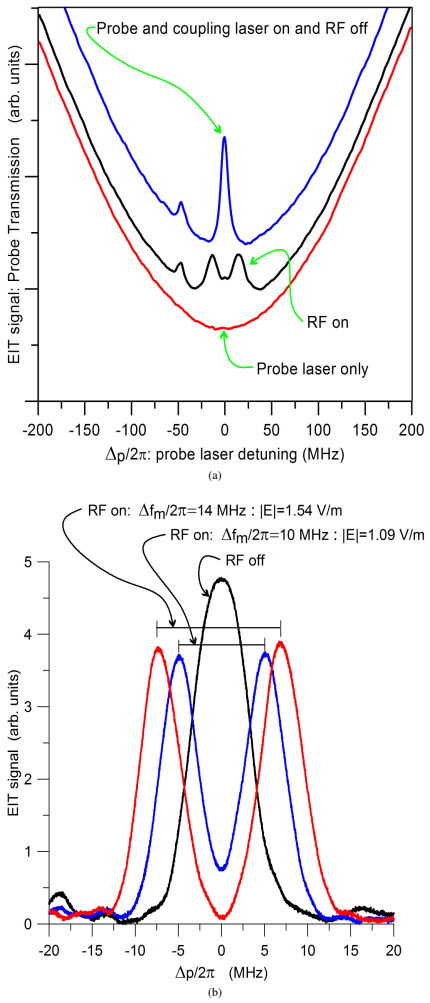


Fig. 3. EIT illustration: (a) with the Doppler ground and (b) after the lock-in is used. These experiments where performed with a vapor cell filled with ^{133}Cs and with a RF source at 9.22 GHz. The RF source couples Rydberg states $43D_{5/2}$-$44P_{3/2}$ for the E-field measurement.

We consider this type of measurement of the E-field strength a direct SI-traceable, self-calibrated measurement in that it is related to Planck's constant (which will become an SI-defined quantity by standard bodies in the near future) and only requires a frequency measurement (Δf_m, which is quantum linked and can be measured very accurately). The one unknown in the expression is the atomic dipole moment $\wp$ which can be calculated accurately, see [2], [15], [16].

Fig. 3(b) shows the measured EIT signal for three different RF incident field strengths (0 V/m, 1.09 V/m and 1.54 V/m). In order to estimate the E-field strength, we first measured Δf_m, then we used eq. (1) to determine $|E|$.

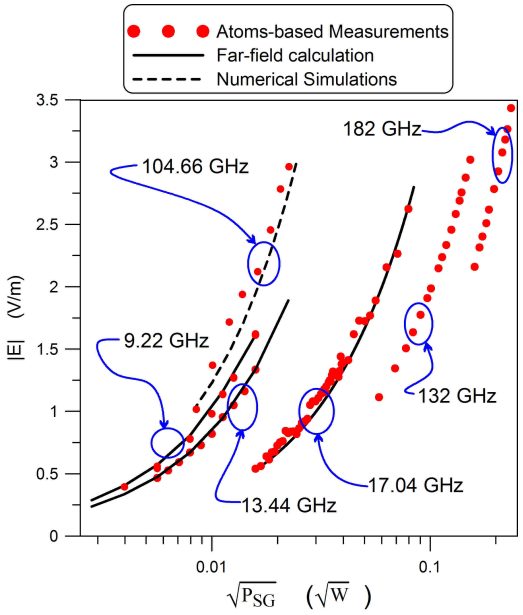


Fig. 4. A comparison of the measured E-field (obtained from the atom-based approach) to results obtained from far-field calculations and from a full-wave numerical simulation. Comparing experimental atom-based data to both numerical simulations and to far-field calculations for various frequencies from 9 GHz to 182 GHz helps to validate this technique. PSG is the signal-generator power level feeding the antennas (through either a cable or a waveguide). Note that a log scale is used because the data sets cover different $\sqrt{P_{SG}}$ ranges.

III. Experimental Results

A. Far-field Comparisons

Utilizing the experiment setup shown in Fig. 1 we can measure the E-field strength in the far-field at various frequencies. Fig. 4 shows measurement for six different frequencies ranging from 9.22 GHz to 182 GHz using two different atomic species (^{85}Rb and ^{133}Cs). In this figure we have compared the estimated E-field obtained for this atom-based approach to both far-field calculations and to numerical simulations. We see that the atom-based approach correlates very well to both the far-field calculations and to numerical simulations.

These results show the wide bandwidth measurement capacity of this approach. With one experimental setup, it is possible to measure E-field strength from 10's MHz into the THz frequency range. Measurements above 110 GHz have been demonstrated here and in [1], [5]. The possibility of performing calibrated measurements above 110 GHz is one of the interesting and possibly, one of the major benefits of this new technique, since it can provide calibration above 110 GHz (which is currently not possible).

B. Movable Probe: Fiber Coupled Probe

While various international metrology organizations groups around the world are beginning to investigate this new approach as a possible new international standard for E-field

Gordon, Joshua; Holloway, Christopher; Simons, Matthew.
"Development of A New Atom-Based SI Traceable Electric-Field Metrology Technique."
Paper presented at Antenna Measurements Techniques Association: 39th Annual Symposium, Atlanta, GA, United States. October 15, 2017 - October 20, 2017.

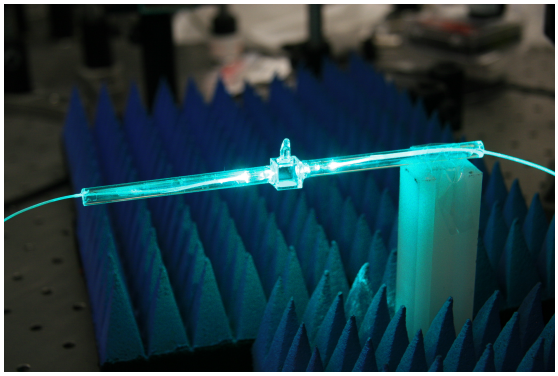


Fig. 5. Photo of first fiber-coupled vapor cell probe for self-calibrated E-field measurements over a large frequency band including 500 MHz to 500 GHz (and possibly up to 1 THz and down to tens of megahertz).

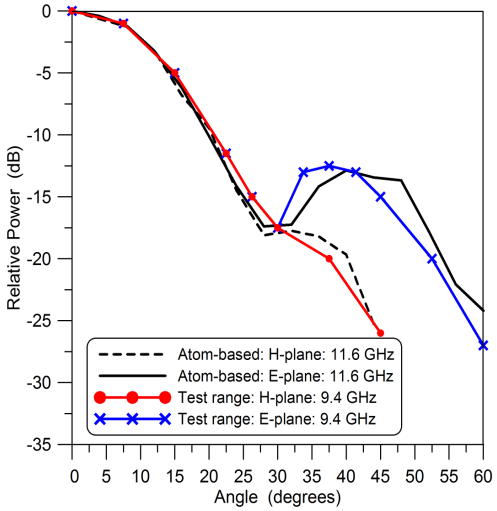


Fig. 6. Fiber-coupled probe measurement for the E-plane and H-plane antenna pattern for a Narda 640 standard gain horn at 11.6 GHz. Also shown are measured results obtained from an antenna range [17] at 9.4 GHz.

measurements and calibrations, all these investigations and measurements have been confined to an optical table. This confinement is a result of the fact that this technique requires the two lasers (probe and coupling laser) to overlap inside the vapor cell. In order to overcome these issues, we have developed the first fiber-coupled vapor cell, where the counter propagating probe and pump fields are overlapped inside the vapor cell while it is moved off the optical bench. Moving the probe off the optical table allows measurements to be performed in free space, and in other standard RF metrology environments. The new probe consists of a 10-mm cubic vapor cell filled with ^{133}Cs and two optical fibers with lenses attached with UV curing epoxy at either end, made with all dielectric material (see Fig. 5). We have performed various types of measurement in order to illustrate its capability.

1) Antenna Pattern Measurements: We used the fiber-coupled probe to measure the antenna pattern for a Narda 640 standard gain horn antenna (mentioning this product does not imply an endorsement, but serves to clarify the antenna used). In these measurements, the fiber-coupled probe was placed in the far-field of the horn antenna and scanned from bore-site to an angle of 60^o. The horn antenna was scanned in both E-plane and H-plane. Fig. 6 shows the measured antenna patterns for both the E-plane and H-plane at 11.6 GHz. Also shown in this figure are results obtained in an anechoic chamber test range [17] at 9.4 GHz. Good correlation between the two types of measurements is seen. The deviations to the two sets of measurements is due to the fact that our measurement were perform in a laboratory with no RF absorber on the walls and the laboratory had several objects in the room. Thus, our results suffer from some background scattering.

2) Near-Field Imaging: In order to illustrate the near-field imaging capability of the fiber-coupled probe, we imaged the E-field at various heights across the surface of a co-planar waveguide (CPW) line. The CPW has a center strip of 3 mm, gaps of 2 mm, and a substrate ($\epsilon_r \approx 3.5$) of thickness 1.52 mm. Fig. 7 shows the scans at six different heights for a frequency of 11.6 GHz. In order to show the repeatability of this probe we performed three sets of measurements for each height and the error-bars represent all these measurements. These results show the capability for near-field imaging and field-mapping across the surface of printed circuit board structures, which will be used in the future to support calibrated on-wafer measurements of high-speed (high-frequency) integrated circuits. The fiber-coupled probe allows for much finer spatial resolution than is possible with current E-field probes.

IV. Measurement Uncertainties

Knowing the uncertainties of this technique is an important step when establishing a new international measurement standard for an E-field strength and is a necessary step for this method to be accepted as a standard calibration technique. The uncertainties can be grouped into two different categories: (a) quantum based uncertainties and (b) RF based uncertainties. These include, 1) the validity of eq. (1), 2) the accuracy of the atomic dipole moment calculation, and 3) perturbation of the RF field due to internal resonances inside the vapor cell, just to name a few. These three and other various types of uncertainties for this atom-based approach are currently being investigated [1], [3], [7], [14], [18], [19]. The uncertainties of this new atom-based technique can be controlled and reduced to be less than the current E-field measurement uncertainties. In the near future we will develop detailed uncertainties for this new approach.

V. Conclusion

We discussed a fundamentally new atom-based approach for E-field metrology. Several international metrology organizations around the world are beginning to investigate this new approach as a possible new international standard for E-field measurements and calibrations. In this paper we have presented some key examples that show the benefits

Gordon, Joshua; Holloway, Christopher; Simons, Matthew.
"Development of A New Atom-Based SI Traceable Electric-Field Metrology Technique."
Paper presented at Antenna Measurements Techniques Association: 39th Annual Symposium, Atlanta, GA, United States. October 15, 2017 - October 20, 2017.

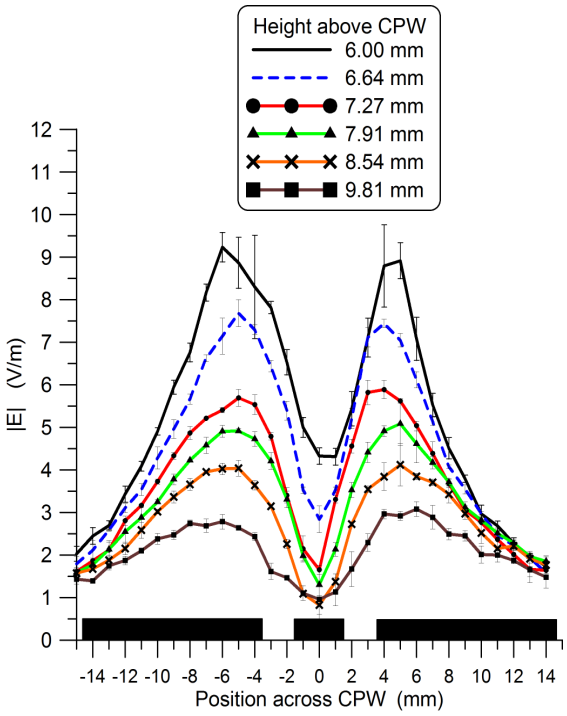


Fig. 7. Measured $|E|$-field distribution across the surface of the CPW line at different heights for 11.6 GHz obtained with the fiber-coupled probe. We also show the CPW geometry in order to illustrate the gap locations.

of the new approach and that show the potential for a new international measurement standard. We have shown results for far-field measurement, results for the broadband nature of this technique, results of antenna pattern measurements, and results of near-field imaging. The fiber-coupled probe is one important advancement of this approach. That is, being able to move the probe off the optical table is an important step when establishing a new international measurement standard for an E-field strength and is a necessary step for this method to be useful and ultimately accepted as a standard calibration technique.

REFERENCES

[1] C.L. Holloway, M.T. Simons, J.A. Gordon, P.F. Wilson, C.M. Cooke, D.A. Anderson, and G. Raithel, "Atom-Based RF Electric Field Metrology: From Self-Calibrated Measurements to Sub-Wavelength and Near-Field Imaging", *IEEE Trans. on Electromagnetic Compat.,* vol. 59, no. 2, 717-728, 2017.

[2] C.L. Holloway, J.A. Gordon, A. Schwarzkopf, D. A. Anderson, S. A. Miller, N. Thaicharoen, and G. Raithel, "Broadband Rydberg Atom-Based Electric-Field Probe for SI-Traceable, Self-Calibrated Measurements," *IEEE Trans. on Antenna and Propag.*, vol. 62, no. 12, 6169-6182, 2014.

[3] J.A. Sedlacek, A. Schwettmann, H. Kübler, R. Löw, T. Pfau and J. P. Shaffer, 'Microwave electrometry with Rydberg atoms in a vapor cell using bright atomic resonances", *Nature Phys.*, vol. 8, 819, 2012.

[4] C.L. Holloway, J.A. Gordon, A. Schwarzkopf, D.A. Anderson, S.A. Miller, N. Thaicharoen, and G. Raithel, "Sub-wavelength imaging and field mapping via electromagnetically induced transparency and Autler-Townes splitting in Rydberg atoms," *Applied Phys. Lett.,* vol. 105, 244102, 2014.

[5] J.A. Gordon, C.L. Holloway, A. Schwarzkopf, D.A. Anderson, S.A. Miller, N. Thaicharoen, and G. Raithel, "Millimeter-wave detection via Autler-Townes splitting in rubidium Rydberg atoms", *Applied Phys. Lett.*, vol. 105, 024104, 2014.

[6] J.A. Sedlacek, A. Schwettmann, H. Kübler, and J.P. Shaffer, "Atom-based vector microwave electrometry using rubidium Rydberg atoms in a vapor cell," *Phys. Rev. Lett.*, vol. 111, 063001, 2013.

[7] H. Fan, S. Kumar, J. Sedlacek, H. Kübler, S. Karimkashi, and J.P. Shaffer, 'Atom based RF electric field sensing," *J. of Phys. B: Atomic, Molecular and Optical Physics*, vol. 48, 202001, 2015.

[8] M. Tanasittikosol, J.D. Pritchard, D. Maxwell, A. Gauguet, K.J. Weatherill, R.M. Potvliege and C.S. Adams, "Microwave dressing of Rydberg dark states," *J. Phys B*, vol. 44, 184020, 2011.

[9] C.G. Wade, N. Sibalic, N.R. de Melo, J.M. Kondo, C.S. Adams, and K.J. Weatherill, "Real-Time Near-Field Terahertz Imaging with Atomic Optical Fluorescence," *Nature Photonics,* **11**, 40-43, 2017.

[10] A.K. Mohapatra, T.R. Jackson, and C.S. Adams, "Coherent optical detection of highly excited Rydberg states using electromagnetically induced transparency," *Phys. Rev. Lett.*, vol. 98, 113003, 2007.

[11] M. Fleischhauer, A. Imamoglu, and J.P. Marangos, "Electromagnetically induced transparency: Optics in coherent media," *Reviews Modern Physics*, vol. 77, pp. 633-673, April, 2005.

[12] K.J. Boller, A. Imamolu, and S.E. Harris, "Observation of electromagnetically induced transparency," *Phys. Rev. Lett.*, vol. 66, no. 20, pp. 2593-2596, May, 1991.

[13] S.H. Aulter and C.H. Townes, "Stark Effect in Rapidly Varying Fields," *Phys. Rev.*, vol. 100, 703-722, October, 1955.

[14] C.L. Holloway, M.A. Simons, J.A. Gordon, A. Dienstfrey, D.A. Anderson, and G. Raithel, "Electric Field Metrology for SI Traceability: Systematic Measurement Uncertainties in Electromagnetically Induced Transparency in Atomic Vapor", *J. of Applied Physics*, vol. 121, 233106, 2017.

[15] I.I. Sobelman, *Atomic Spectra and Radiative Transitions*, (Second Edition): Springer, 1992.

[16] M.A. Simons, J.A. Gordon, and C.L. Holloway, "Simultaneous Use of Cs and Rb Rydberg Atoms for Dipole Moment Assessment and RF Electric Field Measurements via Electromagnetically Induced Transparency", *J. Appl. Phys.*, vol. 102, 123103, 2016.

[17] T.J. Duck, B. Firanski, F.D. Lind, and D. Sipler, "Aircraft-protection radar for use with atmospheric lidars", *Applied Optics*, vol. 44, no. 23, pp. 4937-4945, 2005.

[18] H. Fan, S. Kumar, J. Sheng, J.P. Shaffer, C.L. Holloway and J.A. Gordon, "Effect of Vapor Cell Geometry on Rydberg Atom-based Radio-frequency Electric Field Measurements", *Physical Review Applied*, vol. 4, 044015, November, 2015.

[19] C.L. Holloway, J.A. Gordon, M.T. Simons, H. Fan, S. Kumar, J.P. Shaffer, D.A. Anderson, A. Schwarzkopf, S. A. Miller, N. Thaicharoen, and G. Raithel, "Atom-based RF electric field measurements: an initial investigation of the measurement uncertainties," in Proc. of *Joint IEEE Intern. Symp. on EM and EMC Europe*, Dresden, Germany, pp. 467-472, Aug. 2015.

Gordon, Joshua; Holloway, Christopher; Simons, Matthew.
"Development of A New Atom-Based SI Traceable Electric-Field Metrology Technique."
Paper presented at Antenna Measurements Techniques Association: 39th Annual Symposium, Atlanta, GA, United States. October 15, 2017 - October 20, 2017.

Enhanced Transmission Algorithm for Dynamic Device-to-Device Direct Discovery

Aziza Ben Mosbah, David Griffith and Richard Rouil
National Institute of Standards and Technology, Gaithersburg, Maryland, USA
{*aziza.benmosbah, david.griffith, richard.rouil*}*@nist.gov*

***Abstract*—In order to support the increasing demand for capacity in cellular networks, Long Term Evolution (LTE) introduced Proximity Services (ProSe) enabling Device-to-Device (D2D) communications, defining several services to support such networks. We are interested in the performance in out-of-coverage scenarios of one of these services: direct discovery. As defined in the standard, network and configuration parameters for direct discovery are predefined and do not change over time, which creates an inability to adjust to variations in topologies, number of operating devices, and/or users' mobility during the discovery process. In this paper we propose an enhanced discovery algorithm that, building on previous works, allows users to adapt to potential variations in the discovery group, using optimized transmission probabilities and transmission success probabilities. The performance of this algorithm is evaluated, and we demonstrate gains in the accuracy of the discovery information, and in the time required for discovery.**

***Index Terms*—Long Term Evolution (LTE), Device-to-Device (D2D), D2D Discovery, Proximity Services (ProSe), Simulations, Performance, Algorithm**

I. Introduction

Long Term Evolution (LTE) cellular networks rely on infrastructure nodes, such as Evolved Node B (eNB), Mobility Management Entity (MME), Serving Gateway (SGW), and PDN (Packet Data Network) Gateway (PGW) to manage the communication and network access by the users. This architecture simplifies the administration of the resources and allows for an accurate understanding of the status of the network as a whole by the entities granting access. However, this means that coverage and service quality are dependent on the existence of supporting infrastructure. In order to increase coverage, provide service in areas without access to infrastructure, and improve the quality of service in saturated areas, the Third Generation Partnership Project (3GPP) introduced Proximity Services (ProSe) using different mechanisms (discovery, synchronization, direct communication) to allow devices to communicate directly [1]. Several different operating modes have been defined to account for situations where User Equipment (UE) have access to infrastructure that will arbitrate the in-coverage or out-of-coverage communication, thus allowing the UEs to select the resources used for communication themselves. In the out-of-coverage case, the parameters that the UEs shall use for communicating (e.g. the number of physical resources to use, the length of the period, etc.) are preconfigured in the devices and are not modified during operation, meaning that the devices can not adapt to the actual network conditions to make an efficient use of the available resources.

In this paper, we make the UEs aware of the network conditions (e.g. number of UEs) using the messages from the discovery service. This in turn allows us to improve the use of resources and reduce the time required to complete the discovery of all the UEs in the group. The proposed algorithm allows UEs performing discovery to detect the presence and the withdrawal of other UEs in the discovery group.

This rest of the paper is organized as follows. In Section II, we discuss the related work in D2D discovery. In Section III, we present our proposed transmission algorithm that takes into account success probabilities and recognizes both UE arrivals and departures. Performance evaluation and simulation results are described in Section IV. Finally, we conclude our work in Section V.

II. Related Work

Existing research on D2D Discovery has focused on the modeling and performance of network assisted discovery (that is, D2D discovery for in-coverage scenarios, where the eNB controls the process). For example Madhusudhan et al. study the performance in terms of throughput of network-assisted discovery in [2]. Xenakis et al. [3] study and provide analytical models for the number of UEs and their deployment in a group for discovery to perform optimally. Similarly, Chour et al. in [4] offload the discovery process from the LTE UEs to Vehicular Ad-hoc Network (VANET) nodes (like roadside units), and Albasry and Ahmed in [5] propose power control strategies to minimize interference and noise.

Regarding the D2D discovery process without network intervention (D2D direct discovery), we can find some works in the literature exploring the architecture design: Sharmila et al. in [6] propose an alternate framework to that of 3GPP's that extends the services available for the UEs, and Murzak et al. in [7] look into the potential of direct discovery for interconnecting LTE and 5G networks.

There has been some work on optimizing the D2D direct discovery, in particular the work by Griffith and Lyons [8]. The authors computed the optimal value of the discovery message transmission probability that minimizes the mean number of periods required for all members of a group of UEs to discover each other. Based on this work, an adaptive algorithm is proposed in [9]. The discovery process in LTE D2D out-of-coverage scenarios is improved by dynamically

Ben Mosbah, Aziza; Griffith, David; Rouil, Richard.
"Enhanced Transmission Algorithm for Dynamic Device-to-Device Direct Discovery."
Paper presented at 2018 IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, United States. January 12, 2018 - January 15, 2018.

adjusting the transmission probability to the optimal value as defined in [8]. Therefore, the algorithm gives the UEs the ability to change their transmission probabilities as needed to reduce the time required to discover other UEs. However, that algorithm is able to detect UEs joining the group at any time of the discovery process, but it does not take into account UEs leaving. In this paper, we enhance that algorithm using the probability of a message reception in a given time interval to learn how long a UE should wait before assuming that another UE has left the group, enabling the devices to fully adapt to dynamic scenarios. To the best of our knowledge, this is the only research available that allows the UEs to learn and adapt the size of the discovery group over time, and adapt the transmission parameters accordingly.

III. Enhanced Transmission Algorithm

In Table I, we provide a list of symbols we use in this paper.

TABLE I: List of Symbols

Symbol	Definition
N_f	Number of resource block pairs available for discovery
N_t	Number of subframes available for discovery
N_r	Total number of resources in discovery pool
N_u	Total number of UEs in the scenario
UE_X	Randomly chosen UE
θ_i	Received transmission probability of UE_i
θ_{tx}	Transmission probability of the transmitter UE_{tx}
θ_{rx}	Transmission probability of the receiver UE_{rx}
θ_{ini}	Initial transmission probability for the 3GPP algorithm
n_{min}	Minimum number of periods before assuming a UE is gone
p	Success criteria (i.e. confidence) value
t_i	Time of the last reception of UE_i

A. Optimal Transmission Probability

In the standard, all UEs announce using a preconfigured transmission probability defined in the discovery resource pool for UE-Selected mode. However, based on [8], the use of specific transmission probability values selected according to the size of the group improves the performance of the whole process significantly. The optimal transmission probability θ^* is calculated as shown in Eq. (1), except when Eq. (2) is true, in which case the optimal value of θ^* is 1.

$$\theta^* = \frac{2N_r + N_t(N_u - 1) - \sqrt{4N_r(N_r - N_t) + N_t^2(N_u - 1)^2}}{2N_u} . \quad (1)$$

$$N_u < \frac{N_r(N_t - 2) + N_t}{N_t - 1} , \quad \text{where} \quad N_t > 1 \quad (2)$$

Although the computed value θ^* is not necessarily a multiple of 1/4 (as recommended by 3GPP), it was shown that rounding up to the next allowed value (i.e. 0.25, 0.5, 0.75, 1) does not alter the discovery performance. Therefore, from now on, we will be using θ as the approximation of θ^* to the nearest non-zero multiple of 0.25 less than or equal to 1.

B. Success Probability

A discovery message is successfully received between two UEs if several conditions are satisfied. First, the transmitter UE_{tx} is allowed to announce in the current period after checking its transmission probability θ_{tx}. Secondly, the receiver UE_{rx} should not be announcing at the same time slot (i.e. subframe) or it would miss UE_{tx} discovery message, as the discovery messages are sent over a half-duplex channel, which prevents the UEs from sending and receiving data in the same time slot (half-duplex constraint). Finally, none of the other UEs pick the same resource in the same time slot as the transmitter to avoid any collisions. Accounting for those requirements, the success probability of UE_{rx} discovering UE_{tx} for a **single period** is defined by Eq. (3) for the 3GPP-defined behavior (i.e. static), and by Eq. (4) for the adaptive algorithm (i.e. dynamic) presented in [9].

According to the static 3GPP behavior, all the UEs utilize the initial transmission probability θ_{ini} throughout the whole discovery process. We assume that all UEs have the same θ_{ini}. So, the probability of a discovery message being successfully received is:

$$P_{success_{static}} = \theta_{ini}\left(1 - \frac{\theta_{ini}}{N_t}\right)\left(1 - \frac{\theta_{ini}}{N_r}\right)^{N_u - 2} ; \quad (3)$$

However, using the dynamic adaptive algorithm, we know that each UE_i has its own transmission probability θ_i computed using Eq. (1) and Eq. (2), and from them we derive the probability of success for dynamic values of θ:

$$P_{success_{dynamic}} = \theta_{tx}\left(1 - \frac{\theta_{rx}}{N_t}\right) \prod_{i \neq tx, i \neq rx} \left(1 - \frac{\theta_i}{N_r}\right) ; \quad (4)$$

The resource pool parameters N_r and N_t are known and constant. Knowing that, we use Eq. (3) and Eq. (4) to calculate the probability of a successful reception within n periods, which is:

$$p = 1 - (1 - P_{success})^n ; \quad (5)$$

Eq. (5) allows us to determine the minimum number of periods for a UE to receive an announcement from another UE, given a success criteria equal to p.

$$n_{min} = \frac{\ln(1 - p)}{\ln(1 - P_{success})} ; \quad (6)$$

With these models it is possible for the receiver to know how long it should wait before learning that a transmitter has turned off or moved away, according to the confidence (i.e. the success criteria) on that learning that is desired or required.

C. Redesigned Discovery Message

In order to be able to make use of those analytical models in the discovery process, we need to announce each UE's transmission probability. To do so, we will introduce a minor modification in the discovery message format. The most significant component of the discovery message is the ProSe Application

Ben Mosbah, Aziza; Griffith, David; Rouil, Richard.
"Enhanced Transmission Algorithm for Dynamic Device-to-Device Direct Discovery."
Paper presented at 2018 IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, United States. January 12, 2018 - January 15, 2018.

Code (with a size of 184 bits [10]). This code is allocated per announcing UE and application and has an associated validity timer. Discovery messages are limited in size (only 232 bits) to allow their transmission in a single subframe and a pair of resource blocks, even in bad channel conditions. Increasing its size is not a practical option because that will be resource-consuming and shrink the available bandwidth. To overcome this limitation and to avoid unnecessary overhead, our proposal allocates 2 bits of the ProSe Application ID Name, within ProSe Application Code, to carry the value of the probability of transmission in the form of two coded bits for the four allowed values for θ (i.e. 0.25, 0.5, 0.75, 1).

Using this approach, we maintain the size of the ProSe Application Code, as shown in Fig. 1. A mapping example of the 2-bit values is presented below:

- 0.25: 00
- 0.50: 01
- 0.75: 10
- 1.00: 11

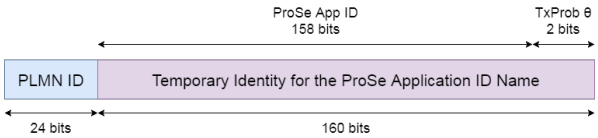


Fig. 1: Modified ProSe Application Code

D. Proposed Algorithm

Given that 3GPP does not define how the detection of departing UEs should happen, we will be testing an implementation similar to the one in our enhanced algorithm. Therefore, the only differences between both implementations (static, i.e. 3GPP defined with our departure detection mechanism, and dynamic, i.e., our proposed enhanced algorithm) will be the use of the optimal theta and keeping track of the individual values of θ.

For any given UE$_X$, the transmission process for D2D direct discovery in UE-Selected mode will follow either Algorithm 1 or Algorithm 2 depending on whether we are using the 3GPP-defined transmission probability or the enhanced algorithm. The discovery period length, the number of subframes and resource blocks dedicated to discovery, and the considered success probability are the inputs to both algorithms.

Using the modified version of the 3GPP algorithm (Algorithm 1), each UE will keep track of UEs it discovered and the time they were discovered, and will update the number of UEs discovered based on both Eq. (3) and Eq. (6). It will be referred to as static configuration because θ does not change throughout the simulation.

For the enhanced algorithm (Algorithm 2), each UE will be able to process the received announcements, check which ones are new or contained a different transmission probability, and compute its own transmission probability after discarding UEs that may have left the discovery group using Eq. (1), Eq. (4), and Eq. (6). It will be referred to as dynamic configuration because of the continuous calculation of θ.

Data: d is the discovery period length in seconds and p is the considered success criteria
for *any given UE$_X$ performing D2D discovery* **do**
  UE$_X$ receives discovery messages from n UEs;
  Record current time as *TimeNow*;
  for i *in* $[1, n]$ **do**
    if *UE$_i$ was never discovered before* **then**
      Create record for UE$_i$;
      Set $t_i = TimeNow$, where t_i is the time of most recent reception from UE$_i$;
    else
      Update UE$_i$'s record: set $t_i = TimeNow$;
    end
  end
  for *each UE$_j$ received so far* **do**
    calculate n_{min} based on all the received transmission probabilities (Eq. (6));
    if $n_{min} < \frac{TimeNow - t_j}{d}$ **then**
      Delete UE$_j$'s record;
  end
end

Algorithm 1: 3GPP transmission algorithm (Static) using success probabilities for D2D Discovery

IV. Simulation and Results

In this section, we provide the scenarios parameters and simulation results. To obtain the results presented here we used the discrete event simulator ns-3 [11] with the LTE D2D models from [12], extended to include our discovery algorithms.

We define arrival and departure scenarios where UEs join and leave the discovery group throughout the simulations. Users are deployed randomly within an area of 200 m $\times$ 200 m. All UEs are able to discover each other. Each UE sends discovery messages by independently choosing a resource from a discovery resource pool using the procedure in [13]. Table II contains a list of simulation parameters and their default values.

Based on this scenario parameters and according to Eq. (1) and Eq. (2), the optimal transmission probability depends on the number of UEs as represented in Fig. 2.

A. Arrival Scenario

First we will look at an scenario with UEs only arriving to the group. With this scenario we will validate that the modifications introduced in the discovery algorithm do not alter the behavior observed in our previous proposal ([9]). We assume that we have X initial UEs in the area. Their number varies from 10 to 90. After 100 seconds (i.e. 306 periods), Y

Ben Mosbah, Aziza; Griffith, David; Rouil, Richard.
"Enhanced Transmission Algorithm for Dynamic Device-to-Device Direct Discovery."
Paper presented at 2018 IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, United States. January 12, 2018 - January 15, 2018.

Data: d is the discovery period length in seconds and p is the considered success criteria
for *any given* UE_X *performing D2D discovery* **do**
  UE_X receives discovery messages from n UEs;
  Record current time as *TimeNow*;
  for i *in* $[1, n]$ **do**
    if UE_i *was never discovered before* **then**
      Create record for UE_i;
      Set $t_i = TimeNow$, where t_i is the time of most recent reception from UE_i;
      Set the transmission probability θ_i;
    else
      Update UE_i's record: set $t_i = TimeNow$ and θ_i;
    end
  end
  $N_u = 1$;
  for *each* UE_j *received so far* **do**
    calculate n_{min} based on all the received transmission probabilities (Eq. (6));
    if $n_{min} < \frac{TimeNow - t_j}{d}$ **then**
      Delete UE_j's record;
    else
      Increment N_u;
    end
  end
  if $N_u > 1$ **then**
    calculate θ based on the new N_u value, and the pool configuration (N_t and N_r) (Eq. (1)) ;
    round θ to the nearest multiple of 0.25;
    add the encoded value to next announcements;
    use the resulting value of θ to announce;
  end
end

Algorithm 2: Enhanced transmission algorithm (Dynamic) using success probabilities for D2D Discovery

TABLE II: Simulation Parameters and Values

Parameters	Values
UE transmission power	23 dBm
Propagation model	Cost231 [14]
Available bandwidth	50 RBs
Carrier frequency	700 MHz
Discovery period d	0.32 s
Number of retransmission	0
Number of repetition	1
Number of resource block pairs N_f	6
Number of subframes N_t	5
Total number of resources N_r	30
Area size	200 m × 200 m
Success criteria p	0.99, 0.95, 0.90
Total simulations per scenario	100

UEs join the group, such as $X + Y = 100$ after 100 seconds (i.e. 306 periods).

Using 3GPP algorithm (i.e. static), the discovery performance varies based on the pre-configured (3GPP-defined) transmission probability used. Using our enhanced algorithm

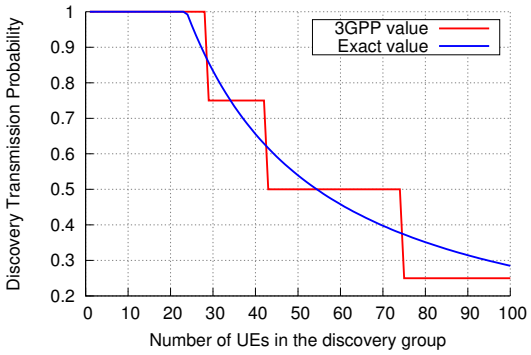


Fig. 2: Optimal transmission probability associated with the number of UEs

(i.e. dynamic), the UEs start announcing using a transmission probability equal to 1 (i.e. 100 %) until they start monitoring discovery messages and use the adaptive algorithm to evaluate the optimal transmission probability value. The second group starts discovery at 100 s (i.e. 306 periods, enough time so that all the UEs in the first group have discovered everyone else in that group). For example, if we have 90 UEs initially, 10 UEs will join later on. We consider a success criteria of 99 % and we compute the number of periods needed for all UEs to discover all other UEs in their own group and the second group, along with a confidence interval of 95 %. Because of the nature of our enhanced algorithm, the number of periods needed to complete discovery is effectively independent of the initial transmission probability used.

We will look at the results of the UEs in group one (1) discovering the UEs in group two (2), and the UEs in group two discovering everyone in Fig. 3 and 4. The legend refers to the number of UEs in group 1. The rest of the cases (UEs in group 1 discovering the UEs in group 1, etc.) are similar to the analyses from our previous paper ([9]), and while we considered and validated that those cases perform similarly, the results are omitted from this paper for brevity.

1) group 1 discovering group 2: We compute the number of periods needed for group 1 to complete the discovery of group 2 in Fig. 3 with 95 % confidence intervals. The UEs in group 1 start discovering UEs in group 2 after 306 periods. The results show that the discovery performance is better when using the enhanced algorithm (dashed lines) in comparison to the 3GPP algorithm (solid lines), independently of the initial transmission probability.

For the 3GPP algorithm, all UEs (from both groups, i.e. 100 UEs) are transmitting at the same initial transmission probability. Based on Fig. 2, the discovery performance is optimal for a transmission probability of 0.25. Therefore, the higher the transmission probability used, the more number of periods needed to finish the discovery. X/Y refers to the scenario where there are X UEs in the first group and Y UEs in the second group. So, we have X UEs discovering Y UEs. The least time needed to finish discovery is when

Ben Mosbah, Aziza; Griffith, David; Rouil, Richard.
"Enhanced Transmission Algorithm for Dynamic Device-to-Device Direct Discovery."
Paper presented at 2018 IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, United States. January 12, 2018 - January 15, 2018.

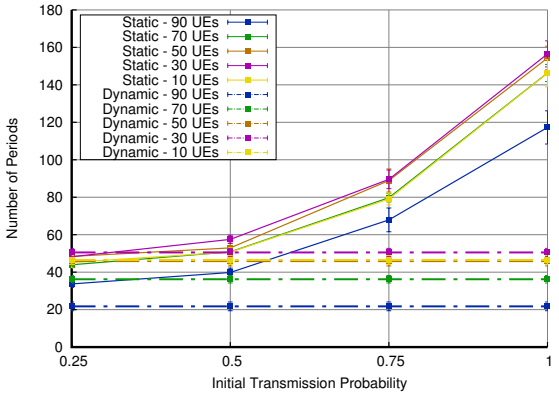


Fig. 3: Number of periods needed for group one to complete discovery of group two, with 95 % confidence intervals

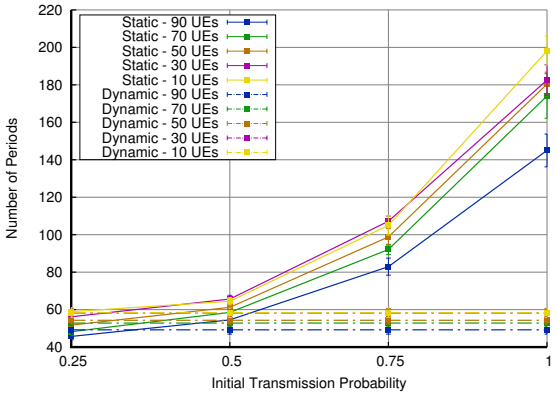


Fig. 4: Number of periods needed to complete discovery by group 2, with 95 % confidence intervals

we have 90/10, i.e. 90 UEs in group 1 and only 10 UEs to be discovered (group 2) by those 90 UEs (blue solid line). However, 70/30 (green solid line) and 10/90 (red solid line) take less time than 50/50 (orange solid line) and 30/70 (purple solid line). In those cases, all UEs in both groups send announcements simultaneously to discover each other, which creates collisions and delays the discovery completion. The delay is related to the number of UEs involved (both the number of UEs performing the discovery and the number of UEs to be discovered). The reason why the 50 and 30 UEs results are worse is that, in the other cases, either the UEs in group 1 were already at a low θ, with a few UEs coming in with $\theta = 1$ initially, which reduced collisions, or the UEs in group 1 were a few UEs with $\theta = 1$, so when group 2 joined, a large number of UEs were discovered in that first period (as everyone is transmitting, all the RBs will be used), and that triggered a quick drop in the value of θ, improving the performance. The results for 50 and 70 UEs show how the "intermediate" values are the ones most likely to be penalized the most by collisions, due to similarly sized groups of UEs having different values of transmission probability (but not 1 or 0.25).

For the enhanced algorithm, as expected, we can see how the initial θ is irrelevant for the results, and only the optimal value of θ for the initial group is a factor that provides different performance. The first group is already using its computed optimal transmission probability. Similarly to the 3GPP algorithm, 90 UEs discovering 10 UEs takes the least number of periods. The UEs in group 2 start using $\theta = 1$ and then adjust it according to the number of UEs they are discovering. Thus the two yellow and green dashed lines are superposed. Finally, the discovery performance for the last two cases (10 UEs and 30 UEs in the first group) is close.

2) group 2 discovering everyone: We compute the number of periods needed to complete the discovery by group 2 in Fig. 4 with 95 % confidence intervals. The enhanced algorithm outperforms the 3GPP algorithm. The UEs in the second group finish discovery (including the first group) later than the first group, because they are simultaneously discovering UEs from the first group and their own group. We notice that, for both algorithms, the number periods needed to finish the discovery for group 2 is inversely proportional to the number of UEs in the first group, as fewer UEs in group 2 means fewer UEs to discover the whole group, thus saving time.

B. Departure Scenario

This second scenario will serve to analyze and validate our algorithms in scenarios with UEs departing the group. We assume that we have a group with 100 UEs. Y (with $0 \leq Y < 100$) UEs start leaving the group after 100 seconds. This value was chosen to allow the UEs to have enough time to complete the discovery. We implemented the departure detection logic as described in Algorithm 1 and Algorithm 2. We first focus on the beginning when the 100 initial UEs start discovering each other. Then, we evaluate the departure process and how UEs react to the changes in the group. We also vary the success criteria (99 %, 95 %, and 90 %) and assess how that affects the number of UEs discovered and the estimate reliability. We also studied 85 %, 80 %, and 75 % success criteria, with the overall trend for their performance being similar to the results presented. However, these results have not been included due to space limitations. From this point onward, the 3GPP results will be about the modified 3GPP, unless explicitly stated otherwise.

1) The discovery process at the beginning: We evaluate the discovery process at the beginning of the simulation, when all UEs are in the group. The results for different success criteria values are represented in Fig. 5.

In the 3GPP case (i.e static), starting with $\theta = 1$ makes the discovery take the most time. For values of 0.25 and 0.5, the performance is close and the best. The enhanced algorithm, although less efficient than starting with the optimal transmission probability, succeeds to catch up with that ideal case, with minimal overhead.

Ben Mosbah, Aziza; Griffith, David; Rouil, Richard.
"Enhanced Transmission Algorithm for Dynamic Device-to-Device Direct Discovery."
Paper presented at 2018 IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, United States. January 12, 2018 - January 15, 2018.

Varying the success criteria value affects the discovery process in different ways. For the enhanced algorithm (i.e. dynamic), the algorithm oscillates at some points of the simulation, as it assumes that some UEs left the group after not hearing from them for a while, due to the fact that most UEs are changing their transmission probabilities simultaneously. This is more obvious for values of the success criteria lower than 99 %. The number of UEs is increasing and the optimal transmission probability is switching from 1 to 0.25. UEs are tuned to wait less according to Eq. 6. When they don't hear from other UEs after the time period they originally computed, they consider them gone and the mean estimated group size decreases.

For low success criteria values, we have an unreliable judgment which impacts the accuracy of the computed wait time. However, those UEs may have changed their own θ values to accommodate the UEs discovered and thus they are announcing less frequently. Once this new information is propagated, the actual number of UEs discovered increases back to what it is expected. We don't observe those oscillations for the 3GPP algorithm because it uses a constant transmission probability (Eq. 3).

For both the enhanced and the 3GPP algorithms we see that, with success criteria lower than 99 %, it is not possible to acknowledge the total number of UEs in the discovery group, with the difference between the "discovered" amount of UEs and the total increasing as the success criteria decreases. This inaccuracy is due to the fact that some UEs do not wait long enough before assuming another UE has departed the group.

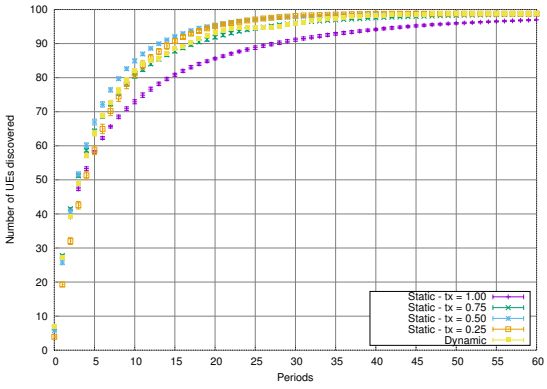
2) The discovery process after 100 seconds: At this time, some UEs leave the group (for simulation purposes, this happens instantly). First, we evaluate the effect of the change of the initial value of θ. Then, we study the impact of different success criteria values on both algorithms. In the following figures, in order to improve clarity of the plots, we have zoomed in at the time at which UEs started leaving the discovery group (i.e. around 300 discovery periods). In Fig. 6, we consider a 3GPP transmission probability of 1 and we vary the success criteria. In this case, the 3GPP algorithm is using $\theta = 1$. Based on Eq. 3 and Eq. 6, the probability of success is low and the UEs wait longer before deciding to discard other UEs from the discovery group because of the potential collisions and the half-duplex feature. The wait time is longer when the accuracy required is high.

For the enhanced algorithm, the departure process starts at $\theta = 0.25$. We notice a stair effect generated by the change of the transmission probability based on the number of UEs discovered over time, which affects the pace of the discovery process as well. The concavity is smoother as the success criteria is smaller.

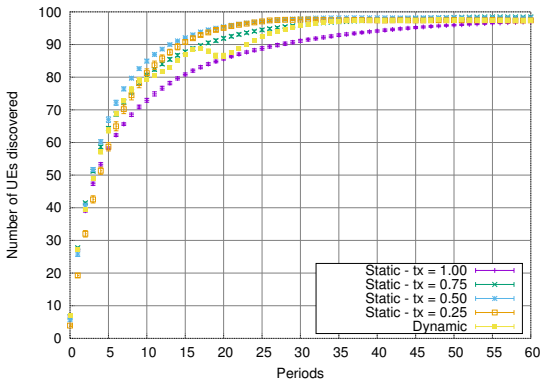
Like the 3GPP algorithm, the enhanced algorithm drops UEs faster and the wait time is reduced for low success criteria values. However, the impact of the success criteria on the reduction of the wait time is more perceptible for the 3GPP algorithm than for the enhanced algorithm.

We also notice how, even though lower success criteria reduces the time required to assume that UEs have left the group, this also makes the algorithms to miscount some UEs as departed, thus reducing estimated total group size. The gap between the computed number of UEs leaving and the "expected" group size gets wider for lower success criteria.

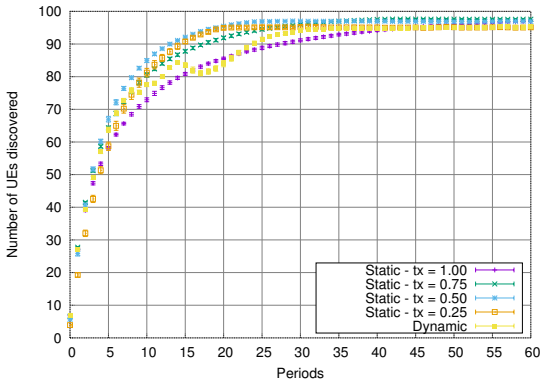
Similar conclusions can be drawn when considering initial transmission probabilities of 0.75, 0.5, and 0.25, with a narrower gap between the performance of both the enhanced and



(a) Success criteria = 99 %



(b) Success criteria = 95 %

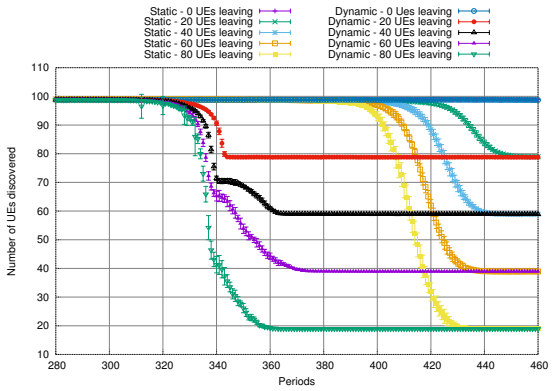


(c) Success criteria = 90 %

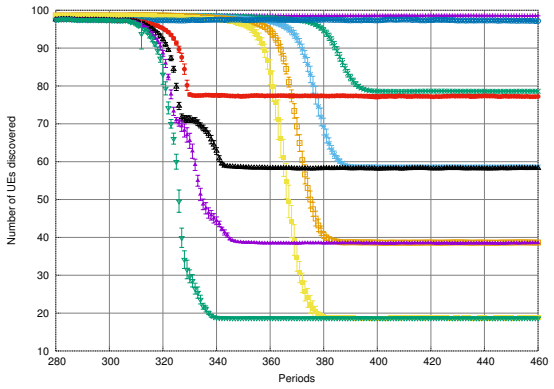
Fig. 5: Number of periods needed for all UEs to discover all other UEs in the group for different success criteria values

Ben Mosbah, Aziza; Griffith, David; Rouil, Richard.
"Enhanced Transmission Algorithm for Dynamic Device-to-Device Direct Discovery."
Paper presented at 2018 IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, United States. January 12, 2018 - January 15, 2018.

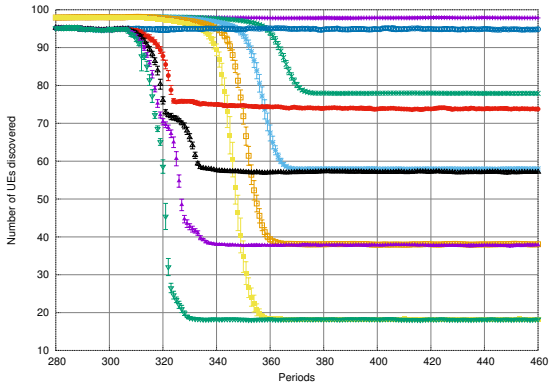
the 3GPP algorithms.



(a) Success criteria = 99 %



(b) Success criteria = 95 %



(c) Success criteria = 90 %

Fig. 6: Number of UEs acknowledged to be in the group over time for different success criteria values (3GPP transmission probability = 1)

In Fig. 7, we fix the success criteria to 99 % while varying the initial transmission probability. The discovery performance and the number of UEs left does not change in the enhanced algorithm case. At that point, the UEs transmit using the optimal θ (i.e. 0.25), and the initial θ value doesn't affect its behavior. The change occurs for the 3GPP case. We notice that it takes less time to start considering some UEs gone. For example, the system reaches a stable state after 330 periods for an initial transmission probability of 0.25, compared to 460 periods for an initial transmission probability of 1. That is when the 3GPP algorithm behaves the worst. The performance penalty is represented as the longer time required to learn that UEs left the group,

For $\theta = 1$ initially, the enhanced algorithm outperforms the 3GPP algorithm and succeeds to reach a stable state faster. For $\theta = 0.75$ initially, the enhanced algorithm reaches a stable state at approximately the same time as the 3GPP algorithm, although the enhanced algorithm drops more UEs over time. Less congestion and contention are recorded, which delays the convergence to the actual number of UEs in the discovery group.

For $\theta = 0.5$ initially, the discovery performance is close to the optimal case. The enhanced algorithm starts detecting more UEs leaving the discovery group at the beginning of the process. But, the 3GPP algorithm succeeds to catch up and reaches a stable state faster.

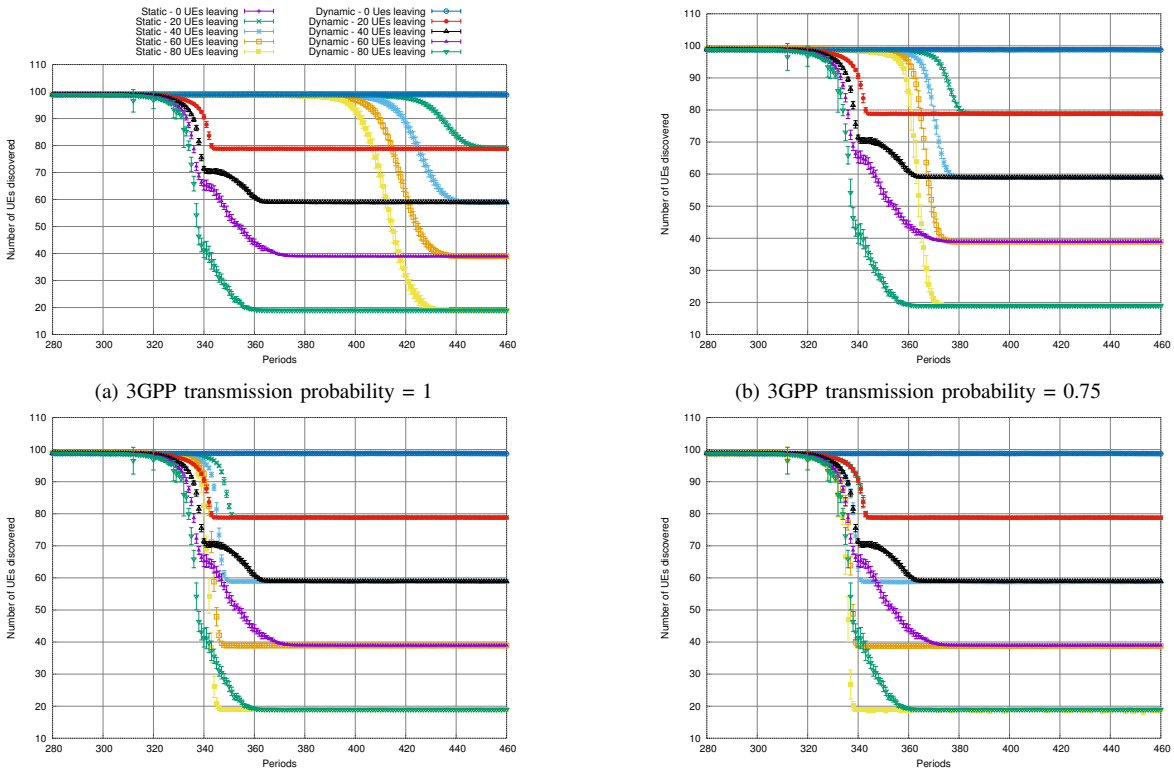
For $\theta = 0.25$ initially, the 3GPP and the enhanced algorithms have the same start point. This is shown through the graphs for the first 20 periods after the actual UEs departure (i.e. 306 periods). However, the 3GPP algorithm drops the number of UEs discovered faster than the enhanced algorithm, because UEs in the enhanced algorithm take time adjusting θ based on the number of UEs.

Although the 3GPP algorithm behaves better than the enhanced algorithm with $\theta = 0.25$, we showed in the previous paper [9] that a low transmission probability with small groups may increase the time required for discovery significantly, making it 3 times longer than needed. That is the strength of the enhanced algorithm: while fixed values of the transmission probability may provide better results for specific group sizes, that knowledge of the group size and channel conditions is generally known a priori, and in that case, the enhanced algorithm consistently provides near-optimal and very consistent results for groups of any size, regardless of UE arrivals and departures.

V. Conclusion and Future Work

In this paper we presented an enhanced discovery algorithm for LTE D2D to be used in out-of-coverage scenarios. The enhanced algorithm builds on previous proposals that identified the optimal transmission probability depending on the group size, and extends them enabling the discovery process to fully be aware and react to dynamic changes in the network. We have shown how the algorithm can be tuned depending on whether the primary concern is fast adaptation or accuracy, making the process more suitable to be used in a wide variety of scenarios. From this contribution we can foresee several research possibilities, such as the automation of the tuning parameters depending on the group size volatility.

Ben Mosbah, Aziza; Griffith, David; Rouil, Richard.
"Enhanced Transmission Algorithm for Dynamic Device-to-Device Direct Discovery."
Paper presented at 2018 IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, United States. January 12, 2018 - January 15, 2018.



(a) 3GPP transmission probability = 1

(b) 3GPP transmission probability = 0.75

(c) 3GPP transmission probability = 0.5

(d) 3GPP transmission probability = 0.25

Fig. 7: Number of UEs acknowledged to be in the group over time for different initial transmission probabilities and a success criteria of 99 %

REFERENCES

[1] 3GPP, "Study on LTE device to device proximity services; Radio aspects," 3rd Generation Partnership Project (3GPP), TR 36.843, 2015. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/36.843.htm

[2] S. Madhusudhan, P. Jatadhar, and P. D. K. Reddy, "Performance evaluation of network-assisted device discovery for lte-based device to device communication system," *Journal of Network Communications and Emerging Technologies (JNCET) www. jncet. org*, vol. 6, no. 8, 2016.

[3] D. Xenakis, M. Kountouris, L. Merakos, N. Passas, and C. Verikoukis, "Performance analysis of network-assisted d2d discovery in random spatial networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 8, pp. 5695–5707, Aug 2016.

[4] H. Chour, Y. Nasser, H. Artail, A. Kachouh, and A. Al-Dubai, "Vanet aided d2d discovery: Delay analysis and performance," *IEEE Transactions on Vehicular Technology*, vol. PP, no. 99, pp. 1–1, 2017.

[5] H. Albasry and Q. Z. Ahmed, "Network-assisted d2d discovery method by using efficient power control strategy," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–5.

[6] K. P. Sharmila, V. Mohan, C. Ramesh, and S. P. Munda, "Proximity services based device-to-device framework design for direct discovery," in *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*, Feb 2016, pp. 499–502.

[7] A. Murkaz, R. Hussain, S. F. Hasan, M. Y. Chung, B. C. Seet, P. H. J. Chong, S. T. Shah, and S. A. Malik, "Architecture and protocols for inter-cell device-to-device communication in 5g networks," in *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, Aug 2016, pp. 489–492.

[8] D. Griffith and F. Lyons, "Optimizing the UE Transmission Probability for D2D Direct Discovery," in *IEEE Global Telecommunications Conference (GLOBECOM 2016)*, Washington D.C., USA, Dec 2016.

[9] A. Ben Mosbah, D. Griffith, and R. Rouil, "A novel adaptive transmission algorithm for Device-to-Device direct discovery," in *IWCMC 2017 Wireless Networking Symposium (IWCMC-Wireless Networks 2017)*, Valencia, Spain, Jun. 2017.

[10] 3GPP, "Numbering, Addressing and Identification," 3rd Generation Partnership Project (3GPP), TS 23.003, 2015. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/23003.htm

[11] NS-3 Documentation, "LTE Module in NS-3," accessed 27-September-2016. [Online]. Available: https://www.nsnam.org/docs/models/html/lte.html

[12] R. Rouil, F. J. Cintrón, A. Ben Mosbah, and S. Gamboa, "Implementation and validation of an lte d2d model for ns-3," in *Proceedings of the Workshop on Ns-3*, ser. WNS3 '17, New York, NY, USA, 2017, pp. 55–62.

[13] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification," 3rd Generation Partnership Project (3GPP), TS 36.321, 2015. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/36321.htm

[14] Commission of the European Communities, "Digital Mobile Radio: COST 231 View on the Evolution Towards 3rd Generation Systems," Luxembourg, 1989, accessed 27-September-2016. [Online]. Available: https://goo.gl/P06OZ7

Ben Mosbah, Aziza; Griffith, David; Rouil, Richard.
"Enhanced Transmission Algorithm for Dynamic Device-to-Device Direct Discovery."
Paper presented at 2018 IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, NV, United States. January 12, 2018 - January 15, 2018.

Channel De-embedding and Measurement System Characterization for MIMO at 75 GHz*

Alexandra E. Curtin, David R. Novotny, Alex J. Yuffa, and Selena Leitner
National Institute of Standards and Technology
Boulder, CO

***Abstract*—We explore the development of a sparse set of measurements for array calibration, relying on coherent multi-channel data acquisition of wideband signals at 75 GHz, and the hardware characterization and post-processing necessary to perform channel de-embedding at an elemental level for a 4x1 system. By characterizing the complete RF chain of our array and the differential skew and phase response of our measurement hardware, we identify crucial quantities for measuring closed commercial systems. In the future, combining these responses with precise elemental location information, will enable us to consider means of de-embedding elemental response and coupling effects that may be compared to conventional single-element calibration information and full-pattern array measurements.**

I. Introduction

As modern antenna array systems for multi-input multi-output (MIMO) and 5G applications are deployed, there is increased demand for measurement techniques for timely calibration, at both research and commercial sites[1]. The desired measurement method must allow for the de-embedding of information about the closed digital signal chain and element alignment, and should be performed in the near-field.

Current means of measuring large arrays cover a variety of methods. Full-array gain and pattern calibration must cover the parameter space of single-element weightings and is time-consuming, to the point where the measurement may take longer than the duration over which the array response is stable[2]. Two other popular methods are the transmission of orthogonal codes and the use of holography to reconstruct a full-array pattern. The first of these methods again requires extremely long measurement time. For an array of N elements and weightings per element W_n, the matrix of orthogonal codes must be of an order greater than NW_n[2][3]. This number varies with the form of W_n depending on whether the array is analog or digital, but in both cases for every desired beam configuration, an order-N encoding matrix must be used. The second method relies on illuminating subsets of elements within an array and reconstructing the full pattern[4]. Each illuminated subset, however, neglects some amount of coupling information inherent to the complete system, making this an imperfect method.

We seek to develop near-field methods for characterizing arrays using a homegrown 4x1 system as a testbed. While the majority of the measurements here are taken using the IF port, we also have access to the 12.5-GHz modulated RF signal prior to upconversion in our RF extender heads. We use spatial metrology techniques developed at NIST as part of the Configurabe RObotic MilliMeter-wave Antenna (CROMMA) facility to locate all antenna elements and align our probe antenna for channel response calibration[5]. We hope to characterize this system fully and use it to develop much-needed near-field techniques to allow for efficient antenna parameter extraction.

The 4-element receive array is a sparse non-uniform linear array where the elemental separation d is much greater than the wavelength. If we were to consider measurements of single 75-GHz antenna elements, then the apertures of our antennas would be on the order of a centimeter; however the aperture of the complete array is much larger due to the large d. This measurement challenge is common to all arrays: the aperture size scales with the dimension of the entire array, rather than with the horn size or number of elements contained within that area. Consequently the Rayleigh distance $\frac{2D^2}{\lambda}$ for our linear array is on the order of 100 m. This means that from a rule-of-thumb perspective, performing a measurement in the far-field of the array within the confines of our optical table is not possible even for this simple array.

The task of finding time-efficient means of characterizing arrays with near-field measurements, then, relies on *a priori* knowledge of the theoretical near-field pattern generated by an array of model dipoles with spacing $\gg \lambda$, knowledge of single-element antenna parameters, knowing the systematic channel response, and knowing the precise element positioning. We will characterize our 4x1 system, eventually sweeping the angle of our transmitting horn and applying phase gradients in hardware and post-processing.

II. Experiment

Our four-element receiver system operates at 75 GHz and is measured at an IF of 279 MHz. The system is comprised of a combination of Rohde&Schwartz[1] WR-10 and WR-15 x6 RF extender heads mounted with four WR-15 8-dB rectangular horns. Two of these elements will have a mismatch at the waveguide transition. For the transmitting side of our setup we use a WR-10 extender head with a WR-10 open-ended waveguide (OEG) for measuring individual antenna

[1]Certain trade names and company products are mentioned in the text. In no case does such identification imply endorsement of these products and equipment by the National Institute of Standards and Technology, nor does it imply that the products are necessarily the best available for the purpose.

Curtin, Alexandra; Novotny, David; Yuffa, Alexey.
"Channel De-embedding and Measurement System Characterization for MIMO at 75 GHz."
Paper presented at Antenna Measurements Techniques Association: 39th Annual Symposium, Atlanta, GA, United States. October 16, 2017 - October 20, 2017.

elements and a high-gain rectangular horn for illuminating the entire array. Figure 1 shows a block diagram for the setup, including the distribution of the clock signal and LO. The five elements (four receive, one transmit) are tied together with a common LO. Measurement of each of the four receive channels is handled by two Tektronix oscilloscopes, each of which receives a reference signal from the transmitting RF head into Channel 1 for triggering. The transmitted signal is generated at 12.5 GHz, mixed with a 2048-symbol pseudo-random noise (PN) code, and upconverted in the extender head to 75 GHz.

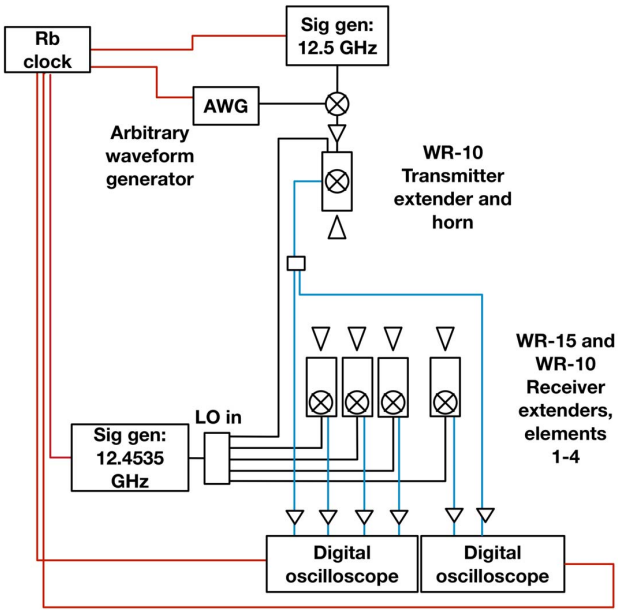


Fig. 1. Block diagram of the setup showing the RF, LO, and IF signals, as well as the Rb clock distribution.

Figure 2 shows the complete experimental setup for the system mounted on the optical table. The transmitting head is mounted on a square tilt plate atop a manual height-adjustable stage, all riding on a flat rail that is mounted to the optical table perpendicular to the direction of signal propagation. Each element of the receive array is mounted to a tubular rail that similarly sits on a second manual height-adjustable stage. The distances between the elements are approximately equal and as short as possible given the dimensions of the extender heads. The two adjusting stages are used in conjunction with the tilt plate and horizontal rail under the transmitter to align the system. Placed throughout the setup are nests for spherical mirror reflectors (SMRs). Visible in the figure are two of the four table landmarks that are used to place all other spatial metrology information in the laboratory frame.

A. *Differential Skew*

In order to understand the phase stability of our system, we began by measuring the differential skew of the oscilloscope and cabling part of our experimental apparatus. This portion of the system characterization bypassed the RF signal and upconversion, and consisted only of a cabled measurement of the PN code at 250 MHz. A marker was input to Channel 1 of each of the two oscilloscopes. The remaining channels sampled the PN code at 5 GS/s. Although each oscilloscope has four channels, Channel 2 on "scope 2" is broken, leaving us with five total receiving channels and five cables, labeled A, B, C, E, and F. Although the cables were nominally the same length, to get a value for the differential skew, we measured each channel, permuting the cable combinations so that we could later extract the channel-to-channel skew separate from cable effects. Skew is defined to be the phase angle between channels, or in this case channel and cable configurations. For every cable configuration, we took multiple acquisitions. A single calculation to remove cable information from the skew between Channel 3 and Channel 4 from scope 1 would look like:

$$\theta_s = \tan^{-1}\left(\frac{\mathfrak{F}(3B_n)/\mathfrak{F}(2A_n)}{\mathfrak{F}(4B_n)/\mathfrak{F}(2A_n)}\right), \tag{1}$$

where $\mathfrak{F}$ is the Fourier transform of the time domain data for each channel and cable. Here, Channel 3 and 4 data are both collected with Cable B and Channel 2 with Cable A is used as a common reference. The range of values shows no dependence on cable arrangement. We then considered differential skew for these many arrangements. The differential skew calculation tells us the phase stability of the skew over time. That is to say, it tells us not only what size phase gradient we can hope to apply and resolve in measurement of the receive array, but also whether we can use a constant channel-to-channel skew value for our measurements.

The differential skew is defined as the slope of the skew between channels $d\theta_s/d\omega$, where:

$$\theta_s = \arctan\left(\frac{\mathfrak{F}(Y_n)/\mathfrak{F}(X_n)}{\mathfrak{F}(Y_0)\ \mathfrak{F}(X_0)}\right). \tag{2}$$

For any two channels X and Y, the numerator defines the skew between the two channels for a file acquisition "n", and the denominator defines the skew between those same two channels for the initial file acquisition "0". For our purposes, X and Y define a cable + channel configuration. Figure 3 shows the differential skew between Channel 2 with Cable A on scope 1 and Channel 3 with Cable B, calculated from (1). We fit the skew to a line over the central 125 MHz of the 250 MHz. The resulting plot shows that we have a slope of roughly 1.3 degrees/GHz and phase noise below 2 degrees for the first 125 MHz of bandwidth. By performing this calculation for every channel + cable pair we find that within a single oscilloscope, the differential skew for each channel spans the range of 1 ± 0.3 degrees/GHz, which gives a timing stability of approximately 3 ps.

Using the same initial "0" file, we calculated differential skew for the channel + cable pair 2A on the first oscilloscope and 4F on the second oscilloscope for five subsequent data acquisitions. Figure 4 shows the resulting data and fitted lines.

Curtin, Alexandra; Novotny, David; Yuffa, Alexey.
"Channel De-embedding and Measurement System Characterization for MIMO at 75 GHz."
Paper presented at Antenna Measurements Techniques Association: 39th Annual Symposium, Atlanta, GA, United States. October 16, 2017 - October 20, 2017.

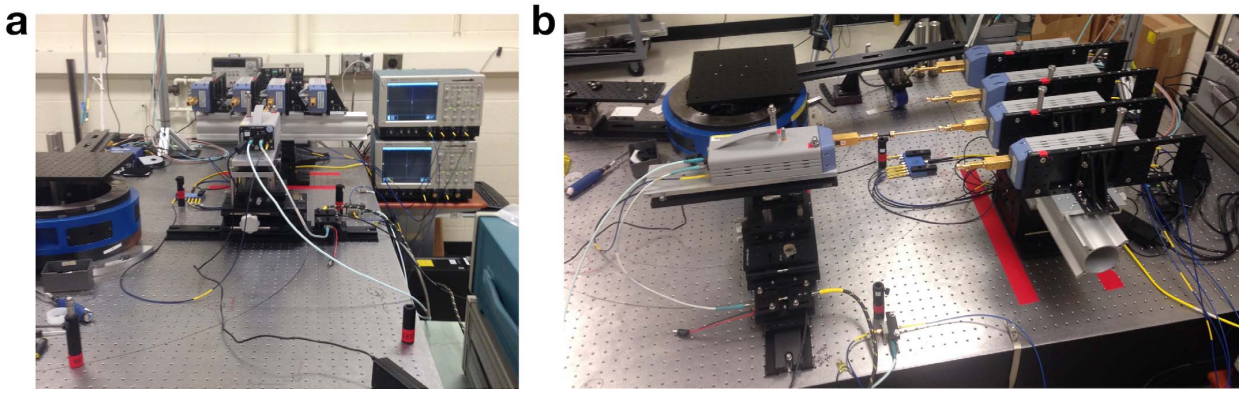


Fig. 2. Photos of the MIMO optical table setup for viewing (a) along the transmitting direction and (b) from above. The extender heads are shown with four 8-dB horns mounted on the receiving elements and an OEG on the transmitting head. Table fiducials are labeled with red tape, while SMR nests are visible on all the heads. More nests are hidden from view on the side of the room facing the laser tracker (not shown, above region in panel b). In these images, amplifiers are not connected for the RF, LO and IF signals.

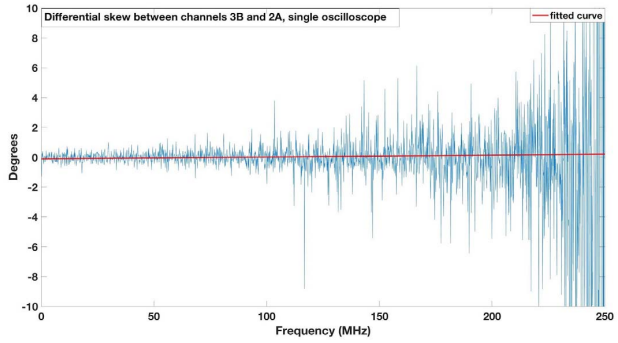


Fig. 3. Differential skew between Channel 3 with Cable B and Channel 2 with Cable A. The 2-A combination was used as a reference for skew calculations. The plot contains two important figures of merit: the slope of the line is roughly 1.3 degrees/GHz, and the noise in the phase data is under 2 degrees for the first 125 MHz of bandwidth.

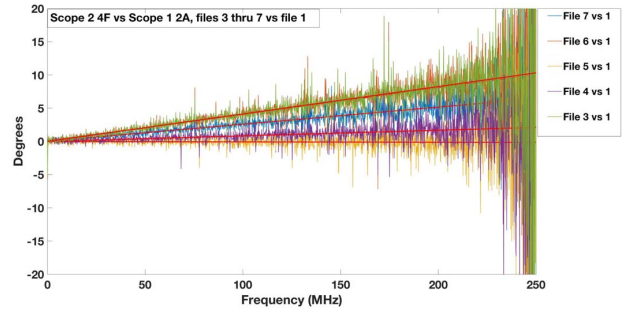


Fig. 4. Differential skew for five pairs of files with the channel+cable configuration 4F and 2A. In all five pairs, file "1" was used for the denominator in (2).This shows that while differential skew is stable among channels on a single oscilloscope, we do not have stable timing between oscilloscopes. Additionally, the differential skew values are much larger, in the tens of degrees per GHz.

The acquisitions were collected approximately two to three minutes apart, although data for later cable configurations not shown here were collected only 30 s apart. The differential skew values for these five lines span a wide range of values up to 45 deg/GHz, showing no phase stability between oscilloscopes. This means that although differential skew between channels within a single oscilloscope is small and stable, values between oscilloscopes are much larger and unstable. For now, this means that the skew of our last receive element relative to the reference signal coming from the transmitting head is small and stable, but we cannot achieve repeatable timing or measure small phase delays on this element relative to the other three. Scope-to-scope skew for a particular measurement will be calculated by applying (1) and (2) to the data collected in Channel 1 of each scope. Nevertheless the large values between the two scopes mean that small applied delays will not be resolvable. For the future, this means a move to sampling equipment with more available coherent channels.

B. Spatial Metrology

The practice of using laser trackers to align antennas was refined at NIST during the development of CROMMA to both align antenna systems and perform iterative path corrections for spherical near-field scans[5]. Here, we use a laser tracker along with a host of SMRs and spatial metrology data acquisition software to map the lab space, the placement of SMR "nests" on the array and transmitting setup, and the location of the five horns relative to these nests.

In order to align the OEG with each element on the receive array, each element was first located within the lab frame by tracking the path of a 0.5" SMR (12.7 mm) in an edge nest over the face of each aperture. The collection of points for each element follows the four sides of the aperture. This point group is easily fit to a plane with a centroid and normal pointing in the direction of propagation. The four point clusters

Curtin, Alexandra; Novotny, David; Yuffa, Alexey.
"Channel De-embedding and Measurement System Characterization for MIMO at 75 GHz."
Paper presented at Antenna Measurements Techniques Association: 39th Annual Symposium, Atlanta, GA, United States. October 16, 2017 - October 20, 2017.

were fit to lines that were then used to find the center of the aperture. Together these constructed data were used to make a coordinate frame at the aperture, with $\hat{\mathbf{x}}$ pointing toward the transmitter and $\hat{\mathbf{z}}$ pointing to the ceiling. A similar mapping was performed on the OEG. A fixed set of four nest positions on the receive array and three nests on the transmitting extender head and associated staging are also measured. An offset frame between the constructed aperture plane and these fixed nests allowed for tracking of element position without physically remeasuring the aperture. This was critical to moving the transmitting head and tracking OEG position for alignment with each of the four stationary receive elements.

The alignments for this system are completely manual. We configure the laser tracker to watch the SMRs on the transmit head and therefore watch the location of the OEG in our spatial metrology software. By tracking the OEG location relative to a working frame with an x-axis pointing out of one of the antenna elements, we can align the OEG with the element for system response calibration. We can also use the data from each antenna element to analyze the planarity of the array and the uniformity of the element separation. We performed the alignment iteratively, using best fit transforms between the repeatedly collected sets of points corresponding to the transmit-head SMR locations.

For illuminating the entire array, we will pull the rail for the transmitting antenna back across the optical table. The fiducial marks T1-T4 on the table and the receive array are not moved, so we may easily measure the change in transmitter location. Using the tilt stage, we can sweep the angle of the high gain horn, measuring this angle between the transmitting aperture and the plane of the entire array.

C. RF Measurements

Similar to our differential skew measurements, the laser tracker-based spatial alignments of the OEG with each receive element enables the calibration of a single element response for the de-embedding of the systematic response of the channel from the free space response of the antennas. Furthermore, if we assume that we can model each element as a dipole, we can back out the relative dipole strength and phase of each element. The channel response is calculated to be:

$$CR = \frac{\mathfrak{F}(channel_n(t))}{\mathfrak{F}(channel_1(t))}, \tag{3}$$

where $channel_1(t)$ is the reference channel from the transmitting head. For each element in the receive array, we can calculate this response relative to the signal in Channel 1 of the corresponding oscilloscope. The phase information needed to inform the array factor model is input relative to the phase of the first element. The phase is calculated from the channel response after extracting the skew between channels.

III. Results and Discussion

A. Alignment

We characterized each receive element by collecting approximately 160-220 data points for each antenna aperture and fit these data points to a plane with a centroid and outward pointing normal. For each element, the y-axis points horizontally along the array, the x-axis points out of each aperture, and the z-axis points vertically. For each antenna element, the rms deviation in data points from the plane ranges from 34.8 μm to 53.6 μm. To locate the center of an element, data points associated with each edge were fit to a line. The rms deviation from the line fit was large, ranging between 120 μm and 258 μm, or roughly $\lambda/20$. From these lines we determine a center point and use the plane to define the normal out of the aperture plane.

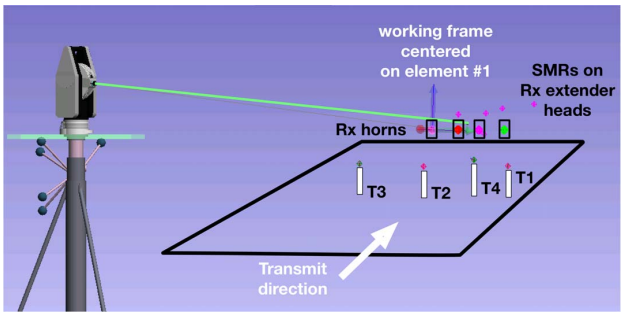


Fig. 5. This screenshot shows the spatial metrology data for the receiving half of our setup. The white and black annotations are added to roughly show the table, the fiducials on the table, and an outline of the rectangular horns (not to scale). The horn apertures appear as point clouds where an edge nest and SMR were used to trace the edges of the each horn. The working frame built from these data on the first element is an example of the frames used to align the transmitting OEG with each element.

Figure 5 shows an annotated screenshot from the spatial metrology software mapping the receiver half of our system. The white and black markings for the optical table, the four SMR fiducials, and the four rectangular horns are added for clarity. The laser-tracker instrument is placed relative to a global lab frame. The data taken around each of the four horns appear as the four clusters of data points aligned with the rectangular horn locations. The farthest left aperture has been used to construct a working frame with which we can align the first placement of the transmitting OEG. After iteratively manually aligning the OEG, we find that we are within 1 degree on-axis with the receive element, with a repeatability of +/- 20 μm in the $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ directions. The repeatability in $\hat{\mathbf{z}}$ is worse by a factor of two due to differences in the stage hardware. The accuracy of the alignment is further dependent on the accuracy of the plane and line-fits performed earlier.

To locate the elements in space for later measurements, we constructed a coordinate frame located at the center of the array. We define every element's location relative to this coordinate frame. The frame will later be used to align our high-gain horn for illuminating the entire array. The elements

Curtin, Alexandra; Novotny, David; Yuffa, Alexey.
"Channel De-embedding and Measurement System Characterization for MIMO at 75 GHz."
Paper presented at Antenna Measurements Techniques Association: 39th Annual Symposium, Atlanta, GA, United States. October 16, 2017 - October 20, 2017.

in the receive array are not perfectly linear. Relative to this frame, we calculated elemental positions shown in Table I:

TABLE I

Element	1	2	3	4
x(mm)	-0.65	0.32	0.76	-0.78
y(mm)	-219.15	-71.60	72.59	219.18
z(mm)	0.95	-0.16	-0.63	1.01

The distance between the OEG and each element was measured as a constant 10.26 mm for each of the first three elements. The fourth element was separated from the OEG by 10.32 mm due to hard limits in the transmit stage motion. This 80 μm +/- 20 μm difference is $\lambda/50$, and is a small path difference relative to the phase differences between elements.

B. RF Results and the Array Factor

The array factor describes how a distribution of antenna elements with a particular elemental response combine to form the radiation pattern for the array. We assume the four receiver elements are identical dipoles with varying dipole strengths ($V_n \in \mathbf{R}$) and phase (ϕ_n). The array factor is given by:

$$S = \sum_{n=1}^{4} V_n e^{ik\mathbf{R}\cdot\mathbf{r}+\phi_n}. \quad (4)$$

Here k is the wave number $2\pi/\lambda$, $\mathbf{R}$ is the vector from the transmit antenna normal to the plane of the receive array, and $\mathbf{r}$ is the vector from locating the element relative to the centerline of the array. The exponent also includes the relative phase delay ϕ_n between elements.

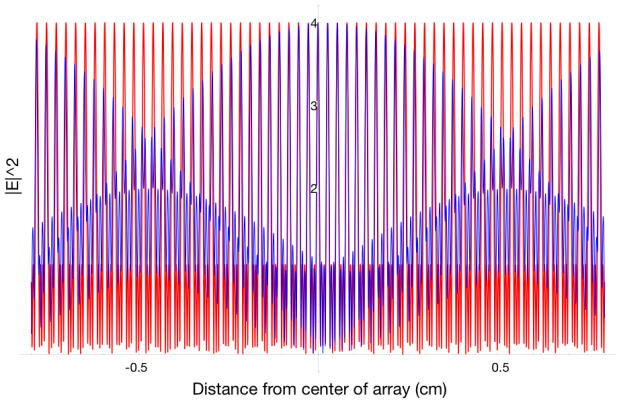


Fig. 6. The dipole approximation of our receive array provides a toy model to demonstrate the effect of spatial distortions on the pattern of a sparse array. The regular red pattern shows the pattern for the 4-element array at 75 GHz with even 12-cm spacing. The blue pattern shows the envelope that a 1-λ dislocation of one edge element produces.

Using this rough dipole model of our setup, operating at 75 GHz with the elements placed approximately 14 cm apart, we produced a theoretical radiation pattern pictured in Figure 6. To clearly show the effect of variation in element location, the blue curve in Figure 6 shows the effect of moving one of the end elements by λ along the y-axis along the array. By varying spatial misalignment with this model, we see that position changes of less than λ still produce an envelope similar to the blue curve, but with a longer period. Changes in position in $\hat{\mathbf{x}}$ and $\hat{\mathbf{z}}$ produce smaller effects more easily seen in the lower side-peaks of our dominant signal.

Figure 7 shows the same array factor plotted with experimental values. For each channel response measurement, we extracted a relative phase ϕ_n by deconvolving the received signal with the measured transmitted signal in Channel 1 of each oscilloscope. We then used our laser tracker data to extract real values for the position of each element relative to the center of the array in the average plane of the apertures, as previously shown in Table I. The location data alone produce the blue curve, with the envelope spread and shifted relative to the single wavelength shift in Figure 6. This model shows the effect of the location perturbations in our array, both along the main axis and smaller effects due to perturbations vertically and in the direction of propagation. The phase between the elements received on different oscilloscopes was estimated by calculating both the delay between Channel 1 and the receiver channel and the skew between Channel 1 on each oscilloscope. By inputting these values into our model along with the spatial data, we produce the orange curve that assumes each element has an identical radiation pattern. Our receive elements are 8-dB standard gain horns and, as such, have higher directivity than our dipole models, likely resulting in a stronger measurable effect due to location. It will be in the scope of future work to simulate and eventually measure precise patterns for these horns, but the dipole approximation provides an adequate simple model for demonstrating our ability to separate spatial distortions from elemental differences with precise spatial metrology.

For the present work, we have set the dipole strength, V_n, of each element to be equal to 1. In calculating the channel response for each element, we observed that differences in the received signal amplitude were dominated by variations in the amplifiers and extender heads. Variations in effective dipole strength for our model will be further measured in future work. Small variations in the dipole strength would create offsets in the amplitude of the signal shown in orange in Figure 7.

IV. Conclusions

We have shown the utility of precise spatial metrology for simulating the pattern of a sparse array. Furthermore, we have shown that the effect of even minor spatial misalignment of array elements on the pattern of the array. This information presents two paths forward: (1) To properly calibrate each element of an array spatial metrology is a vital tool for aligning the probe with the receiving element. (2) For the generation of a particular radiation pattern from an array, elemental phase delays must be chosen and applied not only to form a beam but also to correct for an imperfect grid of elements. Future work will focus on both of these avenues. Additionally, we will experimentally measure the entire array in its near field. We will use a high-gain horn to highlight directionality

Curtin, Alexandra; Novotny, David; Yuffa, Alexey.
"Channel De-embedding and Measurement System Characterization for MIMO at 75 GHz."
Paper presented at Antenna Measurements Techniques Association: 39th Annual Symposium, Atlanta, GA, United States. October 16, 2017 - October 20, 2017.

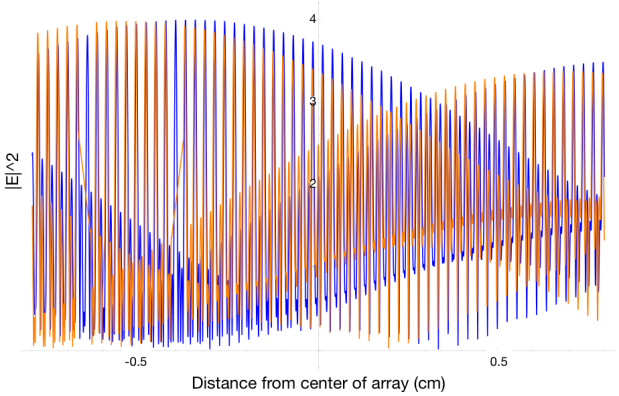


Fig. 7. The dipole simulation of our array is presented again with experimentally-measured location perturbations and element-to-element phase offsets input into the model. The blue curve is the pattern generated by the receive array with only the spatial data inserted into the model. The orange curve additionally has phase offsets input relative to the phase of the first element. Small differences in dipole strength would additionally create an amplitude offset to this curve.

and apply elemental weightings for beam steering. We will also vary these weightings by applying a phase gradient to measured data in post-processing. From this exploration and accompanying simulation, we hope to provide further options for precise and efficient array calibration and characterization with near-field measurements.

REFERENCES

[1] Caleb Fulton, Mark Yeary, Daniel Thompson, John Lake, and Adam Mitchell. Digital phased arrays Challenges and opportunities. *Proceedings of the IEEE*, 104(3):487–503, 2016.

[2] Seth D. Silverstein. Application of orthogonal codes to the calibration of active phased array antennas for communication satellites. *IEEE Transactions on Signal Processing*, 45(1):206–218, 1997.

[3] Erik Lier and Michael Zemlyansky. Phased array calibration and characterization based on orthogonal coding Theory and experimental validation. *2010 IEEE International Symposium on Phased Array Systems and Technology (ARRAY)*, pages 271–278, 2010.

[4] E. N. Grossman, A. Luukanen, and A. J. Miller. Holographic microantenna array metrology. *Proceedings of SPIE, Passive Millimeter-Wave Imaging Technology VIII*, 5789(44), 2005.

[5] Joshua A. Gordon, David R. Novotny, Michael H. Francis, Ronald C. Wittmann, Miranda L. Butler, Alexandra E. Curtin, and Jeffrey R. Guerrieri. Millimeter-wave near-field measurements using coodinated robotics. *IEEE Transactions on Antennas and Propagation*, 63, 2015.

Curtin, Alexandra; Novotny, David; Yuffa, Alexey.
"Channel De-embedding and Measurement System Characterization for MIMO at 75 GHz."
Paper presented at Antenna Measurements Techniques Association: 39th Annual Symposium, Atlanta, GA, United States. October 16, 2017 - October 20, 2017.

Total Break of the SRP Encryption Scheme

Ray Perlner[1], Albrecht Petzoldt[1], and Daniel Smith-Tone[1,2]

[1]National Institute of Standards and Technology,
Gaithersburg, Maryland, USA
[2]Department of Mathematics, University of Louisville,
Louisville, Kentucky, USA

ray.perlner@nist.gov, albrecht.petzoldt@nist.gov, daniel.smith@nist.gov

Abstract. Multivariate Public Key Cryptography (MPKC) is one of the main candidates for secure communication in a post-quantum era. Recently, Yasuda and Sakurai proposed in [7] a new multivariate encryption scheme called SRP, which combines the Square encryption scheme with the Rainbow signature scheme and the Plus modifier.
In this paper we propose a practical key recovery attack against the SRP scheme, which is based on the min-Q-rank property of the system. Our attack is very efficient and allows us to break the parameter sets recommended in [7] within minutes. Our attack shows that combining a weak scheme with a secure one does not automatically increase the security of the weak scheme.

Keywords: Multivariate Cryptography, SRP Encryption Scheme, Cryptanalysis, min-Q-Rank

1 Introduction

Multivariate cryptography is one of the main candidates to guarantee the security of communication in the post-quantum era [1]. Multivariate schemes are in general very fast and require only modest computational resources, which makes them attractive for the use on low cost devices such as RFIDs or smart cards [2, 3]. While there exist many practical multivariate signature schemes such as UOV [4], Rainbow [5] and Gui [6], the number of secure and efficient multivariate public key encryption schemes is quite limited.

At ICISC 2015, Yasuda and Sakurai proposed in [7] a new multivariate encryption scheme called SRP, which combines the Square encryption scheme [8], the Rainbow signature scheme [5] and the Plus method [9]; hence the name SRP. The scheme is very efficient and has a comparably small blow up factor between plain and ciphertext size. In [7] it is claimed that, by the combination of Square and Rainbow into one scheme, several attacks against the single schemes are no longer applicable.

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

In this paper we present a new practical key recovery attack against the SRP encryption scheme, which uses the min-Q-rank property of the system to separate the Square from the Rainbow and Plus polynomials. By doing so, we can easily find (parts of) the linear transformations $\mathcal{T}$ and $\mathcal{U}$ used to hide the structure of the central map $\mathcal{F}$ in the public key. The attack is completed by using the known structure of the Rainbow part of the central map.

Our attack is very efficient and allows us (even with our limited resources) to break the SRP instances proposed in [7] for 80, 112 bit security in 8 minutes and less than three hours respectively. By switching to a larger server we could break the parameters proposed for 160 bit security, too. Our attack therefore shows that this attempt to combine several multivariate schemes into one brings no extra security into the system.

Our paper is organized as follows. In Section 2, we give an overview of the basic concepts of multivariate public key cryptography and introduce the SRP encryption scheme of [7]. In Section 3 we recall the concept of the Q-Rank of a quadratic map, while Section 4 describes the main ideas and results of the Kipnis-Shamir attack on HFE needed for the description of our attack. Section 5 describes our key recovery attack against the SRP scheme in detail, whereas Section 6 deals with the complexity of our attack. In Section 7 we present the results of our computer experiments, and Section 8 concludes the paper.

2 The SRP Encryption Scheme

In this section, we recall the SRP scheme of [7]. Before we come to the description of the scheme itself, we start with a short overview of the basic concepts of multivariate cryptography.

2.1 Multivariate cryptography

The basic objects of multivariate cryptography are systems of multivariate quadratic polynomials over a finite field $\mathbb{F}$. The security of multivariate schemes is based on the *MQ Problem* of solving such a system. The MQ Problem is proven to be NP-Hard even for quadratic polynomials over the field GF(2) [10] and believed to be hard on average (both for classical and quantum computers).

To build a multivariate public key cryptosystem (MPKC), one starts with an easily invertible quadratic map $\mathcal{F} : \mathbb{F}^n \to \mathbb{F}^m$ (*central map*). To hide the structure of $\mathcal{F}$ in the public key, we compose it with two invertible affine (or linear) maps $\mathcal{T} : \mathbb{F}^m \to \mathbb{F}^m$ and $\mathcal{U} : \mathbb{F}^n \to \mathbb{F}^n$. The *public key* of the scheme is given by $\mathcal{P} = \mathcal{T} \circ \mathcal{F} \circ \mathcal{U} : \mathbb{F}^n \to \mathbb{F}^m$. The relation between the easily invertible central map $\mathcal{F}$ and the public key $\mathcal{P}$ is referred to as a morphism of polynomials.

Definition 1 *Two systems of multivariate polynomials $\mathcal{F}$ and $\mathcal{G}$ are said to be related by a* morphism *iff there exist two affine maps $\mathcal{T}$,$\mathcal{U}$ such that $\mathcal{G} = \mathcal{T} \circ \mathcal{F} \circ \mathcal{U}$.*

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

The *private key* consists of the three maps $\mathcal{T}, \mathcal{F}$ and $\mathcal{U}$ and therefore allows to invert the public key.

To encrypt a message $M \in \mathbb{F}^n$, one simply computes $C = \mathcal{P}(M) \in \mathbb{F}^m$.
To decrypt a ciphertext $C \in \mathbb{F}^m$, one computes recursively $\mathbf{x} = \mathcal{T}^{-1}(C) \in \mathbb{F}^m$, $\mathbf{y} = \mathcal{F}^{-1}(\mathbf{x}) \in \mathbb{F}^n$ and $M = \mathcal{U}^{-1}(\mathbf{y})$. $M \in \mathbb{F}^n$ is the plaintext corresponding to the ciphertext C. This process is illustrated in Figure 1.

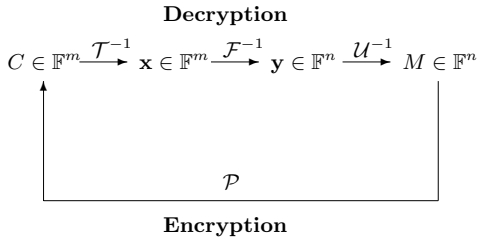


Fig. 1. Encryption and decryption process for multivariate public key encryption schemes

Since, for multivariate encryption schemes, we have $m \geq n$, the pre-image of the vector $\mathbf{x}$ under the central map $\mathcal{F}$ and therefore the decrypted plaintext will (with overwhelming probability) be unique.

2.2 SRP

The SRP encryption scheme was recently proposed by Yasuda and Sakurai in [7] by combining the Square encryption scheme [8], the Rainbow signature scheme [5] and the Plus method [9]. Since both Square and Rainbow are very efficient, the same holds for the SRP scheme. Furthermore, the combination with Rainbow provides an efficient way to distinguish between correct and false solutions of Square. In [7] it is claimed that, by the combination of Square and Rainbow into one scheme, several attacks against the single schemes are no longer applicable.

In this paper, we restrict to variants of SRP in which the Rainbow part is replaced by UOV [4]. Note that the parameter sets proposed in [7] are of this type. However we note that our attack can easily be generalized to variants of SRP which use a Rainbow (and not UOV) map $\mathcal{F}_R$ and that these modifications have no significant effect on the running time of the attack.
We choose a finite field $\mathbb{F} = \mathbb{F}_q$ of odd characteristic with $q \equiv 3 \bmod 4$ and, for an odd integer d, a degree d extension field $\mathbb{E} = \mathbb{F}_{q^d}$. Let $\phi : \mathbb{F}^d \to \mathbb{E}$ be an isomorphism between the vector space $\mathbb{F}^d$ and the field $\mathbb{E}$. Moreover, let o, r, s and l be non-negative integers.

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

Key Generation Let $n = d + o - l$, $n' = d + o$ and $m = d + o + r + s$. The *central map* $\mathcal{F} : \mathbb{F}^{n'} \to \mathbb{F}^m$ of the scheme is the concatenation of three maps $\mathcal{F}_S$, $\mathcal{F}_R$, and $\mathcal{F}_P$. These maps are defined as follows.

(i) The Square part $\mathcal{F}_S : \mathbb{F}^{n'} \to \mathbb{F}^d$ is the composition of the maps

$$\mathbb{F}^{n'} \xrightarrow{\pi_d} \mathbb{F}^d \xrightarrow{\phi} \mathbb{E} \xrightarrow{X \mapsto X^2} \mathbb{E} \xrightarrow{\phi^{-1}} \mathbb{F}^d.$$

Here $\pi_d : \mathbb{F}^{d+o} \to \mathbb{F}^d$ is the projection to the first d coordinates.

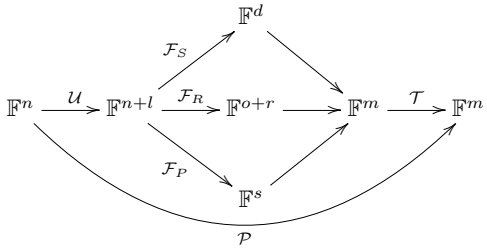
(ii) The UOV (Rainbow) part $\mathcal{F}_R = (f^{(1)}, \ldots, f^{(o+r)}) : \mathbb{F}^{n'} \to \mathbb{F}^{o+r}$ is constructed as the usual UOV signature scheme: let $V = \{1, \ldots, d\}$ and $O = \{d+1, \ldots, d+o\}$. For every $k \in \{1, \ldots, o+r\}$, the quadratic polynomial $f^{(k)}$ is of the form

$$f^{(k)}(x_1, \ldots, x_{n'}) = \sum_{i \in O, j \in V} \alpha_{i,j}^{(k)} x_i x_j + \sum_{i,j \in V, i \leq j} \beta_{i,j}^{(k)} x_i x_j + \sum_{i \in V \cup O} \gamma_i^{(k)} x_i + \eta^{(k)},$$

with $\alpha_{i,j}^{(k)}, \beta_{i,j}^{(k)}, \gamma_i^{(k)}, \eta^{(k)}$ randomly chosen in $\mathbb{F}$. [1]

(iii) The Plus part $\mathcal{F}_P = (g^{(1)}, \ldots, g^{(s)}) : \mathbb{F}^{n'} \to \mathbb{F}^s$ consists of s randomly chosen quadratic polynomials $g^{(1)}, \ldots, g^{(s)}$.

We additionally choose an affine embedding $\mathcal{U} : \mathbb{F}^n \hookrightarrow \mathbb{F}^{n'}$ of full rank and an affine isomorphism $\mathcal{T} : \mathbb{F}^m \to \mathbb{F}^m$. The *public key* is given by $\mathcal{P} = \mathcal{T} \circ \mathcal{F} \circ \mathcal{U} : \mathbb{F}^n \to \mathbb{F}^m$ and the *private key* consists of $\mathcal{T}, \mathcal{F}$ and $\mathcal{U}$.



Encryption: Given a message $M \in \mathbb{F}^n$, the ciphertext C is computed as $C = \mathcal{P}(M) \in \mathbb{F}^m$.

Decryption: Given a ciphertext $C = (c_1, \ldots, c_m) \in \mathbb{F}^m$, the decryption is executed as follows.

(1) Compute $\mathbf{x} = (x_1, \ldots, x_m) = \mathcal{T}^{-1}(C)$.
(2) Compute $X = \phi(x_1, \ldots, x_d) \in \mathbb{E}$.

[1] Note that, while, in the standard UOV signature scheme, we only have o polynomials, the map $\mathcal{F}_R$ consists of $o + r$ polynomials of the Oil and Vinegar type. This fact is needed to reduce the probability of decryption failures (see footnote 3).

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

(3) Compute $R_{1,2} = \pm X^{(q^d+1)/4} \in \mathbb{E}$ and set
$\mathbf{y}^{(i)} = (y_1^{(i)}, \ldots, y_d^{(i)}) = \phi^{-1}(R_i) \in \mathbb{F}^d$ $(i = 1, 2)$. [2]
(4) Given the vinegar values $y_1^{(i)}, \ldots, y_d^{(i)}$ $(i = 1, 2)$, solve the two systems of $o + r$ linear equations in the $n' - d = o$ variables $u_{d+1}, \ldots, u_{n'}$ given by

$$f^{(k)}(y_1^{(i)}, \ldots, y_d^{(i)}, u_{d+1}, \ldots, u_{n'}) = x_{d+k} \quad (i = 1, 2)$$

for $k = 1, \ldots, o + r$. The solution is denoted by $(y_{d+1}, \ldots, y_{n'})$. [3]
(5) Compute the plaintext $M \in \mathbb{F}^n$ by finding the pre-image of $(y_1, \ldots, y_{n'})$ under the affine embedding $\mathcal{U}$.

3 Q-Rank

A critical quantity tied to the security of multivariate BigField schemes is the Q-rank (or more correctly, the min-Q-rank) of the public key.

Definition 2 *Let $\mathbb{E}$ be a degree n extension field of $\mathbb{F}_q$. The Q-rank of a quadratic map $f(\overline{x})$ on $\mathbb{F}_q^n$ is the rank of the quadratic form $\phi \circ f \circ \phi^{-1}$ in $\mathbb{E}[X_0, \ldots, X_{n-1}]$ via the identification $X_i = \phi(\overline{x})^{q^i}$.*

Quadratic form equivalence corresponds to matrix congruence, and thus the definition of the rank of a quadratic form is typically given as the minimum number of variables required to express an equivalent quadratic form. Since congruent matrices have the same rank, this quantity is equal to the rank of the matrix representation of this quadratic form, even in characteristic 2, in which the quadratics x^{2q^i} are additive, but not linear for $q > 2$.

Q-rank is invariant under one-sided isomorphisms $f \mapsto f \circ U$, but is not invariant under isomorphisms of polynomials in general. The quantity that is often meant by the term Q-rank, but more properly called min-Q-rank, is the minimum Q-rank among all nonzero linear images of f. This min-Q-rank is invariant under isomorphisms of polynomials and is the quantity relevant for cryptanalysis.

In particular, min-Q-rank can be defined in circumstances for which Q-rank may make little sense. Specifically, consider the case in which there are more equations than variables, or the case in which we consider an extension field of smaller degree than the number of variables. We may then define min-Q-rank in the following manner.

[2] The fact of $q \equiv 3 \bmod 4$ and d odd allows us to compute the square roots of X by this simple operation. Therefore, the decryption process of both Square and SRP is very efficient.

[3] In [7, Proposition 1] it was shown that the probability of both $(y_1^{(1)}, \ldots, y_d^{(1)})$ and $(y_1^{(2)}, \ldots, y_d^{(2)})$ leading to a solution of the linear system is about $1/q^{-r-1}$. Therefore, with overwhelming probability, one of the two possible solutions is eliminated during this step.

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

Definition 3 *Let $\mathbb{E}$ be a degree $d < n$ extension field of $\mathbb{F}_q$. The min-Q-rank of a quadratic map $f : \mathbb{F}_q^n \to \mathbb{F}_q^m$ over $\mathbb{E}$ is*

$$\textit{min-Q-rank}(f) = \min_{L_1} \max_{L_2} \{ \textit{Q-rank}(L_1 \circ f \circ L_2) \},$$

where $L_1 : \mathbb{F}_q^d \to \mathbb{F}_q^m$ and $L_2 : \mathbb{F}_q^n \to \mathbb{F}_q^d$ are nonzero linear transformations. As above, "Q-rank" computes the rank of its input as a quadratic form over $\mathbb{E}[X_0, \ldots, X_{d-1}]$ via the identification $X_i = \phi(\overline{x})^{q^i}$.

4 The KS Attack and Minors Modeling

The property of low min-Q-rank is a weakness of many BigField schemes and has been exploited in many attacks, see [11–15]. While the attack in [12] exploits the low min-Q-rank property to speed up a direct algebraic attack, the other cryptanalyses use the Kipnis-Shamir (KS) attack of [11] with either the original KS modeling or with the minors modeling approach pioneered in [13].

The KS-attack recovers a related private key for a low min-Q-rank system with codomain isomorphic to a degree n extension field $\mathbb{E}$ by exploiting the fact that a quadratic form embedded in the homogeneous quadratic component of the private key is of low rank, say r. Using polynomial interpolation, the public key can be expressed as a collection of quadratic polynomials G over $\mathbb{E}$, and it is known that there is a linear map N such that $N \circ G$ has rank r as a quadratic form over $\mathbb{E}$; thus, there exists a rank r matrix that is an $\mathbb{E}$-linear combination of the Frobenius powers of G. This turns the task of recovering the transformation N into solving a MinRank problem over $\mathbb{E}$.

Definition 4 (**MinRank Problem(n,r,k)**)*: Given k $n \times n$ matrices $\mathbf{M}_1, \ldots, \mathbf{M}_k \in \mathcal{M}_{n \times n}(\mathbb{E})$, find an $\mathbb{E}$-linear combination $\mathbf{M} = \sum_{i=1}^{k} \alpha_i \cdot \mathbf{M}_i$ satisfying*

$$\textit{Rank}(\mathbf{M}) \leq r.$$

The key recovery attack of [13] revises the KS approach by modeling the low min-Q-rank property differently. The authors show that an $\mathbb{E}$-linear combination of the *public* polynomials has low rank as a quadratic form over $\mathbb{E}$. Setting the unknown coefficients in $\mathbb{E}$ of each of the public polynomials as variables, the polynomials representing $(r+1) \times (r+1)$ minors of such a linear combination, which must be zero due to the rank property, reside in $\mathbb{F}_q[t_{0,0}, \ldots, t_{0,m-1}]$. Thus a Gröbner basis needs to be computed over $\mathbb{F}_q$ and the variety computed over $\mathbb{E}$. This technique is called minors modeling and dramatically improves the efficiency of the KS-attack. The complexity of the KS-attack with minors modeling is asymptotically $\mathcal{O}(n^{(\lceil log_q(D) \rceil + 1)\omega})$, where $2 < \omega \leq 3$ is the linear algebra constant.

One should note that the situation is more complicated when multiple variable

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

types are utilized in a scheme. In the case that there are more variables than the degree of $\mathbb{E}$ over $\mathbb{F}_q$, the dimensions of the matrices do not match the degree of the extension. Still, if there is a central map with low min-Q-rank with a small subspace of the plaintext space as its domain, as it is the case of SRP, it may remain possible to recover a low rank map. Specifically, using fewer variables does not increase the rank of a quadratic form.

5 Key Recovery for SRP

In this section we explain our key recovery attack on SRP in detail. For the purpose of simplicity of exposition, we restrict to the homogeneous quadratic case. The method extends to the general case trivially.

We note that a public key of SRP is isomorphic to an analogous scheme without the embedding as long as $\pi_d \circ \mathcal{U}$ is full rank, which occurs with high probability. In this case, let $\pi'_d : \mathbb{F}_q^n \to \mathbb{F}_q^d$ be the projection onto the first d coordinates and find a projection $\rho : \mathbb{F}_q^{n+l} \to \mathbb{F}_q^n$ such that $\mathcal{U}' = \rho \circ \mathcal{U}$ has full rank and $\pi'_d \circ \mathcal{U}' = \pi_d \circ \mathcal{U}$. Let $\mathcal{F}^* : \mathbb{E} \to \mathbb{E}$ represent the squaring map so that $\mathcal{F}_S = \phi^{-1} \circ \mathcal{F}^* \circ \phi \circ \pi_d$. Then given the central maps $\mathcal{F}'_R = \mathcal{F}_R \circ \mathcal{U} \circ \mathcal{U}'^{-1}$ and $\mathcal{F}'_P = \mathcal{F}_P \circ \mathcal{U} \circ \mathcal{U}'^{-1}$, which are of Rainbow shape and of random shape respectively, one easily checks that

$$\mathcal{T} \circ \begin{bmatrix} \mathcal{F}^* \circ \pi_d \\ \begin{pmatrix} \mathcal{F}_R \\ \mathcal{F}_P \end{pmatrix} \end{bmatrix} \circ \mathcal{U} = \mathcal{T} \circ \begin{bmatrix} \mathcal{F}^* \circ \pi'_d \\ \begin{pmatrix} \mathcal{F}'_R \\ \mathcal{F}'_P \end{pmatrix} \end{bmatrix} \circ \mathcal{U}'.$$

It therefore suffices to consider the scheme with $l = 0$; however, for specificity, we analyze the embedding explicitly in the following discussion.

The attack is broken down into two main steps. The first is finding a related Square component private key. Then we discuss how to systematically solve for the Rainbow and Plus polynomials to complete key recovery.

5.1 The min-Q-Rank of SRP

While it is true that the min-Q-rank of the public key of an instance of SRP over a degree n extension is expected to be high, the public key retains the property that there exists a linear combination of the public forms which is of low Q-rank over the degree d extension used by the Square component. We verify this claim.

Let α be a primitive element of the degree d extension $\mathbb{E}$ of $\mathbb{F}_q$. Fix a vector space isomorphism $\phi : \mathbb{F}_q^d \to \mathbb{E}$ defined by $\phi(\overline{x}) = \sum_{i=0}^{d-1} x_i \alpha^i$. Furthermore, fix a one dimensional representation $\Phi : \mathbb{E} \to \mathbb{A}$ defined by $a \overset{\Phi}{\mapsto} (a, a^q, \ldots, a^{q^{d-1}})$.

Define $\mathcal{M}_d : \mathbb{F}_q^d \to \mathbb{A}$ by $\mathcal{M}_d = \Phi \circ \phi$. We can explicitly represent this map

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

with the matrix

$$\mathbf{M}_d = \begin{bmatrix} \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \alpha & \alpha^q & \cdots & \alpha^{q^{d-1}} \\ \alpha^2 & \alpha^{2q} & \cdots & \alpha^{2q^{d-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^{d-1} & \alpha^{(d-1)q} & \cdots & \alpha^{(d-1)q^{d-1}} \end{pmatrix} \end{bmatrix} \in \mathcal{M}_{d\times d}(\mathbb{E}),$$

acting via right multiplication (so that we may use algebraists' left-to-right composition). Thus we can pass between the two interesting representations of elements in $\mathbb{E}$ of the form $(x_0, \ldots, x_{d-1}) \in \mathbb{F}_q^d$ and $(X, X^q, \ldots, X^{q^{d-1}}) \in \mathbb{A}$ simply by right multiplication by $\mathbf{M}_d$ or $\mathbf{M}_d^{-1}$.

The above map $\mathbf{M}_d$ provides another way of expressing an SRP public key. Note first that any homogeneous $\mathbb{F}_q$-quadratic map from $\mathbb{E}$ to $\mathbb{E}$ induces a quadratic form on $\mathbb{A}$ that can be represented as a $d \times d$ matrix with coefficients in $\mathbb{E}$. Since the maps $\mathcal{F}_R$ and $\mathcal{F}_P$ can be written as vectors of quadratic forms over $\mathbb{F}_q[x_1, \ldots, x_n]$ in matrix form, the entire public key can be expressed as a matrix equation.

To achieve this matrix representation of the public key, we need some additional notation. We blockwise define

$$\widetilde{\mathbf{M}}_d = \begin{bmatrix} \begin{pmatrix} \mathbf{M}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{o+r+s} \end{pmatrix} \end{bmatrix} \in \mathcal{M}_{m\times m}(\mathbb{E})$$

and

$$\widehat{\mathbf{M}}_d = \begin{bmatrix} \begin{pmatrix} \mathbf{M}_d \\ \mathbf{0}_{o\times d} \end{pmatrix} \end{bmatrix} \in \mathcal{M}_{n'\times d}(\mathbb{E}).$$

Note that $\widetilde{\mathbf{M}}_d = \Phi \oplus id_{o+r+s}$ and $\widehat{\mathbf{M}}_d = \Phi \circ \pi_d$. Furthermore, let $\mathbf{F}^{*i}$ be the matrix representation of the quadratic form over $\mathbb{A}$ corresponding to the map $x \mapsto x^{2q^i}$.

Let $(\mathbf{F}_{S,0}, \ldots, \mathbf{F}_{S,d-1}, \mathbf{F}_{R,0}, \ldots, \mathbf{F}_{R,o+r-1}, \mathbf{F}_{P,0}, \ldots, \mathbf{F}_{P,s-1})$ denote the m-dimensional vector of $(d+o)\times(d+o)$ symmetric matrices associated to the private key. The function corresponding to the application of each coordinate of a vector of such quadratic forms followed by the application of a linear map represented by a matrix will be denoted by the right product of the vector by the matrix. Next, note that

$$(\mathbf{F}_{S,0}, \mathbf{F}_{S,1}, \ldots, \mathbf{F}_{S,d-1})\mathbf{M}_d = (\widehat{\mathbf{M}}_d\mathbf{F}^{*0}\widehat{\mathbf{M}}_d^\top, \widehat{\mathbf{M}}_d\mathbf{F}^{*1}\widehat{\mathbf{M}}_d^\top, \ldots, \widehat{\mathbf{M}}_d\mathbf{F}^{*d-1}\widehat{\mathbf{M}}_d^\top),$$

which yields

$$\begin{aligned}(\overline{x}\mathbf{F}_{S,0}\overline{x}^\top, &\overline{x}\mathbf{F}_{S,1}\overline{x}^\top, \ldots, \overline{x}\mathbf{F}_{S,d-1}\overline{x}^\top)\mathbf{M}_d \\ &= (\overline{x}\widehat{\mathbf{M}}_d\mathbf{F}^{*0}\widehat{\mathbf{M}}_d^\top\overline{x}^\top, \overline{x}\widehat{\mathbf{M}}_d\mathbf{F}^{*1}\widehat{\mathbf{M}}_d^\top\overline{x}^\top, \ldots, \overline{x}\widehat{\mathbf{M}}_d\mathbf{F}^{*d-1}\widehat{\mathbf{M}}_d^\top\overline{x}^\top),\end{aligned}$$

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

as functions of $\overline{x}$. Then we obtain the equation

$$\begin{aligned}&(\mathbf{F}_{S,0},\ldots,\mathbf{F}_{S,d-1},\mathbf{F}_{R,0},\ldots,\mathbf{F}_{P,m-1})\widetilde{\mathbf{M}}_d\\&=(\widehat{\mathbf{M}}_d\mathbf{F}^{*0}\widehat{\mathbf{M}}_d^\top,\ldots,\widehat{\mathbf{M}}_d\mathbf{F}^{*d-1}\widehat{\mathbf{M}}_d^\top,\mathbf{F}_{R,0},\ldots,\mathbf{F}_{P,s-1}).\end{aligned}\tag{1}$$

Next, consider the relation between the public key and the central maps of the private key.

$$(\mathbf{P}_0,\ldots,\mathbf{P}_{m-1})\mathbf{T}^{-1}=(\mathbf{U}\mathbf{F}_{S,0}\mathbf{U}^\top,\ldots,\mathbf{U}\mathbf{F}_{P,s-1}\mathbf{U}^\top).$$

By Equation (1), we have

$$\begin{aligned}&(\mathbf{P}_0,\ldots,\mathbf{P}_{m-1})\mathbf{T}^{-1}\widetilde{\mathbf{M}}_d\\&=(\mathbf{U}\widehat{\mathbf{M}}_d\mathbf{F}^{*0}\widehat{\mathbf{M}}_d^\top\mathbf{U}^\top,\ldots,\mathbf{U}\widehat{\mathbf{M}}_d\mathbf{F}^{*d-1}\widehat{\mathbf{M}}_d^\top\mathbf{U}^\top,\mathbf{U}\mathbf{F}_{R,0}\mathbf{U}^\top,\ldots,\mathbf{U}\mathbf{F}_{P,s-1}\mathbf{U}^\top).\end{aligned}$$

Let $\widehat{\mathbf{T}}=\mathbf{T}^{-1}\widetilde{\mathbf{M}}_d=[t_{i,j}]\in\mathcal{M}_{m\times m}(\mathbb{E})$ and let $\mathbf{W}=\mathbf{U}\widehat{\mathbf{M}}_d$. Then we have that

$$\sum_{i=0}^{m-1}t_{i,0}\mathbf{P}_i=\mathbf{W}\mathbf{F}^{*0}\mathbf{W}^\top.\tag{2}$$

Since the rank of $\mathbf{F}^{*i}$ is one for all i, the rank of this $\mathbb{E}$-linear combination of the public matrices is bounded by one. Indeed, if the rank were zero, then $\mathbf{W}=\mathbf{0}$, and the scheme reduces to a weak version of Rainbow+ whose kernel is the vinegar subspace. In particular, for all practical parameters one sets $d>l$, implying $d+o-l>o$, which verifies that $\mathbf{W}\neq\mathbf{0}$ (due to the fact that $\mathbf{U}$ is required to be full rank). Thus we obtain the following:

Theorem 1 *The min-Q-rank of the public key P of SRP(q,d,o,r,s,l) is, with high probability, given by:*

$$\textit{min-Q-rank}(P)=\begin{cases}0 \text{ if } d\leq l \text{ and } \mathbf{U}\widehat{\mathbf{M}}_d=\mathbf{0},\\1 \text{ otherwise.}\end{cases}$$

Proof. If $\mathbf{U}\widehat{\mathbf{M}}_d=\mathbf{0}$, then the span of P is of dimension at most $m-d$, and thus the min-Q-rank of P is zero. Otherwise, with high probability, the public polynomials are linearly independent. In this case, for any choice of L_1, there exists an L_2 such that the Q-rank of the composition $L_1\circ P\circ L_2$ is positive.

Consider, in particular, L_1 to be the $\mathbb{F}_q$-linear transformation defined by the matrix consisting of the first d columns of $\mathbf{T}^{-1}$. Let $L_2:\mathbb{F}_q^d\to\mathbb{F}_q^n$ be linear of full rank. Then

$$\phi\circ L_1\circ\mathcal{P}\circ L_2\circ\phi^{-1}=\mathcal{F}^*\circ\phi\circ\pi_d\circ\mathcal{U}\circ L_2\circ\phi^{-1}.$$

Let $\mathbf{L}_2$ be the $d\times n$ matrix representation of L_2. Then the matrix representation of the above quantity is

$$\mathbf{M}_d^{-1}\mathbf{L}_2\mathbf{U}\widehat{\mathbf{M}}_d\mathbf{F}^{*0}\widehat{\mathbf{M}}_d^\top\mathbf{U}^\top\mathbf{L}_2^\top\mathbf{M}_d^\top.$$

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

Since $\mathbf{F}^{*0}$ is of rank one and the image of $\widehat{\mathbf{M}}_d$ is $\mathbb{A}$, the product is of rank one exactly when $\mathbf{L}_2\mathbf{U}\widehat{\mathbf{M}}_d$ is nonzero, otherwise, the rank of the above matrix is zero. Since L_2 is chosen to maximize rank, the Q-rank is zero exactly when $\mathbf{U}\widehat{\mathbf{M}}_d$ is zero, which necessitates that $d \leq l$.

One may note here that the matrix $\widehat{\mathbf{T}}$ unmixes the Square equations from the Rainbow and Plus polynomials. It further mixes the Rainbow and Plus polynomials, but this is no issue since this phase of the attack is aimed at ultimately recovering a representation of $\mathcal{F}^*$.

5.2 Recovering the Output Transformation with MinRank

As demonstrated in the previous subsection, the recovery of $\widehat{\mathbf{T}}$ begins by solving a MinRank instance over $\mathbb{E}$. This phenomenon is well studied and has been the basis of previous cryptanalyses, see [13–15]. We may use the minors modeling approach to take advantage of the fact that we can compute the Gröbner basis over the small field, $\mathbb{F}_q$.

Due to the extremely low min-Q-rank of the system, the system of minors is homogeneous quadratic. The ideal generated by these minors is one dimensional, so we may set a single variable to a fixed value, say 1. We then recover a system of many quadratic equations in $m-1$ variables. This system is massively overdefined, so a solution can be recovered via linearization.

To accomplish this, we have to compute only as many minors as there are monomials in $m-1$ variables of total degree ≤ 2. There are exactly $\binom{m+1}{2}$ monomials in $m-1$ variables of degree less than or equal to two, so we randomly select $\binom{m+1}{2}$ minors and arrange their coefficients in a $\binom{m+1}{2} \times \binom{m+1}{2}$ matrix. As we will show in Section 6, we expect such a matrix to have full rank with high probability, roughly $\frac{q-1}{q}$ for large n and m. We may then linearly solve, recovering the first column of $\widehat{\mathbf{T}}$.

Once the first column of $\widehat{\mathbf{T}}$ is recovered, the first d columns can be generated by the relation

$$t_{i,j} = t_{i,j-1}^q \text{ for } j = 1, \ldots, d-1.$$

We will return to the issue of computing the remaining columns of $\widehat{\mathbf{T}}$ and separating the Rainbow and Plus polynomials in Subsection 5.5.

5.3 Recovering the Input Transformation

Once the first column of the transformation $\widehat{\mathbf{T}} = [t_{i,j}]$ is discovered, we have access to the rank one matrix

$$\sum_{i=0}^{m-1} t_{i,0}\mathbf{P}_i.$$

This matrix encodes the representation of the squaring map.

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

Theorem 2 *Given the first column of* $\widehat{\mathbf{T}}$*, the recovery of* $\mathbf{W}$ *requires the solution of a linear system of* $d+o-l-1$ *independent equations in* $d+o-l$ *variables.*

Proof. First, note that $\mathbf{W} = [w_{i,j}]$ is of the form $w_{i,j} = w_{i,j-k}^{q^k}$ for all $i \in \{0,1,\ldots,d+o-l\}$ and for all $0 \leq j,k < d$. Thus, it suffices to solve for the first column of $\mathbf{W}$. Let K be the left kernel of the low rank matrix

$$\sum_{i=0}^{m-1} t_{i,0}\mathbf{P}_i.$$

Let $\mathbf{K}$ be the matrix whose rows form a basis of K. By Equation (2), we know that

$$\mathbf{0}_{d+o-l-1\times d+o-l} = \mathbf{K}\mathbf{W}\mathbf{F}^{*0}\mathbf{W}^\top,$$

and since $\mathbf{W}$ is of full rank, it must be the case that

$$\mathbf{K}\mathbf{W}\mathbf{F}^{*0} = \mathbf{0}_{d+o-l-1\times d}.$$

Thus $K\mathbf{W} = ker(\mathbf{F}^{*0})$. In a proper basis the representation of $\mathbf{F}^{*0}$ contains a single nonzero entry in the first row and first column. Thus, the relation that $K\mathbf{W} = ker(\mathbf{F}^{*0})$ is equivalent to the condition that the first column of $\mathbf{W}$ is in the right kernel of $\mathbf{K}$. Since this right kernel is one dimensional, this process recovers all equivalent matrices $\mathbf{W}$.

Recall that we have the relation

$$\mathbf{W} = \mathbf{U}\widehat{\mathbf{M}}_d = \mathbf{U}\begin{bmatrix}\mathbf{M}_d\\ \mathbf{0}_{o\times d}\end{bmatrix}$$

Then multiplying on the right by $\mathbf{M}_d^{-1}$ yields

$$\mathbf{W}\mathbf{M}_d^{-1} = \mathbf{U}\begin{bmatrix}\mathbf{M}_d\\ \mathbf{0}_{o\times d}\end{bmatrix}\mathbf{M}_d^{-1} = \mathbf{U}\begin{bmatrix}\mathbf{I}_d\\ \mathbf{0}_{o\times d}\end{bmatrix} \tag{3}$$

Thus, we obtain the first d columns of $\mathbf{U}$. We may extend this matrix in any manner to obtain a full rank $n\times(d+o)$ matrix. With high probability, a random concatenation of o columns produces a full rank matrix $\mathbf{U}$. For the sake of recovering $\mathcal{F}_S$, we insist that the first n columns of $\mathbf{U}$ form an invertible matrix.

5.4 Recovering the Square Map

We now assume that we have recovered the first column, $[t_{i,0}]$, of $\widehat{\mathbf{T}}$ and that we have recovered $\mathbf{U}$. Let $\widehat{\mathbf{U}}$ represent the matrix consisting of the first $d+o-l$ columns of $\mathbf{U}$. By construction, $\widehat{\mathbf{U}}$ is invertible. We set $\mathbf{U} = \begin{bmatrix}\widehat{\mathbf{U}}\ \widehat{\mathbf{U}}'\end{bmatrix}$.

We can now explicitly compute

$$\sum_{i=0}^{m-1} t_{i,0}\mathbf{P}_i = \mathbf{W}\mathbf{F}^{*0}\mathbf{W}^\top.$$

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

Note that

$$\mathbf{W} = \mathbf{U}\widehat{\mathbf{M}}_d = \left[\widehat{\mathbf{U}}\ \widehat{\mathbf{U}}'\right]\begin{bmatrix}\mathbf{M}_d \\ \mathbf{0}_{o\times d}\end{bmatrix} = \widehat{\mathbf{U}}\begin{bmatrix}\mathbf{M}_d \\ \mathbf{0}_{(o-l)\times d}\end{bmatrix}.$$

Thus we have

$$\sum_{i=0}^{m-1} t_{i,0}\mathbf{P}_i = \widehat{\mathbf{U}}\begin{bmatrix}\mathbf{M}_d \\ \mathbf{0}_{(o-l)\times d}\end{bmatrix}\mathbf{F}^{*0}\left[\mathbf{M}_d^\top\ \mathbf{0}_{d\times(o-l)}\right]\widehat{\mathbf{U}}^\top.$$

Therefore, we may compute

$$\begin{bmatrix}\mathbf{M}_d \\ \mathbf{0}_{(o-l)\times d}\end{bmatrix}\mathbf{F}^{*0}\left[\mathbf{M}_d^\top\ \mathbf{0}_{d\times(o-l)}\right] = \widehat{\mathbf{U}}^{-1}\left(\sum_{i=0}^{m-1} t_{i,0}\mathbf{P}_i\right)\widehat{\mathbf{U}}^{-\top}, \tag{4}$$

Now, by taking the top left $d \times d$ submatrix, we recover $\mathbf{M}_d\mathbf{F}^{*0}\mathbf{M}_d^\top$. Finally, by multiplying on the left by $\mathbf{M}_d^{-1}$ and on the right by $\mathbf{M}_d^{-\top}$, we recover $\mathbf{F}^{*0}$.

5.5 Unmixing the Rainbow and Plus Polynomials

Having identified the vinegar subspace of linear forms on the input variables, we can identify the Rainbow polynomials as those linear combinations of the public polynomials which become linear when their inputs are restricted to the kernel of those linear forms. In other words, we can find the Rainbow polynomials by linearly solving for t_i such that:

$$\left[\mathbf{0}_{(o-l)\times d}\mathbf{I}_{o-l}\right]\widehat{\mathbf{U}}^{-1}\left(\sum_{i=0}^{m-1} t_i\mathbf{P}_i\right)\widehat{\mathbf{U}}^{-\top}\begin{bmatrix}\mathbf{0}_{d\times(o-l)} \\ \mathbf{I}_{o-l}\end{bmatrix} = 0. \tag{5}$$

A basis $t_{i,j}$ of the solution space of this equation forms the columns $d+1$ through $d+o+r$ of $\mathbf{T}^{-1}$. We can place any selection of column vectors in the last s columns of $\mathbf{T}^{-1}$ making it full rank, since no party is concerned with the values of the plus polynomials.

Having recovered the complete transformation $\mathbf{T}^{-1}$, we can compute the Rainbow and Plus part of the central map by

$$\begin{aligned}&(\mathbf{F}_{s,0},\ldots,\mathbf{F}_{S,d-1},\mathbf{F}_{R,0},\ldots,\mathbf{F}_{R,o+r-1},\mathbf{F}_{P,0},\ldots,\mathbf{F}_{P,s-1})\\ &= (\widehat{\mathbf{U}}^{-1}\mathbf{P}_0\widehat{\mathbf{U}}^{-\top},\ldots,\widehat{\mathbf{U}}^{-1}\mathbf{P}_m\widehat{\mathbf{U}}^{-\top})\mathbf{T}^{-1}.\end{aligned} \tag{6}$$

Algorithm 1 shows the process of our attack in algorithmic form. In the appendix of this paper, we illustrate our attack using a toy example.

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

Algorithm 1 Our Key Recovery Attack on SRP

Input: SRP parameters (o, d, r, s, l), SRP public key $\mathcal{P} : \mathbb{F}^{n'} \to \mathbb{F}^m$
Output: equivalent private key $(\mathcal{T}, (\mathcal{F}_S, \mathcal{F}_R, \mathcal{F}_P), \mathcal{U})$

1: Solve a MinRank problem on the m public polynomials with target rank 1. Denote the solution by $v \in \mathbb{F}^m$.
2: Define the elements of the $m \times d$ matrix $\hat{\mathbf{T}}'$ by $\hat{t_{ij}}' = v_i^{q^{j-1}}$ $(j = 1, \ldots, d)$.
3: Compute the first d columns of the matrix $\mathbf{T}^{-1}$ by $\mathbf{T}'^{-1} = \hat{\mathbf{T}} \cdot \mathbf{M}_d^{-1}$.
4: Let $\mathbf{K}$ be the $(n-1) \times n$ matrix representing the left kernel of the low rank matrix $\sum_{i=0}^{m-1} t_{i,0}\mathbf{P}_i$ and choose an element $w \in \mathbb{F}^n$ of its right kernel.
5: Define the elements of the $n \times d$ matrix $\mathbf{W}$ by $w_{ij} = w_i^{q^{j-1}}$ $(j = 1, \ldots, d)$
6: Recover the first d columns of the matrix $\mathbf{U}$ by equation (3).
7: Extend $\mathbf{U}$ to an invertible $n \times n$ matrix $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{U}}$ to a full rank $n \times (d + o)$ matrix $\mathbf{U}$.
8: Recover the map $\mathcal{F}_S$ by equation (4).
9: Compute the columns $d + 1, \ldots, d + o + r$ of the matrix $\mathbf{T}^{-1}$ by solving the linear system of equation (5). Append randomly columns to get an invertible $m \times m$ matrix $\mathbf{T}^{-1}$.
10: Recover the matrices representing the Rainbow and plus polynomials by equation (6).

6 Complexity of Attack

To estimate the complexity of our attack, we compute the Hilbert series of the ideal generated by the 2×2 minors of

$$\sum_{i=0}^{m-1} t_{i,0}\mathbf{P}_i.$$

We can then recover the degree of regularity d_{reg} explicitly.

Theorem 3 *Let $\mathbb{E}[T] = \mathbb{E}[t_{0,0}, \ldots, t_{m-1,0}]$. Let I be the ideal generated by the system of minors arising from the minors modeling variant of the KS-attack on SRP(q, d, o, r, s, l) with $d > l$, $n = d + o - l$ and $m = d + o + r + s$. Then the Hilbert series of I (that is, the Hilbert Series of $\mathbb{E}[T]/I$) is*

$$\text{Hilbertseries}(t) = 1 + mt.$$

Consequently the degree of regularity of the minors system is $d_{reg} = 2$.

Proof. Consider the ideal I generated by the 2×2 minors over $\mathbb{E}[T]$. There are $\binom{n}{2}^2/2$ distinct 2×2 minors in an $n \times n$ symmetric matrix; however, each such minor of the above matrix is a homogeneous quadratic polynomial in m variables. Thus the dimension of the span of the 2×2 minors is $\binom{m}{2} + m = \binom{m+1}{2}$. As a consequence, $\binom{m+1}{2}$ randomly chosen minors should be linearly independent with probability approximately $1 - \frac{1}{q}$.
Since I contains all linear combinations of the minors, I contains all quadratic

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

monomials in $\mathbb{E}[T]$. Thus $\mathbb{E}[T]/I$ contains representatives of exactly all equivalence classes of degree less than two. Therefore, the Hilbert Series of $\mathbb{E}[T]/I$ is

$$HS(t) = 1 + mt.$$

Technically, the ideal I in Theorem 3 is not what we use in the attack. We use $I' = \langle I, t_{0,0} - 1 \rangle$, for example. However, adding polynomials to I cannot increase the degree of regularity; thus, the degree of regularity in the actual attack is still two.

This fact proves that we actually require no Gröbner basis algorithm for the attack. Simple linearization and Gaussian elimination are effective in breaking all parameters.

Specifically, recalling that with one variable fixed we have only $m - 1$ variables, we may use the above calculation to estimate the complexity of recovering the first column of $\widehat{\mathbf{T}}$ using the minors modeling variant of the KS-attack.

Unmixing the Rainbow and plus polynomials only requires $2m$ matrix multiplications of dimension n matrices and solving a linear system in m variables. The complexity of these operations is on the order of $m^{\omega+1}$, and is therefore dominated by the minors modeling step. Thus we obtain the following

Theorem 4 *The complexity of our key recovery attack on SRP(q, d, o, r, s, l) with $d > l$, $n = d+o-l$ and $m = d+o+r+s$ using the minors modeling variant of the KS-attack is*

$$\mathcal{O}\left(\left(\binom{n+1}{2}^{\omega}\right)\left(\right.\right.$$

where $2 < \omega \leq 3$ is the linear algebra constant.

7 Experimental Results

In order to estimate the complexity of our attack in practice, we created a straightforward implementation of the key generation process of SRP and our attack in MAGMA Code. While the experiments were run on large servers with multiple cores, we used, for each of our experiments, only a single core.

Table 1 shows, for different parameter sets, the results of our experiments. The numbers in rows 3 and 10 show the time needed to solve the MinRank problem and to recover the maps $\mathcal{F}_S$ and $\mathcal{U}$ as well as the first d columns of the matrix $\mathbf{T}^{-1}$. The numbers in row 4 and 11 show the time needed to recover the remaining columns of $\mathbf{T}^{-1}$ and the maps $\mathcal{F}_R$ and $\mathcal{F}_P$. The numbers in the fifth and twelfth row show the overall running time of our attack.

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

parameters (q,d,o,r,s,l)	(31,16,16,8,3,8)	(31,24,24,12,4,12)	(31,35,35,15,5,15)
(m, n)	(43,24)	(64,36)	(90,55)
time for recovering $\mathcal{F}_S$ (s)	10.0	74.5	1,295
time for recovering $\mathcal{F}_R$ and $\mathcal{F}_P$ (s)	0.5	2.5	16.5
time (overall) (s)	10.5 [1]	77.1 [1]	1,313 [1]
memory (MB)	354.6	1,970.3	11,867
claimed security level (bit)	80	112	160
parameters(q,d,o,r,s,l)	(31,33,32,16,5,16)	(31,47,47,22,5,22)	(31,71,71,32,5,32)
(m, n)	(86,49)	(121,72)	(179,110)
time for recovering $\mathcal{F}_S$ (s)	487.0	9,705	27,306
time for recovering $\mathcal{F}_P$ and $\mathcal{F}_R$	10.0	69.1	183
time (overall)	497.0 [1]	9,777 [1]	27,494[2]
memory (MB)	8,518.5	47,988	315,407

Table 1. Running time of the proposed attack

[1)] AMD Opteron @ 2.4 GHz, 128GB RAM
[2)] Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz, and 512GB RAM

As the second column of the table shows, doubling the parameters leads to an increase of the running time and memory requirements of our attack by factors of about 50 and 25, which corresponds to our theoretical estimations.[4]
The parameter sets shown in the bottom half of Table 1 are those proposed by the authors of [7] for security levels of 80, 112 and 160 bit respectively. As the table shows, we could (even with our limited resources and poorly optimized attack) break the parameter sets proposed for 80 and 112 bit security in very short time. Since, for a security level of 160 bit, the memory requirements exceeded our possibilities, we had to run these experiments on another server. We want to thank Nadia Heninger for running these experiments.

8 Conclusion

In this paper we propose a practical attack against the SRP encryption scheme of Yasuda and Sakurai [7]. Our attack uses the min-Q-rank property of the scheme to recover parts of the linear transformation $\mathcal{T}$, the transformation $\mathcal{U}$ and the Square part $\mathcal{F}_S$ of the central map. Following this, we use the known structure of the Rainbow polynomials to recover the second half of the map $\mathcal{T}$ as well as the Rainbow and Plus part of the central map. Our attack is very efficient and breaks the SRP instances proposed in [7] in reasonable short time.
Therefore, our attack shows that the security of a weak multivariate scheme like Square is not automatically increased by combining it with another (secure) scheme.

[4] For larger parameters, the memory access time plays a major role in the overall running time. Therefore the corresponding factors are much larger.

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

Acknowledgements

We thank the anonymous reviewers for their comments which helped to improve the paper. Furthermore, we want to thank Nadia Heninger and Cisco for their help with running our experiments.

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

References

1. Bernstein, D.J., Buchmann, J., Dahmen, E., eds.: Post-quantum cryptography. Springer (2009).
2. Chen, A.I.T., Chen, M.S., Chen, T.R., Cheng, C.M., Ding, J., Kuo, E.L.H., Lee, F.Y.S., Yang, B.Y.: SSE implementation of Multivariate PKCs on modern x86 CPUs. CHES 2009, LNCS vol. 5747, pp. 33 -48. Springer (2009).
3. Bogdanov, A., Eisenbarth, T., Rupp, A., Wolf, C.: Time-Area optimized public-key engines: MQ-Cryptosystems as replacement for elliptic curves? CHES 2008, LNCS vol. 5154, pp. 45 -61. Springer (2008).
4. Kipnis, A., Patarin, J., Goubin, L.: Unbalanced oil and vinegar signature schemes. EUROCRYPT 1999, LNCS vol. 1592, pp. 206 - 222. Springer (1999).
5. Ding, J., Schmidt, D.: Rainbow, a new multivariable polynomial signature scheme. ACNS 2005, LNCS vol. 3531, pp. 164 - 175. Springer (2005).
6. Petzoldt, A., Chen, M.S., Yang, B.Y., Tao, C., Ding, J.: Design principles for HFEv- based multivariate signature schemes. ASIACRYPT 2015 (Part 1), LNCS vol. 9742 , pp. 311 -334. Springer (2015).
7. Yasuda, T., Sakurai, K.: A multivariate encryption scheme with Rainbow. ICISC 2015, LNCS vol. 9543, pp. 222 - 236. Springer (2015).
8. Clough, C., Baena, J., Ding, J., Yang, B.Y., Chen, M.S.: Square, a new multivariate encryption scheme. CT-RSA 2009, LNCS vol. 5473, pp. 252 - 264. Springer (2009) .
9. Ding, J., Gower, J.E., Schmidt, D.S.: Multivariate public key cryptosystems. Advances in Information Security vol. 25. Springer (2006).
10. Garey, M.R., Johnson, D.S.: Computers and intractability: A guide to the theory of NP-completeness. A Series of Books in the Mathematical Sciences, W. H. Freeman and Company (1979).
11. Kipnis, A., Shamir, A.: Cryptanalysis of the HFE public key cryptosystem by Relinearization. CRYPTO 1999, LNCS vol. 1666, pp. 19 - 30. Springer (1999).
12. Faugére, J.C.: Algebraic cryptanalysis of Hidden Field Equations (HFE) using Gröbner bases. CRYPTO 2003, LNCS vol. 2729, pp. 44 - 60. Springer (2003).
13. Bettale, L., Faugére, J., Perret, L.: Cryptanalysis of HFE, multi-HFE and variants for odd and even characteristic. Designs Codes and Cryptography 69 (2013), pp. 1 - 52.

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

14. Cabarcas, D., Smith-Tone, D., Verbel, J.A.: Key recovery attack for ZHFE. PQCrypto 2017, LNCS vo. 10346, pp. 289 - 308. Springer (2017).
15. Vates, J., Smith-Tone, D.: Key recovery for all parameters of HFE-. PQCrypto 2017, LNCS 10346, pp. 272 -288. Springer (2017).

A Toy Example

In the following we illustrate our attack using a toy example with small parameters.

A.1 Key Generation

For our toy example we use GF(7) as the underlying field. We choose the parameters of SRP as $(d, o, r, s, l) = (2, 2, 1, 1, 1)$.[5] Therefore our public key consists of six equations in three variables. The Square map is defined over the extension field GF(7)[X] / $\langle X^2 + 6X + 3\rangle$. For simplicity, we restrict to linear maps $\mathcal{T}$ and $\mathcal{U}$ as well as homogeneous quadratic maps $\mathcal{F}_R$ and $\mathcal{F}_P$. By doing so, the public key $\mathcal{P}$ of our scheme will be homogeneous quadratic, too.

Let the linear maps $\mathcal{T}$ and $\mathcal{U}$ be given by the matrices

$$\mathbf{T} = \begin{pmatrix} 1 & 5 & 1 & 6 & 3 & 3 \\ 5 & 3 & 5 & 2 & 2 & 5 \\ 0 & 4 & 0 & 4 & 5 & 0 \\ 0 & 6 & 6 & 2 & 4 & 3 \\ 3 & 3 & 6 & 3 & 6 & 3 \\ 5 & 3 & 5 & 0 & 4 & 6 \end{pmatrix} \in \mathbb{F}^{6\times 6} \quad \text{and} \quad \mathbf{U} = \begin{pmatrix} 6 & 0 & 3 & 2 \\ 2 & 0 & 0 & 4 \\ 4 & 1 & 1 & 0 \end{pmatrix} \in \mathbb{F}^{3\times 4}.$$

The Square map $\mathcal{F}_S(X) = X^2$ is given by the matrix $\mathbf{F} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{F}^{2\times 2}$.

Let the three Rainbow polynomials be given by the 4×4 matrices

$$\mathbf{F}_{R,0} = \begin{pmatrix} 2 & 6 & 2 & 3 \\ 6 & 1 & 6 & 0 \\ 2 & 6 & 0 & 0 \\ 3 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{F}_{R,1} = \begin{pmatrix} 2 & 1 & 5 & 1 \\ 1 & 5 & 0 & 6 \\ 5 & 0 & 0 & 0 \\ 1 & 6 & 0 & 0 \end{pmatrix}, \text{ and } \mathbf{F}_{R,2} = \begin{pmatrix} 5 & 4 & 3 & 0 \\ 4 & 2 & 0 & 1 \\ 3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

The Plus polynomial is given by the 4×4 matrix

$$\mathbf{F}_{P_0} = \begin{pmatrix} 3 & 4 & 3 & 2 \\ 4 & 4 & 0 & 3 \\ 3 & 0 & 5 & 0 \\ 2 & 3 & 0 & 3 \end{pmatrix}.$$

[5] Note that this parameter choice does not meet the description in Section 2.2, where d was required to be odd. However, an odd value of d is only needed for the efficient decryption. The scheme itself can be defined for any value of d.

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

We compute the public key of our scheme by $\mathcal{P} = \mathcal{T} \circ (\mathcal{F}_S, \mathcal{F}_R, \mathcal{F}_P) \circ \mathcal{U}$ and obtain the following 6 3×3 matrices representing $\mathcal{P}$

$$\mathbf{P}_0 = \begin{pmatrix} 6 & 6 & 0 \\ 6 & 6 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \ \mathbf{P}_1 = \begin{pmatrix} 5 & 2 & 5 \\ 2 & 3 & 4 \\ 5 & 4 & 6 \end{pmatrix}, \ \mathbf{P}_2 = \begin{pmatrix} 6 & 4 & 2 \\ 4 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix}$$
$$\mathbf{P}_3 = \begin{pmatrix} 4 & 5 & 3 \\ 5 & 6 & 3 \\ 3 & 3 & 3 \end{pmatrix}, \ \mathbf{P}_4 = \begin{pmatrix} 5 & 1 & 5 \\ 1 & 1 & 4 \\ 5 & 4 & 3 \end{pmatrix}, \text{ and } \mathbf{P}_5 = \begin{pmatrix} 2 & 4 & 6 \\ 4 & 3 & 1 \\ 6 & 1 & 3 \end{pmatrix}.$$

A.2 Recovery of Transformation of Square Polynomials

In the first step of the attack, we have to solve a MinRank problem on the 6 matrices $\mathbf{P}_0, \ldots, \mathbf{P}_5$ with target rank 1. One solution is given by

$$v = (1, b^{19}, b^{13}, b^9, b^{47}, b^9),$$

where b is a generator of the extension field $\mathbb{E}$=GF(7^2).

From this, we obtain the first part of the linear transformation $\mathcal{T}$ which divides the Square part from the remaining polynomials. Let $\widehat{\mathbf{T}}'$ represent the first d columns of $\widehat{\mathbf{T}}$. We may recover the first d columns of $\mathbf{T}^{-1}$ via right multiplication by $\mathbf{M}_d^{-1}$.

$$\widehat{\mathbf{T}}' = \begin{pmatrix} 1 & 1 \\ b^{19} & b^{37} \\ b^{13} & b^{43} \\ b^9 & b^{15} \\ b^{47} & b^{41} \\ b^9 & b^{15} \end{pmatrix}, \ \mathbf{T}^{-1\prime} = \widehat{\mathbf{T}}'\mathbf{M}_d^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 3 & 3 \\ 0 & 3 \\ 5 & 2 \\ 0 & 3 \end{pmatrix}.$$

Note that the entries in the second column of $\widehat{\mathbf{T}}'$ are just the Frobenius powers of the first column entries.

A.3 Recovery of the Input Transformation $\mathcal{U}$

Next we can use the first column, $[t_{i,0}]$, of $\widehat{\mathbf{T}}'$ to recover the first d columns of the matrix representation of the linear transformation $\mathcal{U}$, thus separating the vinegar subspace from the oil subspace. To accomplish this, we construct our rank one solution to the MinRank step

$$L = \sum_{i=0}^{m-1} t_{i,0}\mathbf{P}_i = \begin{pmatrix} b^{45} & b^3 & b^{18} \\ b^3 & b^9 & 6 \\ b^{18} & 6 & b^{39} \end{pmatrix}.$$

Let K be the left kernel of L and construct the reduced row echelon form matrix $\mathbf{K}$ whose rows form a basis of K.

$$\mathbf{K} = \begin{pmatrix} 1 & 0 & b^3 \\ 0 & 1 & b^9 \end{pmatrix}.$$

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

Any element in the right kernel of $\mathbf{K}$ forms the first column of $\mathbf{W}$. The second column is the first Frobenius power of the first. For a random selection we obtain

$$\mathbf{W} = \begin{pmatrix} b^{45} & b^{27} \\ b^{3} & b^{21} \\ b^{18} & b^{30} \end{pmatrix}.$$

We next recover the first $d = 2$ columns of U via the relation

$$\mathbf{W}\mathbf{M}_d^{-1} = \mathbf{U}\begin{bmatrix} \mathbf{I}_d \\ \mathbf{0}_{o\times d} \end{bmatrix} = \begin{pmatrix} 5 & 5 \\ 4 & 5 \\ 1 & 2 \end{pmatrix}.$$

Extending this matrix, we construct the invertible

$$\widehat{\mathbf{U}} = \begin{pmatrix} 5 & 5 & 0 \\ 4 & 5 & 0 \\ 1 & 2 & 1 \end{pmatrix}.$$

We may now extend this matrix to any $n \times n + l$ matrix. The simplest way is to append zeros. This technique is always effective due to the isomorphism described at the beginning of Section 5. Thus we obtain

$$\mathbf{U} = \begin{pmatrix} 5 & 5 & 0 & 0 \\ 4 & 5 & 0 & 0 \\ 1 & 2 & 1 & 0 \end{pmatrix}.$$

A.4 Recovering $\mathcal{F}_S$

Knowing $\mathbf{T}^{-1'}$ and $\widehat{\mathbf{U}}$, we can recover the Square part of the central map. Specifically, we recover the top left 2×2 submatrix of $\widehat{\mathbf{U}}^{-1} L \widehat{\mathbf{U}}^{-\top}$:

$$\mathbf{F}^{*0} = \begin{pmatrix} b^{3} & 0 \\ 0 & 0 \end{pmatrix}.$$

A.5 Recovering $\mathcal{F}_R$ and $\mathcal{F}_P$

We solve the equation

$$\begin{bmatrix} \mathbf{0}_{(o-l)\times d} & \mathbf{I}_{o-l} \end{bmatrix} \widehat{\mathbf{U}}^{-1} \left(\sum_{i=0}^{m-1} t_i \mathbf{P}_i \right) \widehat{\mathbf{U}}^{-\top} \begin{bmatrix} \mathbf{0}_{d\times(o-l)} \\ \mathbf{I}_{o-l} \end{bmatrix}$$

for t_i and append $o + r = 3$ linearly independent solutions as column vectors onto $\mathbf{T}^{-1'}$. The final $s = 1$ column(s) of $\mathbf{T}^{-1}$can be chosen randomly to achieve full rank. Our random selection produces

$$\mathbf{T}^{-1} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 5 \\ 1 & 3 & 0 & 0 & 0 & 6 \\ 3 & 3 & 2 & 6 & 4 & 3 \\ 0 & 3 & 1 & 5 & 4 & 6 \\ 5 & 2 & 2 & 0 & 2 & 1 \\ 0 & 3 & 1 & 0 & 2 & 1 \end{pmatrix}.$$

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

Now with $\mathbf{T}^{-1}$ we can recover explicitly the Rainbow and Plus polynomials. To do so, we compute

$$(\widehat{\mathbf{U}}^{-1}\mathbf{P}_0\widehat{\mathbf{U}}^{-\top}, \ldots, \widehat{\mathbf{U}}^{-1}\mathbf{P}_{m-1}\widehat{\mathbf{U}}^{-\top})\mathbf{T}^{-1}.$$

We may now express the Rainbow and Plus polynomials as quadratic forms in n variables by appending l rows and columns of arbitrary values, since our choice of $\mathbf{U}$ makes these entries obsolete. We obtain

$$\mathbf{F}_{R,0} = \begin{pmatrix} 0 & 5 & 2 & 0 \\ 5 & 4 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \mathbf{F}_{R,1} = \begin{pmatrix} 0 & 0 & 6 & 0 \\ 0 & 2 & 0 & 0 \\ 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \mathbf{F}_{R,2} = \begin{pmatrix} 5 & 4 & 0 & 0 \\ 4 & 4 & 5 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

and

$$\mathbf{F}_{P,0} = \begin{pmatrix} 4 & 5 & 2 & 0 \\ 5 & 4 & 1 & 0 \\ 2 & 1 & 5 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Via composition, one verifies that

$$\mathcal{P} = \mathcal{T} \circ (\mathcal{F}_S, \mathcal{F}_R, \mathcal{F}_P) \circ \mathcal{U}.$$

Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Total Break of the SRP Encryption Scheme."
Paper presented at Selected Areas in Cryptography (SAC 2017), Ottawa, ON, Canada. August 16, 2017 - August 18, 2017.

SDR-Based Experiments for LTE-LAA Based Coexistence Systems with Improved Design

Yao Ma, Ryan Jacobs, Daniel G. Kuester, Jason Coder, and William Young
Communications Technology Laboratory, National Institute of Standards and Technology
325 Broadway, Boulder, Colorado, USA

***Abstract*— To enhance spectrum efficiency in next generation heterogeneous wireless systems, it is important to improve shared spectrum usage between unlicensed long-term evolution (LTE) systems, such as the license-assisted access (LAA), and legacy systems, such as wireless local area networks (WLANs). LTE-LAA uses listen-before-talk (LBT) schemes to enhance coexistence performance. However, available LAA-LBT schemes may incur several problems, such as a slot-jamming effect and a significant slot boundary tracking error, leading to degraded performance. In this paper, we develop an LBT scheme with improved design and software-defined radio (SDR)-based experimental procedure for a performance validation. We program the improved LBT algorithm in the SDR field-programmable gate array (FPGA), and compare the results with the original LBT algorithm. This work also presents an automated testing technique and new SDR functions that modify the LAA parameters in realtime (such as backoff idle slot durations). The experiment results confirm the predicted slot-jamming effect, and show that our improved design enhances LAA throughput. These results provide not only a more robust LAA-LBT design, but also a new method of SDR programming and testing, and insight into the coexistence performance of heterogeneous systems.**

Index Terms: Coexistence; FPGA programming; LTE-LAA; SDR experiments; Test automation; WLAN.

I. Introduction

Experimental design and evaluation are critical for performance validation of spectrum sharing techniques between long-term evolution license-assisted access (LTE-LAA) [1]–[6], [11] systems and incumbent systems, such as IEEE 802.11 wireless local area networks (WLANs). To support LTE system development and coexistence, software-defined radio (SDR) or commercial off-the-shelf (COTS)-based experimental methods have become popular, and various SDR test platforms are reported in [3], [7]–[9], [12].

The 3rd Generation Partnership Project (3GPP) LAA has defined four categories of listen-before-talk (LBT) schemes [1], [2] to improve coexistence with legacy systems. In LAA-based coexistence scenarios, LAA nodes may have a larger backoff slot duration than the WLAN counterpart. In [1], [2], LAA uses an extended clear channel assessment (eCCA) slot in the transmission backoff process. The eCCA duration is flexible, and may be 9 μs to 20 μs, or larger. The WLAN backoff slot duration is 9 μs (which includes CCA time) for several popular physical layer specifications [13]. Various coexistence settings based on LTE-LAA and WLAN transmissions have been evaluated, and laboratory and field test results are reported in [1]–[3], [12]. In particular, SDR devices were used in LAA and WLAN coexistence testing in [12]. However, none of these experimental results report the effect of heterogeneous LAA and WLAN idle slot durations on coexistence performance. In [17], we study the effect of heterogeneous idle slot duration and point out a slot-jamming (SJ) effect of WLAN nodes against LAA nodes in the transmissions backoff process. An anti-SJ-LBT scheme is proposed and its performance is analyzed and verified through computer simulation. Yet, experimental validation of the slot-jamming effect and its mitigation was still absent.

In an LAA transmission cycle, the time slots are divided into idle (backoff), transmission, channel-busy slots, etc. The slot-tracking in different transmission links need to align to maintain a satisfactory performance. In practice, synchronism (or near-synchronism) of slot boundaries has been assumed in the majority of system design, analysis and optimization results [1], [2], [14]–[20], though very often not explicitly discussed. However, the tracking of slot boundaries, such as in the instant when the channel switches from busy to idle, is a nontrivial issue. The LBT process defined in [1], [2] uses an initial CCA (iCCA) duration or a deferred extended CCA (DeCCA) duration to track this channel state change. The iCCA and DeCCA durations are typically suggested to be equal to the WLAN distributed coordination function interframe space (DIFS) duration, which is 34 μs for the IEEE 802.11a, 802.11n and 802.11ac specifications in the 5 GHz industrial, scientific, and medical (ISM) band [13]. To search for the instant that the channel turns from busy to idle, the LBT uses DeCCA duration as search step size for channel status tracking. However, since the DeCCA duration is much larger than the WLAN backoff idle slot duration, the original LBT design may lead to a significant slot-tracking error. Unsatisfactory slot-tracking accuracy can cause increased transmission collisions and reduced throughput. To our knowledge, the effect of this type of slot boundary tracking error has not been investigated in the literature of LAA-based coexistence systems. Experiment-based study and countermeasure design are important to address this problem.

In this paper, we address the SJ and the slot-tracking error problems in the original LBT process, develop an improved LBT scheme with anti-SJ and subslot-tracking features, and

Certain commercial equipment, instruments, or software are identified in this paper in order to adequately specify the experimentation procedure. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the software or equipment identified are necessarily the best available for the purpose.

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao; Young, William.
"SDR-Based Experiments for LTE-LAA Based Coexistence Systems with Improved Design."
Paper presented at 2017 IEEE Globecom, Singapore, Singapore. December 4, 2017 - December 8, 2017.

implement SDR-based experimental validation. In [3], [7]–[9], [12], the capability of changing LAA LBT parameters in realtime for automated testing was neither discussed nor demonstrated. We note that there is a significant technical gap that needs to be bridged to allow flexible system parameter updates in realtime.

For this purpose, we program the improved LAA LBT function and WLAN functions, based on major modifications of available SDR software related to [12]. We implement SDR field-programmable gate array (FPGA) programming to allow flexible LAA LBT functions, where iCCA and eCCA slot durations can be modified in realtime. We also add user datagram protocol (UDP)-based communication functions to switch parameters and data between the host code and SDR-control codes. Then, we implement conducted tests assuming an additive white Gaussian noise (AWGN) channel in the LAA link, the WLAN link, and the LAA-WLAN mutual interference link.

The contributions and novelty of this paper are summarized as follows:

- We develop an improved LBT method to reduce the slot-jamming effect and slot-tracking error. This scheme is compatible with the existing 3GPP LAA LBT framework.
- We develop an FPGA programming method which implements our new LBT algorithm, and allows the updating of many LAA LBT parameters in realtime. This also includes a test-automation method between computer and SDR devices.
- Experimental results confirm the SJ effect, verify our analysis of the anti-SJ-LBT scheme in [17], and demonstrate enhanced throughput performance of the new LBT method relative to the original LBT.

These results address critical problems of LBT SJ effect and slot boundary alignment (tracking) error in system coexistence, and provide a countermeasure scheme. They also provide new SDR programming and demonstrate the capability to implement automated testing of coexisting systems.

II. System Model and Improved Design

Our testing scenario supposes that LTE-LAA nodes utilize unlicensed spectrum in the downlink and share it with incumbent WLAN nodes. The LTE-LAA system uses LBT to track channel status and makes backoff or transmission decisions. Refer to Fig. 1. After the transmission opportunity (TXOP) of an active link (say, a WLAN link) is completed, the transmission node starts the DIFS, followed by random backoff slots with slot duration of 9 μs. Here, for the sake of brevity we do not show the short interframe space (SIFS) and acknowledgement (ACK) signal durations. Note that the active link may also be an LAA link.

Suppose that there are three LAA nodes listening to the channel. To facilitate smooth coexistence, we assume that $T_{\text{DIFS}} = T_{\text{Defer}}$, where T_{DIFS} and T_{Defer} are WLAN DIFS and LAA deferred eCCA durations, respectively. Ideally, after the transmission of the WLAN link is completed, all the listening nodes would immediately start the DeCCA period.

The first link has perfect slot alignment with the WLAN link, and starts the eCCA slot synchronously with the WLAN

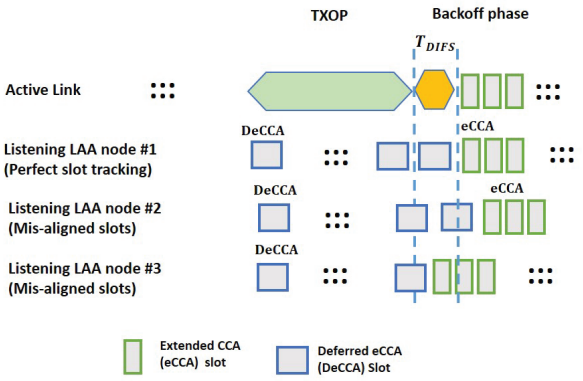


Fig. 1: The iCCA (or DeCCA) slot-tracking errors in the original LAA-LBT process.

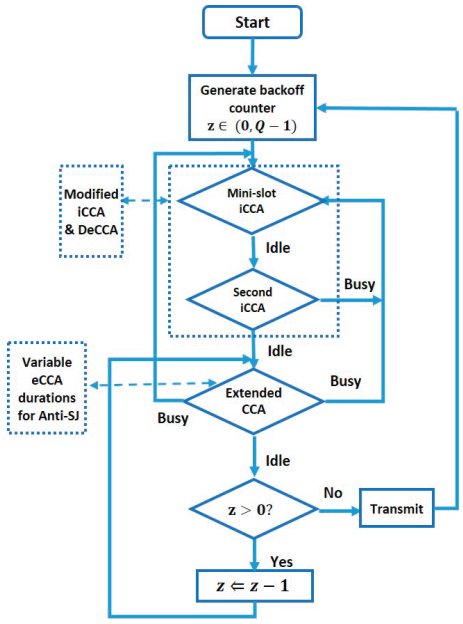


Fig. 2: An improved LAA-LBT scheme that reduces slot-tracking errors and the slot-jamming effect.

link. However, the second and third LAA nodes have major slot-tracking errors with respect to the WLAN link, so that node #2 starts the backoff count-down too late, and node #3 starts it too early.

We explain this phenomenon next. Based on the LBT process in [1], [2], during the channel busy time, the LAA node freezes its backoff process for a DeCCA period (which could be 34 μs as a default value). The energy detection (ED) output of each DeCCA duration at each LAA node is compared with a threshold to declare the channel busy or idle. After the channel status changes from busy to idle, the TXOP stop instant can lie in any range in a DeCCA window with duration 34 μs. This

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao; Young, William.
"SDR-Based Experiments for LTE-LAA Based Coexistence Systems with Improved Design."
Paper presented at 2017 IEEE Globecom, Singapore, Singapore. December 4, 2017 - December 8, 2017.

inherently causes a slot boundary alignment error in the range $(-17, 17)$ μs, or larger. Another weakness of this design is the sensitivity to the ED threshold setting. A low ED threshold may cause the DeCCA slot to be declared as channel-busy, and the LAA node will continue with one additional DeCCA slot, as shown for LAA node #2. A high ED threshold may cause the DeCCA slot be declared as channel-idle, and the LAA node will start the eCCA slot too early, as shown for LAA node #3. In a WLAN network, WLAN nodes may use a network allocation vector (NAV) in the packet header to determine the precise time that a transmission stops. However, we are not aware of an LAA specification that provides a method similar to the NAV for the LAA nodes to track the precise stop time of WLAN transmissions. Thus, the slot boundary tracking error becomes a significant issue in LAA and WLAN coexistence cases.

To address this problem, we propose a modified LBT process with subslot-tracking shown in Fig. 2. We merge the iCCA and DeCCA slots of the LBT for convenience, similar to that in [2]. As proposed in [16], we switched the order of blocks "Extended CCA" and "$z > 0$?" to remove the channel access priority of an LAA node, which just finished a successful transmission. Compared to available LBT schemes, such as those in [1], [2], [16], in Fig. 2 we split the iCCA/DeCCA slot into two sections, namely, "mini-slot iCCA" and "second iCCA" (where the word "DeCCA" is suppressed for convenience of notation). The "mini-slot iCCA" senses the channel with a much shorter duration than a regular iCCA/DeCCA slot, searching for the precise instant when a TXOP is over. Its purpose is to achieve precise slot boundary alignment. When the sensed signal power (or energy) exceeds the ED threshold (channel busy), the mini-slot iCCA is repeated until the channel is declared idle. When the ED result indicates an idle channel, the LBT process moves to the second iCCA sub-slot and senses for a larger time window. The purpose of this part is to be compatible with WLAN DIFS duration. Note that the sum duration of the "mini-slot iCCA" and "second iCCA" in Fig. 2 can be set equal to T_{DIFS} (34 μs), or any iCCA/DeCCA duration specified by LAA network designers.

This improved design may significantly reduce slot-tracking errors. Assuming that channel sensing in the "mini-slot iCCA" is reliable, the slot-tracking error is now bounded by the duration of the mini-slot iCCA (for example, 4 μs), instead of the regular DeCCA duration. This design is compatible with available 3GPP LBT design procedures, because it only involves a change of internal functions in the iCCA and DeCCA blocks. We programmed this method in FPGA code and loaded it onto SDR devices for verification.

Additionally, we experimentally study the impact of LBT SJ effect and the anti-SJ design. We define δ_L and δ_W as the backoff idle slot durations for LAA and WLAN systems, respectively. In [1], [2], the LBT backoff idle slot is called the LBT eCCA slot, with a flexible duration δ_L that can range from 9 μs to 20 μs, or larger. For several popular physical layer specifications [13] in the 5 GHz ISM band, the WLAN backoff slot duration is set to $\delta_W = 9$ μs. Our recent work [17] shows that as the ratio $N_s = \delta_L/\delta_W$ increases, there is an increasing chance the WLAN nodes jam the transmission opportunity of LAA nodes. We call this effect "slot jamming", where the LAA throughput decreases and WLAN throughput increases as N_s increases. We propose an anti-SJ-LBT in [17], which uses a variable eCCA duration based on channel status to enhance LAA transmission opportunity. This feature is shown in Fig. 2. Note that the original LBT uses a fixed eCCA duration, and we call it SJ-LBT in this paper.

III. Experimental Setup

Figure 3 shows a block diagram of the conducted measurement setup. A photo of part of this setup is shown in Fig. 4. The goal of this setup is to experimentally examine the slot-jamming and slot-tracking error effects, and the performance of the improved LBT design.

The software structure is shown in Fig. 5, which consists of host programming and SDR FPGA target programming. The host code runs on the computer, and includes a control interface. The control interface code that we programmed sends parameters to commercial LAA and WLAN codes via one set of UDP ports, reads test results via another set of UDP ports, and saves the results to data files of pre-specified formats on the computer hard drive for record and analysis. On the host computer two software projects run simultaneously and control the two SDR units, respectively. Two SDR devices emulate an LTE-LAA link and an WLAN link, respectively, or two LAA links. The communication channels were implemented with power combiner and splitter, and are AWGN channels, as shown in Fig. 3.

The first SDR device implements some LTE-LAA functions. The RF loopback mode was used: the eNode B (eNB, or base station) transmitted data to the user equipment (UE) via the RF ports and cable connections (downlink only), and did not receive any data from the UE via the RF channel (uplink). Some necessary handshaking signals in the uplink and synchronization functions are implemented internally on the FPGA, because the FPGA on a single SDR device emulates both eNB and UE functions.

The LTE-LAA code used on the SDR was a commercial implementation that we modified for use in these experiments. Several significant modifications to this commercial software included: implementation of new LBT functions, ability to adjust the LBT parameters in realtime, addition of a customized channel reservation signal (CRS) to mitigate LAA signal generation and transmission delay, and the addition of an automated testing ability.

The commercial FPGA code we used included an original LBT function but it was modified to add the improved LBT scheme with subslot-tracking and anti-SJ functions. The original DeCCA (or iCCA) state is split into two sub-states with duration of 4 μs and 30 μs, respectively. The state transitions among many LBT states are updated in FPGA, such as eCCA, mini-slot iCCA, iCCA, DeCCA, counter reduction, and transmission. Also, to implement the anti-SJ-LBT, the eCCA duration on FPGA is updated in realtime based on channel sensing results.

The ability to change the LBT parameters in realtime was achieved by adding several new registers (related to eCCA

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao; Young, William.
"SDR-Based Experiments for LTE-LAA Based Coexistence Systems with Improved Design."
Paper presented at 2017 IEEE Globecom, Singapore, Singapore. December 4, 2017 - December 8, 2017.

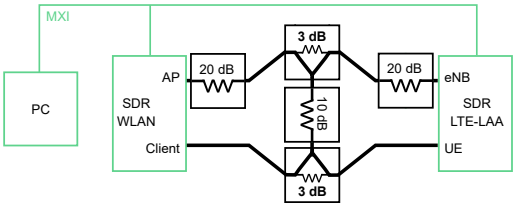


Fig. 3: Conducted SDR test setup for the LAA and WLAN system experiment (with two links).

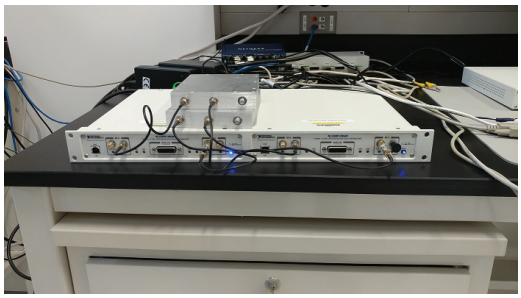


Fig. 4: Photo of part of the conducted SDR test setup.

and iCCA/DeCCA slot durations) to the FPGA code. These registers were also added to the FPGA-host interface.

The SDR software and devices we used had a significant LAA signal generation and transmission delay of about 50∼90 μs. This delay is defined as the duration starting from the instant that the LAA transmission decision is made until the instant that the LAA signal accesses the channel. Unfortunately, this delay can cause significant conflicts in channel access of LAA and WLAN nodes, and lead to increased collisions and degraded throughput. As a mitigation scheme, we added a new CRS generation function before the LAA signal transmission in FPGA code. This customized CRS is generated as soon as the LAA channel access decision is made, and transmitted with a negligible delay. This CRS has an adjustable duration in an host user-interface, and is seamlessly followed by the actual LAA signal transmission. This method basically solved the LAA transmission delay problem.

Finally, an automated testing interface was added to a host user-interface code that we wrote. This host code communicates via UDP with the commercial LAA and WLAN SDR codes that we modified with new FPGA functions. The modified LAA and WLAN codes accept and recognize automated commands and parameters and send back test results in realtime. These commands were not sent through the RF data link and did not cause overhead to the LAA and WLAN transmissions.

The WLAN implementation on the second SDR unit was achieved by use of commercially available software without additional FPGA programming. Additional changes were made to the host portion of the WLAN software, such as UDP-port communication and automated-testing modules. The WLAN communication link also used an RF loopback mode, similar to the case of LAA link. Only downlink WLAN traffic was simulated.

We modified the host software to allow online updating of several parameters in both LAA and WLAN links. The parameters include: transmit power, carrier frequency, ED threshold, throughput type, eCCA, mini-slot iCCA, and iCCA (or DeCCA) durations. However, during the measurements, only the eCCA duration was changed online.

During testing, the host SDR code sends commands to the SDR to update parameters at 0.5-second intervals, and receives test results from the SDR in one-second intervals. The result for each parameter setting is averaged over a specified number of received samples, such as 50. Then, the parameter of interest is updated. After requesting the next update, the software waits for ten seconds to allow the SDR devices to stabilize. Following the ten-second wait, the reading of the next set of samples begins.

This experimental setup enables the examination of coexistence theory and computer simulation in a highly controlled manner. Both the LTE and WLAN implementations used for conducted measurements are SDR representations designed for development purposes and are not commercially functioning networks.

IV. Experimental Results

In this section, we provide experimental results of the coexistence performance (Refer to Table I for values of some of the key parameters). We check the impact of two design features for LBT that may bring performance improvement for LAA links: subslot-tracking and anti-SJ. Note that the original LBT scheme includes neither the anti-SJ nor the subslot-tracking design.

We assume that the WLAN and LAA systems have channels fully overlapped with 20 MHz bandwidth at a center frequency 5.22 GHz in an ISM band. The LAA link uses modulation and coding scheme (MCS) 7 – quaternary phase-shift keying (QPSK) with rate 0.51, and the WLAN link uses QPSK with rate-1/2. When the transmission time efficiency is 100%, the physical layer channel bit rates (CBRs) are $\text{CBR}_L = 12.216$ Mbit/s for the LAA link, and $\text{CBR}_W = 12$ Mbit/s for the WLAN link. The LAA TXOP payload duration is given by $T_{P,L} = 2$ ms. The WLAN payload duration is computed by $T_{P,W} = 2048{\times}8/\text{CBR}_W = 1.4$ ms, where 2048 is the payload size in bytes per WLAN TXOP. The payload here may include physical layer headers and frame check sequence. The SDR-measured throughput for an LAA (or WLAN) link refers to correctly-decoded average data rate at receiver.

For the case of only one active LAA link (the WLAN link is not active), the baseline LAA throughput with Category-3 LBT is derived as

$$S_{L,only} = \frac{T_{p,L}\text{CBR}_L}{T_{p,L} + T_{\text{iCCA}} + \frac{Z_0-1}{2}\delta_L}, \quad (1)$$

where Z_0 is the LBT contention window (CW) size, and T_{iCCA} is the duration of initial CCA, set to be equal to LBT DeCCA duration and the WLAN DIFS duration (34 μs). Fig. 6 provides a comparison between the theoretical result in (1) and an SDR measurement result for the MCS-7 scheme when

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao; Young, William.
"SDR-Based Experiments for LTE-LAA Based Coexistence Systems with Improved Design."
Paper presented at 2017 IEEE Globecom, Singapore, Singapore. December 4, 2017 - December 8, 2017.

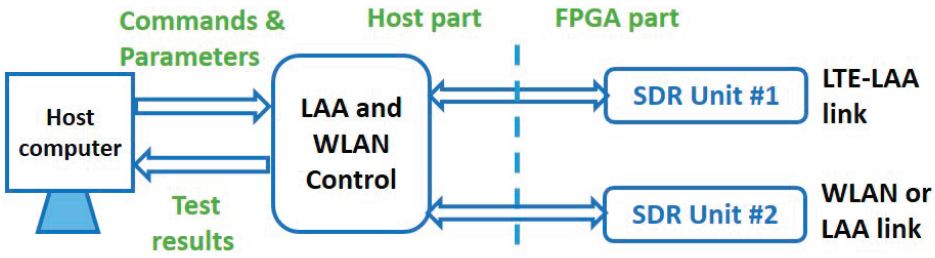


Fig. 5: Software control structure of the LAA and WLAN coexistence performance evaluation in conducted testings.

TABLE I: LTE-LAA and WLAN Parameters in Experimentation.

LAA parameters

Parameter	Value
Payload duration per transmission	2 ms
Transmit power	15 dBm
CCA ED threshold	-70 dBm
LBT defer period: T_{Defer} (=T_{DIFS})	34 μs
LBT eCCA period: T_{eCCA} (=$N_s\delta_W$)	$N_s \times 9$ μs
CW size Z_0	16

WLAN parameters

Parameter	Value
Payload duration per transmission	1.4 ms
Transmit power	10 dBm
CCA ED threshold	-72 dBm
T_{DIFS}	34 μs
Idle slot duration δ_W	9 μs
CW size W_0	16

only the LAA link is active. A good match between analytical and measurement results is observed, with relative error of no more than 2%. We have also verified baseline throughput for one active 802.11 WLAN link (while the LAA link is not active) on an SDR device, which approximately matches with the result calculated by use of a method given in [21]. The detail is omitted here for brevity.

Next, we check two cases of LAA-based coexistence: one LAA link with one WLAN link (shown in Figs. 7 and 8), and two LAA links (shown in Fig. 9). In Fig. 7, we show the throughput of LAA (with the original SJ-LBT) and WLAN links. The WLAN idle slot duration is fixed at 9 μs. We observe that as the eCCA duration increases from 9 μs to 45 μs, the LAA throughput decreases while WLAN throughput increases significantly. Also, the LAA throughput is enhanced substantially by using our proposed subslot-tracking method compared to the original LBT which has no feature of subslot-tracking. For example, at $\delta_L = 9\mu s$, subslot-tracking brings about 0.9 Mbit/s enhancement to the LAA throughput than the case without it.

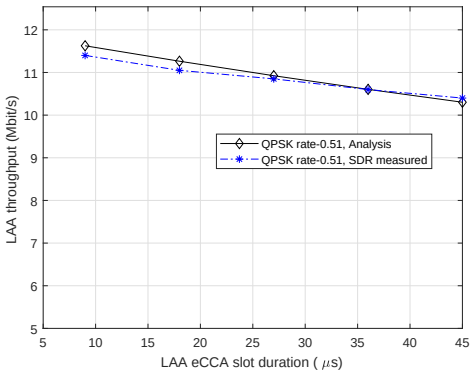


Fig. 6: Analytical and SDR-measured throughput of a single LTE-LAA link vs. eCCA slot duration.

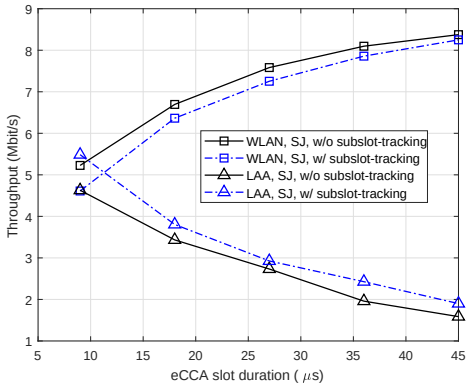


Fig. 7: SDR-measured throughput of the LAA with slot-jamming LBT and WLAN links vs. eCCA slot durations, either with or without subslot-tracking.

In Fig. 8, the throughput of LAA with the anti-SJ-LBT is provided. Comparison between Figs. 8 and 7 demonstrates that the anti-SJ-LBT provides a significantly higher throughput than the original SJ-LBT, either with or without subslot tracking. For example, with subslot-tracking, when $\delta_L = 45$ μs, the anti-SJ-LBT provides a throughput of about 3.2 Mbit/s and the

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao; Young, William.
"SDR-Based Experiments for LTE-LAA Based Coexistence Systems with Improved Design."
Paper presented at 2017 IEEE Globecom, Singapore, Singapore. December 4, 2017 - December 8, 2017.

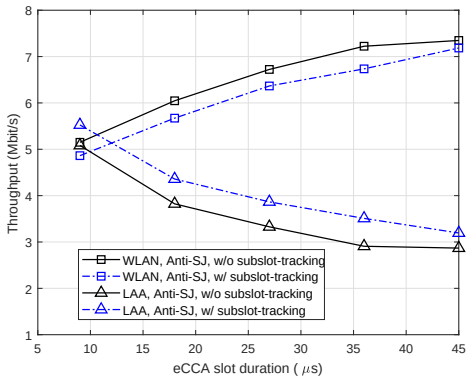


Fig. 8: SDR-measured throughput of the LAA with anti-slot-jamming LBT and WLAN links vs. eCCA slot durations, either with or without subslot-tracking.

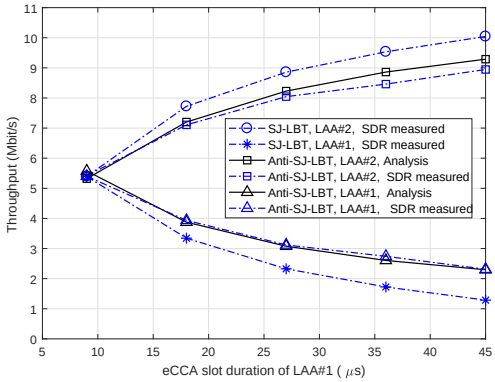


Fig. 9: Analytical and SDR-measured throughput of two LAA links with a fixed and a variable eCCA slot durations, respectively, with subslot-tracking.

original SJ-LBT has a throughput of about 2.0 Mbit/s. Results in Figs. 7 and 8 confirm the effectiveness of subslot-tracking and anti-SJ methods on enhancing the LAA throughput.

Finally, we show throughput of two LAA links in Fig. 9. Link #2 has a fixed eCCA duration (9 μs), and link #1 has a variable eCCA duration (from 9 μs to 45 μs), respectively. For the anti-SJ-LBT, both analytical result [17] and SDR measured result on LAA links are provided. The SJ-LBT scheme has not been analyzed, and only its SDR-measured result is provided. The result in Fig. 9 confirms an approximate match between the analysis given in [17] and SDR measurement result, and demonstrates the effectiveness of the anti-SJ design.

V. Conclusion

In this paper, we have described a slot boundary tracking error problem and a slot-jamming effect in the LTE-LAA LBT scheme, and provided an improved LBT design with subslot-tracking and anti-slot-jamming features. We have programmed the new LBT scheme in an FPGA, and implemented testing on coexisting LAA and WLAN links using SDR devices. Our developed testing method allows LBT parameters to be changed in realtime on the FPGA, and includes an automated testing procedure, which updates many system parameters in realtime. The results demonstrate the effectiveness of subslot-tracking and anti-SJ design methods, and confirm a reasonably good match between our analysis and experimental results. In future work, the effects of various fading channel models will be studied, and radiated testing will be implemented to further evaluate the coexistence performance of LAA-based systems.

References

[1] 3GPP TSG RAN, "Study On Licensed-Assisted Access To Unlicensed Spectrum", 3GPP TR 36.889 V13.0.0, Jun. 2015.

[2] Ericsson, "Discussion on LBT protocols," 3GPP Tech. Rep. R1-151996, Apr. 2015.

[3] LTE-U Forum, "Coexistence study for LTE-U SDL", LTE-U Technical Report, V1.0, Feb. 2015.

[4] R. Zhang, M. Wang, L. X. Cai, Z. Zheng, and X. Shen, "LTE-unlicensed: the future of spectrum aggregation for cellular networks," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 150–159, Jun. 2015.

[5] T. Tao, F. Han and Y. Liu, "Enhanced LBT algorithm for LTE-LAA in unlicensed band," in *Proc. IEEE PIMRC,* Hong Kong, 2015, pp. 1907-1911.

[6] A. Bazzi, B. M. Masini and A. Zanella, "Performance analysis of V2V beaconing using LTE in direct mode with full duplex radios," *IEEE Wireless Commun. Lett.*, vol. 4, no. 6, pp. 685-688, Dec. 2015.

[7] I. Gomez-Miguelez, et al., "srsLTE: An Open-source Platform for LTE Evolution and Experimentation," in *Proc. 10th ACM WiNTECH*, New York, 2016, pp. 25–32.

[8] B. Kouassi et al., "Design and implementation of spatial interweave LTE-TDD cognitive radio communication on an experimental platform," *IEEE Wireless Commun.*, vol. 20, no. 2, pp. 60-67, April 2013.

[9] V. Marojevic, D. Chheda, R. M. Rao, R. Nealy, J. Park, and J. Reed, "Software-defined LTE evolution testbed enabling rapid prototyping and controlled experimentation," in *Proc. IEEE WCNC*, San Francisco, USA, Mar. 2017.

[10] S. Bhattarai, J. Park, B. Gao, K. Bian, and W. Lehr, "An overview of dynamic spectrum sharing: ongoing initiatives, challenges, and a roadmap for future research," *IEEE Trans. Cognitive Commun. Net.*, Vol. 2, No. 2, June 2016, pp. 110-128.

[11] A. Mukherjee et al., "Licensed-assisted access LTE: coexistence with IEEE 802.11 and the evolution toward 5G," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 50-57, Jun. 2016.

[12] National Instrument Inc. white paper, "Real-time LTE/Wi-Fi Coexistence Testbed," Feb. 16, 2016. Available: http://www.ni.com/white-paper/53044/en/

[13] IEEE LAN/MAN Standards Committee, IEEE Std 802.11-2012, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Feb. 2012.

[14] C. Chen, R. Ratasuk, and A. Ghosh, "Downlink performance analysis of LTE and WiFi coexistence in unlicensed bands with a simple listen-before-talk scheme," in *Proc. IEEE VTC*, May 2015, pp. 1–5.

[15] Y. Song, K. W. Sung, and Y. Han, "Coexistence of Wi-Fi and cellular with listen-before-talk in unlicensed spectrum," *IEEE Commun. Lett.*, vol. 20, no. 1, pp. 161–164, Jan. 2016.

[16] Y. Ma and D. G. Kuester, "MAC-Layer coexistence analysis of LTE and WLAN systems via listen-before-talk," in *Proc. IEEE CCNC*, Las Vegas, USA, Jan. 2017.

[17] Y. Ma, D. G. Kuester, J. Coder and W. Young, "Coexistence analysis of LTE and WLAN systems with heterogenous backoff slot durations," in *Proc. IEEE ICC*, Paris, France, 2017, pp. 1-7.

[18] S. Han, Y. C. Liang, Q. Chen and B. H. Soong, "Licensed-assisted access for LTE in unlicensed spectrum: A MAC protocol design," in *Proc. IEEE ICC*, 2016, pp. 1–6.

[19] R. Yin, G. Yu, A. Maaref, and G. Y. Li, "A framework for co-channel interference and collision probability tradeoff in LTE licensed-assisted access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6078–6090, Sept. 2016.

[20] Y. Gao, X. Sun and L. Dai, "IEEE 802.11e Std EDCA networks: modeling, differentiation and optimization," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3863–3879, July 2014.

[21] J. Jun, P. Peddabachagari, and M. Sichitiu, "Theoretical Maximum Throughput of IEEE 802.11 and its Applications," in *Proc. IEEE NCA*, 2003.

Coder, Jason; Jacobs, Ryan; Kuester, Daniel; Ma, Yao; Young, William.
"SDR-Based Experiments for LTE-LAA Based Coexistence Systems with Improved Design."
Paper presented at 2017 IEEE Globecom, Singapore, Singapore. December 4, 2017 - December 8, 2017.

A Layered Graphical Model for Mission Attack Impact Analysis

Changwei Liu
Department of Computer Science
George Mason University
Fairfax VA 22030 USA
cliu6@gmu.edu

Anoop Singhal
National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg MD 20899 USA
anoop.singhal@nist.gov

Duminda Wijesekera
Department of Computer Science
George Mason University
Fairfax VA 22030 USA
dwijesek@gmu.edu

***Abstract*—In this paper, we describe a layered graphical model to analyze the mission impacts of attacks for forensic investigation. Our model has three layers: the upper layer models operational tasks and their dependencies; the middle layer reconstructs attack scenarios by using a forensic tool to find the causality between items of evidence; the lower level reconstructs potentially missing attack steps due to missing evidence from the middle layer. Based on the graphs produced from the three layers, our model computes mission impacts by using NIST National Vulnerability Database(NVD) scores or forensic investigators' estimates. The case study shows our layered graphical model can be used for both forensic analysis and hardening an enterprise infrastructure.**

I. Introduction

In this paper, the mission for an enterprise is a set of business processes that provide a set of services. For example, the mission of a travel management system is to provide a set of business processes to support airline and hotel reservation. Organizational missions enabled by networked infrastructure can be impacted by cyber-attacks. Quantifying the impacts of cyber-attacks is of importance to mission planners. Mission impact evaluation approaches and tools provide a way to estimate the impacts of cyber-attacks on missions [1], of which NIST NVD Common Vulnerability Scoring System (CVSS) [2] provides impact estimates of exploitable vulnerabilities on IT systems. Researchers have expanded NIST NVD CVSS to multi-step attacks to predict the impacts of such attacks on missions by considering all possible vulnerabilities that construct all possible attack paths [1], [3], [4]. However, evaluating all paths is infeasible for forensic analysis to assess damages due to combinatorial explosion caused from considering all paths and vulnerabilities.

Because artifacts obtained from forensic investigations carried out after cyber-attacks provide information that can be used to analyze the attacks, we propose to create a layered graphical model that uses such information to quantify the attacks' impacts on missions. As far as we know, there is no such an integrated forensic analysis framework that quantifies the mission impacts of multi-step attacks in a complex enterprise infrastructure including cloud-based services.

The rest of the paper is organized as follows. Section II presents the three-layered graphical model. Section III uses a case study to show how the model computes mission impacts of attacks. Section IV discusses the related work, and Section V concludes this paper.

II. Our Three-layered Graphical Model

Figure 1 shows our model with three layers. The lower layers reconstruct attack paths so that the attacks can be mapped to tasks and missions in the upper layer for mission impact computation.

A. The Upper Layer

The upper layer models tasks and missions as business processes. We model business processes using a Business Process Diagram (BPD). Our BPDs are specified using the Business Process Modeling Notation (BPMN). BPMN models composite tasks by composing so-called atomic tasks using constructors hierarchically. The current version of BPMN (v 2.0) uses about 100 graphical elements, covering the description of many categories of tasks, events, errors, areas of responsibility, and general annotations [5]. A study of real-world BPMN usage found that the most used subset, labeled as the core subset of BPMN, consists of only eight elements including tasks, start/end events, exclusive decision gateways, parallel gateways, sequence flows, message flows and pools, which are the base of the definition of a BPD [6]. In this paper, we use BPDs constructed from those elements, formalized in Definition 1 (Originally defined in [6]).

Definition 1 (Business Process Diagram-BPD): In BPMN notation, a BPD is a graph BPD=(N, F, P, pool, L, lab), where:

1) $N \subseteq T \cup E \cup G$, in which tasks T are the basic actions performed as part of a business process, events $E \subseteq E^S \cup E^E$ consist of two disjoint sets E^S and E^E representing start and end events, gateways $G \subseteq G^M \cup G^F \cup G^D$ are the disjoint sets G^M, G^F and G^D representing parallel merge, parallel fork and exclusive decision gateways.
2) F $\subseteq$ S $\cup$ M is a set of flow relations, in which sequence flows $S \subseteq N \times N$ relate nodes including tasks, events, exclusive decision and parallel gateways to each other, message passing $M \subseteq T \times G^M$ is a relation between task nodes and parallel merge gateways, employed to pass messages between separate processes.
3) $P \subset P(N)$ is a set of disjoint pools.

Liu, Changwei; Singhal, Anoop; Wijesekera, Duminda.
"A Layered Graphical Model for Mission Attack Impact Analysis."
Paper presented at 2017 IEEE Conference on Communications and Network Security (CNS), Las Vegas, NV, United States. October 9, 2017 - October 11, 2017.

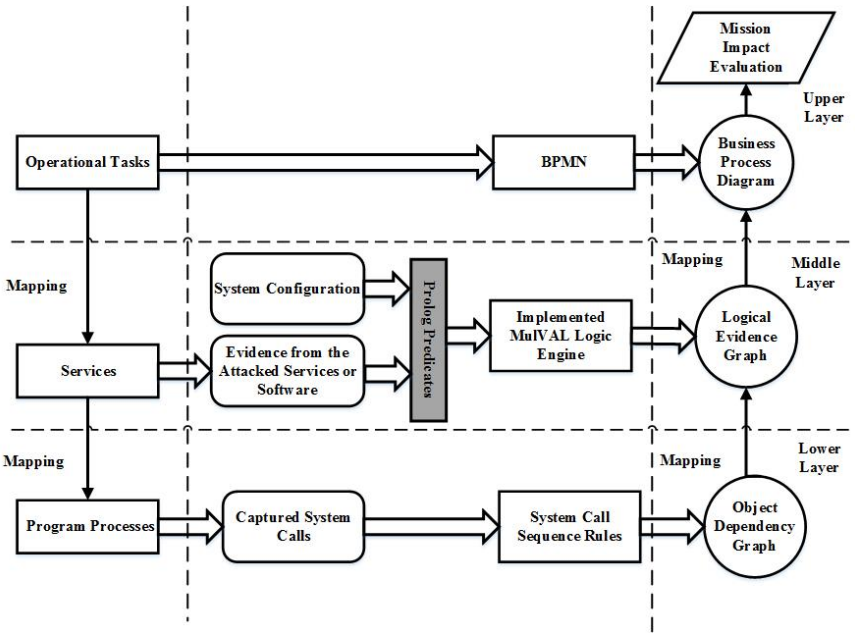


Fig. 1. The three-layered graph model for mission impact evaluation

4) Pool: $N \rightarrow P$ maps nodes to a pool $p \in P$. A pool is a basic BPMN element that sets the boundaries of a business process, which contains at most one business process.
5) L is a set of labels.
6) Lab: $F \rightarrow L$ is a labelling function that assigns labels to flows.

B. The Middle Layer

The middle layer creates attack scenarios using evidence available from system logs, intrusion detection system (IDS) alerts. Our objective is to map these attack scenarios to missions modeled as BPDs. Because we only consider attack scenarios substantiated using available evidence, the created attack paths do not include all possible attack paths and vulnerabilities that are infeasible for forensic analysis. However, sometimes, we may not be able to find all evidence to reconstruct all attack scenarios in this layer, which will be addressed in the lower layer.

We reconstruct attack scenarios by using a forensic analysis tool created in our previous work [7]. This tool uses rules to create directed graphs of available evidence and correlate them together as witnesses of attacks. Because rules are used to create these graphs, they are called logical evidence graphs (LEGs) formalized in Definition 2 (originally defined in [7]).

Definition 2 (Logical Evidence Graph-LEG): A LEG=(N_r, N_f, N_c, E, L, G) is said to be a logical evidence graph (LEG), where N_f, N_r and N_c are three sets of disjoint nodes in the graph (they are called fact, rule, and consequence fact nodes respectively), E $\subseteq ((N_f \cup N_c) \times N_r) \cup (N_r \times N_c)$, L is the mapping from a node to its labels, and G $\subseteq N_c$ are the observed attack events. Every rule node has a consequence fact node as its single child and one or more fact or consequence fact nodes from prior attack steps as its parents. The labels of nodes consist of instantiations of rules or sets of predicates specified as follows:

1) A node in N_f is an instantiation of predicates that codify system states including access privileges, network topology consisting of interconnectivity information, or known vulnerabilities associated with host computers in the system. We use the following predicates:
   a) "hasAccount(_principal,_host, _account)", "canAccessFile(_host, _user, _access, _path)" and etc. to model access privileges.
   b) "attackerLocated(_host)" and "hacl(_src, _dst, _prot, _port)" to model network topology, namely, the attacker's location and network reachability information.
   c) "vulExists(_host, _vulID, _program)" and "vulProperty(_vulID, _range, _consequence)" to model vulnerabilities exhibited by nodes.
2) A node in N_c represents a predicate that codifies the post attack state as the consequence of an attack step. We use predicates "execCode(_host,_user)" and "netAccess(_machine,_protocol, _port)" to model the attacker's capability after an attack step. Valid instantiations of these predicates after an attack will update valid instantiations of the predicates listed in (1).
3) A node in N_r consists of a single rule in the form $p \leftarrow p_1 \wedge p_2, \cdots, \wedge p_n$. p as the child node of N_r is an instantiation of predicates from N_c. All p_i for $i \in$

Liu, Changwei; Singhal, Anoop; Wijesekera, Duminda.
"A Layered Graphical Model for Mission Attack Impact Analysis."
Paper presented at 2017 IEEE Conference on Communications and Network Security (CNS), Las Vegas, NV, United States. October 9, 2017 - October 11, 2017.

$\{1 \ldots n\}$ as the parent nodes of N_r are the collection of all predicate instantiations of N_f from the current step and N_c from the prior attack step.

C. The Lower Layer

This layer uses instances of interactions between services and the execution environment to obtain evidence unavailable from systems logs and IDS alerts. We obtain such interaction instances from systems call logs. This usage is based on our postulate of missing evidence due to the attackers' using anti-forensic techniques, limitation of forensic tools, or zero-day attacks. Because there are many system calls, we use those used in [8], listed in the right-hand column of Table 1. Our abstraction of them appear in the left-hand column of Table 1. A process making system calls creates dependencies between itself and other processes, files, or sockets for network connection. We model these dependencies as graphs that we call object dependency graphs (ODGs), formalized in Definition 3.

Definition 3 (Object Dependency Graph-ODG): The reflexive transitive closure of " $\rightarrow$ " defined in Table 1 is an object dependency graph. We use the notation ODG=(V_O, V_E, E) to represent an object dependency graph, where V_O is the set of vertexes that are composed of objects including Processes P, Files F or Sockets S; V_E is the set of textual event descriptions listed in the middle column; and E is the set of dependency edges listed in the left-hand column of Table 1.

D. The Mapping between the Three Layers

The left-hand and right-hand columns in Figure 1 show the system resource mapping and graph mapping of our model. We use the resource mapping obtained from the infrastructure configuration and software deployment to map graphs. To do so, we add the attacked services as attributes to the corresponding nodes(vertexes) in LEGs and ODGs, so that the mapping can match between node attributes of source and destination graphs. These nodes are either processes, infrastructure services or tasks that are the entities shown in the left-column of Figure 1. A LEG is easily mapped to a BPD by matching the attacked services to the corresponding tasks. We use depth first search (DFS) method to map an ODG to a LEG, which is explained in Algorithm 1.

In Algorithm 1, the first *for* loop colors all object nodes in an ODG white, marking them as *having not been checked*. The second *for* loop repeatedly calls $Find(V_O, LEG)$ function, where, for a given white node V_O, the algorithm attempts to find the matching Node N_{c1} by checking if the attacked service in the LEG is equal to the attacked service in the ODG(the pseudo code in $Find(V_O, LEG)$ function is $N_c.service == V_O.service$). If such a post attack status node N_{c1} is found, the algorithm checks if the attack step between Node V_O and its parent node *parent(V_O)* in the ODG has a mapping attack step between Node N_{c1} and its parent node(s) *parent(N_{c1})* in the LEG. If there is no such a mapping attack step, the attack step is added to the LEG. If there is no any matching post attack status Node N_{c1} for node V_O, one is added to the

Input: An ODG=(V, V_E, E) and a LEG=(N_r, N_f, N_c, E, L, G).
Output: A LEG integrated with attack paths from the ODG.

```
//Color all nodes in ODG WHITE
for each node V_O in ODG do
    color[V_O] ← WHITE
end
//Go through each object in ODG
for each node V_O in ODG do
    if V_O == WHITE then
        //Initialize all nodes in LEG white
        for each node N_c in LEG do
            color[N_c] ← WHITE
        end
        //Search for the corresponding N_c1 in LEG
        N_c1 = Find(V_O, LEG)
        //If there is such a matching N_c1
        if N_c1 ≠ ∅ then
            color(V_O) ← BLACK
            //See if the object's parent matches
            //corresponding N_c1' s parent
            N_c2=Find(parent(V_O), LEG)
            //If not matching parents,
            //add the missing attack step from ODG to
              LEG
            if N_c2 ≠ parent(N_C1) then
                LEG ← Flow(N_c1, V_E)
                LEG ← Flow(V_E, N_c2)
            end
        end
        else
            //If there is no such a matching N_c1 in LEG
            //Add the new object to LEG
            LEG ← V_O
            color [V_O]=GRAY
        end
        V_O =child(V_O)
    end
end
Function Find(V_O, LEG)
    //Go through each N_c in LEG
    for each post attack status N_c from LEG do
        //Check if there is any matching N_c for V_O
        if N_c.service == V_O.service AND
        color[N_c] == white then
            color[N_c] ← BLACK
            return N_c
        end
        else
            color[N_c] ← GRAY
            N_c ← the child post attack status node of N_c
        end
    end
    return ∅
```

Algorithm 1: Mapping an ODG to a LEG

Liu, Changwei; Singhal, Anoop; Wijesekera, Duminda.
"A Layered Graphical Model for Mission Attack Impact Analysis."
Paper presented at 2017 IEEE Conference on Communications and Network Security (CNS), Las Vegas, NV, United States. October 9, 2017 - October 11, 2017.

TABLE I
DEPENDENCIES ARISING OUT OF SYSTEMS CALLS

Dependency	Event Description	Unix System Calls
$process \rightarrow file$	Process modifies file	write, pwrite64, rename, mkdir, linkat, link, symlinkat, etc
$file \rightarrow process$	Process reads file	stat64, lstat6e, fsat64, open, read, pread64, execve, etc.
$process \leftrightarrow file$	Process uses/modifies file	open, rename, mount, mmap2, mprotect etc.
$process1 \rightarrow process2$	Process1 creates/terminates Process2	vfork, fork, kill, etc.
$process \rightarrow socket$	process writes socket	write, pwrite64, etc.
$socket \rightarrow process$	process checks/reads socket	fstat64, read, pread64, etc.
$process \leftrightarrow socket$	Process reads/writes/checks socket	mount, connect, accept, bind, sendto, send, sendmsg, etc.
socket $\leftrightarrow$ socket	process reads/writes socket	connect, accept, sendto, sendmsg, recvfrom, recvmsg

LEG and the search continues until all nodes in the ODG are checked(colored).

E. Mission Impact Computation

We propose to use the interval [0,1] to quantify a mission impact of an attack, computed by using the following steps.

- **Compute the impact scores of attacks in a LEG**
  In a LEG, we use P(a) to represent the impact of attacks on a service deployed on a host computer. NIST NVD CVSS published reported vulnerabilities with assigned impact scores, which we propose to use for each P(a) if an attack *a* can be found in NIST NVD. If the attack *a* cannot be found in NIST NVD, we suggest using expert knowledge to assign an impact score to P(a). We use our previous work [9] to compute a cumulative impact score of an attack as follows.

$$P(a) = P(a_1) \cup P(a_2) \tag{1}$$

  In Equation 1, a_1 and a_2 are two attacks on the same service. $P(a_1) \cup P(a_2) = P(a_1) + P(a_2) - P(a_1) \times P(a_2)$.
- **Assign weight to tasks/missions**
  A value between [0,1] is proposed as the weight of mission impacts, indicating the importance of the corresponding tasks/missions.
- **Map LEGs to BPDs**
  We map LEGs integrated with missing attack steps to BPDs so that the mission impact of attacks I(B) on a business process B is computed using Equation 2.

$$I(B) = weight \times P(B) \tag{2}$$

  In Equation 2, P(B) is the impact of attacks on a business process *B* in a BPD. It is computed by using Equation 3 and Equation 4 respectively, since the mapping from attacks(represented by *a*, a_1, a_2) in a LEG to a business process(represented by *B*) in a BPD has two relationships including one-to-one(Equation 3) or many-to-one(Equation 4).

$$P(B) = P(a) \tag{3}$$

$$P(B) = P(a_1) \cup P(a_2) \tag{4}$$

- **Compute the cumulative mission impact**
  Mission impact of attacks on each business process(task) can be computed using Equation 2, Equation 3 and Equation 4. However, in some cases, the cumulative mission impact for the final mission is required to estimate the overall damage, which is computed using the *Max* function. We use *M* to represent the cumulative mission impact. Correspondingly, in the flow relationships including *sequence* and *message passing* as defined in Definition 1, M is computed as follows.
  - If the tasks $B_1, B_2, \ldots, B_n$ composing of the final mission B have a sequence relationship.

$$M(B) = Max(I(B_1), I(B_2), \ldots, I(B_n)) \tag{5}$$

  - If, in the tasks $B_1, B_2, \ldots, B_n$ composing of the final mission B, there are tasks, say B_2, B_3, which have exclusive decision relationship with the predecessor task B_1 and the successor tasks $B_4, \ldots, B_n$.

$$M(B) = Max(I(B_1), I(B_2), I(B_4), \ldots, I(B_n))$$
$$or$$
$$M(B) = Max(I(B_1), I(B_3), I(B_4), \ldots, I(B_n)) \tag{6}$$

  - If, in the tasks $B_1, B_2, \ldots, B_n$ composing of the final mission B, there is message passing relationship between a task $B\prime$ from another pool to tasks in this pool.

$$M(B) = Max(I(B_1), I(B_2), \ldots, I(B_n), I(B\prime)) \tag{7}$$

III. THE CASE STUDY

This section describes our case used to determine the utility of our model. Figure 2 shows our experimental network configured to manage the customers' medical records and their health insurance policy files. Customers' medical records are stored in a MySQL database server deployed in a private cloud.

Liu, Changwei; Singhal, Anoop; Wijesekera, Duminda.
"A Layered Graphical Model for Mission Attack Impact Analysis."
Paper presented at 2017 IEEE Conference on Communications and Network Security (CNS), Las Vegas, NV, United States. October 9, 2017 - October 11, 2017.

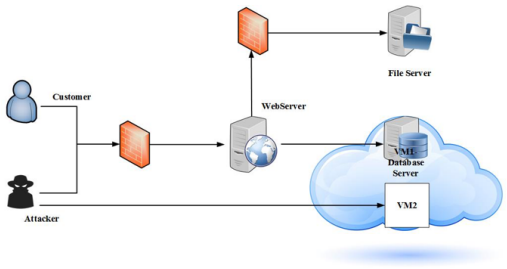


Fig. 2. The network example and corresponding attacks

Customers can query the medical records and the policy files using a web application on a webserver, using valid (username, password) pairs as an access control mechanism. We built the cloud on OpenStack, a free and open source cloud system.

We assume that an attacker's objective is to steal customers' medical records or prevent service availability. The attacker can probe deployed web and cloud services looking to find vulnerabilities that can be exploited to satisfy his/her objective. Our case study used two such vulnerabilities. The first vulnerability was the web application not sanitizing user input, named CVE-89 that created a SQL injection attack to access customers' medical records. We played the attacker role that created the SQL injection query *select * from profile where name='Alice' and (password='alice' or '1' = '1')*, where *profile* was the database name, and *'1'='1'* was the payload that made the query bypass the password check. The second vulnerability is named CVE-2015-3241 that allows authenticated users to prevent service availability by first resizing and then deleting virtual machine instances of OpenStack Compute (Nova) versions 2015.1 through 2015.1.1, 2014.2.3 and earlier. We, playing the attacker role, exploited this vulnerability as a privileged IaaS user by repeatedly resizing and deleting VM2 that co-resided in the same physical machine as the database server residing on VM1, which bypassed user quota enforcement to deplete available disk space. The three kind of graphs, including a BPD, a LEG and an ODG, generated by using our experimental example are as follows.

A. The Business Process Diagram

Figure 3 is the constructed BPD. In this BPD, there is only one pool that has start, end events and two missions *return customer records* and *return files*. The two missions are fulfilled by tasks *visit web application, request customer records, request files, verify username and password, query SQL database, query a file, data available* and *file available* that are represented by boxes. Figure 3 shows parallel fork and exclusive or gateways represented by diamonds. The exclusion or gateways have *yes* and *no* choices.

B. The Logical Evidence Graph

Table II shows evidence of the SQL injection attack including Snort(the IDS we deployed in the experimental network) alerts and database server access logs. Using timestamps, corresponding alert content and MySQL general query logs, we asserted that the attacker used a typical SQL injection with payload *'1'='1'*. Our IDS failed in capturing the DoS attack launched by exploiting the vulnerability CVE-2015-3241 in OpenStack Nova services. Because OpenStack API logs provide users' operations of running instances, we used them to conclude that the user admin (the attacker in our experiment) was trying to resize and delete the instance VM2 that co-resided in the same physical machine as the database server (VM1).

We converted the system configuration and the evidence to Prolog predicates as shown in Figure 4 and Figure 5 as input files to our LEG reconstruction tool. During the system runtime, the input files instantiated the rules representing the generic attack techniques in this tool to correlate constant predicates representing different items of evidence or system configuration, forming the LEGs as shown in Figure 6 and Figure 7 respectively (the notation can be found in Table III and Table IV). The two LEGs could not be grouped together, because the attacker's locations were different. Consider an attack step (Nodes $3, 7, 8 \rightarrow 2 \rightarrow 1$ in Figure 7) as an example. Facts of LEGs are shown in boxes (Nodes 7, 8), representing network configurations and vulnerabilities as the evidence prior to the attack step. The consequence fact node is shown in a diamond (Node 1), representing the evidence of the post attack status that is derived by applying a rule to the parent facts (Nodes 7, 8) and the parent consequence facts (Node 3 obtained from a prior attack step as the attacker's stepping-stone to the current attack step). The rule node is shown in an ellipse representing the attack and connects pre-attack system status (Nodes 3, 7, 8) and the post attack status (Node 1).

C. The Object Dependency Graph

To show how to use system call sequences to construct an ODG, we simulated an attack without triggering IDS alerts in our experimental network. We assume the attacker used a social engineering attack to obtain a legitimate user's (username, password) pair to log into the file server using ssh. In our experiment, the legitimate user's name is gmu. The corresponding server log showed that the attacker stole the user gmu's credentials. We used the right column in Table 1 to filter system calls and used dependency rules listed in the left column of Table 1 to construct an ODG as shown in Figure 8, showing the attacker modified the policy file in the file server. In this figure, the two objects *attacker* and *file in fileserver* are represented by boxes, and the rule *modify* is represented by an ellipse. Figure 8 was mapped the LEG in Figure 6, showing the attacker from the Internet could attack the file server by using stolen credentials and attack the database server by using a SQL injection attack (figure omitted due to space limitations).

D. Mission Impact Computation in Our Case Study

The impact score of each attack step in Figure 6, 7 and 8 is shown in Table V, where two impact scores of CWE-89 and CVE-2015-3241 are obtained from NIST NVD CVSS. The

Liu, Changwei; Singhal, Anoop; Wijesekera, Duminda.
"A Layered Graphical Model for Mission Attack Impact Analysis."
Paper presented at 2017 IEEE Conference on Communications and Network Security (CNS), Las Vegas, NV, United States. October 9, 2017 - October 11, 2017.

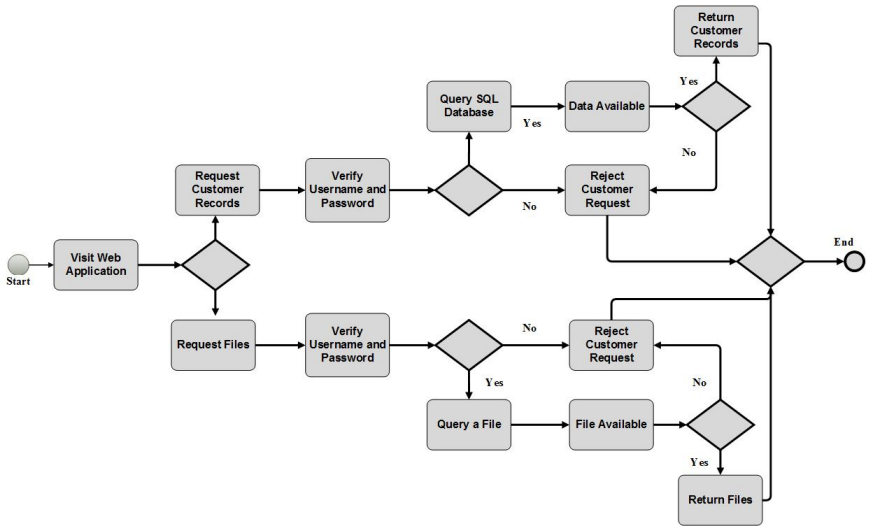


Fig. 3. The BPD of customers' retrieving their medical records and policy files

TABLE II
THE SNORT ALERT AND DATABASE SERVER LOG OF SQL INJECTION ATTACK

Time Stamp	Machine	IP Address/Port	Snort Alert and Database Server Access Log
	Attacker	129.174.124.122	
06/13-14:37:27	web server	129.174.124.184	SQL injection attack(CWE-89)
13/Jun/2017:14:37:34	Database server	129.174.124.35	Access from 129.174.124.184

TABLE III
THE NOTATION OF ALL NODES IN FIGURE 6

No.	Notation of all nodes
1	execCode(database,_)
2	RULE 2 (remote exploit of a server program)
3	netAccess(database,tcp,3306)
4	RULE 5 (multi-hop access)
5	hacl(webServer,database,tcp,3306)
6	execCode(webServer,apache)
7	RULE 2 (remote exploit of a server program)
8	netAccess(webServer,tcp,80)
9	RULE 6 (direct network access)
10	hacl(internet,webServer,tcp,80)
11	attackerLocated(internet)
12	networkServiceInfo(webServer,httpd,tcp,80,apache)
13	vulExists(webServer,'directAccess',httpd, remoteExploit,privEscalation)
14	networkServiceInfo(database,httpd,tcp,3306,_)
15	vulExists(database,'CWE-89',httpd, remoteExploit, privEscalation)

```
/* the initial attack location and final attack status*/
attackerLocated(internet).
attackGoal(execCode(database,user)).

/* the network access configuration*/
hacl(internet, webServer, tcp, 80).
hacl(webServer, database, tcp, 3306).

/* configuration information of webServer */
vulExists(webServer, 'directAccess', httpd).
vulProperty('directAccess', remoteExploit, privEscalation).
networkServiceInfo(webServer , httpd, tcp , 80 , apache).

/* the vulnerability of the web application */
vulExists(database, 'CWE-89', httpd).
vulProperty('CWE-89', remoteExploit, privEscalation).
networkServiceInfo(database , httpd, tcp , 3306, user).
```

Fig. 4. Prolog predicates for SQL injection

impact scores in NIST NVD CVSS are based on a [0, 10] scale, which we converted to a [0,1] interval scale. Because our example does not have services attackable using multiple methods, equation (1) is not used in our LEGs.

We mapped all attacks shown in Figures 6, 7 and 8 to the

/* the initial attack status of being an iaas user and the final attack status*/
attackerLocated(iaas).
attackGoal(execCode(nova,admin)).

/*the cloud configuration, the "_" represents any protocol and port*/
hacl(iaas,nova,_, _).

/* the vulnerability in nova */
vulExists(nova, 'CVE-2015-3241', 'REST').
vulProperty('CVE-2015-3241',remoteExploit, privEscalation).
networkServiceInfo(nova , 'REST', http, _, admin).

Fig. 5. Prolog predicates for DoS attack

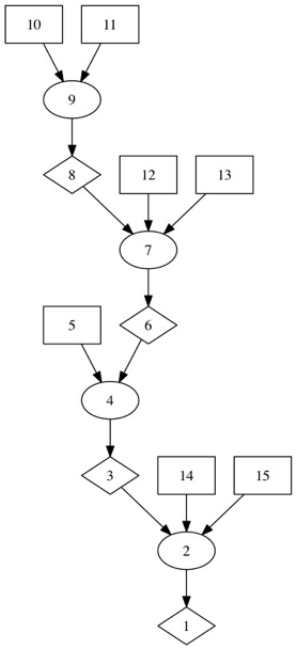


Fig. 6. The LEG of SQL injection attack toward the database

TABLE IV
THE NOTATION OF ALL NODES IN FIGURE 7

No.	Notation of all nodes
1	execCode(nova,admin)
2	RULE 2 (remote exploit of a server program)
3	netAccess(nova,http,_)
4	RULE 6 (direct network access)
5	hacl(cloud,nova,http,_)
6	attackerLocated(cloud)
7	networkServiceInfo(nova,'REST',http,_,admin)
8	vulExists(nova,'CVE-2015-3241', 'REST', remoteExploit,privEscalation)

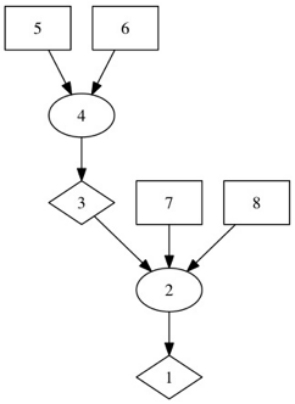


Fig. 7. The LEG of DoS attack toward the database server

Fig. 8. The attackers using social engineering attack to modify a file in the fileserver

BPD in Figure 3, and calculated the mission impact of the three attacks as shown in Table VI, where different weights were given respectively. Based on Table VI and the BPD in Figure 3, the cumulative mission impacts were also calculated and listed in Table VII. Table VI shows that SQL attack on the task of *verifying username and password* in the database server is considered to have a higher mission impact than the DoS attack toward the database server(on the task of *data available*) and using social engineering to modify policy files in the file server(on the task of *verify username and password in the file server*). Table VII shows that the mission impact of attacks on the customers' medical records is higher than the attack on the policy files.

IV. RELATED WORK

Modern-day attackers tend to use multi-step, multi-stage attacks to impact important services protected using complex mechanism. Researchers have proposed and designed models to estimate the mission impacts of such attacks. Sun et al. proposed using a multi-layered impact evaluation model to estimate the mission impacts [10]. In this multi-layered model consisting of four layers, a lower vulnerability layer is mapped to an asset layer, and then to a service layer, which finally maps to the mission layer, where the mission impacts are calculated by using vulnerabilities' CVSS scores and the relationships between missions to the lower level assets, services and vulnerabilities. Another group of researchers, Sun et al., combined mission dependency graphs with attack graphs generated by an attack graph generation tool MulVAL [11] to estimate the attack mission impacts in the clouds [4]. Noel at el. designed a cyber-mission impact assessment framework by leveraging BPMN and their attack graph generation tool

Liu, Changwei; Singhal, Anoop; Wijesekera, Duminda.
"A Layered Graphical Model for Mission Attack Impact Analysis."
Paper presented at 2017 IEEE Conference on Communications and Network Security (CNS), Las Vegas, NV, United States. October 9, 2017 - October 11, 2017.

TABLE V
THE CVSS IMPACT SCORES

Symbol Representation	Attack Step	Attack	Attack Impact
N_1	Figure 6: (3,14,15)$\rightarrow$ 2 $\rightarrow$ 1	CWE-89	0.9 (from NIST NVD)
$N_1\prime$	Figure 7: (3,7,8)$\rightarrow$ 2$\rightarrow$1	CVE-2015-3241	0.69 (from NIST NVD)
N_s	Figure 8: 1 $\rightarrow$ 2 $\rightarrow$ 3	Social Engineering	0.5 (expert knowledge)

TABLE VI
THE MISSION IMPACT SCORES

Symbol Representation	Server	Task	Weight	Mission Impact
A	Database Server	Verify Username and Password	1	$I(A) = 1 \times P(N_1) = 1 \times 0.9 = 0.9$
B	Database Server	Data Available	0.9	$I(B) = 0.9 \times P(N_1\prime) = 0.9 \times 0.69 = 0.621$
C	File Server	Verify Username and Password	1	$I(C) = 1 \times P(N_s) = 1 \times 0.5 = 0.5$

TABLE VII
THE CUMULATIVE MISSION IMPACT

Symbol Representation	Server	Mission	Cumulative Mission Impact
D	Database Server	Return Customer Records	M(D)=Max(I(A),I(B))=Max(0.9,0.621)=0.9
E	File Server	Return Files	M(E)=Max(I(C)) =Max(0.5)= 0.5

named Topological Vulnerability Analysis (TVA) [12] that combines an exploit knowledge base and a remote network scanner, analyzing all potential attack paths leading to attack goals to evaluate potential mission impacts [3], [4]. However, these approaches use vulnerabilities collected from the bug-report community such as NIST NVD to assess the impacts of attacks. These do not scale to large infrastructures or zero-day attacks.

Forensics researchers have proposed using reasoning on collected evidence from attacked infrastructures using evidence correlation rules to reconstruct the attack scenarios. The objective of this work has been to reconstruct criminal or unauthorized actions shown to be disruptive to missions [7], [13]. To reconstruct attack scenarios that have legal standing, we integrated a Prolog logic tool, MulVAL, with two databases, including a vulnerability database and an anti-forensic database, to ascertain the admissibility of evidence and explain missing evidence due to attackers' using anti-forensics [8]. We also expanded their work by using system calls to reconstruct the missing attack steps due to missing evidence in the higher application levels, and using Bayesian Network to estimate the experts' belief on the reconstructed attack scenarios [14]. However, no researchers have proposed any method to estimate the mission impacts of attacks launched toward an enterprise's infrastructure, which leaves a gap between the mission impact analysis and forensic analysis.

V. CONCLUSION

In this preliminary work we proposed a three-layered graphical model to quantify mission impacts of cyber-attacks computable using forensic techniques. We did so by reconstructing attacks based on available evidence from attack logs and system call sequences when logs did not have requisite evidence for attack steps. We used attack impact scores published in the NIST NVD CVSS and expert opinions when such numbers are unavailable. We then mapped the attacks to higher-level business processes and considered their importance weight for business processes to compute the impacts of cyber-attacks on missions.

DISCLAIMER

This paper is not subject to copyright in the United States. Commercial products are identified in order to adequately specify certain procedures. In no case does such an identification imply a recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the identified products are necessarily the best available for the purpose.

REFERENCES

[1] S. Musman and A. Temin, A cyber mission impact assessment tool, In Technologies for Homeland Security (HST), 2015 IEEE International Symposium on 2015 Apr 14 (pp. 1-7).

[2] NIST National Vulnerability Database Common Vulnerability Scoring System, available at https://nvd.nist.gov/vuln-metrics/cvss.

[3] S. Noel, J. Ludwig, P. Jain, D. Johnson, R. K. Thomas, J. McFarland, B. King, S. Webster and B. Tello, Analyzing mission impacts of cyber actions (AMICA), In NATO IST-128 Workshop on Cyber Attack Detection, Forensics and Attribution for Assessment of Mission Impact, Istanbul, Turkey, 2015.

[4] X. Sun, A. Singhal, P. Liu, Towards Actionable Mission Impact Assessment in the Context of Cloud Computing, In Livraga G., Zhu S. (eds) Data and Applications Security and Privacy XXXI. DBSec 2017. Lecture Notes in Computer Science, vol 10359.

[5] Online Resource for Markup Language Technologies, retrieved from http://xml.coverpages.org/bpm.html#bpmi.

[6] L. Herbert, Specification, Verification and Optimization of Business Processes, A Unified Framework, Technical University of Denmark (2014).

Liu, Changwei; Singhal, Anoop; Wijesekera, Duminda.
"A Layered Graphical Model for Mission Attack Impact Analysis."
Paper presented at 2017 IEEE Conference on Communications and Network Security (CNS), Las Vegas, NV, United States. October 9, 2017 - October 11, 2017.

[7] C. Liu, A. Singhal and D. Wijesekara, A logic-based network forensic model for evidence analysis, in Advances in Digital Forensics XI, G. Peterson and S. Shenoi (Eds.), Springer, Heidelberg, Germany, pp. 129-145, 2015.
[8] X. Sun, J. Dai, A. Singhal, P. Liu and J. Yen, Towards Probabilistic Identification of Zero-day Attack Paths, Accepted for IEEE Conference on Communication and Network Security, Philadelphia, October 17th 19th, 2016.
[9] C. Liu, A. Singhal and D. Wijesekera, Mapping evidence graphs to attack graphs, In Information Forensics and Security (WIFS), 2012 IEEE International Workshop on (pp. 121-126). IEEE.
[10] Y. Sun, T. Y. Wu, X. Liu, X. and M.S. Obaidat, Multilayered Impact Evaluation Model for Attacking Missions, IEEE Systems Journal, 10(4), pp.1304-1315, 2016.
[11] X. Ou, S. Govindavajhala, S. and A. W. Appel, MulVAL: A Logic-based Network Security Analyzer, In USENIX Security Symposium (pp. 8-8), July 2005.
[12] S. Jajodia and S. Noel, Topological vulnerability analysis, In Cyber situational awareness, pp. 139-154. Springer US, 2010.
[13] W. Wang, E.D. Thomas, A graph based approach toward network forensics analysis, ACM Transactions on Information and Systems Security 12 (1) 2008.
[14] C. Liu, A. Singhal and D. Wijesekera, A Probabilistic Network Forensic Model for Evidence Analysis, IFIP International Conference on Digital Forensics. Springer International Publishing, 2016.

Liu, Changwei; Singhal, Anoop; Wijesekera, Duminda.
"A Layered Graphical Model for Mission Attack Impact Analysis."
Paper presented at 2017 IEEE Conference on Communications and Network Security (CNS), Las Vegas, NV, United States. October 9, 2017 - October 11, 2017.

Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness

Peter Mell[1], John Kelsey[12], and James Shook[1]

[1] National Institute of Standards and Technology, Gaithersburg MD, USA
[2] Department of Electrical Engineering, ESAT/COSIC, KU Leuven, Belgium

Abstract. Most modern electronic devices can produce a random number. However, it is difficult to see how a group of mutually distrusting entities can have confidence in any such hardware-produced stream of random numbers, since the producer could control the output to their gain. In this work, we use public and immutable cryptocurrency smart contracts, along with a set of potentially malicious randomness providers, to produce a trustworthy stream of timestamped public random numbers. Our contract eliminates the ability of a producer to predict or control the generated random numbers, including the stored history of random numbers. We consider and mitigate the threat of collusion between the randomness providers and miners in a second, more complex contract.

1 Introduction

Most modern computing devices can produce secure random numbers. However, there are applications which require that many parties share and trust some source of random numbers. For example, running a lottery requires some trustworthy source of *public random numbers*. In the rest of the paper, we define a lottery abstractly as any mechanism that randomly picks a proper subset of elements from some larger set. It is necessary to ensure that the chosen subset cannot be predicted (before some published time), controlled (deliberately set), or influenced (biased toward values that are more desirable for some party). The interesting research question is: *how can we get trustworthy public random numbers sampled from a uniform distribution, especially when the producer of random numbers has a financial incentive to cheat?*

Currently an individual 'beacon' service, a public producer of randomness, may use specialized hardware setups and cryptography to reduce the possibility of the numbers to be compromised [3]. However, the ability to control the numbers (by the beacon owner or some attacker that has compromised the beacon) may remain. What is needed is a consensus protocol for a set of mutually distrusting entities to collaborate to produce a trustworthy stream of publicly available random numbers.

Our solution is to create an Ethereum[3] [22] smart contract, called a *lighthouse*, which implements a beacon service while taking as input random num-

[3] Any mention of commercial products is for information only; it does not imply recommendation or endorsement by NIST.

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

bers from one or more external and potentially malicious randomness producers. To produce the lighthouse output, we combine producer input with blockchain hashes while forcing producers to commit to future values. In creating the distributed consensus protocol, we leverage the security capabilities associated with smart contracts and blockchains along with a novel commitment system we call Merlin chains (which mitigates a vulnerability common in other systems). Our lighthouse service's timestamped random outputs are published on the Ethereum blockchain, which ensures their immutability and their public visibility. This merging of beacons, smart contracts, and blockchains enables the production of public random numbers at an extremely high level of security, even when assuming the presence of powerful malicious actors in the system (as long as all participating actors aren't malicious).

We provide two main proposed designs:

1. A **single-producer contract** which provides security against control or influence from the randomness producer *or* a large coalition of miners competing in the digital currency system, but not against both.
2. A **multiple-producer contract** which provides security against control or influence from all k of the randomness providers colluding, or a large coalition of miners conspiring with $k-1$ of the randomness providers.

Both designs publish random numbers along with a time before which the random number could not have been predicted by any entity, thus eliminating prediction attacks. With these designs, we have provided a solution for the trustworthy public production of streams of immutable public random numbers. Finally, we create such a contract and empirically test it on the Ethereum test network using both the single and multiple producer models.

Usage of lighthouse services can greatly benefit any public lottery so that selection of random numbers is no longer done behind closed doors, where the public has to trust that no cheating is taking place. Lotteries enable a limited set of resources to be fairly chosen for, or distributed to, a set of customers. Among many other areas, their uses include school placements, dorm rooms allocations, gambling, military drafts, jury duty, immigration applications, election site auditing, and large public financial games run by governments. The utility of a beacon extends far beyond lotteries, but a complete discussion of those applications is outside the scope of this paper.

Different types of public lotteries are more or less sensitive to the three attack types mentioned previously: prediction, control, and influence. For example, with election site auditing an attacker primarily wants to ensure that the election sites chosen for auditing do not correspond to the compromised sites. The attacker then primarily wants *influence* to change the sites chosen for audit if the unmodified result is going to include a compromised site. However, in a gambling scenario, the attacker probably wants to *predict* the winning number or, even better, *control* the result. Our approach must mitigate all three types of attack.

The rest of this paper is organized as follows. Section 2 discusses previous and related work. Section 3 discusses background information. Section 4 provides

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

partial solutions that build towards our final solution. Section 5 describes our design for a single producer contract. Section 6 describes our multiple producer contract; Sect. 7 discusses our empirical work; and Sect. 8 concludes.

2 Previous and Related Work

The original idea of a beacon (a public service that publishes signed, timestamped random numbers) comes from Rabin [16]. More recently, in [11], Fischer et. al. propose the usefulness of a beacon service, and describe the NIST beacon. They also propose a general protocol to allow many beacons to be used together to decrease required trust in a single TTP/point of failure, and describe some practical applications for a beacon service. There have also been many attempts to find verifiable public random numbers for use in other applications, such as election auditing [10] and the choice of parameters in cryptographic standards [7].

The simplest way to build a beacon is to simply set up a trusted machine, which generates and signs timestamped random numbers. Existing services such as the NIST Beacon [3] and the beacon-like random.org [6] follow this approach. For many applications of a beacon, this provides sufficient practical security. However, it has a single point of failure – the owner of the beacon (or anyone who compromises the trusted machine on which the beacon is running) can influence or predict future random numbers[4].

2.1 Entropy from the Environment

In order to avoid a single point of failure or trust, many people have tried to use unpredictable data from the world to generate public random numbers. In order to be useful, these numbers need to be public, widely-attested, and not under anyone's control.

In [10], the authors consider using financial data as a source of randomness, particularly for election auditing, and use existing tools from finance to estimate the entropy and difficulty of influencing these numbers. [7] considers the use of public financial lotteries to generate random numbers (intended for use in defining cryptographic standards). [8] uses the hash of a block from the Bitcoin blockchain and analyzes the cost of exerting influence on these random numbers by bribing miners to discard inconvenient mined blocks. Our approach uses block hashes in a related way and we have to consider similar attacks.

2.2 Combining Randomness from Multiple Parties

Still another approach is to combine random values from multiple sources, with the goal of getting a trustworthy public random number if enough of the con-

[4] The NIST Beacon's published format includes features to mitigate some attacks–for example, the beacon operator cannot directly control the beacon outputs, as they're the result of a SHA512 hash. However, he *can* predict and influence future random numbers.

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

tributors are honest. This may be done by first collecting *commitments* from participants[5], and then asking each participant to *reveal* their commitments.

For example, if Alice and Bob want to each furnish a part of a shared random number, Alice generates random number R_A and publishes $\texttt{hash}(R_A)$, while Bob generates R_B and publishes $\texttt{hash}(R_B)$. After both commitments are published, Alice and Bob reveal their random numbers, and agree to use $R_A \oplus R_B$ as their shared random number. (This is referred to as a commit-then-reveal protocol.) The generic attack against this kind of scheme is for Alice to wait until Bob has published R_B, and then decide whether she likes the resulting random number or not. If not, she can "hit the reset button," claiming to have suffered a system failure that caused her to lose R_A. If this leads to the shared random number being generated again in an actual random way (even in a way that excludes Alice), she has now exerted some influence on the shared random number.

Commit-Then-Reveal Approaches The new NIST Beacon format [12] has a pre-commitment field intended to allow for combining of beacons using a commit-then-reveal protocol. However, preventing the 'hit the reset button' attack is left to be handled by reputation–a beacon that skips providing an output often will get a reputation for unreliability. The Randao [4] is an Ethereum service that tries to solve this problem by requiring each party that contributes a commitment to also post a performance bond. Anyone who refuses to reveal their random number forfeits the bond. [19] describes an elaborate set of protocols to use verifiable secret sharing and Byzantine agreement to generate public random numbers from $3k$ independent participants, so that the shared random numbers will be trustworthy (and impossible to prevent from being published) so long as at least $k + 1$ participants are trustworthy.

Variants Using Slow Computations [13] takes a different approach to combining contributions from multiple parties. Contributions from the public as well as environmental inputs from a public video camera are hashed together and the hash is published. The inputs are fed into an inherently sequential computationally slow hash function, and much later after the hash is computed the result is published. Since nobody could have known the result of the slow hash function when the inputs were hashed and published, nobody could have influenced the output by deciding what or whether to send an input in. A related approach is considered in [9], in which a computationally slow function is used to produce shared random numbers from Bitcoin or Ethereum block hashes while preventing miners from influencing the resulting random numbers. The same paper describes a set of protocols for ensuring that the computationally slow function is correctly computed, and considers the necessary financial rewards for incentivizing participants to keep verifying the correctness of the computation. Another related possibility to prevent an attacker "hitting the reset button" is to use time-lock puzzles, as described in [17]. If Alice publishes $TL(R_A)$, where

[5] Without these commitments, Alice can always wait for Bob to publish a random number, and then choose hers to control the resulting shared value.

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

$TL()$ is a time-lock scheme with a minimum time to unlock of one hour, and then five minutes later all parties reveal their random numbers, the attack is prevented. Even if Alice wants to hit the reset button (refuse to publish her number to stop the beacon from publishing), she can only delay knowledge of the shared random number for one hour.

Merlin Chains In this paper, we describe still another approach, called a Merlin chain, to address this problem by giving participants a way to credibly commit to being able to recover their 'lost' random numbers after hitting the reset button. This is an example of a common situation, in which a party in a protocol becomes more capable by restricting its future freedom of action[6].

3 Background

Beacons are entities that produce a stream of random numbers [16] (see [3] for a currently-operating example). Each time a beacon releases a random number, it is called a 'pulse'. Beacons have three properties:

1. A beacon will put a random number R, unpredictable to anyone outside the beacon itself, in each message.
2. A beacon will never release a signed random number with a timestamp T before time T (so nobody outside the beacon could have known the random number earlier than that time).
3. A beacon will emit only one random number for each timestamp T.

In order to be useful, the outputs from a beacon must be publicly available and must be immutable. A beacon pulse may have many fields, but only two are really essential: the random number, R, and the timestamp, T.

Blockchains are immutable digital ledger systems and were first used for digital cash with Bitcoin [15]. Each 'block' contains a set of transactions as well as the hash of the previous block (thus forming the 'chain'). They can be implemented in a distributed fashion (without any central authority) and enable a community of users to record transactions in an immutable public ledger. This technology has undergirded the emergence of cryptocurrencies where digital transfers of money take place in distributed systems; it has enabled the success of currencies such as Bitcoin [15] and Litecoin [2]. In such systems, a community of 'miners' maintain a blockchain by competing to solve a mathematical puzzle. The solution is evidence that the miner is performing computation, and for this reason such system are called 'proof-of-work' systems. The 'miner' that solves the current puzzle can then publish the next 'block' which contains recent digital cash transactions. The winning miner receives a block award and may receive fees from included transactions, both in terms of the applicable electronic currency. Some blockchains use other techniques, such as consensus among trusted nodes,

[6] A more general version of this idea appears in [18], applied to many real-world situations that can be modeled by game theory.

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

proof-of-stake, or proof-of-storage. Without modification, our protocol will work only with 'proof-of-work' systems.

Ethereum [22] is a blockchain-based cryptocurrency that supports 'smart contracts'. Contracts are programs whose code and state exist on the public blockchain and they can both send and receive funds while performing arbitrary computations. They can act as a trusted third party in financial transactions, since the code is public but immutable. The programming language used for contract transactions, Solidity [5], is limited in functionality but is Turing Complete [20]. Ethereum charges a fee for contract execution, called 'gas'. The originator of any transaction must pay this fee or the transaction aborts. There is a maximum gas limit, currently 3 000 000, to prevent computationally expensive programs from being submitted to the Ethereum miners (since each miner will execute each transaction in parallel).

3.1 Merlin Chains

In the rest of this paper, we use a sequence of unpredictable numbers we call a *Merlin chain*[7]. This is a (usually long) sequence of values where every value V_x is the hash of the value with the next higher index V_{x+1} (i.e., $V_x = \mathrm{SHA3}(V_{x+1})$). This use of a hash function then provides a series of random values taken from a uniform distribution but where each value is related to the previous value (because the current value is created by hashing the previous value).

A Merlin chain has three important properties:

1. An attacker who has seen all previous entries ($V_{0,1,2,\ldots,j-1}$) in the Merlin chain cannot predict anything about the next entry (V_j).
2. Each entry in the chain works as a *commitment* to the next entry in the chain. Once an entity has revealed V_0, it has no valid choice except to follow this with V_1, then V_2, and so on.
3. By storing V_n offsite, the entity revealing the chain entries can guarantee that even a catastrophic hardware or software failure will not prohibit the production of chain values (as would happen were the chain data lost).

The most important feature of the Merlin chain is that it takes away the choices of the entity using it, while still allowing that entity to produce numbers (unpredictable to everyone else). For the user of the Merlin chain, "Everything not forbidden is compulsory"[21].

4 Preliminary Approaches

In this section, we describe some plausible-sounding strategies to make a beacon. These approaches don't work but will build towards our proposed solution, thus motivating our design choices in the rest of the paper.

[7] The Merlin Chain is named after the character of Merlin in T.H. White's *The Once and Future King*[21], who lives his life backwards in time.

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

4.1 Block hashes

Each block in the Ethereum blockchain is hashed using 256-bit SHA3 and this result is published on the blockchain along with a timestamp. This meets our definition of a beacon in Sect. 3 and one might consider using these hashes as a source of public randomness. However, in this case it turns out that it is possible for the Ethereum miners to influence the beacon results. Consider the situation where a coalition controlling a fraction F of all the processing power of the Ethereum miners is working to predict, control, or influence a block hash. Predicting the block hash would require knowing all transactions to be included in the blockchain up to and including the block whose hash will be used for a random number. Thus, prediction a very short time in advance is sometimes possible for a coalition of miners but prediction far in advance would require control of the whole mining pool and a very visible-to-the-world denial of service attack on the transactions submitted to Ethereum. With respect to control, it's clear that even when $F = 100\,\%$, there is no way for the coalition to control the value of the block hash, since it's the output of a hash function.

However, influencing the block hash is quite feasible. Consider a coalition controlling $F\,\%$ of the total mining power, which wants to force a single bit of the block hash to be a one. The coalition members attempt to mine the next block, but when they reach a valid proof of work (so that they've successfully mined a block) they check to see whether the resulting block hash has the desired bit set. If not, they simply throw the block hash away and keep trying to mine the next block. Table 4.1 shows the result of simulating this attack, for various fractions of mining power controlled by the coalition.

Table 1. Extent to which a coalition of miners can influence one bit of the block hash

Fraction of processing power in coalition	Bias in targeted bit
5 %	0.01
10 %	0.03
20 %	0.06
30 %	0.09
40 %	0.13
50 %	0.17

As the table shows, even a coalition with only 10 % of the miners' processing power can impose a potentially significant amount of bias on a selected bit of the block hash, causing the selected bit to have probability 0.53 of being a one.

4.2 Adding a Producer of Randomness

The above analysis demonstrates why the block hash alone cannot be used as a public source of randomness. We now consider adding an external producer of

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

randomness, moving us closer to a useful solution. The producer sends a random number V, and then the contract produces an output $R = \text{SHA3}(H \parallel V)$, where H is the block hash of the previous block. If the producer does not reveal V until the block hash is calculated, the miners no longer can exert any influence over R. However, in this scenario the producer can choose V after H is generated and thus influence R. In addition, this influence is greater since it is very easy for the producer to compute many R values by simply changing the V input (it is much harder for the miners because to compute a new candidate R value they must create a blockchain block that wins the current block competition).

Our solution to these residual security issues is for the contract to require the producer to generate V prior to H being computed. It does this by requiring that the producer submit the hash of V before it records the value of H to be used. Then only after H is computed by the miners, the producer submits V to the contract. The contract can check that this is the value the producer committed to upfront by simply hashing V. The miners can't influence R because they don't know V when computing the block hash. The producer can't influence R because it can't know the block hash when initially committing to a V value (when it sends the hash of V to the contract). The next sections more formally present this approach and handle a variety of security issues that arise (including the possibility that the producer and miners might collaborate to circumvent the security architecture).

5 Single Producer Contract

In this section we present a contract whose input comes from a single producer and whose output is a beacon. It is designed to produce a 32-byte random number on the blockchain with a maximum frequency of about once every 30 seconds (more precisely once every other Ethereum block). To maximize the usability of the provided beacon service, we recommend that the producer provide input to the contract at some fixed interval greater than 30 seconds.

The producer will provide unpredictable values from a Merlin chain, and so must pre-compute all inputs that will be provided to the contract for its lifetime. Let n represent the chosen number of input values. The value V_n is chosen randomly, $V_{n-1} = \text{SHA3}(V_n)$, $V_{n-2} = \text{SHA3}(V_{n-1})$ and so on until the computation of V_1. The Merlin values are released to the contract starting with V_1 (the reverse of the order in which they were generated).

The function `B`() will provide the block number in which some input or output is processed by the contract. The function `BH`() provides the block hash of some block number. Lastly, the function `timestamp`() provides the Ethereum timestamp for some block.

The producer will periodically provide the contract some message containing a V_x value along with a timestamp U_x. The contract in response may produce a random value R_x and a timestamp T_x (note that in certain circumstances the contract may not publish an R_x value). T_x will be the time before which

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

no entity could have predicted R_x, including the producer (usually this will be about 30 seconds prior to R_x being publicly released).

The core idea is that for each message (containing some V_x) received from the producer, the contract will attempt to generate R_x using as input both an Ethereum block hash and V_x. The block hash used will be one that was generated after V_{x-1} was submitted to the contract but before V_x was submitted. This way the miners can't know V_x when the relevant block hash is created and they can't then influence R_x (assuming that the producer and a group of miners are not colluding). Likewise, the producer can't influence R_x because V_x was predetermined by the submission of V_{x-1} and this was done before the relevant block hash was generated. T_x is then generated by taking the minimum of U_{x-1} and the Ethereum timestamp for the block in which V_{x-1} was submitted (taking the minimum eliminates malicious producers from being able claim a Merlin value was revealed later than it was revealed). The actual protocol is slightly more complicated (to account for unexpected input, messages submitted too early, and Ethereum implementation issues). It is outlined below.

5.1 Single Producer Protocol

For each message, with associated V_x and U_x values, the contract checks the following prior to accepting the input:

1. The message must come from the Ethereum address registered in the contract as the one pertaining to the producer.
2. V_x must be the next value on the producer's Merlin chain (i.e., $V_{x-1} = \text{SHA3}(V_x)$). This ensures that the producer can't influence R_x.

However, V_x is not considered 'valid' for producing a random number, R_x, and a timestamp, U_x, unless the following hold (assume that R_y is the last produced R value, usually R_{x-1}):

1. The block number in which V_x is processed by the contract must be at least 2 more than the block number where the last valid V value was processed by the contract[8] (i.e., $\texttt{B}(V_x) \geq \texttt{B}(R_y)+2$). This ensures that the miners can't use the block hash to influence R_x (since miners can discard a block after computing the block hash).
2. The contract must have access to $\texttt{BH}(\texttt{B}(R_y)+1)$. The contract will retrieve this given any activity (either from the producer or any customer retrieving random numbers) but Ethereum only provides access to the blockhashes for the last 256 blocks. If this is not available[9], the contract will output a public

[8] The producer can ensure this is always true by verifying that it doesn't send the next (V_x, U_x) message until it has seen at least one block go past on the blockchain since the last random output.

[9] This availability could be ensured by setting up another provider which does nothing except send a message to the lighthouse contract once every 256 blocks (since blockhashes produced more than 256 blocks in the past are irretrievable in the Ethereum system).

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

error log message and reset the block hash used to be the one from the next Ethereum block (i.e., $\texttt{BH}(\texttt{B}(V_x)+1)$.)

If these conditions are satisfied, R_x and T_x are generated according to the following formulas:

$$R_x = \text{SHA3}(V_x \parallel \texttt{BH}(\texttt{B}(R_y) + 1)) \tag{1}$$

$$T_x = \min(\texttt{timestamp}(\texttt{B}(R_y)), U_x) \tag{2}$$

Figure 5.1 provides an example of two valid messages arriving to the contract and shows how the contract uses them to generate R and T values. In the figure, we use b_x to represent the block number at which some V_x arrived to the contract.

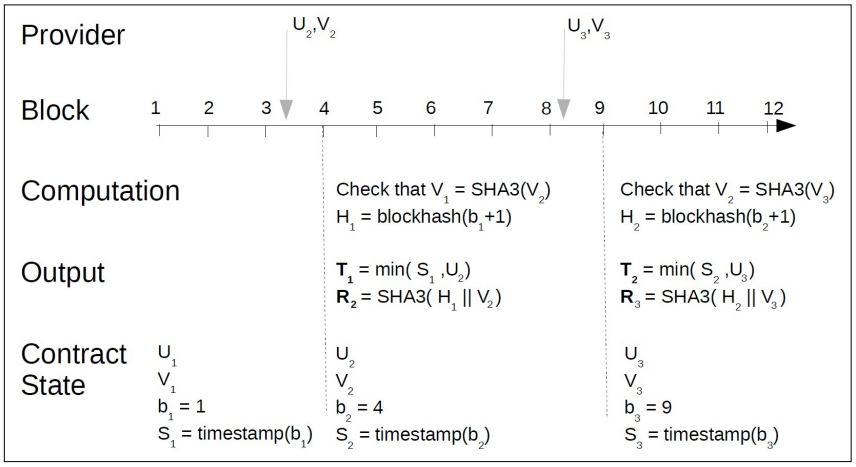


Fig. 1. The Single Producer Protocol

5.2 Mitigated Security Flaws

We now analyze different attack scenarios and discuss how they are mitigated:

1. The producer might try to use V_x to influence R_x. However, this won't work because V_x is fixed based on V_{x-1} and the block hash used was generated after V_{x-1} was revealed.
2. The producer might try to delay sending V_x to influence R_x. This was possible in earlier designs where the block hash used for R_x was the one prior to V_x. In this case, the producer could watch the block hashes being produced and then quickly issue a pulse after a desirable block hash was published on the blockchain. We mitigated this by fixing the block hash to be used to be $\texttt{BH}(\texttt{B}(R_y)+1)$.

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

3. A producer could purposefully submit a message too early. However, the message is rejected as invalid and this simply updates the Merlin value V stored on the contract (which is fine since the relevant block hash has not yet been generated).
4. Because of a design limitation in the Ethereum Solidity language, the contract is only able to retrieve up to the last 256 block hashes (about 68 minutes of blockchain operation). The threat is that prior to revealing V_x, a producer might calculate R_x and find it undesirable. The producer may then wait 256 blocks prior to releasing V_x so that the correct blockhash can't be retrieved. This effectively changes the result since the contract can no longer retrieve the block hash $\texttt{BH}(\texttt{B}(R_y)+1)$. We mitigate this by enabling the contract to retrieve the block hash during any transaction (including customer retrieval of V values). Thus, even if the producer waits, other activity will enable the contract to retrieve the needed value within the period of availability. If this does not happen, the contract emits an error log and resets the block hash used to be one not yet generated. To strongly mitigate this problem for little used beacons, the contract owner should arrange for some party to access the contract at least every 256 blocks to ensure that the block hash is retrieved within the time constraints.
5. Miners (not collaborating with the producer) may try to affect R_x by throwing out discovered blocks that have block hashes that will produce undesirable random numbers. However, miners must compute the block hash to be used, $\texttt{BH}(\texttt{B}(R_y)+1)$, prior to V_x being revealed and thus this won't work. This is why the block number in which V_x is processed by the contract must be at least 2 more than the block number where the last valid V value was processed by the contract. Note that a separate vulnerability arises if one uses the block hash of the block where the last V value was processed and so that was not available as an option.
6. The contract owner has only the ability to register and de-register the producer. De-registration only occurs after a set number of blocks (eliminating the possibility of the contract owner seeing a revealed V_x value message and trying to remove the producer before the contract processes it). With respect to registering a producer, its first message is used only to set the initial V_x Merlin value and so registration can't be used to influence or control the V values.
7. An attacker could compromise the producer but they would still have to produce the values on the pre-determined Merlin chain. To influence the results they would have to collaborate with a group of miners (this attack is discussed in the next section).
8. The producer who has sent some V_x can predict an R_{x+1} after the next block hash has been calculated. Our mitigation of this is for the contract to publish T_{x+1} which indicates at what time the producer could have predicted R_{x+1} (this is usually less than a minute in the past).
9. Since the producer can predict the next R value, it may not send some V_x because revealing it will generate an R_x that is deemed undesirable (e.g., the producer made a bet on the outcome). However, then it must stop producing

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

any values because the contract will wait for V_x. We mitigate this by requiring producers to keep an offsite backup copy of their Merlin chain. This does not stop a producer from refusing to reveal V_x. However, it does eliminate their ability to claim an inability to reveal due to a hardware failure or natural disaster. This weakness could be more strongly mitigated in future work by requiring the producer to submit a timelock puzzle [17] along with each V value. Such puzzles would allow contract customers to perform an expensive computation on a V_{x-1} to reveal any V_x withheld by the producer. The producer couldn't lie at the right moment because they can't predict an R_x when sending in a V_{x-1} (and lying in general is easy to detect by solving the timelock puzzle).

5.3 Residual Security Flaw

The remaining security flaw is that the producer (or an attacker that has compromised the producer) may collaborate with a set of miners to attempt to influence, but not control, R_x. The malicious producer would provide the collaborating miners the value V_x, enabling them to compute a candidate R_x if they successfully mine block $\mathtt{B}(R_y)+1$. If this is a desirable outcome, they publish the completed block to the mining community. If not, they discard the completed block and lose the associated block reward and transaction fee funds. We mitigate this attack with our multiple producer contract.

6 Multiple Producer Contract

The multiple producer contract permits multiple producers to submit values to mitigate the possibility of a single producer collaborating with a group of miners. Each producer is handled independently using the single producer protocol from Sect. 5.1 (with some exceptions) and the contract maintains a beacon independently for each producer. When all beacons have pulsed, the contract pulses R and T values derived from the combination of beacon pulses. We call this combined output a lighthouse pulse. We change our notation to handle multiple producers as follows. We identify each producer with an integer, add this as a subscript to each variable, and let each variable refer to its most recent value. Thus, R_1 references the most recent R value for producer 1. We use R_L and T_L to refer to the most recent lighthouse output.

The contract handles each producer using the single producer protocol from Sect. 5.1 with the following exceptions (that force the beacons to progress in a lockstep manner):

1. Once pulsed, beacons are not allowed to pulse again until the lighthouse pulses. If a producer sends additional messages prior to the lighthouse pulse, they are marked as invalid.
2. The 'R_y' references in Sect. 5.1 now correspond to the R_L values produced by the lighthouse (not the particular producer's beacon). This causes all beacons to use the same block hash for each beacon pulse.

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

Once all beacons have pulsed, the lighthouse pulses as follows:

$$R_L = R_1 \oplus R_2 \oplus ... \oplus R_m \tag{3}$$

where $\oplus$ is exclusive or (XOR) and m represents the number of participating beacons. This has the convenient feature that the lighthouse output using only a single producer is identical to that producer's beacon output.

$$T_L = \max(T_1, T_2, ..., T_m) \tag{4}$$

While not necessary, the lighthouse will work more efficiently if all producers synchronize their time (e.g., using the Network Time Protocol [14]) and issue messages at some agreed upon interval.

Each producer's beacon follows the single producer protocol and thus has the same security advantages. The small exceptions to the protocol in Sect. 6 do not affect the per beacon security analysis. Each beacon is still secure unless both the producer and a group of miners collude. The small exceptions cause the beacons to produce in lockstep. Due to the common block hash used, no beacon can predict the lighthouse output until after the block hash has been calculated (at which point the potentially malicious beacon has already committed to its next value).

This leaves open the possibility that a set of t malicious producers could collaborate on which will refuse to reveal in order to try to manipulate 2^t bits. However, any such activity will be publicly viewable, will cause the lighthouse to stop production, and cause the contract owner to deregister any such producers. The producers can't claim technical failures because they are required to keep a backup copy of their Merlin chains.

The only way to influence the R_L values then is for all producers to collaborate with each other and also with a group of miners. They can then throw out successfully mined but undesirable blocks (those that would produce an unwanted R_L value). In no situation can the R_L value be controlled (i.e., directly chosen).

However, there is one remaining weakness that must be addressed. If all producers colluded when initially creating their Merlin chains then they could use the same V value making the beacons all pulse the same value. If there are an even number of producers, this will force R_L to be 0 since it used XOR. To mitigate this, our contract simply refuses to pulse an R_L value equal to 0. This obviously reduces the output state space by 1.

7 Empirical Work

We implemented our multiple producer contract using the Solidity language [5] and deployed it to the Ethereum test network. The test network is identical to the production network except the Ether has no real world value. Given that our system does not rely on the transfer of digital assets, the test network works just as well for our lighthouse as the real Ethereum network. We also created

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

distributed application (DApp) software to enable producers to submit pulses to the contract and for customers to retrieve R values. We used multiple producers and tested the contract's ability to generate the independent beacon values as well as the lighthouse values.

We found that coding our contracts in Solidity was rather straightforward. The main challenges were that we easily ran out of gas (performed too much computation) or ran out the very limited stack space for individual functions. However, creating the beacon software that submitted pulses to the contract was much more difficult since very little documentation exists on how to enable a program outside of Ethereum to communicate with an Ethereum contract.

We didn't use the main Ethereum network for our empirical testing because the current contract execution prices made it too expensive (due to Ether currency speculation). The price of Ethereum has risen from $8.00 per Ether to $358 per Ether in six months [1] (as of June 20, 2017) and the gas fees have not dropped accordingly although Ethereum has a mechanism to do so. Table 2 shows the costs of the main functions in terms of Ether, USD on January 2017, and USD on June 2017.

If a producer pulses once a minute, the cost using June 2017 prices would be $673,000 USD per year. Using January 2017 prices, it would be $17,870 USD (which the authors believe to still be excessively high).

Table 2. Approximate Ether and USD Costs of Lighthouse Functions as of 2017-06-15

Request Type	Gas	Ether	USD (2017-06-20)	USD (2017-01-01)
Contract Deployment	1.9M	.0399	$14.29	$0.32
Register Producer	205k	.0043	$1.54	$0.035
Producer Pulse	200k	.0042	$1.50	$0.034
Retrieve Output	22k	.000462	$0.17	$0.0037

Due to these cost issues, future implementations of our contract may use an alternate to Ethereum or a private Ethereum network. This latter approach is fully supported by the Ethereum development tools and would be privately managed but publicly accessible. Another option is to design the system so that the users of the system pay the cost by charging a small fee for each delivered random number.

8 Conclusion

It is possible to use cryptocurrency smart contracts to create a distributed consensus protocol to publicly produce a stream of trustworthy random numbers. Our contract design eliminates both prediction and control attacks. Neither is it possible for any entity to change the values once published. What is possible is that the output might be indirectly influenced without being directly controlled but this can be mitigated by registering multiple producers.

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

References

1. Ethereumprice, https://ethereumprice.org/, accessed: 2017-06-27
2. Litecoin, https://litecoin.org/, accessed: 2017-06-16
3. National Institute of Standards and Technology Beacon Program, https://beacon.nist.gov/home, accessed: 2017-06-16
4. Randao, https://github.com/randao/randao, accessed: 2017-07-10
5. Solidity language, https://solidity.readthedocs.io/en/develop/, accessed: 2017-06-16
6. www.random.org, https://www.random.org/, accessed: 2017-07-10
7. Baignères, T., Delerablée, C., Finiasz, M., Goubin, L., Lepoint, T., Rivain, M.: Trap me if you can - million dollar curve. IACR Cryptology ePrint Archive 2015, 1249 (2015)
8. Bonneau, J., Clark, J., Goldfeder, S.: On bitcoin as a public randomness source. IACR Cryptology ePrint Archive 2015, 1015 (2015)
9. Bunz, Goldfeder, B.: Proofs-of-delay and randomness beacons in ethereum. IEEE Security & Privacy on the Blockchain (2017), http://www.jbonneau.com/publications.html
10. Clark, J., Hengartner, U.: On the use of financial data as a random beacon. IACR Cryptology ePrint Archive 2010, 361 (2010), http://eprint.iacr.org/2010/361
11. Fischer, M.J., Iorga, M., Peralta, R.: A public randomness service. In: Security and Cryptography (SECRYPT), 2011 Proceedings of the International Conference on. pp. 434–438. IEEE (2011)
12. Kelsey, J.: The new nist beacon protocol and combining beacons (2017)
13. Lenstra, A.K., Wesolowski, B.: A random zoo: sloth, unicorn, and trx. IACR Cryptology ePrint Archive 2015, 366 (2015)
14. Mills, D., Martin, J., Burbank, J., Kasch, W.: RFC 5905: Network Time Protocol Version 4: Protocol and Algorithms Specification. Internet Engineering Task Force (IETF), 2010. tools. ietf. org/html/rfc5905
15. Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system (2008)
16. Rabin, M.O.: Transaction protection by beacons. Journal of Computer and System Sciences 27(2), 256–267 (1983)
17. Rivest, R.L., Shamir, A., Wagner, D.A.: Time-lock puzzles and timed-release crypto (1996)
18. Schelling, T.C.: The Strategy of Conflict. Oxford University Press (1960)
19. Syta, E., Jovanovic, P., Kokoris-Kogias, E., Gailly, N., Gasser, L., Khoffi, I., Fischer, M.J., Ford, B.: Scalable bias-resistant distributed randomness. In: 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017. pp. 444–460 (2017), https://doi.org/10.1109/SP.2017.45
20. Turing, A.M.: On computable numbers, with an application to the entscheidungsproblem. Proceedings of the London mathematical society 2(1), 230–265 (1937)
21. White, T.H.: The Once and Future King. Ace Books (1987)
22. Wood, G.: Ethereum: A secure decentralised generalised transaction ledger. Ethereum Project Yellow Paper 151 (2014)

Kelsey, John; Mell, Peter; Shook, James.
"Cryptocurrency Smart Contracts for Distributed Consensus of Public Randomness."
Paper presented at 19th International Symposium on Stabilization, Safety, and Security of Distributed Systems, Boston, MA, United States.
November 6, 2017 - November 8, 2017.

Electron Microscopy (Big and Small) Data Analysis With the Open Source Software Package HyperSpy

Francisco de la Peña[1,11,13], Tomas Ostasevicius[1], Vidar Tonaas Fauske[2], Pierre Burdet[1], Petras Jokubauskas[3], Magnus Nord[2,4], Mike Sarahan[5], Eric Prestat[6], Duncan N. Johnstone[1], Joshua Taillon[7], Jan Caron[8], Tom Furnival[1], Katherine E. MacArthur[8], Alberto Eljarrat[9], Stefano Mazzucco[7], Vadim Migunov[8], Thomas Aarholt[10], Michael Walls[11], Florian Winkler[8], Gaël Donval[12,13], Ben Martineau[1], Andreas Garmannslund[2], Luiz-Fernando Zagonel[14] and Ilya Iyengar[1]

[1.] Department of Materials Science and Metallurgy, University of Cambridge, Cambridge, U.K.
[2.] Department of Physics, NTNU, Trondheim, Norway
[3.] Institute of Geochemistry, Mineralogy and Petrology, University of Warsaw, Poland
[4.] School of Physics and Astronomy, University of Glasgow, Glasgow, U.K.
[5.] SuperSTEM, STFC Daresbury Laboratories, Warrington, U.K.
[6.] School of Materials, The University of Manchester, U.K.
[7.] Material Measurement Laboratory, NIST, Gaithersburg, MD, U.S.A.
[8.] Forschungszentrum Jülich GmbH, Jülich, Germany
[9.] Laboratory of Electron NanoScopies, Universitat de Barcelona, Barcelone, Spain
[9.] Department of Materials, University of Oxford, Oxford, United Kingdom
[11.] Laboratoire de Physique des Solides, Paris-Sud University, France
[12.] Institut des Matériaux Jean Rouxel (IMN), Université de Nantes, Nantes, France
[13.] CEA Grenoble, Grenoble, France
[14.] Brazilian Nanotechnology National Laboratory, CNPEM, Campinas, Brazil
[15.] Simula Research Laboratory, Lysaker, Norway

Advances in scientific instrumentation, driven by the evolving needs of science and technology, are dramatically increasing the amount of experimental data generated. Microscopy is following this trend across all domains, posing common challenges for data analysis. Developing new methods of data analysis to harness the wealth of physical insight contained in this data is now the key to maximising the advancement of scientific understanding using these tools. Here, the focus is on scanning transmission electron microscopy, where modern microscopes enable multiple signals with dimensions of energy, reciprocal space and time to be acquired across real space length scales from ~0.1 nm to 10 μm.

Electron microscopy data analysis is mostly performed using dedicated proprietary software at present. These software packages typically offer a familiar and easy to use Graphical User Interface (GUI) for performing common data analysis tasks. However, they are often limited in comparison with the potential to develop innovative data analysis methods based on the platform of modern scientific programming languages. The success of ImageJ [1]—a program for scientific image processing— has shown that open source development of microscopy software can lead to high quality data analysis tools often where no proprietary alternatives exist.

HyperSpy [2] is an open source Python software package for multi-dimensional data analysis that includes a wide range of features for electron microscopy. Python is increasingly recognised as the *lingua franca* of scientific computing. It offers numerous technical advantages as a programming language but the primary advantage for microscopy data analysis is the availability of an outstanding range of high-quality scientific libraries on which to draw. HyperSpy (and its GUI, HyperSpyUI [3]) is

Aarholt, Thomas; Burdet, Pierre; Caron, Jan; Donval, Gael; Eljarrat, Alberto; Fauske, Vidar; Furnival, Tom; Garmannslund, Andreas; Iyengar, Ilya; Jokubauskas, Petras; MacArthur, Katharine; Martineau, Ben; Mazzucco, Stefano; Migunov, Vadim; Nord, Magnus; Ostasevicius, Tomas; Prestat, Eric; Sarahan, Mike; Taillon, Joshua; Walls, Michael; Winkler, Florian; Zagonel, Luiz-Fernando; de la Pena, Francisco.
"Electron Microscopy (Big and Small) Data Analysis With the Open Source Software Package HyperSpy."
Paper presented at Microscopy & Microanalysis 2017, St. Louis, MO, United States. August 6, 2017 - August 10, 2017.

available for Linux, macOS and Windows. Under development since 2007, it has a large community of users and developers in the electron microscopy community and beyond. The main development of HyperSpy is an elegant yet powerful syntax for visualizing, analyzing, accessing and storing multi-dimensional datasets (regardless of their size). HyperSpy complements the scientific Python ecosystem by significantly simplifying the usage of external scientific libraries when operating on multi-dimensional data. The aim is to foster innovation in the field by lowering the entry barrier to advanced interactive data analysis and providing a quality peer-reviewed channel to distribute innovative algorithms. Indeed, many of the HyperSpy developers started as users and, in time, they started contributing new features that they had developed for their research.

Features specific to electron microscopy include support for most common file formats along with functions for analysing electron energy-loss spectra (EELS), energy dispersive X-ray (EDX) spectra and performing holography. Further, HyperSpy provides easy access to principal component analysis (PCA), blind source separation and multi-dimensional model fitting in 1 and 2 dimensions. Most of its routines (as well as external functions) can be easily run in parallel and using out-of-memory computational schemes enabling seamless big data analysis.

[1] M. Abràmoff *et al.*, Biophotonics international **11.7** (2004), p. 36.
[2] F. de la Peña *et al.* HyperSpy v.1.1.2 (2016). doi:10.5281/zenodo.60697. http://hyperspy.org
[3] V.T. Fauske, HyperSpyUI. http://hyperspy.org/hyperspyUI/
[4] The HyperSpy project has not received direct funding and all authors are grateful for support within their respective research groups to make the data analysis tools that they have developed available open-source.

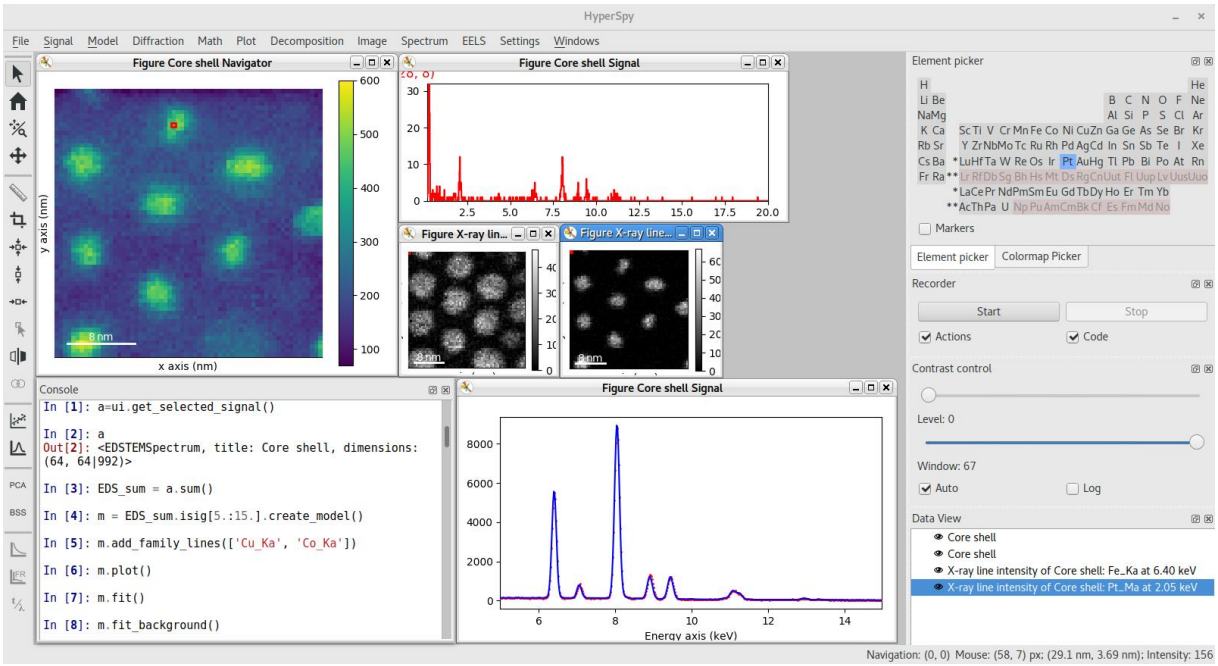


Figure 1. Screenshot of HyperSpyUI [3] performing quantification by curve fitting of an EDX spectrum image of FePt core-shell nanoparticles. The embedded IPython console (bottom-left) enables interactive Python scripting.

Thermodynamic Analysis of Classical and Quantum Search Algorithms

Ray Perlner[1] and Yi-Kai Liu[1,2]

[1]National Institute of Standards and Technology (NIST)
Gaithersburg, MD, USA
[2]Joint Center for Quantum Information and Computer Science (QuICS)
University of Maryland, College Park, MD, USA

Abstract. We analyze the performance of classical and quantum search algorithms from a thermodynamic perspective, focusing on resources such as time, energy, and memory size. We consider two examples that are relevant to post-quantum cryptography: Grover's search algorithm, and the quantum algorithm for collision-finding. Using Bennett's "Brownian" model of low-power reversible computation, we show classical algorithms that have the *same* asymptotic energy consumption as these quantum algorithms. Thus, the quantum advantage in query complexity does not imply a reduction in these thermodynamic resource costs. In addition, we present realistic estimates of the resource costs of quantum and classical search, for near-future computing technologies. We find that, if memory is cheap, classical exhaustive search can be surprisingly competitive with Grover's algorithm.

Key words: Quantum algorithms, thermodynamics of computation, reversible computation, quantum cryptanalysis, Grover search, collision finding

1 Introduction

1.1 Motivation

Quantum computers are believed to solve a number of important problems, in areas such as number theory, physical simulation and combinatorial search, asymptotically faster than classical computers [17]. How large is this quantum speedup? There is now a rich body of literature that analyzes different quantum speedups, using idealized models of computation, such as the quantum circuit model, and quantum oracle models (i.e., quantum query complexity). In most of this work, the models of computation are intentionally made simple, in order to allow rigorous analyses; for instance, in quantum query complexity, one only accounts for the number of oracle queries made by the algorithm, while disregarding the actual time-complexity of the algorithm.

In order to obtain more realistic analyses of quantum speedups, one must consider more realistic models of quantum computation, which take into account time-complexity (not just query complexity), as well as possible time-space tradeoffs due to the use of parallel processors, and restrictions on the connectivity of the different components of the computer. There have been a few results of this type [8, 5, 10, 4]. Some of these results suggest that practical quantum speedups may fall short of the most idealized theoretical predictions.

In this paper we give a new analysis of some fundamental quantum speedups, from the perspective of thermodynamics. A computer can be viewed as an engine that converts

Liu, Yi-Kai; Perlner, Ray.
"Thermodynamic Analysis of Classical and Quantum Search Algorithms."
Paper presented at Quantum Information Processing (QIP 2018), Delft, Netherlands. January 15, 2018 - January 19, 2018.

energy into computational work; and one may ask how much energy is consumed by running a particular algorithm. To answer this question, one must specify the model of computation. Following Bennett [7], one can consider three different models:

1. "Conventional" computation (as in present-day electronic computers) uses operations that are irreversible and deterministic. Since the operations are irreversible, at temperature T, each operation must dissipate at least $kT(\ln 2)$ of energy, where k is Boltzmann's constant. (This is the "Landauer limit.")
2. "Ballistic" computation (as in billiard ball models [11]) uses operations that are reversible and deterministic. Since the operations are reversible, in principle, they can dissipate zero energy. The computer is assumed to be isolated from all sources of thermal noise, hence energy barriers are not needed to prevent errors.
3. "Brownian" computation (as in DNA computation [6] and adiabatic circuits [3]) uses operations that are reversible and stochastic. Each operation dissipates a small amount of energy ε, which may be less than kT, so that the computation "drifts" forward, even in the presence of strong thermal noise.

By using reversible operations, models (2) and (3) are able to compute using an amount of energy per operation that is below the Landauer limit. However, it has been argued that model (2) is unrealistic, because it cannot be made fault-tolerant, i.e., it is sensitive to small errors when performing a long computation. In this paper, we use model (3), the Brownian model of computation, as our standard.

We consider quantum and classical algorithms for unstructured search and collision finding. Using the Brownian model of computation, we analyze the cost of these algorithms in terms of time, energy consumption, and memory size. The motivation for this study comes from post-quantum cryptography. For instance, an algorithm for unstructured search can be used to recover the secret key of a block cipher, given a sufficient number of plaintext-ciphertext pairs; while an algorithm for collision-finding can be used to compromise the security of a cryptographic hash function. In order to design block ciphers and hash functions that achieve sufficiently high levels of security, one must make detailed estimates of the resources required to carry out both quantum and classical cryptanalytic attacks.

1.2 Our Results

For the problem of collision finding, previous work suggested that quantum algorithms were unlikely to provide an asymptotic advantage in terms of circuit size (despite using fewer oracle queries) [8]. Our thermodynamic analysis leads to a similar conclusion. We compare in detail the classical collision finding algorithm of Van-Oorschot and Wiener [21], and the quantum collision finding algorithm of Brassard, Høyer, and Tapp (BHT [9],) including parallelized generalizations of BHT. We find that the energy consumption required to search for collisions on a range of size N using a memory of size $M < O(N)$ in time t is $O\left(\frac{N}{Mt}\right)$, regardless of the choice of algorithm.

While we focus on the collision finding problem, it should be noted that similar analysis may also be applied to the Claw Finding problem, which seeks to find collisions between two functions with domain sizes N_1 and N_2 . A quantum algorithm was proposed by Tani for this purpose [19]. This may be compared to the algorithm given by Van-Oorschot and Wiener [21]. In this case, the energy consumption required to find a claw using a memory

of size $M < O(N_1 + N_2)$ in time t is $O\left(max\left(\frac{N_1 N_2}{Mt}, \frac{N_1}{t}, \frac{N_2}{t}\right)\right)$ for the quantum algorithm, and $O\left(\frac{(N_1+N_2)^3}{M^2 t}\right)$ for the classical algorithm.

For the problem of unstructured search, it was known previously that Grover's algorithm does achieve a quadratic speedup over classical exhaustive search, both in terms of circuit size, and in terms of oracle queries. Quite surprisingly, we do *not* find a quantum advantage using our thermodynamic analysis. On the contrary, we find that a Brownian implementation of classical random search can achieve the *same* asymptotic performance as Grover's algorithm (up to logarithmic factors), where we measure the performance in terms of running time, memory size and energy consumption.

To show this, we use a variant of the Brownian model, where certain steps in the computation are *unpowered*, in the sense that we set $\varepsilon = 0$, so that no energy is dissipated, and the computation is simply driven by random thermal noise, with equal probability of moving forwards or backwards. Energy consumption is instead dominated by memory initialization costs. This model may be of independent interest. Our analysis shows that, in order to find a preimage within a domain of size N using a memory of size M in time t, both Grover's algorithm and unpowered classical search require an energy consumption of $O(\frac{N}{t})$, regardless of the memory size.

Finally, we turn to a more detailed comparison of Grover's algorithm and (powered and unpowered) classical search. Unlike the case for quantum versus classical collision search, here there are some plausible reasons why Grover's algorithm may be more efficient than classical search in practice. In particular, for unpowered preimage search, the independence of memory size and energy consumption relies on a heuristic assumption of scale invariance. If we remove this assumption, and instead assume that unpowered preimage search can only be efficiently implemented at a fixed temperature scale T, we find that unpowered preimage search is significantly more memory intensive than Grover's algorithm. Additionally, unpowered preimage search is less efficient than Grover's algorithm when oracle queries have a large memory complexity, although this is only a minor problem in the typical scenario where the memory complexity of oracle queries scales logarithmically with N.

1.3 Near-Future Computing Technologies

We end by making some quantitative estimates, based on the hypothetical scaling of near-future computing technologies. This is necessarily somewhat speculative. Nonetheless, we argue that one can draw some rough conclusions regarding applications such as brute-force cryptanalysis of block ciphers, which takes on the order of 2^{80} or 2^{96} operations, and thus is comfortably in the asymptotic regime. By making rough calculations based on asymptotic scaling, one can draw some conclusions that do not depend too much on any particular type of qubit, or any particular scheme for quantum error correction.

Our results suggest that the cost of constructing computing hardware (e.g., memory and CPU's) is a major factor that determines the practical advantage of running Grover's algorithm (as compared to classical brute-force search). If memory and CPU's are cheap, then one can use classical search and still be competitive with Grover's algorithm, simply by building enormous data centers. This seems obvious, on a qualitative level. Our quantitative estimates show that, in fact, classical computing technology can do surprisingly well. In particular, it is possible to envision a possible future state of technology where unpowered classical preimage search could outperform both Grover search and powered

classical search. Such a scenario could occur if memory costs could be brought very close to fundamental thermodynamic limits, but quantum computers could not be implemented without very low operating temperatures and expensive error correction.

This paper is organized as follows. In Sections 2 and 3, we describe the model of Brownian computation, and its unpowered variant. In Sections 4 and 5, we give simple analyses of quantum and classical algorithms for collision search and preimage search, focusing on energy costs and time-space tradeoffs. In Sections 6 and 7, we investigate the cost of preimage search at constant temperature and power, and we estimate the cost of implementing the oracle that checks each solution. Finally, in Section 8, we make some detailed estimates of quantum-versus-classical speedups for some hypothetical future computing technologies, and in Section 9, we conclude.

2 Powered Brownian Computation

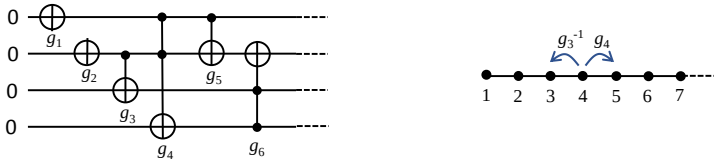


Fig. 1. A Brownian computation is described by a logically reversible circuit, whose gates are to be applied in some specified order. The dynamics of the Brownian computer are described by a random walk along a 1-dimensional chain.

Brownian computers are assumed to operate near thermal equilibrium at a finite temperature, T. A program for a Brownian computer consists of a logically reversible circuit, whose gates (denoted $g_1, g_2, \ldots, g_m$) are to be applied in some specified sequential order. The time-evolution of the Brownian computer may be described as a random walk on a 1-dimensional chain, where the i'th vertex corresponds to the state of the computation after the first $i-1$ gates $g_1, g_2, \ldots, g_{i-1}$ have been applied. A forward step from vertex i to vertex $i+1$ corresponds to applying the i'th gate g_i, and a backwards step from vertex $i+1$ to vertex i corresponds to undoing the i'th gate g_i. (See Figure 1.) In the absence of any driving force, forward and backward steps occur with equal probability. In order for a computation to proceed forward at a nonzero rate, a driving force, dissipating an energy of ε per gate, is imposed. This causes forward steps to occur with $e^{\frac{\varepsilon}{kT}}$ times greater probability than backward steps, resulting in a net forward computation rate proportional to $\frac{\varepsilon}{kT}$, for ε small compared to kT.

Note that the Brownian model of computation can be generalized to work with quantum circuits, provided that all operations are unitary, and all measurements are deferred to the

Liu, Yi-Kai; Perlner, Ray.
"Thermodynamic Analysis of Classical and Quantum Search Algorithms."
Paper presented at Quantum Information Processing (QIP 2018), Delft, Netherlands. January 15, 2018 - January 19, 2018.

end of the computation. Essentially, this amounts to running a quantum computer with a classical controller that behaves in a Brownian fashion.

2.1 Energy Consumption and Running Time

For the Brownian model, we can derive an asymptotic scaling for the relation between per-gate time and per-gate energy, assuming a fixed temperature T. Strictly speaking this analysis can only be extended to a range of possible temperatures under an assumption that physics is scale invariant within that range. However, we can give a somewhat heuristic argument specifying a lower bound, independent of temperature, on the per-gate energy ε required to perform G sequential operations in time t. Briefly, if we assume that Brownian motion within a reversible circuit can be modeled as a series of ballistic motions, each with typical energy scale kT and each completing $O(1)$ gates, then we can apply the Margolus-Levitin theorem [16] to bound the total rate at which forward and backward transitions occur in the circuit. This suggests that the rate at which gates are traversed due to Brownian motion is no more than $\frac{4kT}{h}$. Combining this with the expectation that approximately $G \cdot \frac{kT}{\varepsilon}$ total transitions are required to complete G sequential operations, we obtain the bound:

$$t > G \cdot \frac{kT}{\varepsilon} \cdot \frac{h}{4kT} = \frac{hG}{4\varepsilon}$$

or equivalently:

$$\varepsilon > \frac{hG}{4t}.$$

As the above argument is somewhat heuristic, for the remainder of this paper we will ignore the small unitless factors in the above formula, and simply use $\varepsilon \sim \frac{\hbar G}{t}$ when we need to give a concrete estimate of the dissipation energy required to perform a serial computation at a desired rate.

2.2 Fault Tolerance

Some additional costs are required in order for Brownian computation to be achieved fault-tolerantly. Energy barriers must be imposed to prevent transitions to physical states outside the reversible circuit, representing the prescribed computation path. In order to suppress the probability of such undesirable transitions so that a circuit of size G can be completed with high probability, the size of these energy barriers must at least be on the order of $kTln(\frac{kT}{\varepsilon} \cdot G)$. Additionally, dissipating a "latching" energy of about $kTln(\frac{kT}{\varepsilon})$ during the computation's final step is required to suppress backwards transitions once the computation has reached its halting state. These costs are described in detail in [7].

The above costs may, however, be assumed to be negligible in a number of important cases: In particular, the latching energy will be negligible when $\frac{\varepsilon}{kT}$ is at least logarithmically more than $\frac{1}{G}$. Additionally, establishing energy barriers to non-computational paths is likely to be a negligible cost when a description of the circuit can be expressed in a physically compact form, for example, using looping constructs. More precisely, if we assume that the circuit can be compressed into a program with memory requirement m_0 (including both the memory required to store the program and the data it acts on), then

the cost of imposing energy barriers should be on the order of $m_0 \cdot kTln(\frac{kT}{\varepsilon} \cdot G)$. This cost is negligible as long as $\frac{\varepsilon}{kT}$ is significantly larger than $e^{-\frac{G}{m_0}}$.

In fact, the initialization cost may be less than this, since it may be more proper to think of the initialization process as rearranging the energy barriers already present in the available raw materials for constructing our computer. The cost is then determined by the Landauer limit and the information content of the circuit, including appropriately large energy barriers. Since the size of these barriers does not need to be precislely specified, but merely bounded above $kTln(\frac{kT}{\varepsilon} \cdot G)$, the information content of the circuit may grow sublogarithmically with G. All we can say with confidence is that the information content of the circuit is at least m_0, and therefore the initialization energy is at least on the order of $m_0 \cdot kT$.

Finally, it is worth commenting on the feasibility of Brownian computation for quantum computers. Brownian computation was originally proposed as a way to improve the thermodynamic efficiency of classical computation. It should be noted that many of the techniques that have been proposed for fault tolerance in quantum compuation are thermodynamically irreversible, in particular, syndrome measurement and magic state preparation. These techniques cannot be used in a Brownian mode of computation. However, there are some proposed techniques, such as the use of Fibonacci anyons for universal quantum computation [20], that may be able to achieve fault tolerance without requiring significant thermodynamic irreversibility (although even in such cases, the cost of fault tolerance is believed to be polylogarithmic in the size of the circuit[1].) We will therefore optimistically assume that quantum operations can be implemented in a Brownian fashion.

3 Unpowered Brownian Computation

For some of our results, we will use a variant of the Brownian model of computation, where the intermediate steps in the computation are *unpowered*. More precisely, we dissipate energy when initializing the state of the computer, and when reading the final output; but for the intermediate steps in the computation, we set $\varepsilon = 0$, so that no energy is dissipated, and the computation has equal probability of moving forwards or backwards, driven by random thermal noise. (In other words, the computation is a random walk without any "forward drift.") We now describe this in more detail.

Formally, the computation is described by a random walk on a graph $G = (V, E)$, together with a marked vertex v_{start}, a set of marked vertices $V_{\text{finish}} \subset V$, and an energy threshold $\varepsilon_{\text{th}} > 0$. The computation proceeds as follows:

1. The computer initializes its memory. (This dissipates $\varepsilon_{\text{th}} s_{\max}$ units of energy, where $s_{\max}$ is the size of the computer's memory.) Then the random walk begins at v_{start}.
2. At every step, the walk moves from its current position v to a neighboring vertex $w \in \Gamma(v)$ chosen uniformly at random.
3. When the walk reaches a vertex v that belongs to the set V_{finish}, we say that the computation has returned a result, which consists of the vertex v. (To read out this result, the computer dissipates $\varepsilon_{\text{th}} \log_2 |V|$ units of energy.)

[1] In addition to the need for logarithmic-size energy barriers, shared with the classical case, only a discrete subset of the continuous space of quantum gates can be implemented fault tolerantly in proposed systems. The remaining gates must be approximated, and the cost of doing this is believed to be logarithmic in the inverse of the approximation error. See e.g. [15]

Liu, Yi-Kai; Perlner, Ray.
"Thermodynamic Analysis of Classical and Quantum Search Algorithms."
Paper presented at Quantum Information Processing (QIP 2018), Delft, Netherlands. January 15, 2018 - January 19, 2018.

We assume that the graph G, and the energy threshold ε_{th}, have a few specific properties. Then computations of this type can be implemented by the same physical mechanisms as in the usual Brownian model. Specifically, we make the following assumptions:

1. We assume that the graph G has constant degree (say, at most 10), so that every step in the random walk can be implemented in constant time, by coupling the computer to a noisy environment.
2. We assume that the energy threshold ε_{th} is large enough so that auxiliary data stored in the computer's memory will remain stable for the duration of the computation. (In particular, this ensures that, once the random walk reaches a vertex in the set V_{finish}, it will stay there for the remainder of the computation.)

To illustrate this model of unpowered Brownian computation, we now consider some representative examples, and we analyze their energy consumption and running time.

3.1 Energy Consumption

First, consider an unpowered Brownian computation that uses a memory of size $s_{\max}$, and runs in $\tau_{\max}$ steps. We argue that the total energy consumption (call this E) grows almost linearly with $s_{\max}$, but only *logarithmically* with $\tau_{\max}$. This is in contrast to powered Brownian computation, where E grows linearly with the running time,[2] since energy is dissipated at every step.

This can be seen as follows: Note that we have $E = O(s_{\max}\varepsilon_{\text{th}})$. We need to choose ε_{th} large enough so that errors will not occur in the computer's memory while it is running. We can estimate the probability of having an error (call this p_{err}) as follows: suppose that at temperature T, errors occur independently on each bit of the memory, at each step of the computation, with probability $\exp(-\varepsilon_{\text{th}}/kT)$. Then p_{err} can be bounded by

$$p_{\text{err}} \leq \tau_{\max} s_{\max} \exp(-\varepsilon_{\text{th}}/kT). \tag{1}$$

For example, for any $c_0 > 0$, if we set the energy threshold ε_{th} to be

$$\varepsilon_{\text{th}} = kT(\ln(\tau_{\max} s_{\max}) + c_0), \tag{2}$$

then we have that $p_{\text{err}} \leq \exp(-c_0)$. Thus, in order to make p_{err} small, it is sufficient to set ε_{th} to grow logarithmically with $\tau_{\max}$ and $s_{\max}$. Furthermore, by increasing ε_{th}, we can force p_{err} to drop exponentially.

3.2 Running Time

We now consider three examples of unpowered Brownian computation (see Figure 2). We will compute the expected running times of these computations; that is, we let τ_{finish} be the first time when the random walk reaches a vertex in V_{finish}, and we compute the expected value $\mathbb{E}(\tau_{\text{finish}})$, averaging over the coin flips of the random walk.

In general, we expect that unpowered Brownian computation will be slower than powered Brownian computation. However, this slow-down varies depending on the structure of the computation. In particular, we will show that sequential computations have a quadratic

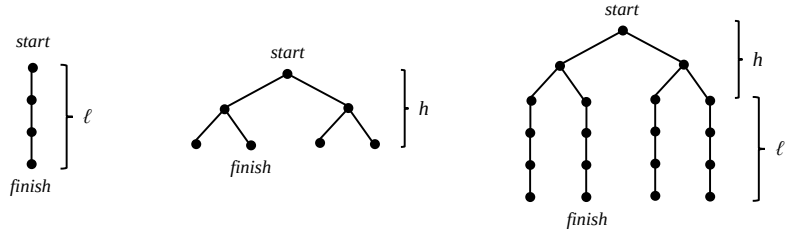


Fig. 2. Examples of unpowered Brownian computations: *(left)* a sequential computation of length ℓ, *(middle)* a branching computation of depth h, *(right)* a branching computation, followed by a sequential computation on each leaf.

slow-down, whereas branching computations incur a slow-down that is only a logarithmic factor.

First, we consider a sequential computation of length ℓ. Here, the graph G is a chain of $\ell + 1$ vertices. The expected running time of the computation is the expected time for a random walk to travel from one end of the chain to the other. A straightforward calculation [2] gives

$$\mathbb{E}(\tau_{\text{finish}}) = \ell^2. \tag{3}$$

This is consistent with the intuition that a random walk on a 1-D chain will take $\sim r^2$ steps to move a distance r. Thus, for sequential computation, unpowered Brownian computation is quadratically slower than powered Brownian computation.

Second, we consider a branching computation, which begins at the root of a binary tree of height h, and finishes when it hits a particular marked leaf of the tree. The expected running time can be upper-bounded as follows (Example 5.14 in [2]):

$$\mathbb{E}(\tau_{\text{finish}}) \leq 2(|V| - 1)h < 4 \cdot 2^h h. \tag{4}$$

For comparison, a deterministic search of the tree would take time $O(2^h)$. Hence, for branching computation, unpowered Brownian computation is only slightly slower than powered Brownian computation (they differ by a factor of $O(h)$, which is logarithmic in the total number of vertices).

Finally, we will consider a computation that consists of h branching steps, followed by ℓ sequential steps. Such a computation can be used to perform brute-force search: the computation first branches to select one of 2^h possible candidate solutions, then it does ℓ sequential operations to check whether that solution is correct. The expected running time can be upper-bounded as follows (Theorem 5.20 in [2]):

$$\mathbb{E}(\tau_{\text{finish}}) \leq 2(|V| - 1)(h + \ell) < 2 \cdot 2^h(\ell + 2)(h + \ell). \tag{5}$$

For comparison, a deterministic search of the tree would take time $O(2^h \ell)$. Hence, for brute-force search, this shows that unpowered Brownian computation is only slightly slower than

[2] Note, however, that powered Brownian computation may have a shorter running time than unpowered Brownian computation; we will discuss this in the next section.

powered Brownian computation, provided that $\ell \ll 2^h$ (i.e., one can quickly check whether a candidate solution is correct).

These observations suggest that branching computation (in the unpowered Brownian model) can be a useful tool for solving search problems. By using many parallel processors, one can make a natural time-space tradeoff. Moreover, in this situation, an unpowered Brownian algorithm can beat a powered Brownian algorithm, because its running time is not much worse (since most of the computation is branching rather than sequential), and its energy cost can be quadratically better (since the energy consumption scales linearly with space, but only logarithmically with time). As we will see in Section 5, this can lead to classical search algorithms whose energy cost is competitive with Grover's algorithm.

4 Collision Search

The best known classical algorithm for finding collisions in a random function is the parallel collision search algorithm of Van Oorschot and Wiener [21]. If the range of the function is of size N, then, given M parallel processes each with memory $O(1)$ the algorithm can find a collision in expected serial depth $O(\frac{\sqrt{N}}{M})$. The communication cost between threads is negligible compared to overall computational costs as long as M is smaller than $\sqrt{N}$ by at least a logarithmic factor.

An improvement over classical collision search (and claw finding) has been claimed by Brassard, Høyer, and Tapp (BHT) [9]. Their algorithm is a serial process consisting of $O(N^{\frac{1}{3}})$ operations and requires a memory of size $N^{\frac{1}{3}}$. This can be generalized to arbitrary memory size, $M < O(N^{\frac{1}{3}})$, giving a serial complexity of $O\left(\sqrt{\frac{N}{M}}\right)$. The BHT algorithm may be further generalized to a parallel algorithm involving p parallel processors and a shared memory M, where $p < M < O\left((Np)^{\frac{1}{3}}\right)$.[3] in this case, the serial complexity is $O\left(\sqrt{\frac{N}{Mp}}\right)$.

Bernstein [8] has observed that the BHT algorithm, even if parallelized, does not improve upon the Van Oorschot - Wiener algorithm, when measured in terms of memory and serial depth. Since the BHT algorithm also requires $O\left(\sqrt{\frac{N}{M}}\right)$ random access queries to a memory of size M, each requiring $O(M)$ gates, it also does not improve upon Van Oorschot Weiner algorithm when evaluated in terms of circuit size and depth (See Beals et al. [5] for a more thorough analysis.) However, BHT does represent an improvement over all classical algorithms in terms of query complexity. Furthermore, the Quantum RAM model of Giovanetti et al. [12] gives a theoretical argument that despite their large gate complexity, quantum memory access operations can be performed at logarithmic energy cost. A question therefore remains whether there exists a physically realistic model of computation where BHT is actually cheaper than the classical algorithms for the same problem. However, if there is such a model, it is not the Brownian model of computation, as we proceed to show:

[3] Note this also implies that $M < O\left(\sqrt{N}\right)$. The constraint arises from the requirement that the serial complexity, $O\left(\frac{M}{p}\right)$, of filling a table of size M with oracle values does not exceed the serial complexity $O\left(\sqrt{\frac{N}{Mp}}\right)$ of Grover search.

We first analyze the quantum algorithm, calculating the total energy required to perform a collision search, given a maximum time limit t and a maximum memory size M. (Here, we assume, following the quantum RAM model, that the energy complexity of the BHT is dominated by oracle queries rather than memory access): The per operation energy ε scales with the serial complexity divided by t, i.e.:

$$\varepsilon_{\text{quant}} = O\left(\frac{\sqrt{\frac{N}{Mp}}}{t}\right). \tag{6}$$

The total energy E is then the product of the parallelism, the serial complexity, and the per operation energy, i.e.:

$$E_{\text{quant}} = O\left(p \cdot \sqrt{\frac{N}{Mp}} \cdot \frac{\sqrt{\frac{N}{Mp}}}{t}\right) = O\left(\frac{N}{Mt}\right). \tag{7}$$

Now, we analyze the classical algorithm: The per operation energy again scales with the serial complexity, i.e.:

$$\varepsilon_{\text{cl}} = O\left(\frac{\sqrt{N}}{Mt}\right). \tag{8}$$

The total energy E is again the product of the parallelism (In this case $p = O(M)$), the serial complexity, and the per operation energy, i.e.:

$$E_{\text{cl}} = O\left(M \cdot \frac{\sqrt{N}}{M} \cdot \frac{\sqrt{N}}{Mt}\right) = O\left(\frac{N}{Mt}\right). \tag{9}$$

Thus, even under optimistic assumptions within the Brownian model of computation, we find that *quantum computers provide no advantage in terms of energy, memory, or time, for solving the collision search problem.*

5 Preimage Search

Grover's algorithm finds preimages in a function with domain size N in serial complexity $O(\sqrt{N})$. Grover's algorithm can be generalized to take advantage of M parallel processes each with memory $O(1)$, in which case the serial complexity is reduced to $O\left(\sqrt{\frac{N}{M}}\right)$. This serial complexity was shown to be optimal by Zalka [22]. If we implement Grover's algorithm in a Brownian fashion, we find that

$$\varepsilon_{\text{quant}} = O\left(\frac{\sqrt{\frac{N}{M}}}{t}\right), \tag{10}$$

and,

$$E_{\text{quant}} = O\left(M \cdot \sqrt{\frac{N}{M}} \cdot \frac{\sqrt{\frac{N}{M}}}{t}\right) = O\left(\frac{N}{t}\right). \tag{11}$$

A naïve Brownian implementation for classical search would divide the key space among M parallel processes, each of which would deterministically step through $\frac{N}{M}$ keys searching for the correct one. Such a deterministic classical algorithm would require,

$$\varepsilon_{\text{det}} = O\left(\frac{N}{Mt}\right), \tag{12}$$

and,

$$E_{\text{det}} = O\left(M \cdot \frac{N}{M} \cdot \frac{N}{Mt}\right) = O\left(\frac{N^2}{Mt}\right). \tag{13}$$

This already allows us to compete with Grover's algorithm if we allow ourselves a memory of size $O(N)$. However, we can exploit the structure, or rather the lack of structure, of the search problem to improve upon this figure. In particular, rather than deterministically stepping through the keys, dissipating a driving energy each time, we can simply allow Brownian motion to drive the system on a random walk through the keyspace. (That is to say, we can use the unpowered Brownian computation model described in detail in section 3.) We will still require a latching energy to end the computation, once the correct key has been found, and an initialization energy to create the necessary energy barriers to prevent unwanted transitions from occuring.

If the search is implemented by M parallel processes, each of size $O(1)$, then each process must reach $\frac{N}{M}$ keys. This requires the processes to operate at a temperature:

$$kT = O\left(\frac{N}{Mt}\right). \tag{14}$$

The initialization energy should be of order MkT i.e.:

$$E_{\text{init}} = O\left(M \cdot \frac{N}{Mt}\right) = O\left(\frac{N}{t}\right).$$

This is identical to the energy required by a Brownian implementation of Grover's algorithm. All that remains is to show that the latching energy is negligible. Indeed, we find that the energy required to suppress backwards transitions from the final state for a time of order t is $O\left(kTln(tkT)\right) = O\left(\frac{N}{Mt} ln\left(\frac{N}{M}\right)\right)$. This is negligible as long as M is at least logarithmic in N.

Thus, as with collision search, *the quantum and classical algorithms for preimage search appear to offer the same tradeoffs between time, energy and space*:

$$E_{\text{cl}} = O\left(\frac{N}{t}\right) \text{ and } E_{\text{quant}} = O\left(\frac{N}{t}\right). \tag{15}$$

6 Preimage Search at Constant Power and Temperature

In contrast to the collision search case, matching the time/ memory/ energy tradeoffs of Grover's algorithm with a classical search requires a somewhat unrealistic assumption. We assume that if a computational process can be accomplished at a temperature T in a time t, then an isomorphic computation can also be accomplished at a temperature αT in a time $\frac{T}{\alpha}$. This would be true if physics were scale invariant, but the physics of the real world is

almost certainly not scale invariant. A more realistic model would therefore restrict the range of temperatures where a given computation is considered feasible. We will therefore repeat the analysis of the previous section assuming a fixed temperature T. For added realism, in addition to memory M, and time t, we will express the resources required for search in terms of power, $P = \frac{E}{t}$, rather than energy, since a fixed power budget is a more common limitation than a fixed energy budget.

From Equation (15) we find:

$$N = O\left(Pt^2\right).$$

Plugging this into Equation (14) gives us:

$$M = O\left(\frac{Pt}{T}\right).$$

We can now calculate time and memory requirements in terms of T, P, and N:

$$t_{\text{cl}} = O\left(\sqrt{\frac{N}{P}}\right); \tag{16}$$

$$M_{\text{cl}} = O\left(\frac{\sqrt{NP}}{T}\right). \tag{17}$$

A similar analysis may be done in the quantum case. Here we use Equation (10) as a lower bound for T. If the per gate energy ϵ exceeds kT, we enter the thermodynamic regime of irreversible computing, as opposed to Brownian computing, at which point the time per gate not only fails to further decrease with increasing ε, but must in fact increase to prevent the waste heat from heating the computing system to a temperature higher than T. Combining this bound with Equation (11) then yields the following time and memory requirements for Grover search at fixed power and temperature:

$$t_{\text{quant}} = O\left(\sqrt{\frac{N}{P}}\right); \tag{18}$$

$$M_{\text{quant}} = O\left(\frac{P}{T^2}\right). \tag{19}$$

Thus, fixing power and temperature, *we find that our classical search strategy recovers the square-root time scaling of Grover's algorithm.* However, unlike Grover's algorithm, whose space requirement is determined only by the power budget and maximum operating temperature, *the classical algorithm also requires memory* that scales, like the time, with the square root of the size of the search space.

7 The Cost of Oracle Queries

The asymptotic complexities given in previous sections ignore the computational complexity of individual oracle queries. Most of the results of previous sections remain substantively similar if these factors are included. We will model each oracle query as a circuit with depth d_0, width m_0, and total gates g_0.

In the case of powered Brownian computation, the effect of these factors is fairly straightforward. The memory imposed limit on parallelism (and number of table entries in the case of BHT) is now $p_{\max} = O\left(\frac{M}{m_0}\right)$. Likewise, if t_0 is the time per query required to complete the computation in time t, we will now require an energy per gate of $\varepsilon = O\left(\frac{d_0}{t_0}\right)$. We must also ensure that all the bits or qubits in the circuit advance through it roughly synchronously. This can be done, for example, by associating a clock state of size $O(\log(d_0))$ to each bit or qubit in the oracle circuit, and imposing a restoring potential proportional to the squared difference of the clock states of neighboring qubits. This will tend to couple the clock states of nearby qubits, but will not dissipate any net energy. As with other energy barriers ensuring correct computation, this potential need only extend logarithmically far from the equilibrium point, relative to the total size of the computation. We will generally ignore the logarithmic memory cost of the clock state and the logarithmic computational costs associated with creating interactions between the clock state, but in more detailed models, they may be subsumed into m_0 and g_0 respectively. Finally, we must take into account the number of gates required to perform an oracle query, g_0.

7.1 Collision search

Making these substitutions into equations (7) and (9) gives the following energy costs for quantum and classical collision search:

$$E_{\text{quant}} = O\left(p \cdot g_0 \sqrt{\frac{m_0 N}{Mp}} \cdot d_0 \frac{\sqrt{\frac{m_0 N}{Mp}}}{t}\right) = O\left(\frac{g_0 m_0 d_0 N}{Mt}\right); \tag{20}$$

$$E_{\text{cl}} = O\left(\frac{M}{m_0} \cdot g_0 \frac{m_0\sqrt{N}}{M} \cdot \frac{m_0 d_0 \sqrt{N}}{Mt}\right) = O\left(\frac{g_0 m_0 d_0 N}{Mt}\right). \tag{21}$$

Again, we find the classical and quantum complexities to be identical, up to constant factors. In both cases, the useful memory size is bounded above by $O\left(m_0\sqrt{N}\right)$.

7.2 Preimage search

Similarly, we may make the same substitutions in equations (10) and (11) to include these factors in the per-gate and total energy cost of Grover search:

$$\varepsilon_{\text{quant}} = O\left(\frac{d_0\sqrt{\frac{m_0 N}{M}}}{t}\right); \tag{22}$$

$$E_{\text{quant}} = O\left(\frac{M}{m_0} \cdot g_0 \sqrt{\frac{m_0 N}{M}} \cdot \frac{d_0\sqrt{\frac{m_0 N}{M}}}{t}\right) = O\left(\frac{g_0 d_0 N}{t}\right). \tag{23}$$

In the case of unpowered Brownian computation, we must calculate the temperature T required for random Brownian motion to power the traversal of an oracle circuit of depth d_0 and containing g_0 gates in time t_0. To do this, we create a random variable, x indicating

the total number of gates that have been completed at a time t. We expect that x will obey the usual formula for Brownian motion, $\langle x^2 \rangle = Dt$, for some D, which will depend on T, g_0, and d_0. We will then require $Dt_0 = O\left(g_0^2\right)$. It remains to determine the scaling of D. Note that at any given time, on average $O\left(\frac{g_0}{d_0}\right)$ gates will be exposed to activation by thermal noise. (The remaining gates will be disallowed by the clock states associated with their input/output bits.) Each of these gates is expected to contribute $O(Tdt)$ to $d\langle x^2 \rangle$. The coupling potential between neighboring clock states will also drive the activation of individual gates, but it should have no net effect on x, since every gate driven forward by the coupling potential will be counterbalanced by another gate driven backwards. Thus we find that $D = O\left(\frac{Tg_0}{d_0}\right)$ and therefore $T = O\left(\frac{g_0 d_0}{t_0}\right)$.

We may now apply this analysis to equations (14) and (15). Since, in order to complete a preimage search of size N in time, t with memory M, we need $t_0 = \frac{m_0 N}{Mt}$, we find that:

$$T_{\text{cl}} = O\left(\frac{g_0 m_0 d_0 N}{Mt}\right), \tag{24}$$

and,

$$E_{\text{cl}} = O\left(M \cdot \frac{g_0 m_0 d_0 N}{Mt}\right) = O\left(\frac{g_0 m_0 d_0 N}{t}\right) = O\left(m_0 E_{\text{quant}}\right). \tag{25}$$

Note that, when we include cost factors associated with the size and computational complexity of oracle queries, *the mostly unpowered randomized preimage search is more energy intensive than Grover's algorithm by a factor of* $O(m_0)$. Nonetheless, this factor is generally expected to be logarithmic in N, and may easily be overwhelmed by the various costs associated with implementing fault tolerant quantum computation.

Finally, we may also consider the fixed power and temperature scenario discussed in Section 6. In this case, equations (16), (17), (18) and (19) become:

$$t_{\text{cl}} = O\left(\sqrt{\frac{g_0 m_0 d_0 N}{P}}\right), \tag{26}$$

$$M_{\text{cl}} = O\left(\frac{\sqrt{g_0 m_0 d_0 N P}}{T}\right), \tag{27}$$

and,

$$t_{\text{quant}} = O\left(\sqrt{\frac{g_0 d_0 N}{P}}\right), \tag{28}$$

$$M_{\text{quant}} = O\left(\frac{m_0 d_0 P}{g_0 T^2}\right). \tag{29}$$

For completeness, we will also consider the case of powered preimage search. Adapting equations (12) and (13) gives us:

$$\varepsilon_{\text{det}} = O\left(\frac{m_0 d_0 N}{Mt}\right), \tag{30}$$

and,

$$E_{\text{det}} = O\left(\frac{M}{m_0} \cdot g_0 \frac{m_0 N}{M} \cdot \frac{m_0 d_0 N}{Mt}\right) = O\left(\frac{g_0 m_0 d_0 N^2}{Mt}\right). \tag{31}$$

Note that, by comparing equations (31) and (21), we can see that the cost of powered preimage search with a domain of size N is identical to the cost of collision search on a range of size N^2.

8 Near-Future Computing Technologies and the Grover Speedup

We are now in a position to estimate the practical relevance of Grover's algorithm and its classical counterpart, unpowered Brownian preimage search. The particular questions we intend to answer are the following: Is it reasonable, given Grover's algorithm (and its classical counterpart), to treat finding a preimage within a domain of size N as an easier problem than finding a collision within a range of size N^2? How much easier? How do the answers to these questions depend upon the present and future state of technology – in particular how do they depend upon the various ways that present and future technology may fall significantly short of thermodynamically ideal behavior?

The technology-dependent costs we will consider are:

1. The cost of memory. For example, if we assume that power costs 10 cents per kWh and memory costs $100 per TB, then the cost of a bit of memory is on the order of $memcost = 10^{15} kT$, where the temperature, T is taken to be on the order of 300K.
2. The increase in physical quantum circuit depth and gate count due to quantum error correction. Based on [14] we roughly estimate that near-future quantum error correction may increase memory requirements by a factor of $\frac{m_{\text{quant}}}{m_0} = 10^5$, effective circuit depth by a factor of $\frac{d_{\text{quant}}}{d_0} = 10^3$, and effective gate count by a factor of $\frac{g_{\text{quant}}}{g_0} = 10^8$.
3. The need for various quantum computing technologies to operate at extremely low temperatures. In addition to placing a lower limit on gate times, such low temperatures impose an energy cost due to the fact that any energy dissipated as heat at the lower temperature must eventually be removed and expelled to a heat bath, which typically must be at a much higher temperature, e.g. 300K. Moving heat from a sytem at a low temperature T_{quant} to a system at a higher temperature T increases energy consumption by a factor of $\frac{T}{T_{\text{quant}}}$.

In the remainder of this section, we will study the relative cost of *Grover's algorithm*, *unpowered classical search*, and *powered classical search*, when the above-mentioned technology-dependent cost factors take on our estimated current and near future values, and as they approach unity. For concreteness, we will also need to set values for the memory, m_0, circuit depth, d_0, and total gate count g_0 involved in oracle queries. Based on [13] we will estimate typical ranges for these values as $m_0 = 10^3$, $d_0 = 10^5$ and $g_0 = 3 \times 10^6$.

For each of the three algorithms considered, we will express the efficiency of the algorithm based on the maximum value of N such that the search problem can be solved given an energy budget E and a time budget t. We will denote this by N_{quant}, N_{det} and N_{cl}, for Grover's algorithm, powered classical search, and unpowered classical search, respectively.

We will take the time budget to be 1 year. As we will calculate relative efficiencies (i.e. $\frac{N_{\text{quant}}}{N_{\text{det}}}$ and $\frac{N_{\text{cl}}}{N_{\text{det}}}$), and in all cases N will scale with E, we will not need to set a concrete

value for E. Rather than providing an explicit memory budget, we will assume that the memory budget, M is set so that $memcost \cdot M \leq E$.[4] By taking the log base 2 of these ratios, we can calculate Grover speedups in terms of the "bits of security" metric typically used to evaluate cryptographic hardness.

8.1 Powered classical search

We will first evaluate powered classical search (either for a collision within a range of size N_{det}^2 or for a preimage within a domain of size N_{det}.) We first solve for M by setting $memcost \cdot M \leq E$, where the relation between E and N_{det} is given by Equation (31). We find that

$$M = O\left(N_{\text{det}} \cdot \sqrt{\frac{\hbar g_0 m_0 d_0}{kTt}} \cdot \left(\frac{memcost}{kT}\right)^{-\frac{1}{2}}\right). \tag{32}$$

Plugging M back into Equation (31), and solving for N_{det}, we get:

$$N_{\text{det}} = O\left(E \cdot \sqrt{\frac{t}{\hbar g_0 m_0 d_0 kT}} \cdot \left(\frac{memcost}{kT}\right)^{-\frac{1}{2}}\right) \tag{33}$$

Note that the above computations assume a Brownian model of computation. To check that this is reasonable we must verify, that $\varepsilon \leq kT$ for the range of values we're interested in. We will check this using equation (30) with M given by equation (32). Here we find:

$$\frac{\varepsilon}{kT} = O\left(\frac{\hbar m_0 d_0 N}{kTMt}\right) = O\left(\sqrt{\frac{\hbar m_0 d_0}{g_0 kTt}} \cdot \left(\frac{memcost}{kT}\right)^{\frac{1}{2}}\right).$$

If we use $\frac{memcost}{kT} = 10^{15}$ along with our other estimated values: $m_0 = 10^3$; $d_0 = 10^5$; $g_0 = 3 \times 10^6$; $T = 300K$; $t = 1$ year, we get

$$\frac{\varepsilon}{kT} \approx 5 \times 10^{-2}.$$

Letting $\frac{memcost}{kT}$ approach unity yields

$$\frac{\varepsilon}{kT} \approx 2 \times 10^{-9}.$$

As both values are less than 1, this indicates that the Brownian model of computation should yield a plausible, if slightly optimistic, approximation of N_{det} accross the range of values of $memcost$ of interest to us. While these estimates of $\frac{\varepsilon}{kT}$ indicate that there may be advantages to using reversible computing even with current memory costs, it is not surprising that these advantages have not yet been realized, even for applications like bitcoin mining, which seems like a good fit but nonetheless continues to use standard irreversible computing technology at the time of writing. 5×10^{-2} does not differ from 1 by many orders of magnitude, and it could easily be overwhelmed by engineering costs

[4] Strictly speaking a larger memory budget is possible, since memory costs can be amortized accross multiple computations using the same hardware in series, but we judge that 1 year is a long enough time window to make this cost savings of only minor consideration.

not included in our model (fixed overhead associated with using reversible logic, extra gates required to synchronize different parts of a non-serial reversible circuit, different gate technology etc.) Nonetheless, it seems likely that if memory costs continue to fall, a broadly Brownian approach to computing will become cost effective. 2×10^{-9} differs from 1 by a significantly larger amount, and is less likely to be overcome by these sorts of overheads.

8.2 Grover's algorithm

We will now evaluate Grover's algorithm. Here all we need to do is modify Equation (11) to account for technology-dependent cost factors and solve for N_{quant}.

$$N_{\text{quant}} = O\left(E \cdot \frac{t}{\hbar g_0 d_0} \cdot \frac{T_{\text{quant}}}{T} \cdot \frac{g_0}{g_{\text{quant}}} \cdot \frac{d_0}{d_{\text{quant}}}\right). \tag{34}$$

We compare Grover's algorithm with powered classical search, and we compute a speedup factor:

$$\frac{N_{\text{quant}}}{N_{det}} = O\left(\sqrt{\frac{m_0 kTt}{\hbar g_0 d_0}} \cdot \left(\frac{memcost}{kT}\right)^{\frac{1}{2}} \cdot \frac{T_{\text{quant}}}{T} \cdot \frac{g_0}{g_{\text{quant}}} \cdot \frac{d_0}{d_{\text{quant}}}\right). \tag{35}$$

Using our estimated near future values for the technology dependent cost factors: $\frac{memcost}{kT} = 10^{15}$; $\frac{m_{\text{quant}}}{m_0} = 10^5$; $\frac{d_{\text{quant}}}{d_0} = 10^3$; and $\frac{g_{\text{quant}}}{g_0} = 10^8$ along with our other estimated values: $m_0 = 10^3$; $d_0 = 10^5$; $g_0 = 3 \times 10^6$; $T = 300K$; $t = 1$ year, we get

$$\log_2\left(\frac{N_{\text{quant}}}{N_{\text{det}}}\right) \approx 4,$$

indicating that Grover's algorithm will provide little, if any, advantage over classical search in the near future. Setting all technology-dependent cost factors to unity, yields a somewhat larger, but still modest, advantage of

$$\log_2\left(\frac{N_{\text{quant}}}{N_{\text{det}}}\right) \approx 21.$$

In order to envision a larger advantage for Grover's algorithm, we must instead envision a scenario where classical memory remains as expensive as it is today, but all technology-dependent cost factors associated with quantum computers are eliminated. In this case, we get

$$\log_2\left(\frac{N_{\text{quant}}}{N_{\text{det}}}\right) \approx 46.$$

In the above analysis, we have ignored memory initialization costs for Grover's algorithm. To demonstrate that memory initialization costs do not necessarily overwhelm computation costs, we evaluate the memory requirements for Grover's algorithm at $T_{quant} = 10mK$, assuming that physical qubits can be manufactured as cheaply as classical bits today (i.e., $memcost = 10^{15}kT$, for $T = 300K$). For fault-tolerance related cost factors, we make no special assumptions other than that error correction does not appreciably change circuit density (i.e., $\frac{g_{\text{quant}}}{m_{\text{quant}} d_{\text{quant}}} \approx \frac{g_0}{m_0 d_0}$). Suitably modifying equation (22) gives us:

$$\frac{memcost \cdot M_{\text{quant}}}{E} = O\left(\frac{\hbar m_0 d_0}{g_0 kTt} \cdot \frac{m_{\text{quant}}}{m_0} \cdot \frac{d_{\text{quant}}}{d_0} \cdot \frac{g_0}{g_{\text{quant}}} \cdot \frac{T}{T_{\text{quant}}} \cdot \frac{memcost}{kT}\right) \approx 1.$$

We also note that, while we have considered the possibility that quantum fault tolerance might be implemented in a fashion that avoids irreversible operations like measurement, if it cannot, this does not affect the above analysis. The energy cost of Grover's algorithm, aside from initialization, does not depend on memory within the Brownian computation regime. Thus, in order to minimize the memory requirement it is optimal to set $\varepsilon \approx kT$, at which point Brownian computation exhibits essentially the same energy costs as irreversible computation.

8.3 Unpowered classical search

Finally, we evaluate unpowered classical preimage search. Modifying Equation (15) gives us:

$$N_{\mathrm{cl}} = O\left(E \cdot \frac{t}{\hbar g_0 m_0 d_0} \cdot \left(\frac{memcost}{kT}\right)^{-1}\right). \tag{36}$$

Again, we can compare this with powered classical search, by computing a speedup factor:

$$\frac{N_{\mathrm{cl}}}{N_{\mathrm{det}}} = O\left(\sqrt{\frac{kTt}{\hbar m_0 g_0 d_0}} \cdot \left(\frac{memcost}{kT}\right)^{-\frac{1}{2}}\right) \tag{37}$$

As before, we may use $\frac{memcost}{kT} = 10^{15}$ along with our other estimated values: $m_0 = 10^3$; $d_0 = 10^5$; $g_0 = 3 \times 10^6$; $T = 300K$; $t = 1$ year. With these values, we get

$$\log_2\left(\frac{N_{\mathrm{cl}}}{N_{\mathrm{det}}}\right) \approx -14,$$

that is to say we find that unpowered classical search has a modest disadvantage over powered classical search assuming near future memory costs. However, this can be turned into a modest advantage of

$$\log_2\left(\frac{N_{\mathrm{cl}}}{N_{\mathrm{det}}}\right) \approx 11,$$

if $\frac{memcost}{kT}$ goes to unity. A somewhat larger advantage may also be possible if we consider computations that last significantly longer than a year or scenarios where the ratio $\frac{memcost}{kT}$ reaches its optimum value at a temperature higher than $300K$.

In considering extremely optimistic scenarios for unpowered classical preimage search, it is worth noting that memory costs may be determined by the scarcity of matter rather than energy. We may estimate the total energy budget of the Earth, based on the total solar irradiance received by the Earth's atmosphere, which has been estimated at 174 PW [18]. This translates to approximately $2 \times 10^{45} kT$ per year at 300K. In contrast, we may give an estimated matter budget based on the total number of atoms in the earth, which has been estimated around 10^{50} [1]. As 10^{50} is a few orders of magnitude larger than 2×10^{45}, it remains plausible that energy could be the limiting factor in determining memory requirements, even given an extremely energy efficient manufacturing process, but the numbers are quite similar (especially if we are limited to only use atoms in the Earth's crust, for example.)

Liu, Yi-Kai; Perlner, Ray.
"Thermodynamic Analysis of Classical and Quantum Search Algorithms."
Paper presented at Quantum Information Processing (QIP 2018), Delft, Netherlands. January 15, 2018 - January 19, 2018.

9 Conclusion

The development of quantum computing has created a great deal of excitement, particularly due to the discovery of quantum algorithms, such as Shor's algorithm, that perform exponentially better than the best known classical algorithm. Nonetheless, a large body of research concerns quantum algorithms, such as Grover's algorithm, that have only demonstrated a polynomial improvement over the best known classical algorithm with respect to metrics, such as query complexity, that bear an uncertain relationship to the real physical costs of computation.

We argue that, in order to assess the impact of such algorithms, we need a more explicitly physical model of computation. We also feel that, in order to fairly compare classical algorithms to their future quantum counterparts, we need to take into account, not just the current state of classical computing technology, but possible future developments, such as low-power reversible computing. For example, it certainly does not seem reasonable to consider extremely low cost and low power quantum memories, without assuming similar advances in classical computing technology. To this end, we have developed the Brownian computation model of Bennett, and given extensive analysis of the costs of classical and quantum algorithms for collision and preimage search.

In the case of collision search, our analysis suggests that despite their lower query complexity, quantum collision-finding algorithms do not offer a substantial, physically plausible advantage over their classical counterparts.

The case of preimage search is more delicate. In our analysis, we have developed a novel variant of Brownian computation, namely unpowered Brownian computation. It is interesting to note that, using this model of computation, we can perform a randomized classical search with the same asymptotic thermodynamic costs as Grover's algorithm. This is certainly of theoretical interest. But the practical significance of this result is somewhat less clear than in the case of collision search, since there are plausible reasons for thinking Grover's algorithm may indeed turn out to be more efficient than unpowered classical search in practice, although it should be noted that there are plausible scenarios where the reverse might hold. (As a further point of contrast, in the present state of technology, powered classical search appears to be more efficient than both approaches in finding preimages.)

We analyze in detail the technological costs which may affect the true advantage of Grover's algorithm over powered and unpowered classical preimage search. Aside from the various unique challenges involved in building fault tolerant quantum computing hardware, a key metric which appears to be relevant here is the cost of memory (or perhaps more accurately, the cost of hardware in general). As the cost of memory falls, thermodynamically reversible computing becomes more attractive to relative to current (non-reversible) computing technology. Our estimates indicate that the cost of semiconductor hardware is fairly close to the point at which reversible computing would begin to offer a real advantage. If the cost of hardware continues to fall, we would expect to see reversible computing developed for very computationally expensive tasks such as proof of work, or for very low power devices.

As the cost of memory falls further, Grover's algorithm looks less attractive, because the efficiency of classical powered search improves relative to the efficiency of Grover's algorithm, and the efficiency of unpowered classical search improves relative to the efficiency of powered classical search. Nonetheless, even without the assumption of very low hardware costs, we find that the potential advantage provided by Grover's algorithm is significantly smaller than is often assumed. Even in scenarios that are simultaneously ex-

Liu, Yi-Kai; Perlner, Ray.
"Thermodynamic Analysis of Classical and Quantum Search Algorithms."
Paper presented at Quantum Information Processing (QIP 2018), Delft, Netherlands. January 15, 2018 - January 19, 2018.

tremely optimistic with respect to quantum computing and extremely pessimistic with regard to classical computing, Grover's algorithm will only extend the reach of classical search by a factor of 2^{46}.

This analysis can be used to give guidance for post-quantum cryptography, in particular, for choosing key lengths for block ciphers. This analysis suggests that doubling the key sizes is likely unnecessary to provide protection against quantum computers, and that a smaller increase, from 128 to 192 bits for example, is likely sufficient.

Note: Contributions to this work by NIST, an agency of the US government, are not subject to US copyright. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s), and do not necessarily reflect the views of NIST.

References

1. Physics questions people ask Fermilab. `http://www.fnal.gov/pub/science/inquiring/questions/atoms.html`. Accessed: 2017-09-28.
2. David Aldous and James Allen Fill. Reversible Markov chains and random walks on graphs, 2002. Unfinished monograph, recompiled 2014, available at `http://www.stat.berkeley.edu/~aldous/RWG/book.html`.
3. Muhammad Arsalan and Maitham Shams. Asynchronous adiabatic logic. In *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, pages 3720–3723. IEEE, 2007.
4. Gustavo Banegas and Daniel J. Bernstein. Low-communication parallel quantum multi-target preimage search. Cryptology ePrint Archive, Report 2017/789, 2017. `http://eprint.iacr.org/2017/789`.
5. Robert Beals, Stephen Brierley, Oliver Gray, Aram W Harrow, Samuel Kutin, Noah Linden, Dan Shepherd, and Mark Stather. Efficient distributed quantum computing. In *Proc. R. Soc. A*, volume 469, page 20120686. The Royal Society, 2013.
6. C. H. Bennett. Logical reversibility of computation. *IBM J. Res. Dev.*, 17(6):525–532, November 1973.
7. Charles H Bennett. The thermodynamics of computation — a review. *International Journal of Theoretical Physics*, 21(12):905–940, 1982.
8. Daniel J Bernstein. Cost analysis of hash collisions: Will quantum computers make SHARCS obsolete. *SHARCS09 Special-purpose Hardware for Attacking Cryptographic Systems*, page 105.
9. Gilles Brassard, Peter Høyer, and Alain Tapp. *Quantum cryptanalysis of hash and claw-free functions*, pages 163–169. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.
10. Scott Fluhrer. Reassessing grover's algorithm. Cryptology ePrint Archive, Report 2017/811, 2017. `http://eprint.iacr.org/2017/811`.
11. Edward Fredkin and Tommaso Toffoli. Conservative logic. *International Journal of Theoretical Physics*, 21(3):219–253, Apr 1982.
12. Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Quantum random access memory. *Physical Review Letters*, 100(16):160501, 2008.
13. Markus Grassl, Brandon Langenberg, Martin Roetteler, and Rainer Steinwandt. *Applying Grover's Algorithm to AES: Quantum Resource Estimates*, pages 29–43. Springer International Publishing, Cham, 2016.
14. N Cody Jones, Rodney Van Meter, Austin G Fowler, Peter L McMahon, Jungsang Kim, Thaddeus D Ladd, and Yoshihisa Yamamoto. Layered architecture for quantum computing. *Physical Review X*, 2(3):031007, 2012.

15. Vadym Kliuchnikov, Alex Bocharov, and Krysta M. Svore. Asymptotically optimal topological quantum compiling. *Phys. Rev. Lett.*, 112:140504, Apr 2014.
16. Norman Margolus and Lev B. Levitin. The maximum speed of dynamical evolution. *Physica D: Nonlinear Phenomena*, 120(1):188 – 195, 1998. Proceedings of the Fourth Workshop on Physics and Consumption.
17. Michael Nielsen and Isaac Chuang. *Quantum Computation and Quantum Information*. Cambridge Univ. Press, 2001.
18. Vaclav Smil. *General Energetics: Energy in the Biosphere and Civilization*. Wiley, 1991.
19. Seiichiro Tani. *An Improved Claw Finding Algorithm Using Quantum Walk*, pages 536–547. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.
20. Simon Trebst, Matthias Troyer, Zhenghan Wang, and Andreas WW Ludwig. A short introduction to Fibonacci anyon models. *Progress of Theoretical Physics Supplement*, 176:384–407, 2008.
21. Paul C. van Oorschot and Michael J. Wiener. Parallel collision search with cryptanalytic applications. *Journal of Cryptology*, 12(1):1–28, Jan 1999.
22. Christof Zalka. Grover's quantum searching algorithm is optimal. *Phys. Rev. A*, 60:2746–2751, Oct 1999.

Assessing Coverage and Throughput for D2D Communication

Jian Wang and Richard Rouil
Wireless Networks Division, Communications Technology Laboratory
National Institute of Standards and Technology
Emails:{jian.wang, richard.rouil}@nist.gov

***Abstract*—In this paper, we access the performance of a Device-to-Device (D2D) communication link. Particularly, we design a framework to evaluate performance with respect to coverage probability and average throughput. Our modeling framework considers a variety of D2D deployment scenarios (i.e., Outdoor-to-Outdoor, Outdoor-to-Indoor, and Indoor-to-Indoor), channel effects (i.e., path loss, shadowing, and small-scale fading), and system parameters (i.e., resource block size, fixed Modulation and Coding Scheme (MCS) versus adaptive MCS, and transmit power). For the defined scenarios, we derive mathematical expressions for the coverage probability and average throughput as a function of the distance between two D2D user equipments. Based on the designed framework, we conduct an extensive performance evaluation of the D2D communication link under various D2D deployment scenarios with varying channel effects. We also evaluate the impact of system parameters, including Physical Resource Block (PRB) size, fixed/adaptive MCS, and transmit power, on the performance of D2D communication links. Our results demonstrate the expected performance that can be achieved in practical D2D scenarios.**

***Keywords*—D2D Communication, Public Safety Applications, Performance Measurements.**

I. Introduction

In order to support mission critical applications for public safety, reliable communication is critical [1]. For example, when a large incident occurs, the network infrastructure could be either damaged or overloaded. In either case, continuous communication between public safety personnel is required to successfully carry out their missions. To this end, Device-to-Device (D2D) communication provides a viable solution, as it is capable of providing a direct communication between the sender and the receiver without relying on the network infrastructure. In addition, when a first responder is out of cell coverage, an in-cell user equipment (UE) can be used as a relay node to extend the cell coverage using the D2D link. In doing so, the connection of the first responder to the cellular network can be established.

The LTE-based D2D communication was introduced by 3^{rd} Generation Partnership Project (3GPP) as the Proximity Service (ProSe) starting in Release 12 [2], providing both D2D communication and D2D discovery services. To support the LTE-based D2D communication, several new physical channels were introduced, including the Physical Sidelink Control Channel (PSCCH) and the Physical Sidelink Shared Channel (PSSCH). PSSCH is the D2D data channel, while PSCCH is used to transmit control information to the receiver, so that the message transmitted on PSSCH can be decoded. Notice that successfully decoding PSSCH depends on whether the control information transmitted over PSCCH can be received correctly. We define PSCCH coverage as the event that the receiver can correctly decode the control messages conveyed on PSCCH. For PSSCH, the coverage probability is defined as the probability that the receiver can correctly decode both the message transmitted on PSCCH and on PSSCH.

There have been a number of research efforts devoted to D2D communication [3], [4]. It is generally agreed that D2D requires UEs to be spatially located close to one another. Thus, the following fundamental issues remain unresolved: *How far apart can D2D UEs be located and still maintain a reliable communication given channel conditions? What is the average throughput that can be achieved?* To address these issues, in this paper we conduct a comprehensive and practical study by considering different channel effects for different deployment scenarios.

To summarize, our contributions are three-fold, outlined as follows: (i) We design a generic framework to assess the performance of a D2D communication link with respect to its coverage probability and average throughput. Our framework considers a variety of D2D deployment scenarios (i.e., Outdoor-to-Outdoor (O2O), Outdoor-to-Indoor (O2I), and Indoor-to-Indoor (I2I) and different channel effects (i.e., path loss, log-normal shadowing, and Nakagami-m small-scale fading), as well as various system parameters (i.e., the Physical Resource Block (PRB) size, fixed/adaptive Modulation and Coding Scheme (MCS), and the transmit power). (ii) Based upon the defined scenarios, we derive mathematical expressions for the coverage probability and average throughput of a D2D communication link. (iii) We conduct extensive performance evaluation to show the performance of a D2D communication link using the framework defined. We also measure the impact of system parameters (PRB size, fixed/adaptive MCS, and transmit power) on the performance of D2D communication links. Monte Carlo simulation results are also presented to support the analytic computation outcomes.

The remainder of this paper is organized as follows: In Section II, we describe channel models used in our analysis. In Section III, we introduce our approach in detail. In Section IV, we show the performance evaluation results. Finally, we conclude the paper in Section V.

Rouil, Richard; Wang, Jian.
"Assessing Coverage and Throughput for D2D Communication."
Paper presented at 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, United States. May 20, 2018 - May 24, 2018.

TABLE I: Channel Model [5], [6]

Scenario	Path loss (dB)	Log-normal Shadowing (dB)
O2O LOS [5]	$PL_{LOS\_O2O} = 40\log_{10}(d) + 7.56 - 17.3\log_{10}(h^{'}_{BS}) - 17.3\log_{10}(h^{'}_{MS}) + 2.7\log_{10}(f_c) + 20\log_{10}(f_C/f_{REF})$	7 dB
O2O NLOS [5]	$PL_{NLOS\_O2O} = (44.9 - 6.55\log_{10}(h_{BS}))\log_{10}(d) + 5.83\log_{10}(h_{BS}) + 18.38 + 23\log_{10}(f_C)$	7 dB
O2I LOS [5]	$PL_{LOS\_O2I} = PL_{LOS\_O2O} + 20 + 0.5d_{in}$	7 dB
O2I NLOS [5]	$PL_{NLOS\_O2I} = PL_{NLOS\_O2O} + 20 + 0.5d_{in} - 0.8h_{MS}$	7 dB
I2I (different building) [5]	$PL_{I2I_{DB}} = \max(131.1 + 42.8\log_{10}(\frac{d}{1000}), 147.4 + 43.3\log_{10}(\frac{d}{1000})) + 40 + 20\log_{10}(\frac{f_c}{f_{REF}})$	10 dB
I2I (same building) [6]	$PL_{I2I_{SB}} = PL_0 + \begin{cases} 20\log_{10} d & 1 < d < 10\,m \\ 20 + 30\log_{10}\frac{d}{10} & 10 < d < 20\,m \\ 29 + 60\log_{10}\frac{d}{20} & 20 < d < 40\,m \\ 47 + 120\log_{10}\frac{d}{40} & d > 40\,m \end{cases}$	4 dB

II. Channel Model

An accurate channel model is critical in the assessment of D2D communication link performance. To evaluate the D2D performance accurately, in our study we consider the following three aspects of the channel model: (i) *Small-scale fading*, which represents the rapid signal power fluctuation over small distance, (ii) *Large-scale fading*, which is also denoted as shadowing, and is obtained via averaging over a small distance to smooth out the rapid fluctuation of the received power, reflecting the slow variation of the local mean of the received power, and (iii) *Path loss*, which represents an area mean obtained by averaging the received power over a large distance to smooth out the shadowing effect.

To characterize path loss, we adopt the 3GPP path loss models for different D2D scenarios [5], which define the path loss for three different deployment scenarios: O2O, O2I, and I2I. Notice that the 3GPP D2D channel models were adapted from the IMT-advanced channel model defined in the ITU document [7]. The ITU I2I model is based on the channel model of indoor RRH/Hotzone [8]. The indoor RRH/Hotzone scenario in [8] considers a single floor of the building, and two hotzones deployed in the middle of the hallway, which is a very specific indoor case. To evaluate the I2I in same building case in a more generic way, we use the Distance-Partitioned model [9], which is obtained through measurements in a multi-story building. In this model, the distance is divided into four regions, and a different path loss exponent is used in each region to account for the wall and floor attenuations in the indoor environment.

Table I summarizes the path loss and the shadowing models used in our evaluation [5], [6]. From the table, d is the distance between the transmitter and the receiver, and $h^{'}_{BS}$ and $h^{'}_{MS}$ are effective antenna heights in meters for the transmitter and the receiver, respectively, f_c is the carrier frequency in GHz, f_{REF} is reference frequency in GHz, and h_{BS} and h_{MS} are the transmitter and the receiver antenna heights in meters. For O2I, d_{in} is the distance from the wall to the indoor UE in meters. For I2I same building case, PL_0 is the free space path loss at 1 m distance.

To derive the average performance, we need to consider both LOS and NLOS cases. The probability of LOS for the O2O scenario, denoted as P_{LOS_O2O}, as a function of the distance between the transmitter and the receiver, can be presented by [5]

$$P_{LOS\_O2O} = \begin{cases} 1, & \text{if } d \le 2.5\,m \\ 1 - 0.9(1 - (1.24 - 0.61\log_{10}(d))^3)^{\frac{1}{3}}, & \text{if } d > 2.5\,m \end{cases} \quad (1)$$

For the probability of the O2I LOS scenario, denoted as P_{LOS_O2I}, we have [5]: $P_{LOS\_O2I} = \min(\frac{18}{d}, 1)(1 - \exp(\frac{-d}{36})) + \exp(\frac{-d}{36})$, where d is the distance between the transmitter and the receiver.

In our study, the key parameters in the aforementioned equations are set as follows: carrier frequency $f_c = 700\,MHz$, reference frequency $f_{REF} = 2\,GHz$, and $h_{BS} = h_{MS} = 1.5\,m$ to match the average height of a human. The effective transmitter and receiver height $h^{'}_{BS} = h^{'}_{MS}$ are both set to $0.8\,m$, and the break point distance $d^{'}_{BP} = 4h^{'}_{BS}h^{'}_{MS}\frac{f_c}{c} = 5.97\,m$ [5]. It is worth mentioning that up to the break point, the free space path model can be used to compute the path loss. In our analysis, we evaluate D2D communication link performance after the break point, because we care about how far the D2D signals can propagate. For large-scale fading, we use the 3GPP defined Log-normal shadowing model [8], which uses 7 dB standard deviation for O2O and O2I, 4 dB for I2I in same build cases [6] and 10 dB for I2I in different building cases, respectively.

In order to make our analysis more practical, we consider small-scale fading as well. Particularly, we select Nakagami-m fading [10] as our fading model,

$$f(g) = \frac{2m^m g^{2m-1}}{\Gamma(m)\Omega^m}\exp(-\frac{mg^2}{\Omega}), g > 0, \quad (2)$$

where $f(g)$ is the probability density function (PDF) of the magnitude of the received signal envelope due to the fading, m is the fading parameter, Ω is the average received power, and

Rouil, Richard; Wang, Jian.
"Assessing Coverage and Throughput for D2D Communication."
Paper presented at 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, United States. May 20, 2018 - May 24, 2018.

Γ is the Gamma function, such that $\Gamma(m) = \int_0^\infty x^{m-1}e^{-m}\,dx$ and $\Omega = E(g^2)$.

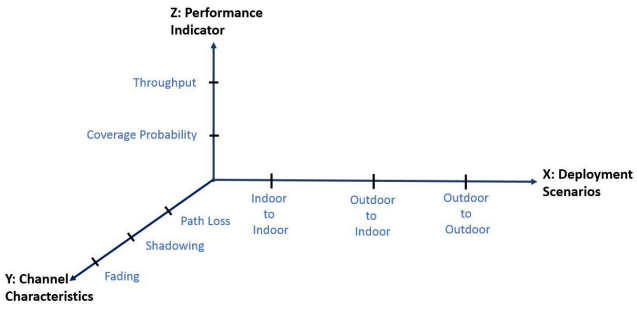


Fig. 1: Problem Space

Here, by varying m, we can simulate channels with different severities of small-scale fading. A smaller m means more severe fading. When m equals 1, Nakagami fading becomes Rayleigh fading.

III. Our Approach

In the following, we first present the problem space, and then show the deriving the mathematical expressions for both coverage probability and average throughput.

A. Problem Space

Figure 1 illustrates the overall problem space, which consists of the following three dimensions: X-axis as deployment scenarios, Y-axis as channel characteristics, and Z-axis as performance indicators. For the deployment scenarios, we consider I2I, I2O, and O2O communications. For I2I, we further consider two cases: (i) both the transmitter and the receiver are in the same building, and (ii) the transmitter and the receiver are located in different buildings. For the channel characteristics, we consider all three aspects of channel effects: the path loss, shadowing, and fading. For the performance indicator, we consider two metrics. One is the coverage probability, and the other is the average throughput, which will be defined later.

Based on the defined problem space, we conduct the performance analysis of the D2D communication link in different scenarios. In our analysis, in addition to the distance between the transmitter and the receiver, we investigate the performance impact of other parameters (e.g., PRB size, fixed/adaptive MCS, and transmit power) on the performance of D2D communication channels.

B. Coverage Probability

Coverage Probability is defined as the probability of the event Pr_c that the received signal power Pow_{RX} is greater than a given receiver sensitivity Pow_{th} (i.e., $Pr_c = Pr\{Pow_{RX} > Pow_{th}\}$). Thus, how to determine the receiver sensitivity Pow_{th}, and derive received signal power Pow_{RX} is the key to our analysis.

In determining the receiver sensitivity Pow_{th}, we consider the following three factors: the thermal noise floor, the device noise figure, and the SNR margin. The thermal noise floor is a function of channel bandwidth with a spectral density of $-174\,dBm/Hz$. We use $9\,dB$ as the noise figure of the receiver (i.e., a UE). The SNR margin is referred to as the minimum required SNR to maintain a reliable D2D communication. The SNR margin depends on the targeted BLER after the 4^{th} transmission, which is the maximum number of HARQ transmissions, along with the adopted MCS [2], [11]. The receiver sensitivity Pow_{th} will be the summation of the noise floor, UE noise figure, and the SNR margin in dB scale.

To derive the received signal power, we consider the following three channel effects: the path gain (inverse of the path loss), shadowing, and small-scale fading. In the linear scale, the received power Pow_{Rx} can be derived by $Pow_{Rx} = Pow_{Tx}\cdot G_{PG}\cdot G_S\cdot G_F$, where Pow_{Tx} is the transmitter power, G_{PG} is the path gain, G_S is the shadowing gain, and G_F is the fading gain. Here, G_S follows the Log-normal distribution, and G_F follows Nakagami-m fading distribution. In our analysis, we compute the coverage probability Pr_c for a given distance d between the transmitter and the receiver.

The PDF of the Log-normal shadowing [12] is $f(\Omega) = \frac{10/\ln 10}{\sqrt{2\pi}\sigma\Omega}\exp(-\frac{(10\log_{10}\Omega-\mu)^2}{2\sigma^2})$, where $\Omega > 0$, σ is the standard deviation of the Log-normal shadowing, and μ is the received power after considering the path loss in dBm.

The PDF of the Nakagami-m fading can be found in Equation (2). Thus, by considering both the Log-normal shadowing and the Nakagami-m fading, the composite PDF of received power can be derived as [12],

$$f(x) = \int_0^\infty (\frac{m}{w})^m \frac{x^{m-1}}{\Gamma(m)}\exp(-\frac{mx}{w})\frac{10}{\sqrt{2\pi}w\sigma ln10} \exp(-\frac{(10\log_{10}w-\mu)^2}{2\sigma^2})\mathrm{d}w, \tag{3}$$

where w is the local mean of received power. In addition, $\Gamma(m)$ is the $\Gamma(.)$ function.

With the PDF of the received SNR in Equation (3) and a given outage threshold γ, we have

$$\begin{aligned} Pr(x>\gamma) &= \int_\gamma^\infty f(x)\,dx \\ &= \frac{1}{\sqrt{\pi}\Gamma(m)}\int_{-\infty}^{+\infty}\exp(-x^2)\Gamma(m, \frac{m}{10^{\frac{\sqrt{(2)}\Omega x+\mu}{10}}})\,dx. \end{aligned} \tag{4}$$

Notice that the integral $\int_{-\infty}^{+\infty}\exp(-x^2)\Gamma(m, \frac{m}{10^{\frac{\sqrt{2}\omega x+\mu}{10}}})\,dx$ can be numerically approximated using Gaussian-Hermite Quadrature [12]. In our analysis, we use first 20 Hermite polynomials for the approximation and have $\int_{-\infty}^{+\infty}\exp(-x^2)g(x) = \sum_{i=1}^{i=20} w_i g(x_i)$.

C. Average Throughput

The average throughput is defined as the average number of bits successfully received by the receiver, divided by the time taken to transmit those bits, considering various channel conditions at a given distance between the transmitter and the receiver. To accurately evaluate the average throughput of a D2D communication link, we leverage the D2D BLER performance curve obtained though the link layer simulation

Rouil, Richard; Wang, Jian.
"Assessing Coverage and Throughput for D2D Communication."
Paper presented at 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, United States. May 20, 2018 - May 24, 2018.

in our prior study [11]. We analyze the link level throughput by considering the Adaptive Modulation and Coding (AMC) scheme and the PDF of the received signal power, as well as considering both the shadowing and the Nakagami-m small-scale fading.

To use AMC scheme, it is critical to determine the MCS switching points, which are determined based on the simulated BLER versus SNR curves. We consider the BLER versus SNR curves for the first transmission to find the SNRs corresponding to the $10\,\%$ BLERs. For example, based on our prior results [11], 3.1 dB and 4.3 dB SNRs are required for MCS10 and MCS11 to achieve $10\,\%$ BLER for the first transmission. Thus, if the SNR falls between 3.1 and 4.3 dB, we assume MCS10 is used in the transmission. Notice that, in the throughput analysis, we assume the perfect channel quality feedback so that the AMC scheme can be used.

Before computing the average throughput, we first use an exponential function to curve-fit the PSSCH BLER curves for the sake of computational efficiency. Thus, we have $f_{MCS_i}(\gamma) = a_i e^{b_i \gamma}$, where $f_{MCS_i}(\gamma)$ is the block error rate as a function of SNR γ for a given MCS i ($i \in [0\,20]$), a_i and b_i are the parameters associated with fitted curves. Using the fitted block error rate function, we can obtain the throughput $g_{MCS}(\gamma)$ corresponding to a given SNR γ for a given MCS i by $g_{MCS_i}(\gamma) = S_{TB}(1 - a_i e^{b_i \gamma})$, where S_{TB} is the transport block size.

Using the PDF of the received signal power as Equation (3), we can compute the average throughput by $E(g(\gamma)) = \int_0^{t_1} g_{MCS_0}(\gamma) f(\gamma)\, d\gamma + \sum_{i=1}^{i=20} \int_{t_i}^{t_{i+1}} g_{MCS_i}(\gamma) f(\gamma)\, d\gamma + \int_{t_{20}}^{\infty} g_{MCS_{20}}(\gamma) f(\gamma)\, d\gamma$. Here, t_i is the linear SNR value corresponding to the MCS_i switching point, which is defined as the $10\,\%$ BLER of the 1^{st} transmission. To compute $E(g(\gamma))$, the second term on right side can be computed by $\int_{t_i}^{t_{i+1}} g_{MCS_i}(\gamma) f(\gamma)\, d\gamma = \int_{t_i}^{\infty} g_{MCS_i}(\gamma) f(\gamma)\, d\gamma - \int_{t_{i+1}}^{\infty} g_{MCS_i}(\gamma) f(\gamma)\, d\gamma$, where $\int_{t_i}^{\infty} g_{MCS_i}(\gamma) f(\gamma)\, d\gamma = \int_{t_i}^{\infty} T_{BS}(1 - a e^{b\gamma}) f(\gamma)\, d\gamma = T_{BS}(\int_{t_i}^{\infty} f(\gamma)\, d\gamma - \int_{t_i}^{\infty} a e^{b\gamma} f(\gamma)\, d\gamma)$. The first term can be computed as Equation (4) and to compute the second term, we have

$$\begin{aligned} &\int_{t_i}^{\infty} a e^{bx} f(x)\, dx \\ &= \frac{a m^m}{\Gamma(m)\sqrt{\pi}} \int_{-\infty}^{\infty} \exp(-x^2) \frac{1}{(m - b 10^{\frac{\sqrt{2}\sigma x + \mu}{10}})^m} \\ &\Gamma(m, (\frac{m}{10^{\frac{\sqrt{2}\sigma x + \mu\omega}{10}}} - b) t_i)\, dx. \end{aligned} \tag{5}$$

Equation (5) can also be solved using Hermite approximation. Hence, the average throughput can be computed via the aforementioned mathematic derivation. Notice that, to derive average throughput for a fixed MCS_i, we have $E(g_{MCS_i}(\gamma)) = \int_0^{\infty} g_{MCS_i}(\gamma) f(\gamma)\, d\gamma$.

IV. Performance Evaluation

To evaluate the performance of D2D communication links, we select the coverage probability and average throughput, defined in Section III, as the two key metrics. To validate the analytic results, we also conducted Monte Carlo simulation. We generate log-normal random variable to simulate the shadowing effect and Gamma random variable with its mean following log-normal distribution to simulate combined shading and fading effect on the received signal power. Both numerical analysis and simulation were implemented in Matlab[1]. The Monte Carlo simulation results match well with the analytic results and they are presented together in each figure.

To successfully receive messages on PSSCH, PSCCH has to be correctly decoded first. In our evaluation, we assume the channel is semi-static (i.e., the received signal power is unchanged for the message transmission duration), including sending channel configuration on PSCCH and delivering the message over PSSCH. Thus, we define the coverage probability of the PSSCH as the probability that the received signal power is above the maximum value of the PSCCH receiver sensitivity threshold and the PSSCH receiver sensitivity threshold.

To evaluate the performance of the D2D communication link, we design the following two sets of experiments: (i) In the first set of experiments, we evaluate the performance of D2D communication links with respect to coverage probability and average throughput by varying the distance between the transmitter and the receiver. In this evaluation, we fix the system parameters. The PRB size is set to 6 and TX transmit power is set to 23 dBm. To evaluate the coverage probability, we use the lowest MCS (i.e., MCS0) to accommodate the worst channel condition. We also set the reliable threshold as 1 % BLER, meaning that when the transport block error is over 1 %, the communication becomes unreliable, and the communication system experiences outage. Notice that the set of parameters used in this evaluation is for demonstration purposes; the system parameters are easily configurable in our performance evaluation system. (ii) In the second set of experiments, we evaluate the performance of D2D communication links with respect to coverage probability and average throughput by varying system parameters, including PRB size, transmit power, and fixed MCS. In doing so, we can evaluate how these parameters can affect the performance of D2D communication links.

Coverage Probability: We use Equation (4) to compute the coverage probability for LOS and NLOS for a PSCCH channel, respectively. We can derive the average value by

$$Pr_c = Pr_{LOS_c} Pr_{LOS} + Pr_{NLOS_c}(1 - Pr_{LOS}). \tag{6}$$

Here, Pr_c is the average coverage probability, Pr_{LOS_c} is the coverage probability of a LOS link, Pr_{NLOS_c} is the coverage probability of a NLOS link, and Pr_{LOS} is the probability that the transmitter and the receiver have a LOS connection. Notice

[1]Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Rouil, Richard; Wang, Jian.
"Assessing Coverage and Throughput for D2D Communication."
Paper presented at 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, United States. May 20, 2018 - May 24, 2018.

Coverage Prob.	O2O (m)	O2I (m)	I2I SB (m)	I2I DB (m)
99 %	255 - 374	91 - 138	86 - 108	22 - 33
98 %	284 - 433	109 - 155	89 - 111	27 - 38
95 %	374 - 493	132 - 178	97 - 114	36 - 48
90 %	433 - 552	161 - 207	103 - 117	47 - 58

TABLE II: PSCCH Coverage Probability vs. TX-RX Separation (Distance) for Normal Power UE (23 dBm)

Coverage Prob.	O2O (m)	O2I (m)	I2I SB (m)	I2I DB (m)
99 %	374 - 582	138 - 207	100 - 125	35 - 50
98 %	433 - 642	161 - 231	106 - 128	41 - 58
95 %	552 - 761	198 - 271	114 - 134	55 - 72
90 %	672 - 850	242 - 300	122 - 136	69 - 89

TABLE III: PSCCH Coverage Probability vs. TX-RX Separation (Distance) for High Power UE (31 dBm)

that $1 - Pr_{LOS}$ is the probability of transmitter and receiver having a NLOS connection.

For I2I scenarios, we consider two cases, shown in Table I. The first case is that both UEs are in the same building, and the second case is that the two UEs are located in different buildings.

Figure 2 shows the coverage probability versus the distance between the transmitter and the receiver, in which both Log-normal shadowing and Nakagami-m small-scale fading are considered for O2O scenario. As we can see from the figure, the fading generally decreases the coverage probability. In addition, when the severity of fading increases (e.g., fading parameter m decreases), the transmitter coverage becomes smaller.

For O2I and I2I scenarios, we observe the same trend, and the simulation results are summarized in Table II. Table II lists the D2D communication ranges, which correspond to the distances between no fading and Rayleigh fading cases, to maintain given coverage probabilities with respect to UE with normal power. Notice that I2I SB means the same building for I2I scenario and I2I DB means the different buildings for I2I scenario.

To evaluate the coverage probability of the PSSCH, the coverage probability of the PSCCH shall be considered. Successfully receiving channel configurations on PSCCH is required for decoding PSSCH. PSSCH performance depends on the channel configuration, such as PRBs occupied by the channel. Thus, we observe that PSSCH outperforms PSCCH in certain cases, as shown in Figure 3. In addition, PSSCH with 6 PRBs and MCS0 has a slightly higher coverage probability than PSCCH, which is due to the fact that PSCCH uses the same modulation scheme as PSSCH MCS0 with a slightly higher coding rate. Furthermore, PSSCH has four Hybrid Automatic Repeat reQuest (HARQ) transmissions while PSCCH only transmits two identical copies. Thus, PSSCH performance is limited by PSCCH in this case. When the used PRB size is larger or the adopted MCS is higher, PSSCH performance will become the limiting factor for the coverage performance, as shown in Figures 4 and 5.

Average Throughput: Based on the analysis in Section III-C, we compute the numerical results for the average throughput. Similar to the evaluation of outage probability, since both LOS and NLOS scenarios are considered in the O2O and O2I channel models, we compute the average throughput for LOS and NLOS models separately, and then combine these two results using LOS probability.

With the fixed system parameters, and varying the distance between the transmitter and the receiver, we obtain the average throughput for the O2O scenario if AMC is used, as shown in Figure 6. In this figure, we compare the average throughput for path loss and shadowing only and path loss, shadowing, and Nakagami-m fading with different levels of fading severity ($m = 1, 2, 3$). As shown in the figure, when TX-RX separation (distance) increases, the average throughput declines. Also, fading reduces the average throughput compared to the non-fading case. The maximum throughput of the AMC depends on the allocated PRBs, and can be achieved by using the highest MCS. For O2I and I2I scenarios, we have similar observations to those of the O2O scenarios.

Parameter Sensitivity: In the following, we vary system parameters (PRB size, fixed MCS, and transmitter power), and study how these parameters affect the performance of D2D communication links. Without the loss of generality, all results in the following are based on the O2O scenario, and similar observations can be applied to O2I and I2I scenarios as well.

For UEs to communicate with each other, besides PSCCH, PSSCHs need to be decoded. Since PSSCH can occupy different sizes of PRBs, we evaluate how PRB size can affect the PSSCH coverage probability in Figure 7. In this case, the shadowing and Rayleigh fading are considered. As can be seen in the figure, with the increase of the PRB size, the distance between the transmitter and the receiver declines significantly in order to maintain a given coverage probability. This is because the wider the bandwidth, the higher the thermal noise floor, leading to an increased receiver sensitivity threshold.

In D2D communication, especially in group communication, it can be hard to obtain channel state information. In this case, the fixed resource allocation of MCS and PRB needs to be used. Figure 8 shows the throughput for a fixed MCS of MCS10 and 2 PRBs in an O2O scenario, and TX power of 23 dBm. From the figure, we observe that AMC achieves a higher spectrum efficiency compared with fixed MCS, especially when the transmitter and the receiver are relatively close to each other, where channel conditions are good.

The impact of transmit power on the outage probability for O2O scenario is shown in Figure 9, and Table III lists the D2D communication range to maintain given coverage probability for high power UE.

V. Conclusion

In this paper, we developed a modeling and simulation framework to assess the performance of D2D communication

Rouil, Richard; Wang, Jian.
"Assessing Coverage and Throughput for D2D Communication."
Paper presented at 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, United States. May 20, 2018 - May 24, 2018.

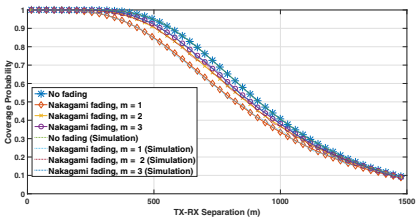


Fig. 2: PSCCH O2O Coverage Probability

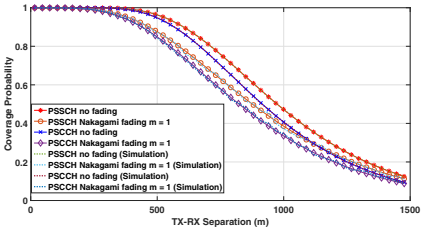


Fig. 3: O2O Coverage Probability of PSCCH vs. PSSCH (PRB = 6, AMC)

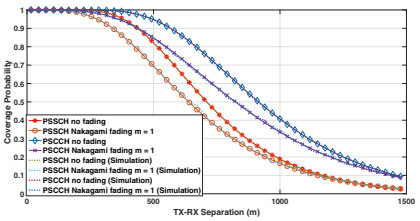


Fig. 4: O2O Coverage Probability of PSCCH vs. PSSCH (PRB = 25, AMC)

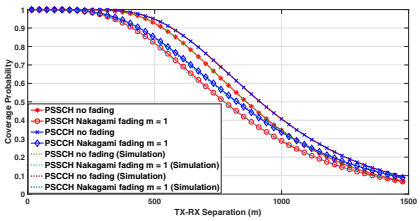


Fig. 5: O2O Coverage Probability of PSCCH vs. PSSCH (PRB = 2, MCS = 10)

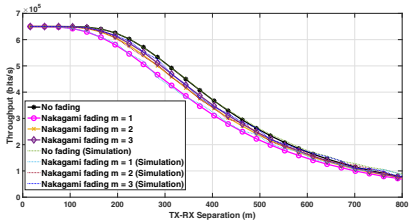


Fig. 6: O2O Throughput (PRB = 6, AMC)

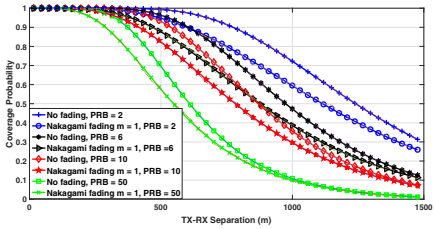


Fig. 7: Coverage Probability vs. PRB Size in PSSCH using AMC for O2O

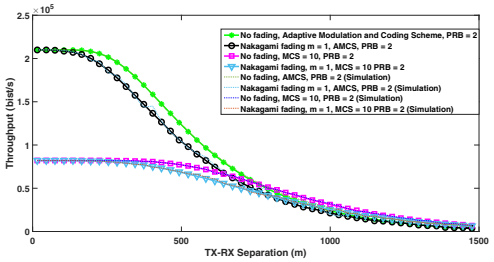


Fig. 8: Throughput of Fixed MCS vs. AMC in PSSCH for O2O Scenario

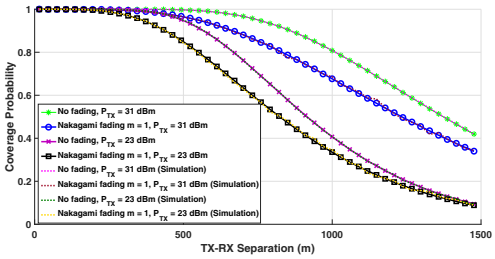


Fig. 9: Coverage Probability vs. Transmit Power in PSCCH

links with respect to coverage probability and average throughput. In our study, we considered various D2D deployment scenarios and channel effects. Based on the framework, we designed different outdoor and indoor scenarios, and derived mathematical expressions for the performance in different scenarios. We conducted an extensive performance evaluation to show the performance of the D2D communication link under various D2D deployment scenarios with different channel effects. We also evaluated the impact of system parameters (PRB size, fixed/adaptive MCS, and transmit power) on the performance of D2D communication links. Our designed framework is generic and can be extended to consider new deployment scenarios via adding or modifying its channel models, as well as reconfiguring system parameters and taking interference into account.

REFERENCES

[1] R. A. Rouil, A. I. Manzanares, M. R. Souryal, C. A. Gentile, D. W. Griffith, and N. T. Golmie, "Modeling a nationwide public safety broadband network," *IEEE Transactions on Vehicular Technology*, vol. 8, no. 2, pp. 83–91, 2013.

[2] 3GPP, "Technical Specification Group Services and System Aspects; Proximity-based services (ProSe); Stage 2 v.12.7.0," 3rd Generation Partnership Project (3GPP), TS 23.303, 2015.

[3] A. Asadi, Q. Wang, and V. Mancuso, "A survey on Device-to-Device communication in cellular networks," *IEEE Communications Surveys Tutorials*, vol. 16, no. 4, pp. 1801–1819, Fourthquarter 2014.

[4] J. Liu, N. Kato, J. Ma, and N. Kadowaki, "Device-to-device communication in lte-advanced networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 1923–1940, Fourthquarter 2015.

[5] 3GPP, "Technical Specification Group Radio Access Network; Study on LTE Device to Device Proximity Services; Radio Aspects v.12.0.1 ," 3rd Generation Partnership Project (3GPP), TR 36.843, 2014.

[6] D. Akerberg, "Properties of a TDMA pico cellular office communication system," in *Proc. of IEEE GLOBECOMM*, 1988.

[7] ITU, "Guildelines for evaluation of radio interface technologies for imt-advanced," International Telecommunication Union (ITU), TS M.2135-1, 2009.

[8] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA): Further Advancements for E-UTRA Physical Layer Aspects v9.0.0," 3rd Generation Partnership Project (3GPP), TS 36.814, 2010.

[9] K. Pahlavan and A. H. Levesque, *Wireless Information Networks*. John Wiley & Sons, Inc., 2005.

[10] N. C. Beaulieu and C. Cheng, "Efficient Nakagami-m fading channel simulation," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 2, pp. 413 – 424, 2005.

[11] J. Wang and R. A. Rouil, "BLER performance evaluation of LTE Device-to-Device communications," *National Institute of Standards Technology (NIST)*, vol. NISTIR 8157, 2016.

[12] G. L. Stuber, *Principles of Mobile Communication (2nd Edition)*. Springer, 2000.

Rouil, Richard; Wang, Jian.
"Assessing Coverage and Throughput for D2D Communication."
Paper presented at 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, United States. May 20, 2018 - May 24, 2018.

Modeling and Simulation Analysis of the Physical Sidelink Shared Channel (PSSCH)

David Griffith, Fernando Cintrón, Aneta Galazka, Timothy Hall, and Richard Rouil
National Institute of Standards and Technology
Gaithersburg, Maryland 20899
Contact Email: david.griffith@nist.gov

***Abstract*—This paper examines the performance of the Long Term Evolution (LTE) Physical Sidelink Shared Channel (PSSCH) in out-of-coverage (OOC) device-to-device (D2D) communication scenarios. We develop a closed form expression for the distribution of the number of User Equipments (UEs) that successfully decode a message sent on the PSSCH, given the number of UEs that received the transmitter's Sidelink Channel Information (SCI) message over the Physical Sidelink Control Channel (PSCCH). We validate our results using Monte Carlo simulations of the PSSCH and network simulations in ns3, and discuss some of the effects of system parameters on performance.**

I. Introduction

Device to device (D2D) communications was developed by the 3rd Generation Partnership Project (3GPP) to provide Proximity Services (ProSe) for LTE networks, and was added to the LTE standard in Release 12 [1]. Communications between D2D User Equipments (UEs) go over a *sidelink* rather than from the source UE to a base station via an uplink and then via a downlink from the base station to the destination UE. D2D communications will be used by network operators to provide new services and to offload intra-cell traffic, reducing load on the base station.

D2D communications is also an important component of the Nationwide Public Safety Broadband Network (NPSBN) [2]. The ProSe standard includes support for out-of-coverage (OOC) D2D communications, although it is for public safety agencies only. A motivation for OOC communications is the clear need for public safety personnel to be able to communicate when no base station is available. Example cases include operations in remote areas, loss of network infrastructure (e.g., due to hurricanes or wildfires), or operating inside buildings with severe structural penetration loss.

ProSe defines various sidelink channels that use resource pools consisting of groups of Physical Resource Blocks (PRBs); these channels carry data and control messages to support various D2D functions. Resource pools do not have to be contiguous in either the time or frequency domains, but they recur periodically in the time domain. Such pools exist to support device discovery, synchronizing clocks among of groups of devices, and communication between devices.

Disclaimer: Certain commercial products are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the commercial products identified are necessarily the best available for the purpose.

A. Background

A UE that intends to send data to other UEs over the sidelink uses the Physical Sidelink Shared Channel (PSSCH). The UE must first advertise the pending transmission using the Physical Sidelink Control Channel (PSCCH) to send a Sidelink Control Information (SCI) message, which tells other UEs which PSSCH resources the transmission will occupy, in addition to other information such as the Modulation and Coding Scheme (MCS) that will be used [3, Clause 5.14]. OOC UEs choose PSCCH resources randomly; each PSCCH resource corresponds to a pair of PRBs. The ProSe standard specifies the mapping from resource index numbers to PRB locations in the control channel resource pool [4, Clause 14.2.1.1]. If two or more UEs choose the same PSCCH resource index, their SCI messages will interfere with each other and will be unintelligible[1]. An additional source of message loss is the half-duplex nature of UE transmissions. A UE can miss an SCI message from another UE if it sends its own SCI in the same pair of subframes. In previous work, we modeled the PSCCH and we showed that if the PSCCH resource pool is properly dimensioned, the only cause of missed advertisements is collisions [6]. Whether SCIs are missed due to collisions or the half-duplex effect, UEs that miss advertisements will not be able to receive the corresponding data that is sent during the subsequent occurrence of the PSSCH.

The PSSCH consists of a set of periodically repeating PRBs that occur after the PSCCH in the time domain. The band of PRBs spanned by the PSSCH in the frequency domain is divided into N_{sb} sub-bands, while the set of subframes spanned by the PSSCH in the time domain is divided into multiple Time Resource Patterns (TRPs); each TRP spans N_{TRP} subframes. An OOC UE with data to send chooses a sub-band at random and also randomly chooses a set of k_{TRP} out of N_{TRP} subframes to use to transmit data in each TRP; the UE uses the same set of subframes in each TRP. The chosen pattern of subframes is called the UE's TRP mask.

[1]If the Signal to Interference and Noise Ratio (SINR) at the receiver is high enough, it may be possible to decode one of the interfering messages.

Cintron, Fernando; Griffith, David; Hall, Timothy; Rouil, Richard.
"Modeling and Simulation Analysis of the Physical Sidelink Shared Channel (PSSCH)."
Paper presented at 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, United States. May 20, 2018 - May 24, 2018.

UEs choose resources in the PSSCH randomly, so there is a risk that they can interfere with each other. For example, if two UEs choose the same sub-band and also choose TRP masks that partially overlap, then some of their transmissions will collide, causing interference. In addition, a UE that transmits in a given subframe will be unable to receive transmissions from other UEs in the same subframe if UEs cannot transmit and receive simultaneously, due to the half-duplex effect.

ProSe uses the Hybrid Automatic Repeat Request (HARQ) mechanism to mitigate the impact of collisions. UEs do not provide feedback over the sidelink for each HARQ transmission [4, Clause 14.1.1]. A transmitting UE sends four Redundant Versions (RVs) of data over the PSSCH; each RV is composed of information and error correction bits [5]. For this paper, we assume that each HARQ transmission takes up all the available PRBs in a subframe; i.e., it fills the chosen sub-band. For example, a set of four HARQ transmissions with $k_{\text{TRP}} = 1$ would be sent over the course of four TRPs, as shown in the top row of Fig. 1. If $k_{\text{TRP}} = 2$, then a UE can send two messages, with the first occupying the first two TRPs and the next occupying the last two TRPs; if $k_{\text{TRP}} = 4$, 4 RVs can be sent during every TRP.

B. Purpose of this work

In this paper, we develop a performance metric for the PSSCH. The analytical model underlying this metric incorporates the major features of the PSSCH: random sub-band selection, random TRP mask selection, the half duplex effect, and HARQ. We define our metric with respect to a single UE of interest in a group of N_u UEs, which we call UE_0. We define $\mathcal{R}_\rho^C$ to be the event, "ρ UEs decode UE_0's SCI," and $\mathcal{R}_\delta^S$ to be the event, "δ UEs decode UE_0's data on the PSSCH." Our performance metric is the conditional probability distribution of the number of UEs that decode UE_0's data: $\mathsf{Pr}\{\mathcal{R}_\delta^S \mid \mathcal{R}_\rho^C\}$. We can remove the condition on the number of UEs that decode the SCI by using the distribution of this number that we derived previously [6], giving

$$\mathsf{Pr}\{\mathcal{R}_\delta^S\} = \sum_{\rho=0}^{N_u-1} \mathsf{Pr}\{\mathcal{R}_\delta^S \mid \mathcal{R}_\rho^C\}\mathsf{Pr}\{\mathcal{R}_\rho^C\}. \tag{1}$$

This metric will allow a network operator to characterize the performance of an OOC group of UEs, and thus dimension the PSSCH to optimize the Quality of Experience (QoE) for public safety users. The metric can also be used to determine what input parameters (number of UEs, PSSCH pool dimensions, etc.) have the greatest impact on performance.

The rest of this paper is organized as follows. We briefly survey related work in Section II. In Section III, we develop the conditional distribution of the number of UEs that decode a transmitted message given that ρ of them decoded the corresponding SCI message. We validate our model in Section IV, and also examine the sensitivity of the metric to various input parameters. We summarize our discussion in Section V.

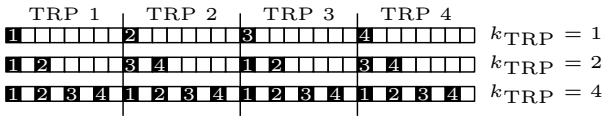


Fig. 1. Examples of transmission occurences of the four HARQ repetitions during a PSSCH period consisting of four TRPs, for various values of k_{TRP}.

II. Survey of Other Work

In [6], we developed a closed-form expression for the distribution of UEs that receive an SCI over the PSCCH, and we showed that the half-duplex effect can be eliminated by properly sizing the control resource pool. To the best of our knowledge, ours is the first work that models the PSSCH in this fashion, although some other works have considered D2D communications. Yoon, Park, and Choi developed a feedback scheme for the sidelink that uses a feedback pool that follows the PSSCH in time and has the same "shape" as the PSCCH [7]. Park et al. developed a resource selection scheme for the PSCCH that aims to improve performance by using measured interference to inform the selection process, rather than using only random selection [8]. Shih et al. developed an autonomous resource selection algorithm for out of coverage UEs [9]. Their approach partitions the control channel resource pool so that UEs sense the energy in the first PRB that they choose and then pick a different resource if they sense a collision. They use a collision analysis for the SCI and the transmitted data, but their performance analysis does not account for the half duplex effect in the PSCCH or the PSSCH, and does not account for the effect of HARQ in the PSSCH. The analysis also does not include the 3GPP mapping from resource index to PSCCH PRBs.

III. Theoretical Analysis

A. Model description

For the following model description, we provide a list of variable definitions in Table I. Consider a group of OOC UEs and let N_u be the number of UEs in the group. We assume that all UEs have data to send. We focus on a single transmitting UE of interest, which we call UE_0. All UEs contend for PSCCH resources to send their SCI messages. In this analysis, we assume that a subset of ρ UEs successfully decoded UE_0's SCI message.

To model the HARQ function, we define ψ_i to be the probability that a UE successfully decodes UE_0's message given that it received i HARQ transmissions, where $i \in \{0, 1, 2, 3, 4\}$. The number of messages that a UE can send during a period depends on k_{TRP}, as shown in Fig. 1. The rows in Fig. 1 each represent a period consisting of four TRPs, and correspond to the cases $k_{\text{TRP}} = 1, 2, 4$. Example subframe masks are shown in black in each row.

From the example masks in Fig. 1, we can determine which decoding probabilities ψ_i to use for a given value of k_{TRP}. Since a message requires four HARQ transmissions, and since each UE uses a single mask for all TRPs in a

Cintron, Fernando; Griffith, David; Hall, Timothy; Rouil, Richard.
"Modeling and Simulation Analysis of the Physical Sidelink Shared Channel (PSSCH)."
Paper presented at 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, United States. May 20, 2018 - May 24, 2018.

TABLE I
LIST OF VARIABLES

Symbol	Definition
N_u	UE group size
UE_0	Randomly chosen UE of interest
$\mathcal{R}_\rho^C$	Event where $\rho \leq (N_u - 1)$ UEs decode UE_0's SCI
$\mathcal{R}_\delta^S$	Event where $\delta \leq \rho$ UEs decode UE_0's data on the PSSCH
N_{TRP}	Number of subframes per TRP
k_{TRP}	Number of subframes per TRP used by UEs to send data
Y	Number of subframes per TRP used by UE_0 not affected by collisions with other UEs operating in the same sub-band
N_{sb}	Number of sub-bands partitioning the PSSCH in the frequency domain
ρ	Number of UEs that decode UE_0's SCI message
ι	Number of UEs that do not decode UE_0's SCI message
s	Number of UEs that transmit in UE_0's sub-band
d	Number of UEs that transmit in sub-bands different than UE_0's sub-band
s''	Number of same-sub-band (SSB) UEs that decode UE_0's SCI
d''	Number of other-sub-band (OSB) UEs that decode UE_0's SCI
ψ_i	Probability that a UE decodes UE_0's SCI given that it received i transmissions
S_ρ	Number of receiver UEs that transmit in UE_0's sub-band
S_ι	Number of interferer UEs that transmit in UE_0's sub-band
S	Total number of UEs that transmit in UE_0's sub-band
D_ρ	Number of receiver UEs that transmit in different sub-bands
D_ι	Number of interferer UEs that transmit in different sub-bands
s'	Value taken by S_ρ
σ	Value taken by S_ι
n	Value taken by Y
ω_n	$\Pr\{\text{SSB receiver UE decodes } \text{UE}_0\text{'s SCI} \mid Y = n\}$
ϕ_n	$\Pr\{\text{OSB receiver UE decodes } \text{UE}_0\text{'s SCI} \mid Y = n\}$
N_{run}	Number of Monte Carlo runs per validation simulation
N_{trials}	Number of trials per validation run
T	Number of UEs decoding UE_0's message per trial

TABLE II
VALUES OF ψ_n VERSUS n FOR VARIOUS VALUES OF k_{TRP}

n	k_{TRP}		
	1	2	4
0	ψ_0	ψ_0	ψ_0
1	ψ_4	ψ_2	ψ_1
2	–	ψ_4	ψ_2
3	–	–	ψ_3
4	–	–	ψ_4

given PSSCH period, the decoding probability depends on the number of UE_0's subframes that are not subject to interference from collisions and that other UEs can receive because they are not transmitting during those subframes. If $k_{\text{TRP}} = 1$, UE_0's message requires four TRPs to transmit, and a collision will impact all four transmissions, and all receivers decode UE_0's message with probability ψ_0; conversely, if there is no collision, then all four of UE_0's transmissions can be received (provided that a receiver in another sub-band is not using UE_0's mask) with probability ψ_4. By similar reasoning, we can construct the table of message decoding probabilities shown in Table II for other values of k_{TRP}. We will use these values in the table in the model development in Section III-C.

B. Constructing the model

By examining the selection of resources by different UEs in a particular sequence, which does not affect the fidelity of the model, we can determine subframe choices of UEs that operate in UE_0's sub-band affect outcomes for UEs that are not in UE_0's sub-band.

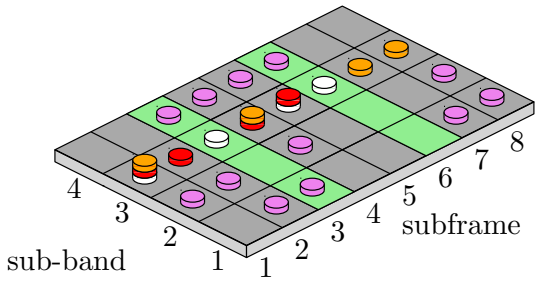


Fig. 2. An example of sub-band and TRP mask choices by UE_0 (white disks), UEs that choose UE_0's sub-band (red and orange disks), and UEs that choose other sub-bands (purple disks).

In the example shown in Fig. 2, $N_{\text{TRP}} = 8$ subframes, $k_{\text{TRP}} = 4$ subframes, and $N_{\text{sb}} = 4$ sub-bands. The figure shows an example outcome of UE_0's choice of a subframe mask, and a random sub-band. We show UE_0's choice with white tokens placed in spaces corresponding to subframes $1, 3, 5, 6$ in sub-band 3. Fig. 2 also shows transmissions by two UEs that have chosen UE_0's sub-band. We show these two UEs' choices using red and orange tokens placed in spaces corresponding respectively to subframes $1, 2, 4, 5$ and $1, 4, 7, 8$ in sub-band 3. Because the two other UEs' transmissions overlap some of those of UE_0, UE_0 has two collided and two non-collided transmissions. Thus the two UEs that transmit in UE_0's subframe decode UE_0's message with probability ψ_2.

Finally, Fig. 2 shows the effect of transmissions by UEs that have chosen sub-bands other than UE_0's sub-band. For clarity, only one UE is shown per sub-band; we show the other-sub-band (OSB) UEs' choices with purple tokens placed in spaces corresponding to the masks that they have chosen. The UE that chose sub-band 1 chose a mask that overlaps one of UE_0's two uncollided transmissions in subframe 3, so this UE decodes UE_0's message with probability ψ_1. The UE that chose sub-band 2 has a mask that overlaps neither of UE_0's uncollided transmissions; this UE decodes UE_0's message with probability ψ_2. Finally, the mask chosen by the UE that chose sub-band 4 overlaps both of UE_0's uncollided transmissions, so this UE decodes UE_0's message with probability ψ_0.

C. Development of $\Pr\{\mathcal{R}_\delta^S \mid \mathcal{R}_\rho^C\}$

Our model uses Jordan's formula [11, Eqs. (3,4)], which we briefly describe here. Given a probability space Ω with equally likely outcomes $\{\omega \in \Omega\}$, define n events $A_1, \ldots, A_n$, each of which corresponds to a subset of Ω; the probability of an event A_i is $\Pr\{A_i\} = \mathcal{N}(A_i)/\mathcal{N}(\Omega)$, where $\mathcal{N}(A)$ is the number of elements in the set A. Let the random variable ν be the number of events that occur due to an outcome ω. The probability that exactly k out of n events occur (i.e., $\nu = k$)

Cintron, Fernando; Griffith, David; Hall, Timothy; Rouil, Richard.
"Modeling and Simulation Analysis of the Physical Sidelink Shared Channel (PSSCH)."
Paper presented at 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, United States. May 20, 2018 - May 24, 2018.

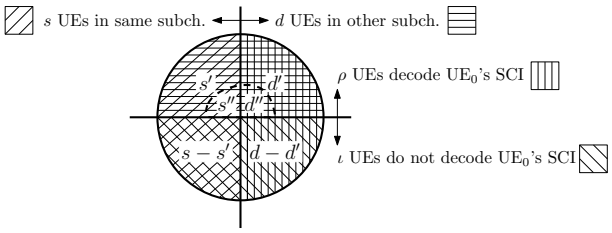


Fig. 3. Partitioning of the set of $(N_u - 1)$ UEs with respect to the UE of interest, UE_0.

is

$$\mathsf{Pr}\{k \text{ events occur}\} = \sum_{r=k}^{n} (-1)^{r-k} \binom{r}{k} \mathsf{E}\left\{ \binom{\nu}{r} \right\} \tag{2}$$

where

$$\mathsf{E}\left\{ \binom{\nu}{r} \right\} = \sum_{1 \le i_1 < i_2 < \cdots < i_r \le n} \mathsf{Pr}\{\cap_{j=1}^{r} A_{i_j}\} \tag{3}$$

is the rth binomial moment of ν.

To derive $\mathsf{Pr}\{\mathcal{R}_\delta^S \,|\, \mathcal{R}_\rho^C\}$, we partition the set of $(N_u - 1)$ UEs in UE_0's group based on whether they received UE_0's SCI and whether they randomly select the sub-band chosen by UE_0. We show this partitioning in Fig. 3. We call the ρ UEs that received UE_0's SCI, "receivers;" and the remaining $\iota = (N_u - 1 - \rho)$ UEs, "interferers." We define the random variables S_ρ and S_ι to be, respectively, the number of receivers and interferers that use UE_0's sub-band. Let $S = S_\rho + S_\iota$ be the number of UEs in UE_0's sub-band. We define s' and σ to be values taken by S_ρ and S_ι, respectively, and s to be the value taken by S, so that $s = \sigma + s'$. In addition, we define D_ρ and D_ι to be random variables that are the number of receivers and interferers that use sub-bands other than UE_0's; d' is the specific value taken by D_ρ.

Next, we condition on the number of receivers and interferers that choose UE_0's sub-band, which gives us

$$\begin{aligned}\mathsf{Pr}\{\mathcal{R}_\delta^S \,|\, \mathcal{R}_\rho^C\} = \sum_{s'=0}^{\rho} \sum_{\sigma=0}^{\iota} \mathsf{Pr}\{\mathcal{R}_\delta^S \,|\, \mathcal{R}_\rho^C,\, S_\rho = s',\, S_\iota = \sigma\} \\ \times \mathsf{Pr}\{S_\rho = s',\, S_\iota = \sigma\}.\end{aligned} \tag{4}$$

The probability that a given UE picks UE_0's sub-band is $1/N_{\text{sb}}$. Thus the probability that s' out of ρ receivers and $\sigma = s - s'$ out of ι interferers pick UE_0's sub-band, which is $\mathsf{Pr}\{S_\rho = s',\, S_\iota = \sigma\}$ in Eq. (4), is

$$\begin{aligned}&\mathsf{Pr}\{S_\rho = s',\, S_\iota = \sigma\} \\ &\quad = \tbinom{\rho}{s'} \left(\tfrac{1}{N_{\text{sb}}}\right)^{s'} \left(1 - \tfrac{1}{N_{\text{sb}}}\right)^{\rho - s'} \tbinom{\iota}{\sigma} \left(\tfrac{1}{N_{\text{sb}}}\right)^{\sigma} \left(1 - \tfrac{1}{N_{\text{sb}}}\right)^{\iota - \sigma} \\ &\quad = \tbinom{\rho}{s'} \tbinom{\iota}{\sigma} \left(\tfrac{1}{N_{\text{sb}}}\right)^{s} \left(1 - \tfrac{1}{N_{\text{sb}}}\right)^{N_u - 1 - s}.\end{aligned} \tag{5}$$

Let the random variable Y be the number of UE_0's transmitted subframes in a TRP that do not experience interference from other UEs in UE_0's sub-band. By conditioning on the value of Y, we can expand $\mathsf{Pr}\{\mathcal{R}_\delta^S \,|\, \mathcal{R}_\rho^C,\, S_\rho = s',\, S_\iota = \sigma\}$ from Eq. (4) as follows:

$$\begin{aligned}&\mathsf{Pr}\{\mathcal{R}_\delta^S \,|\, \mathcal{R}_\rho^C,\, S_\rho = s',\, S_\iota = \sigma\} \\ &\quad = \sum_{n=0}^{k_{\text{TRP}}} \mathsf{Pr}\{\mathcal{R}_\delta^S \,|\, \mathcal{R}_\rho^C,\, S_\rho = s',\, S_\iota = \sigma,\, Y = n\} \\ &\qquad \times \mathsf{Pr}\{Y = n \,|\, S = s' + \sigma\}.\end{aligned} \tag{6}$$

The probability that Y takes the value n is

$$\mathsf{Pr}\{Y = n \,|\, S = s\} = \sum_{\ell=n}^{k_{\text{TRP}}} (-1)^{\ell - n} \binom{\ell}{n} \mathsf{E}\left\{ \binom{Y}{\ell} \,\middle|\, S = s \right\} \tag{7}$$

where

$$\mathsf{E}\left\{ \binom{Y}{\ell} \,\middle|\, S = s \right\} = \sum_{1 \le i_1 < \cdots < i_\ell \le k_{\text{TRP}}} \mathsf{Pr}\{\cap_{j=1}^{\ell} A_{i_j} \,|\, S = s\}, \tag{8}$$

and we define the set of events $\{A_i\}_{i=1}^{N_{\text{TRP}}}$ as follows. Event A_i occurs when UE_0 chooses subframe i and no other UE transmitting in UE_0's sub-band chooses that subframe. Thus $\mathsf{Pr}\{\cap_{j=1}^{\ell} A_{i_j} \,|\, S = s\}$ is the probability that the ℓ subframes chosen by UE_0 whose indices are $i_1, i_2, \ldots, i_\ell$ are not used by s other UEs. We compute this conditional probability by taking the ratio of the number of ways that a single UE can pick a mask such that it does not use subframes $i_1, i_2, \ldots, i_\ell$, to the total number of ways that a UE can pick a mask. To avoid picking subframes $i_1, i_2, \ldots, i_\ell$, a UE has $(N_{\text{TRP}} - \ell)$ subframes from which to choose k_{TRP} subframes for its mask, and there are $\binom{N_{\text{TRP}} - \ell}{k_{\text{TRP}}}$ ways to do this. The total number of masks that any UE can pick is $\binom{N_{\text{TRP}}}{k_{\text{TRP}}}$, so the probability that a UE chooses a mask so that it does not overlap subframes $i_1, i_2, \ldots, i_\ell$ is $\binom{N_{\text{TRP}} - \ell}{k_{\text{TRP}}} / \binom{N_{\text{TRP}}}{k_{\text{TRP}}}$ [2]. UEs choose masks independently, so if the number of other UEs that choose UE_0's sub-band is $S = s$, then

$$\mathsf{Pr}\{\cap_{j=1}^{\ell} A_{i_j} \,|\, S = s\} = \left[\frac{\binom{N_{\text{TRP}} - \ell}{k_{\text{TRP}}}}{\binom{N_{\text{TRP}}}{k_{\text{TRP}}}} \right]^{s}, \; s = 0, 1, \ldots, N_u - 1. \tag{9}$$

All of the $\mathsf{Pr}\{\cap_{j=1}^{\ell} A_{i_j} \,|\, S = s\}$ terms in the sum in Eq. (8) are identical and are given by Eq. (9). There are $\binom{k_{\text{TRP}}}{\ell}$ terms in the sum in Eq. (8), because this is the number of ways to pick ℓ subframes from the set of k_{TRP} subframes that make up UE_0's mask. Thus, if $S = s = s' + \sigma$,

$$\begin{aligned}&\mathsf{Pr}\{Y = n \,|\, S = s' + \sigma\} \\ &\quad = \sum_{\ell=n}^{k_{\text{TRP}}} (-1)^{\ell - n} \binom{\ell}{n} \binom{k_{\text{TRP}}}{\ell} \left[\frac{\binom{N_{\text{TRP}} - \ell}{k_{\text{TRP}}}}{\binom{N_{\text{TRP}}}{k_{\text{TRP}}}} \right]^{s' + \sigma}.\end{aligned} \tag{10}$$

To get $\mathsf{Pr}\{\mathcal{R}_\delta^S \,|\, \mathcal{R}_\rho^C,\, S_\rho = s',\, S_\iota = \sigma,\, Y = n\}$ in Eq. (6), we define s'' to be number of same-sub-band (SSB) UEs that decode UE_0's message, and d'' to be the number of OSB UEs that decode UE_0's message. It follows that $s'' \le s'$ and $d'' \le$

[2] Note that $\binom{N_{\text{TRP}} - \ell}{k_{\text{TRP}}} / \binom{N_{\text{TRP}}}{k_{\text{TRP}}} = 0$ if $N_{\text{TRP}} - \ell < k_{\text{TRP}}$.

Cintron, Fernando; Griffith, David; Hall, Timothy; Rouil, Richard.
"Modeling and Simulation Analysis of the Physical Sidelink Shared Channel (PSSCH)."
Paper presented at 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, United States. May 20, 2018 - May 24, 2018.

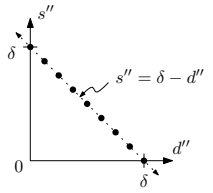


Fig. 4. Illustration of the relationship between d'' and s''.

d', as shown in Fig. 3. If δ is the total number of UEs that decode UE_0's message, then $s'' + d'' = \delta$, as shown in Fig. 4, and we let s'' vary from 0 to δ and then set $d'' = \delta - s''$.

Conditioning on the number of UEs in UE_0's sub-band that decoded the transmitted message gives

$$\begin{aligned}&\mathsf{Pr}\{\mathcal{R}_\delta^S \,|\, \mathcal{R}_\rho^C,\, S_\rho = s',\, S_\iota = \sigma,\, Y = n\} \\ &\quad = \sum_{s''=0}^{\delta} \mathsf{Pr}\{\mathcal{R}_\delta^S \,|\, \mathcal{R}_\rho^C,\, S_\rho = s',\, S_\iota = \sigma,\, Y = n,\, S = s''\} \\ &\qquad \times \mathsf{Pr}\{S = s''\}. \end{aligned} \tag{11}$$

Given $Y = n$, the probability that s'' receivers out of the s' receivers in UE_0's sub-band decode the message is

$$\mathsf{Pr}\{S = s''\} = \binom{s'}{s''} \omega_n^{s''} (1 - \omega_n)^{s' - s''}. \tag{12}$$

where $\omega_n = \psi_{4n/k_{\text{TRP}}}$ is the probability that a SSB UE decodes UE_0's message given that it received n transmissions, as given in Table II.

Let ϕ_n be the probability that an OSB receiver decodes UE_0's message given $Y = n$. We condition on the value of m, the number of unblocked subframes that the OSB receiver can access, given n unblocked subframes, and use the same approach that we used to develop Eq. (9) and Eq. (10) (with $s = 1$ in this case):

$$\begin{aligned}\phi_n &= \sum_{m=0}^{n} \psi_{4m/k_{\text{TRP}}} \mathsf{Pr}\{\text{receive } m \text{ of } n \text{ unblocked subframes}\} \\ &= \sum_{m=0}^{n} \psi_{4m/k_{\text{TRP}}} \left[\sum_{\ell=m}^{n} (-1)^{\ell - m} \binom{\ell}{m} \binom{n}{\ell} \frac{\binom{N_{\text{TRP}} - \ell}{k_{\text{TRP}}}}{\binom{N_{\text{TRP}}}{k_{\text{TRP}}}} \right]. \end{aligned} \tag{13}$$

Then the probability that d'' out of d' OSB receivers decode UE_0's message is

$$\begin{aligned}&\mathsf{Pr}\{\mathcal{R}_\delta^S \,|\, \mathcal{R}_\rho^C,\, S_\rho = s',\, S_\iota = \sigma,\, Y = n,\, S = s''\} \\ &\quad = \binom{d'}{d''} \phi^{d''} (1 - \phi)^{d' - d''} \\ &\quad = \binom{\rho - s'}{\delta - s''} \phi^{\delta - s''} (1 - \phi)^{(\rho - s') - (\delta - s'')}. \end{aligned} \tag{14}$$

Note that we have to have $s'' \le s'$ for Eq. (12) to be non-zero and $\delta - s'' \le \rho - s'$ for Eq. (14) to be non-zero; thus the series in Eq. (11) runs from $s'' = \max(0, s' - \rho + \delta)$ to $s'' = \min(\delta, s'')$.

To obtain the final form for the conditional distribution, $\mathsf{Pr}\{\mathcal{R}_\delta^S \,|\, \mathcal{R}_\rho^C\}$, we substitute Eq. (12) and Eq. (14) into Eq. (11), then combine the resulting expression with Eq. (10) in Eq. (6), and insert this result into Eq. (4) along with Eq. (5). The result is the expression in Eq. (15).

IV. Numerical Results

A. Monte Carlo simulations

To validate the theoretical model, we implemented a set of Monte Carlo simulations in Matlab whose output is an empirical conditional probability mass function of the number of UEs that decode transmitted data from a randomly chosen peer. For each set of input parameters, we performed $N_{\text{runs}} = 5$ runs, with N_{trials} 10 000 trials per run.

In each trial, we use the following procedure. We initialize T, the tally of UEs that decoded UE_0's message, to zero, and we randomly assign a sub-band and mask to UE_0. Next, we assign sub-bands and masks to the other $(N_u - 1)$ UEs in the group, and we determine n, the number of UE_0's subframes that are not blocked by transmissions from other UEs in UE_0's sub-band. For each receiver UE in UE_0's sub-band, we generate U, a $U[0, 1]$ random variate; if $U \le \psi_n$, we increment T. For each receiver in other sub-bands, we determine k, the number of UE_0's un-collided subframes that do not overlap with subframes chosen by the OSB receiver UE. Then we generate U for that UE and if $U \le \psi_k$, we increment T. After processing the last OSB UE, we record the value of T for the trial, and move on to the next trial in the run.

Once all the runs were complete, we computed an empirical probability mass function (PMF) for each run. For the rth run, for $0 \le m \le N_u - 1$, we computed $\widehat{p}_r(m)$, which is the number of times that $T = m$ divided by N_{trials}. Then we generated $\widehat{p}(m) = \sum_{r=1}^{N_{\text{runs}}} \widehat{p}_r(m) / N_{\text{runs}}$. We also estimated the standard deviation of $\widehat{p}_r(m)$, $\varsigma(m)$. The approximate 95 % confidence interval for the PMF is $\widehat{p}(m) \pm 1.96\, \varsigma(m) / \sqrt{N_{\text{runs}}}$.

B. NS3 simulations

We simulated a group of N_u UEs using the ns3 simulation tool as an additional check of the model. In order to fix the number of UEs that successfully receive a given UE's SCI message to have a given value of ρ during each simulated period, we set the SINR threshold for SCI messages so that all transmitted SCI messages will be received by all other UEs, even if collisions occur in the PSCCH. We performed 5 simulation runs for each set of input parameters; for each run, we simulated 400 s of activity. Each instance of the PSSCH contained 4 TRPs whose duration was 8 subframes each, and the PSCCH's duration was 8 subframes. This results in a period duration of 40 ms, so we simulated 10 000 periods per run. We picked $(\rho+1)$ UEs to send SCIs that were received by all UEs in the group. We randomly picked one of the $(\rho+1)$ SCI-transmitting UEs to be UE_0 and all N_u UEs in the group then sent data over the PSSCH. We counted the number of UEs that were able to decode UE_0's data and we saved this tally as the result for the run. We combined the results from the

Cintron, Fernando; Griffith, David; Hall, Timothy; Rouil, Richard.
"Modeling and Simulation Analysis of the Physical Sidelink Shared Channel (PSSCH)."
Paper presented at 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, United States. May 20, 2018 - May 24, 2018.

$$\mathsf{Pr}\{\mathcal{R}_\delta^S \mid \mathcal{R}_\rho^C\} = \sum_{s'=0}^{\rho}\sum_{\sigma=0}^{\iota}\binom{\rho}{s'}\binom{\iota}{\sigma}\left(\frac{1}{N_{\rm sb}}\right)^{s'+\sigma}\left(1-\frac{1}{N_{\rm sb}}\right)^{(N_u-1)-(s'+\sigma)} \times\left(\sum_{n=0}^{k_{\rm TRP}}\left[\sum_{\ell=n}^{k_{\rm TRP}}(-1)^{\ell-n}\binom{\ell}{n}\binom{k_{\rm TRP}}{\ell}\left[\frac{\binom{N_{\rm TRP}-\ell}{k_{\rm TRP}}}{\binom{N_{\rm TRP}}{k_{\rm TRP}}}\right]^{s'+\sigma}\right]\right. \times\left.\left[\sum_{s''=\max(0,s'-\rho+\delta)}^{\min(\delta,s')}\binom{s'}{s''}\omega_n^{s''}(1-\omega_n)^{s'-s''}\binom{\rho-s'}{\delta-s''}\phi_n^{\delta-s''}(1-\phi_n)^{(\rho-s')-(\delta-s'')}\right]\right) \quad (15)$$

set of runs in the same manner as in Section IV-A to produce an empirical PMF with 95 % confidence intervals.

C. Discussion of results

To generate our results, shown in Fig. 5, we used $N_u = 11$ UEs, and the PSSCH bandwidth was 50 PRBs, or 10 MHz. We examined two partitions of the PSSCH bandwidth: $N_{\rm sb} = 8$ sub-bands and $N_{\rm sb} = 16$ sub-bands, and we considered two mask types: $k_{\rm TRP} = 2$ and $k_{\rm TRP} = 4$. We also investigated two values for the number of UEs that received UE_0's SCI: $\rho = 3$ UEs and $\rho = 9$ UEs. We validated our model for other values of N_u and ρ, and for $k_{\rm TRP} = 1$, but we do not show these results due to space limitations.

In all four sub-figures, we observe excellent agreement between the theoretical results and the empirical results from both sets of simulations, which is strong evidence for the accuracy of our model. All four sub-figures also show that increasing the number of sub-bands shifts the mass of the conditional distribution to the right, i.e., we see a reduction in the probability that no receiver UEs decode UE_0's while we also see an increase in decoding probabilities for the greatest number of UEs. However, we note that the price of increasing $N_{\rm sb}$ is a reduction in the number of PRBs per subframe that a UE is able to use to send data; the trade-off that results in the maximum throughput is a topic for further study. In addition, we see that while increasing $k_{\rm TRP}$ increases the number of possible masks (70 when $k_{\rm TRP} = 4$ vs. 28 when $k_{\rm TRP} = 2$), this does not produce higher decoding probabilities; masks that cover more subframes create greater chances for interference or loss of transmissions due to the half duplex effect. The extreme example of this effect occurs when $k_{\rm TRP} = 8$; all UEs will block or miss each others' transmissions with probability one, since they transmit in every subframe. However, reducing $k_{\rm TRP}$ also reduces the number of PRBs available for a UE to use, and so there is a second trade-off with respect to throughput.

To further illustrate the trends shown in Fig. 5, we show the theoretical values of the conditional means and variances, $\mathsf{E}\{\mathcal{R}_\delta^S \mid \mathcal{R}_\rho^C\}$ and $\mathsf{Var}\{\mathcal{R}_\delta^S \mid \mathcal{R}_\rho^C\}$, as ordered pairs in Table III for the cases plotted in the figure. We note from the table that if the number of UEs that decoded UE_0's SCI message is small, then varying other parameters such as ρ or $k_{\rm TRP}$ or $N_{\rm sb}$ does not significantly affect the conditional statistics associated with the number of these UEs that decode UE_0's message on the PSSCH. If the PSCCH is configured to allow most UEs to decode the SCI message, then the design of the PSSCH has more impact, with the greater impact coming from the mask design (i.e., the value of $k_{\rm TRP}$).

TABLE III
($\mathsf{E}\{\mathcal{R}_\delta^S \mid \mathcal{R}_\rho^C\}$, $\mathsf{Var}\{\mathcal{R}_\delta^S \mid \mathcal{R}_\rho^C\}$) FOR PARAMETERS IN FIG. 5

	$N_{\rm sb} = 8$		$N_{\rm sb} = 16$	
	$k_{\rm TRP} = 2$	$k_{\rm TRP} = 4$	$k_{\rm TRP} = 2$	$k_{\rm TRP} = 4$
$\rho = 3$	(1.69 , 1.11)	(0.84 , 0.78)	(1.95 , 0.91)	(1.13 , 0.85)
$\rho = 9$	(5.07 , 6.72)	(2.52 , 3.96)	(5.86 , 4.83)	(3.38 , 3.83)

V. Summary and Conclusions

In this paper, we developed a mathematical model to characterize the performance of the PSSCH for out-of-coverage D2D communications. The model produces the distribution of the number of devices that receive a UE's data transmission given the number of UEs that received the corresponding SCI message. This result can be combined with existing models of the PSCCH. We validated our model using two sets of simulations: a simple Monte Carlo model and a full simulation in ns3 of an OOC group of UEs using D2D communications. The results that we obtained show that increasing the number of sub-bands in the PSSCH improves the likelihood of a UE's decoding a transmitted message, although this is at the expense of throughput, which we intend to quantify in future work. We also showed that the value of $k_{\rm TRP}$ has a significant impact on performance, with lower values resulting in a greater likelihood that a receiver UE decodes the message on the PSSCH, but that this also reduces throughput. In future work, we will discuss how to use this model to maximize the throughput on the sidelink.

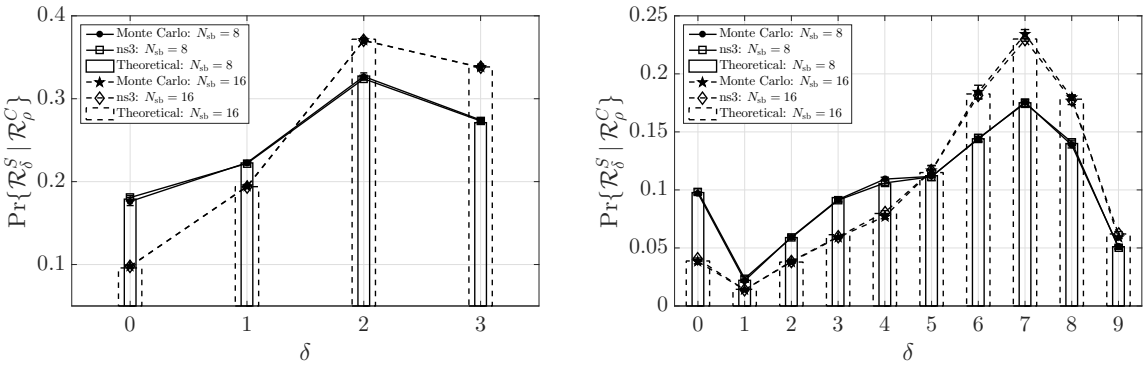
Acknowledgment

The authors would like to thank the NIST Summer Undergraduate Research Fellowship (SURF) program, which made it possible for Aneta Galazka to work on this project.

References

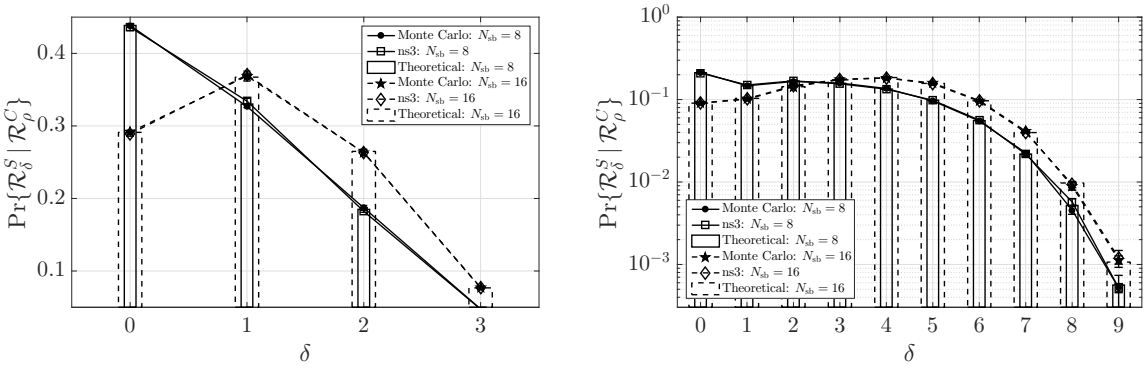
[1] D. Astely, E. Dahlman, G. Fodor, S. Parkvall, and J. Sachs, "LTE release 12 and beyond [accepted from open call]," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 154-160, July 2013.

Cintron, Fernando; Griffith, David; Hall, Timothy; Rouil, Richard.
"Modeling and Simulation Analysis of the Physical Sidelink Shared Channel (PSSCH)."
Paper presented at 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, United States. May 20, 2018 - May 24, 2018.



(a) $k_{\text{TRP}} = 2$, $\rho = 3$

(b) $k_{\text{TRP}} = 2$, $\rho = 9$



(c) $k_{\text{TRP}} = 4$, $\rho = 3$

(d) $k_{\text{TRP}} = 4$, $\rho = 9$

Fig. 5. Bar graphs of theoretical PMF values from Eq. (15) with scatter plots of empirical PMF values from Monte Carlo simulations using Matlab and ns3, with 95 % confidence intervals shown for all simulation results.

[2] G. Fodor, S. Parkvall, S. Sorrentino, P. Wallentin, Q. Lu and N. Brahmi, "Device-to-Device Communications for National Security and Public Safety," *IEEE Access*, vol. 2, pp. 1510-1520, 2014.

[3] 3GPP, "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification; TS 36.321," 3rd Generation Partnership Project (3GPP), Tech. Rep., 2015. [Online]. Available: http://www.3gpp.org/DynaReport/36321.htm

[4] 3GPP, "Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures; TS 36.213," 3rd Generation Partnership Project (3GPP), Tech. Rep., 2016. [Online]. Available: http://www.3gpp.org/DynaReport/36213.htm

[5] J. Wang and R. Rouil, "BLER Performance Evaluation of LTE Device-to-Device Communications," NISTIR 8157, National Institute of Standards and Technology, US Department of Commerce, November 2016, Available: https://doi.org/10.6028/NIST.IR.8157.

[6] D. W. Griffith, F. J. Cintrón and R. A. Rouil, "Physical Sidelink Control Channel (PSCCH) in Mode 2: Performance analysis," 2017 IEEE Int. Conf. Communications (ICC), Paris, 2017.

[7] H. Yoon, S. Park and S. Choi, "Efficient feedback mechanism and rate adaptation for LTE-based D2D communication," 2017 IEEE 18th Int. Symp. A World of Wireless, Mobile and Multimedia Networks (WoWMoM), Macau, 2017.

[8] Seung-Hoon Park, Jun Suk Kim, and Min Young Chung, "Resource selection scheme for the transmission of scheduling assignment in Device-to-Device communications," Wireless Personal Communications, Aug 2017.

[9] M.J. Shih, H.H. Liu, W.D. Shen and H.Y. Wei, "UE autonomous resource selection for D2D communications: Explicit vs. implicit approaches," 2016 IEEE Conf. Standards for Communications and Networking (CSCN), Berlin, 2016.

[10] S.H. Sun, J.L. Hu, Y. Peng, X.M. Pan, L. Zhao and J.Y. Fang, "Support for vehicle-to-everything services based on LTE," *IEEE Wireless Commun.*, vol. 23, no. 3, pp. 4-8, June 2016.

[11] L. Takács, "Charles Jordan, 1871-1959," *Ann. Math. Stat.*, vol. 32, no. 1, pp. 1–11, 1961.

[12] A. Préopka, E. Boros, and K. W. Lih, "The Use of Binomial Moments for Bounding Network Reliability," in *Reliability of Computer and Communication Networks: Proceedings of a DIMACS Workshop, December 2-4, 1989*, American Mathematical Society, 1991.

Cintron, Fernando; Griffith, David; Hall, Timothy; Rouil, Richard.
"Modeling and Simulation Analysis of the Physical Sidelink Shared Channel (PSSCH)."
Paper presented at 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, United States. May 20, 2018 - May 24, 2018.

A 3D Topology Optimization Scheme for M2M Communications

Yalong Wu, Wei Yu, Jin Zhang
Department of Computer and Information Sciences
Towson University
Maryland, USA
Emails: ywu11, jzhang13 @students.towson.edu, wyu@towson.edu

David Griffith, Nada Golmie
Wireless Networks Division
National Institute of Standards and Technology
Maryland, USA
Emails: david.griffith, nada.golmie @nist.gov

Chao Lu
Department of Computer and Information Sciences
Towson University
Maryland, USA
Email: clu@towson.edu

***Abstract*—Communication networking leverages emerging network technologies such as topology management schemes to satisfy the demand of exponentially increasing devices and associated network traffic. Particularly, without efficient topology management, Machine-to-Machine (M2M) communications will likely asymmetrically congest gateways and eNodeBs in 3rd Generation Partnership Project (3GPP) Long-Term Evolution (LTE) and Long-Term Evolution Advanced (LTE-A) networks, especially when M2M devices are massively deployed to support diverse applications. To address this issue, in this paper, we propose a 3D Topology Optimization (3D-TO) scheme to obtain the optimal placement of gateways and eNodeBs for M2M communications. By taking advantage of the fact that most M2M devices rarely move, 3D-TO can specify optimal gateway positions for each M2M application, which consists of multiple M2M devices. This is achieved through global optimization, based on the distances between gateways and M2M devices. Utilizing the optimization process, 3D-TO likewise determines optimal eNodeB positions for each M2M application, based on the distances between eNodeBs and optimal M2M gateways. Our experimental results demonstrate the effectiveness of our proposed 3D-TO scheme towards M2M communications, with regard to throughput, delay, path loss, and packet loss ratio.**

***Keywords*—M2M communications, 3GPP LTE/LTE-A, Topology optimization, M2M applications**

I. Introduction

Unlike Human-to-Human (H2H) communication that highly relies on human intervention, Machine-to-Machine (M2M) communication, also known as Machine Type Communication (MTC), enforces connectivity between massive devices independently [1], [15]. M2M communication has become the skeleton of Internet-of-Things (IoT) communication and enables a myriad of smart applications in the realms of public safety, smart grid, smart transportation, smart health, smart city, etc. [3], [10], [16]–[18], [20].

Nonetheless, 3rd Generation Partnership Project (3GPP) Long-Term Evolution (LTE) and Long-Term Evolution Advanced (LTE-A) network infrastructures are unprecedentedly challenged with the explosion of M2M devices. Then, the development of intelligent networking techniques to satisfy the demand of exponentially increasing M2M devices and their traffic becomes critical. This calls for developing effective topology optimization to support M2M communications.

In this paper, we propose a 3D topology optimization (3D-TO) scheme for M2M communications in 3GPP LTE/LTE-A networks, which can efficiently improve M2M communication performance with respect to throughput, delay, path loss, and packet loss ratio. 3D topology optimization focuses on the repositioning of M2M gateways and eNodeBs, such that load-balanced network topology can be achieved. In this scheme, within a particular M2M application, 3D-TO first identifies an optimal position for each M2M gateway in the feasible deployment space. The optimization procedure is performed on the basis of minimizing the summation of the distances between each M2M gateway and its associated M2M devices. Utilizing the same optimization process, 3D-TO also specifies optimal eNodeB positions, but instead considers the distances between eNodeBs and their connected M2M devices' associated optimal M2M gateways. Under the minimization of distance between devices, 3D-TO can largely reduce relay times and path loss in data transmission. Through experimental simulation, we validate the effectiveness of our proposed 3D-TO scheme in terms of throughput, delay, path loss, and packet loss ratio.

The remainder of this paper is organized as follows: In Section II, we introduce the system model. In Section III, we introduce our approach in detail. In Section IV, we present experimental results to validate the effectiveness of our approach. In Section V, we review related works. Finally, in Section VI, we conclude the paper.

II. System Model

In 3GPP LTE/LTE-A networks [13], M2M devices are deployed with a MTC server and *Evolved Packet Core* (EPC), which consists of *Mobility Management Entity* (MME), *Serving Gateway* (S-GW), and *Packet Data Network Gateway* (P-GW). Particularly, MME is engaged in the control plane, performing activities such as roaming, handover and security management, and also selects S-GW and P-GW for M2M devices and user equipment (UEs). S-GW operates in the user plane to enable data transmission towards eNodeBs (eNBs) and P-GW, while P-GW establishes secure connections between M2M devices (UEs). All three EPC components are connected to eNodeBs via interface, and eNodeBs

Golmie, Nada; Griffith, David; Wu, Yalong; Yu, Wei; Zhang, Jin.
"A 3D Topology Optimization Scheme for M2M Communications."
Paper presented at 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2018), Busan, Republic of Korea. June 27, 2018 - June 29, 2018.

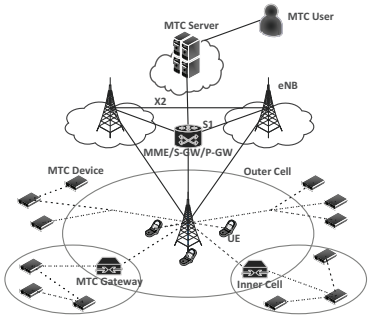


Fig. 1. System Model of M2M Communications in LTE/LTE-A Networks

communicate with each other through interface. The MTC gateway enables a mixture of diverse access methods (wireless LAN, WiMAX, ZigBee, etc.) to EPC, while the MTC server is accessed by the MTC users (smart home, logistic service, remote surveillance, etc.) to exploit diverse Internet of Things applications supported by massive M2M devices through some *Application Programming Interface* (API) provided by the network operator [19].

Within each M2M application, as shown in Fig. 1, we consider large networks organized by eNodeBs as outer cells, and small networks managed by MTC gateways as inner cells. To be specific, massively deployed M2M devices in outer cells can be bonded to eNodeBs either directly via LTE/LTE-A link, or indirectly via MTC gateways, under which the device-to-device communications between M2M devices might be enabled by distinct wireless network protocols other than LTE/LTE-A. MTC gateways in inner cells are able to not only optimally select transmission paths between M2M devices, effectively balancing their energy consumption, but also facilitate a connection to back-hauling. For the simplicity of our topology optimization analysis in 3D-TO, we assume that M2M devices do not move over time, but instead have fixed positions. Notice, however, that our proposed topology optimization mechanism can be extended to mobile device scenarios, as well as mixed mobile/stationary devices scenarios. All notations used in this paper are shown in Table I.

TABLE I
NOTATION

	Feasible and infeasible deployment space.
	Number of M2M applications.
	M2M application.
	Number of eNodeBs in application .
	eNodeB in application .
	Number of devices associated with indirectly.
	Number of gateways connected to .
	gateway connected to .
	Number of devices associated with .
	device associated with .
	-axis of .
	-axis of .
	-axis of .
	Coordinate of stationary point regarding .
	Hessian matrix regarding .
	Coordinate of position closest to .
	-axis of optimal .
	-axis of optimal .
	-axis of optimal .
	-axis of optimal .
	-axis of optimal .
	-axis of optimal .
	device associated with .
	-axis of .
	-axis of .
	-axis of .
	Coordinate of stationary point regarding .
	Hessian matrix regarding .
	Coordinate of position closest to .

III. OUR APPROACH

In 3D-TO, M2M device positions are considered, such that a more efficient network topology can be obtained. Within a particular M2M application, optimal gateway positions can be identified in the feasible deployment space via a minimization of the distances between gateways and their connected M2M devices. Consequently, optimal feasible eNodeB positions for the same M2M application are specified by applying the same minimization mechanism to the distances between eNodeBs and their connected M2M devices optimal gateways.

3D topology optimization is used to specify optimal gateway and eNodeB positions within the feasible deployment space, denoted as , for each M2M application. Then, the infeasible deployment space (lakes, swamps, volcanoes, etc.) can be avoided. With a given space ( ), assume that the number of M2M applications is . In application , where , denote the number of eNodeBs as . As to each eNodeB in , where , refer to the number of M2M devices connected to it not via gateway as , the number of gateways associated to it as , and the number of M2M devices correlated to gateway , where , as .

Based on those deployed devices, our *topology optimization* mechanism comprises the following two steps: (i) *Step 1. Optimal gateway position*, and (ii) *Step 2. Optimal eNodeB position*.

Step 1. Optimal Gateway Position: In this stage, 3D-TO identifies optimal gateway positions in for each M2M application, such that the distance between each gateway and its associated M2M devices is minimized, resulting in

Golmie, Nada; Griffith, David; Wu, Yalong; Yu, Wei; Zhang, Jin.
"A 3D Topology Optimization Scheme for M2M Communications."
Paper presented at 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2018), Busan, Republic of Korea. June 27, 2018 - June 29, 2018.

reduced relay times and path loss. Intuitively, if the connectivity of devices can be guaranteed over long distances between devices, it can more likely be assured over short distances resulting from our optimal gateway positioning. Assume that the coordinate of M2M device , which is connected to gateway towards eNodeB in M2M application through one or multiple hops, as . If the summation of the squares of distances between devices is minimized, the summation of distances between devices will likely be minimized as well. For the simplicity of our analysis, we use the square functions of the distances between gateway and its associated M2M devices:

(1)

In order to identify the optimal position that minimizes the distance between devices, the stationary points of Equation (1) must be assessed, because extreme values can only occur at stationary points. In this regard, we obtain the first derivatives of the distance function from Equation (1) as follows:

(2)

By leveraging Equation (2), there exists only one stationary point. As the stationary points of the device distance function cause its first derivatives to be zeros, if we assume that the coordinates of the only stationary point of Equation (1) are , then we can have the coordinate of -axis as , it is similar for the coordinate of -axis to be derived as , along with the coordinate of -axis as .

Notice that is also the center-of-gravity of all M2M devices associated with gateway . As the only stationary point makes its corresponding extreme value either smallest or largest, the next step is to further identify whether it is the one that makes the distance between a gateway and its associated M2M devices smallest. To this end, we derive the second derivatives of Equation (1), which form a *Hessian Matrix* [14] :

It is observed that H is a positive definite matrix, meaning that is the minimum point of the square distance function, represented by Equation (1), which then is a strictly convex function [2]. Thus, if falls into the feasible deployment space, it will be the optimal gateway position. Otherwise, the optimal gateway position has to be the one, denoted as , which is closest to , in the feasible deployment space, according to the concavity of Equation (1). Recall that is the center-of-gravity of M2M devices, which means gateways are placed in or closest to the center. This implies that our optimal gateway position also considers the density of M2M devices by orientating gateways near to the area with high M2M device density, and father from the area with low M2M density. Assume that the coordinate of the optimal gateway position that we are looking for as , then it will be , if . If , the coordinate will be .

Regarding the specification of , the procedure is exactly the same as that demonstrated for identifying , with the exception that the stationary point in becomes the point of interest. We need simply to move in over to its closest spot in , and the concavity of Equation ensures that is the minimum point of Equation (1) in . After finalizing the optimal gateway position for , 3D-TO then determines the optimal eNodeB position for based on their associated M2M devices and the generated optimal gateway positions.

Step 2. Optimal eNodeB Position: In this stage, the optimal eNodeB position, denoted as , for each is confirmed, and the device connectivity is also guaranteed. As to each M2M application , 3D-TO determines the optimal eNodeB positions via a minimization of the distances between eNodeBs and their associated M2M devices optimal gateways generated in *Step 1*.

Assume that the number of M2M devices, which are connected to eNodeB in application through one or multiple hops as . If we represent the coordinate of M2M device , as . Then, we can also have the square function of distance between and ,

Golmie, Nada; Griffith, David; Wu, Yalong; Yu, Wei; Zhang, Jin.
"A 3D Topology Optimization Scheme for M2M Communications."
Paper presented at 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2018), Busan, Republic of Korea. June 27, 2018 - June 29, 2018.

as:

(3)

With the illustration of *Step 1*, we also enumerate below the first derivatives of Equation (3) to identify the stationary points. There is one which can minimize the distance between and , .

(4)

Assume that the coordinates of the only stationary point of Equation (3) are . By setting each derivative in Equation (4) to be zero, we can have the coordinate of x-axis as , and the coordinate of y-axis as , along with the coordinate of z-axis as .

Notice that is the center-of-gravity of all the M2M devices and optimal gateways, which are associated with . Next, we find the second derivatives of Equation (3) to be the following *Hessian Matrix* .

is also a positive definite matrix, which implies to be a point minimizing strictly convex distance function, represented by Equation (3), similar to that illustrated in *Step 1*. Thus, the optimal eNodeB position will be either , or the closest point to it that lies in the feasible deployment space, denoted as , if is determined to lie in the infeasible deployment space. As the center-of-gravity (or nearest point to the center-of-gravity) of M2M devices and optimal gateways, or also takes the density of M2M devices and optimal gateways into account. The coordinate of the optimal eNodeB position that we are looking for will be , if . If , the coordinates will be .

If relocating the optimal eNodeB is needed, the affirmation of will follow the same process in *Step 1*. Thus, the optimized topology for M2M communications in 3GPP LTE/LTE-A networks is accomplished.

Recall that 3D-TO finds optimal gateway and eNodeB positions, based on a procedure of minimization over distances between gateways, eNodeBs and their associated M2M devices, with the consideration of restriction from the infeasible deployment space. This also leads the optimal positions of gateways and eNodeBs to be the centers of gravity of their associated M2M devices, which validates the device density attention property of 3D-TO.

IV. Performance Evaluation

In our performance evaluation, we first implement 3D-TO in MATLAB to numerically demonstrate its effectiveness, and then deploy certain numeric results from MATLAB into NS-3 to further assess the performance of 3D-TO in a real network simulation environment[1]. In our evaluation, the comparison baseline against 3D-TO is the normal case without any topology optimization towards M2M gateways and eNodeBs.

First, we conduct our numeric demonstration over the average and variance of distances between devices of each M2M application in MATLAB, with the consideration of a number of coexisting M2M applications enabled by massive M2M devices. Thus, we set the number of M2M applications to 2000. Within each M2M application, there are 2 eNodeBs, 6 M2M gateways, and 1000 M2M devices. The initialized three-dimensional positions of all devices are generated randomly,

[1]Certain commercial equipment, instruments, or materials are identified in this chapter in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Golmie, Nada; Griffith, David; Wu, Yalong; Yu, Wei; Zhang, Jin.
"A 3D Topology Optimization Scheme for M2M Communications."
Paper presented at 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2018), Busan, Republic of Korea. June 27, 2018 - June 29, 2018.

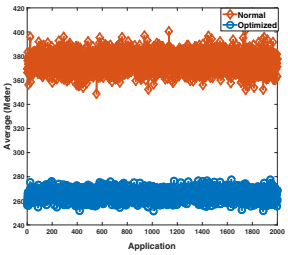


Fig. 2. Distance Average

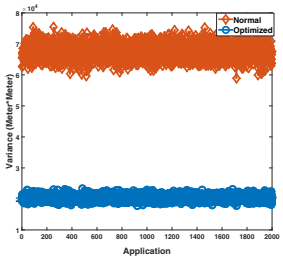


Fig. 3. Distance Variance

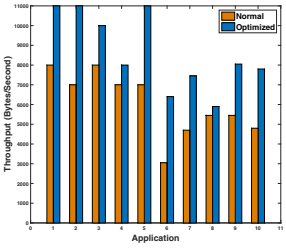


Fig. 4. Throughput

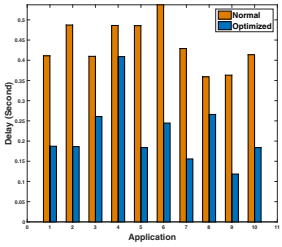


Fig. 5. Delay

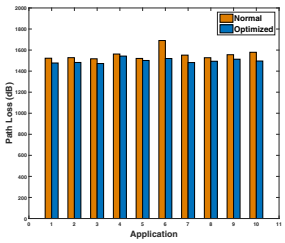


Fig. 6. Path Loss

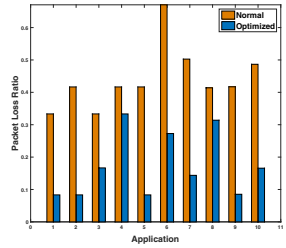


Fig. 7. Packet Loss Ratio

and one single eNodeB is directly connected to 250 M2M devices and 3 M2M gateways, which are directly or indirectly associated with another 250 M2M devices via one or multiple hops. In addition, we assume that M2M devices separated by distance less than the transmission range are also paired up via either LTE/LTE-A link or other kinds of wireless links (WLAN, WiMAX, ZigBee, etc.). The network connectivity is guaranteed by the cooperation among the transmission ranges of eNodeB, M2M gateway, and M2M device, which are 1000 m, 500 m, and 50 m, respectively.

Second, for simulations in NS-3, we shrink the network size due to the capacity limitation of our computing hardware. Thus in NS-3, we implement 10 M2M applications, consisting of 1 macro eNB, 2 home eNBs (working as M2M gateways), and 12 UEs (serving as M2M devices with fixed positions) in each application. The macro eNB is directly connected to 6 UEs and 2 home eNBs, and each home eNB is associated with 3 UEs. All device positions in NS-3 are determined according to the results generated from MATLAB by applying 3D-TO. For simplicity, communications between devices are all enabled under LTE standard, with UDP as the transmission protocol.

3GPP has standardized the arrival traffic of M2M communications as a *Beta* distribution with a small data transmission feature [3], [4]. Thus, we set up a *Beta* distribution for each M2M application and randomly generate integers from 1 to 4 as , , and the function range (in seconds). The effectiveness of 3D-TO is assessed based on the following metrics: (i) *Throughput* is computed based on the entire data transmissions of all M2M devices in each M2M application, (ii) *Delay* is computed based on the average time of M2M devices to finish transmitting data in each M2M application, (iii) *Path Loss* is examined based on the overall power attenuation of all M2M devices in each M2M application, and (iv) *Packet Loss Ratio* is defined as the ratio of lost data packets over the total transmission data packets associated with all M2M devices in each M2M application. For generality, we repeat the experiment 10 times, taking the average of all 10 iterations as the data in the figures.

Fig. 2 and Fig. 3 show the performance comparison of 3D-TO and the baseline normal case (i.e., without topology optimization), with respect to distance average and variance. As we can see from the figures, the average and variance of distances between devices of each M2M application running 3D-TO is much lower than those of M2M applications without 3D-TO. For instance, 3D-TO reduces the average distances between devices of all M2M applications to below 280 m, seen in Fig. 2, while the average distances between devices of almost all M2M applications in the normal case are above 400 m. In Fig. 3, the variances of distances between devices of all M2M applications with 3D-TO are lower than 24 000 , but reach around 70 000 for most baseline M2M applications without 3D-TO.

Fig. 4 illustrates the average throughput of each M2M application in the normal and optimized cases, respectively. As we can see from the figure, our proposed 3D-TO scheme performs better than the normal case without topology optimization in all 10 M2M applications. This implies that 3D-TO can significantly improve LTE network performance in terms of throughput. For example, certain M2M applications, running 3D-TO in Fig. 4, achieve a throughput as high as 11,000 bytes/second, and even the application with the worst performance has a throughput above 5,700 bytes/second. In contrast, only 2 M2M applications without topology optimization reach a throughput at 8,000 bytes/second, while most maintain a throughput under 5,600 bytes/second.

Fig. 5 highlights the comparison of 3D-TO and the normal case with respect to delay. As we can see from the figure, the average delay performance for each M2M application with 3D-

Golmie, Nada; Griffith, David; Wu, Yalong; Yu, Wei; Zhang, Jin.
"A 3D Topology Optimization Scheme for M2M Communications."
Paper presented at 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2018), Busan, Republic of Korea. June 27, 2018 - June 29, 2018.

TO is much better than that of the M2M applications without 3D-TO. For example, 3D-TO maintains a delay under 0.27 s for almost all M2M applications in Fig. 5, and some even reach as low as 0.12 s. Nonetheless, in the normal case, without topology optimization, the delay of all M2M applications is above 0.35 s.

Fig. 6 shows an evident decrease for each M2M application with respect to path loss by applying 3D-TO. Fig. 7 illustrates the packet loss ratio of M2M applications in both the normal case without topology optimization and optimized case with 3D-TO. As we can see from the figure, our proposed 3D-TO scheme outperforms the normal case in each M2M application. For example, all M2M applications running the normal case in Fig. 4, have a packet loss ratio above 0.3, even reaching as high as 0.68. In comparison, the packet loss ratio of almost all applications in the optimized case with 3D-TO is below 0.3, with some M2M applications with packet loss ratios below 0.1.

V. Related Works

In order to improve M2M communications in 3GPP LTE/LTE-A networks, a number of research efforts have been devoted to M2M topology optimization [5]–[9], [11], [12].

Topology optimization has been generally adopted in M2M communication networks to obtain interference reduction, energy economy, and the extension of operating lifetimes. Primarily, topology optimization consists of topology construction and topology maintenance, which are responsible for initialization optimization and connectivity preservation, respectively [5], [7]–[9], [11]. For instance, Lee *et al.* in [7] proposed a distributed energy-efficient topology control algorithm to establish a best-parent based new topology in the construction phase, and a signal topology reconstruction by monitoring energy status of neighbors in the maintenance phase. The algorithm delivered significant energy efficiency and prolonged lifetime. Li *et al.* in [8] designed a network flow theory-based topology adaption algorithm with low time complexity. The authors analyzed the heterogeneity property of M2M networks and identified an optimal solution for energy efficient topology control.

Unlike the existing schemes highlighted, our proposed 3D-TO (3D Topology Optimization) scheme considers not only the identification of optimal device positions, but also the avoidance of obstruction areas. Our work consists of a thorough theoretical modeling of topology construction and maintenance based on distance minimization. We have also conducted experiments in both Matlab and NS-3 to demonstrate the performance of our proposed approach.

VI. Conclusion

In this paper, we proposed a 3D topology optimization (3D-TO) scheme that can optimally construct network topology for M2M applications in 3GPP LTE/LTE-A networks. Particularly, 3D-TO applies the first derivative to extract the stationary point of the distance between devices, and then utilizes the second derivative to identify it as the one that minimizes the distance. Our proposed scheme is capable of leveraging M2M device positions to obtain optimal M2M gateway and eNodeB placement. The results of our extensive experimentation validate that 3D-TO obtains better network performance in throughput, delay, path loss, and packet loss ratio than the baseline configuration.

References

[1] D. Boswarthick, O. Elloumi, and O. Hersent. *M2M communications: a systems approach*. John Wiley & Sons, 2012.

[2] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[3] F. Ghavimi and H.-H. Chen. M2M communications in 3GPP LTE/LTE-A networks: architectures, service requirements, challenges, and applications. *IEEE Communications Surveys & Tutorials*, 17(2):525–549, 2015.

[4] X. Jian, X. Zeng, J. Huang, Y. Jia, and Y. Zhou. Statistical description and analysis of the concurrent data transmission from massive mtc devices. *International Journal of Smart Home*, 8(4):139–150, 2014.

[5] E.-J. Kim, S.-P. Heo, H.-J. Chong, and H.-W. Jung. Integrated hybrid MAC and topology control scheme for M2M area networks. In *Network Operations and Management Symposium (APNOMS), 2012 14th Asia-Pacific*, pages 1–5. IEEE, 2012.

[6] J. Kim, J. Lee, J. Kim, and J. Yun. M2M service platforms: Survey, issues, and enabling technologies. *IEEE Communications Surveys and Tutorials*, 16(1):61–76, 2014.

[7] C.-Y. Lee and C.-S. Yang. Distributed energy-efficient topology control algorithm in home M2M networks. *International Journal of Distributed Sensor Networks*, 2012, 2012.

[8] X. Li, J. Cai, and H. Zhang. Topology control for heterogeneous connectivity requirements to sink in M2M networks. In *Communications and Networking in China (CHINACOM), 2013 8th International ICST Conference on*, pages 575–580. IEEE, 2013.

[9] S.-Y. Lien, K.-C. Chen, and Y. Lin. Toward ubiquitous massive accesses in 3GPP machine-to-machine communications. *IEEE Communications Magazine*, 49(4):66–74, 2011.

[10] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao. A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications. *IEEE Internet of Things Journal*, 4(5):1125–1142, Oct 2017.

[11] A. Paul. Graph based M2M optimization in an IoT environment. In *Proceedings of the 2013 Research in Adaptive and Convergent Systems*, pages 45–46. ACM, 2013.

[12] J. Swetina, G. Lu, P. Jacobs, F. Ennesser, and J. Song. Toward a standardized common M2M service layer platform: Introduction to oneM2M. *IEEE Wireless Communications*, 21(3):20–26, 2014.

[13] T. Taleb and A. Kunz. Machine type communications in 3gpp networks: potential, challenges, and solutions. *IEEE Communications Magazine*, 50(3):178–184, 2012.

[14] W. C. Thacker. The role of the hessian matrix in fitting models to measurements. *Journal of Geophysical Research: Oceans*, 94(C5):6177–6196, 1989.

[15] Y. Wu, W. Yu, D. Griffins, and N. Golmie. A dynamic rate adaptation scheme for M2M communications. In *Proc. of IEEE International Conference on Communication (ICC)*, 2018.

[16] Y. Wu, W. Yu, F. Yuan, J. Zhang, C. Lu, J. Nguyen, and D. Ku. A cross-domain optimization scheme for manet communications in heterogeneous networks. *EAI Endorsed Transactions on Wireless Spectrum*, 17(12), 12 2017.

[17] G. Xu, W. Yu, D. W. Griffith, N. T. Golmie, and P. Moulema. Performance evaluation of integrating distributed energy resources and storage devices in the smart grid. Technical report, 2016.

[18] W. Yu, D. An, D. Griffith, Q. Yang, and G. Xu. On statistical modeling and forecasting of energy usage in smart grid. In *Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems*, pages 12–17. ACM, 2014.

[19] W. Yu, F. Liang, X. He, W. G. Hatcher, C. Lu, J. Lin, and X. Yang. A survey on the edge computing for the Internet of Things. *IEEE Access*, 6:6900–6919, 2018.

[20] W. Yu, H. Xu, H. Zhang, D. Griffith, and N. Golmie. Ultra-Dense Networks: Survey of state of the art and future directions. In *Proc. of IEEE 25th International Conference on Computer Communication and Networks (ICCCN)*, 2016.

Golmie, Nada; Griffith, David; Wu, Yalong; Yu, Wei; Zhang, Jin.
"A 3D Topology Optimization Scheme for M2M Communications."
Paper presented at 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2018), Busan, Republic of Korea. June 27, 2018 - June 29, 2018.

G-band Reflectivity Results of a Conical Blackbody for Radiometer Calibration

Derek A. Houtz[1], Dazhen Gu[1]

[1]National Institute of Standards and Technology, Boulder, CO.

Abstract — **Two hollow conical cavities have been developed and built for the National Institute of Standards and Technology (NIST) for use as radiometer calibration sources, or blackbodies. We seek high emissivity, thus low reflectivity, to approximate an ideal blackbody. We present new results on the reflectivity of the smaller conical blackbody in G-band between 130 GHz and 230 GHz. We found monostatic reflectivity, or return loss, no larger than -45 dB at critical remote sensing bands near 165 GHz, 183 GHz, and 229 GHz. We found that use of a thin closed-cell polyethylene insulation layer has a significant impact on reflectivity performance. We compared the reflectivity of the conical blackbody with the reflectivity of a pyramidal absorber array of the type typically used as on-board radiometer calibration sources. The insulated conical blackbody showed an average of 15 dB lower reflectivity than the pyramidal array over the measured band.**

Index Terms — **Free-space reflectivity, materials characterization, millimeter-wave radiometry, radiometer calibration.**

I. Introduction

Microwave and millimeter-wave radiometers measure passively-emitted Planck and spectral radiation, and have principal applications in weather forecasting and environmental remote sensing. Operational weather satellites use radiometers to collect data on tropospheric temperature, sea-surface temperature, cloud moisture and precipitation, sea ice, ocean salinity and soil moisture, to name a few. Because of the lack of long-term stability of active components in microwave radiometers, continuous calibration is a necessity. Satellite-borne, aircraft-borne, and ground based or in-situ sensors commonly have some form of internal calibration source. Whether this be via an internal blackbody or a noise diode, gain and offset drifts in the radiometer's detection hardware must be quantified and accounted for.

Use of unreliable or non-traceable internal and pre-launch blackbody calibration sources across radiometer platforms has created a lack of consistency and has led to offset biases between instruments [1]. At the National Institute of Standards and Technology (NIST), we are developing reliable and accurate microwave brightness temperature sources to act as a traceability standard for microwave radiometers.

We have designed two conical blackbodies to interface with the Advanced Technology Microwave Sounder (ATMS) [2] instrument's pre-launch calibration hardware. The ATMS is an integral part of the Joint Polar Satellite System (JPSS). The ATMS is currently flying on the Suomi-NPP satellite, a copy will fly aboard the JPSS-1 satellite in late 2017, and there are plans for ATMS copies aboard JPSS-2 and JPSS-3.

The two most important characteristics of a radiometer calibration source, or blackbody, are high emissivity and uniform temperature. According to Kirchoff's law of reciprocity, high emissivity can be demonstrated by showing low reflectivity and thus high absorptivity. In this paper, we measure reflectivity in G-band (130 GHz to 230 GHz) to demonstrate the blackbody performance of the small NIST conical blackbody. We also measure a microwave absorber-coated pyramidal array, the type of blackbody typically used as onboard calibration sources for airborne and spaceborne radiometers.

The two conical blackbodies have radii of 6.8 cm and 10.8 cm respectively and have been designed to operate at frequencies between 18 GHz and 230 GHz. The larger of the two devices was characterized and discussed in [3] but, until now, no measurements have been made on the smaller target or above 110 GHz. Both the small and large targets have the same nominal absorber-layering structure consisting of 3 layers of carbonyl-iron-powder (CIP) impregnated epoxy. In the direction from the copper base structure to the air, the layering structure consists of: 1.3 mm pure epoxy, 2.2 mm 50% CIP by volume, 1.7 mm 5% CIP by volume, and a 3 mm layer of HD-80 closed-cell polyethylene foam. The structure has a cone half-angle of 10°. The pyramidal array we measured has a pyramid height-to-base aspect ratio of 3 to 1, with a base length of 1 cm and a 1 mm coating of microwave absorber.

The organization structure of this paper is as follows: in Section II, we discuss the reflectivity measurement technique and setup, and present the raw measurement data. Section III presents the processed results of the reflectivity measurements in G-band. Section IV discusses the results and draws conclusions.

II. Reflectivity Measurement

Measuring the true emissivity of a body requires knowledge of the full bi-directional scattering function, sometimes referred to as the bi-directional reflectance distribution function (BRDF), as discussed in [3]. This is an extremely difficult measurement to make experimentally, and the monostatic reflectance has often been considered sufficient to approximate emissivity at normal incidence. We have also provided simulations to justify this approximation at low reflectance

Gu, Dazhen; Houtz, Derek.
"G-band Reflectivity Results of Conical Blackbody for Radiometer Calibration."
Paper presented at ARFTG 90th Microwave Measurement Conference, Boulder, CO, United States. November 28, 2017 - December 1, 2017.

magnitudes [3]. We measure the monostatic reflectance of the blackbody following the technique developed in [4].

We use a network analyzer with a WR-05 (140 GHz to 220 GHz) frequency-extender head. In this investigation, we expanded the operation down to 130 GHz and up to 230 GHz without introducing spurious modes. Measurements are made at 0.1 GHz steps across the 100 GHz frequency range. The entire measurement was conducted within a small anechoic chamber to attenuate background noise and eliminate reflections from the surrounding laboratory environment. The waveguide flange of the extender head was calibrated with the Short-Offset-Load (SOL) 1-port calibration technique. After calibration of the network analyzer, a pyramidal standard-gain horn was attached to the waveguide flange. The pyramidal horn was aligned to a flat and polished aluminum plate affixed to a linear translation stage. The linear translation stage has one-dimensional repeatability of less than 2 μm, or λ/650 at 230 GHz.

The aluminum plate is stepped across about 5.2 mm, in the direction of propagation, at a step size of 0.0408 mm (λ/32 at 230 GHz) for 128 steps. The IF bandwidth of the network analyzer is set to 10 Hz for maximum sensitivity, though this results in slow measurements, allowing time for the network analyzer to drift. We measure the complex S_{11} one-port scattering parameter at each step, and the standing wave pattern of the aluminum plate is traced out in space. The aluminum plate effectively acts as a short and a multiple offset short as described in the free-space calibration technique of [4]. Next, the conical blackbody is aligned and measured with the same horn and we obtain S_{11} at the same distance steps. The conical blackbody is measured both with and without the HD-80 closed-cell polyethylene layer that was included in the original design to act as an insulator and thermal radiation block. This stepping procedure is then repeated for the pyramidal array. A photograph of the measurement setup with the conical blackbody is shown in Figure 1. Figure 2 shows the setup with the pyramidal array.

Figure 3 shows the measured standing wave pattern for the aluminum plate at 165.5 GHz and 183.3 GHz, two important remote sensing frequencies, one being a window channel and the other a water vapor absorption band. Figure 4 and Figure 5 show the results for the insulated and non-insulated blackbody at the same two frequencies respectively. Due to the relatively high magnitude reflections from the plate, we see a multiple reflection interference pattern at 183.3 GHz. This likely causes a slight overestimation of the blackbody reflectivity because the plate is effectively used as a normalizing scalar in the two-tier calibration. At 183.3 GHz without the polyethylene insulation, the blackbody standing wave is not observed. The standing wave, in theory, should have the same wavelength as seen in the polyethylene insulated case, corresponding to half the excitation wavelength (~0.82 mm in this case). Instead, we see a lower frequency signal likely caused by near-field effects and network analyzer drift.

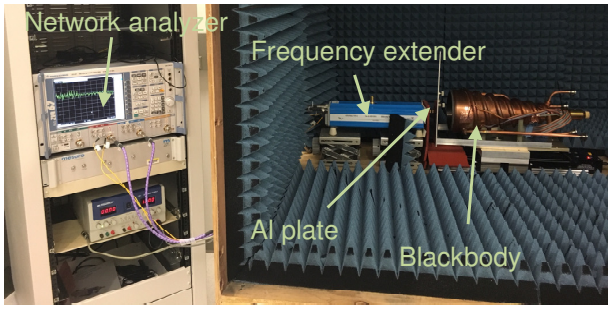


Fig. 1. Photograph of the measurement setup, where components are labeled.

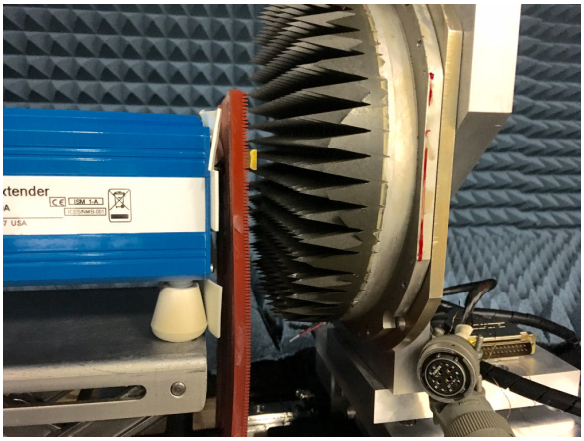


Fig. 2. Photograph of the measurement setup for the pyramidal array.

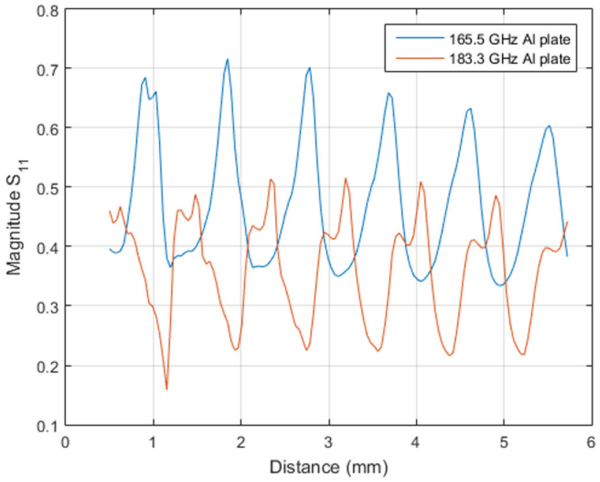


Fig. 3. Plot of S_{11} reflections from aluminum plate versus distance at 165.5 GHz and 183.3 GHz.

Gu, Dazhen; Houtz, Derek.
"G-band Reflectivity Results of Conical Blackbody for Radiometer Calibration."
Paper presented at ARFTG 90th Microwave Measurement Conference, Boulder, CO, United States. November 28, 2017 - December 1, 2017.

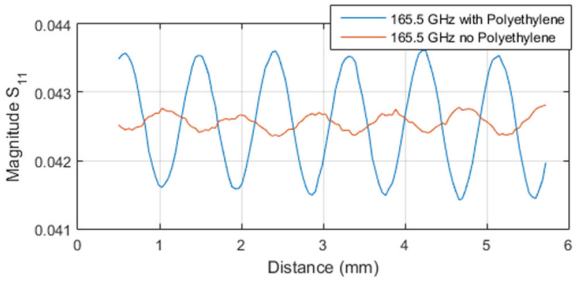


Fig. 4. Plot of S_{11} of the conical blackbody versus distance for the polyethylene insulated and non-insulated cases at 165.5 GHz.

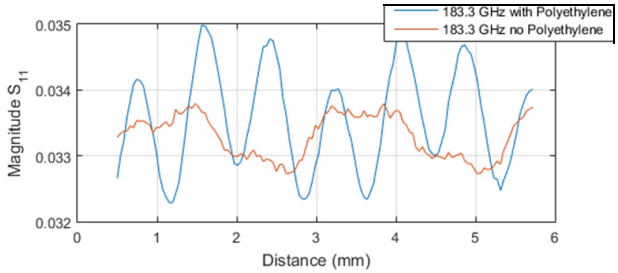


Fig. 5. Plot of S_{11} of the conical blackbody versus distance for the polyethylene insulated and non-insulated cases at 183.3 GHz.

III. Results and Discussion

The data are processed according to the linear fitting procedure outlined in [3] and [4]. This procedure can be thought of as a calibration of single-mode plane-wave scattering matrix in free space. The flat plate acts as a short, or ideal reflector, and is effectively used to normalize the magnitude of the standing wave between the source and the blackbody. We also measured the stationary target in the chamber over the same time span and under similar measurement conditions as the distance-stepped target, and processed these data as a nominal noise-floor estimate. Figure 6 shows the processed reflectivity data and uncertainty for the conical blackbody, pyramidal array, and the noise floor estimate.

In Figure 6 we see much poorer performance, or higher reflectivity, for the insulated blackbody compared to the non-insulated case. This contrasts with our design model which suggested equivalent or slightly better performance from the insulated blackbody. The design model, discussed in [3], assumed non-dispersive dielectric properties of the polyethylene foam, which could have resulted in the design model underestimating the reflectivity for the frequency range considered here. The layer of polyethylene is also not perfectly formed near the apex of the cone and does not form a precise point at the cone apex. There may be some specular reflecting

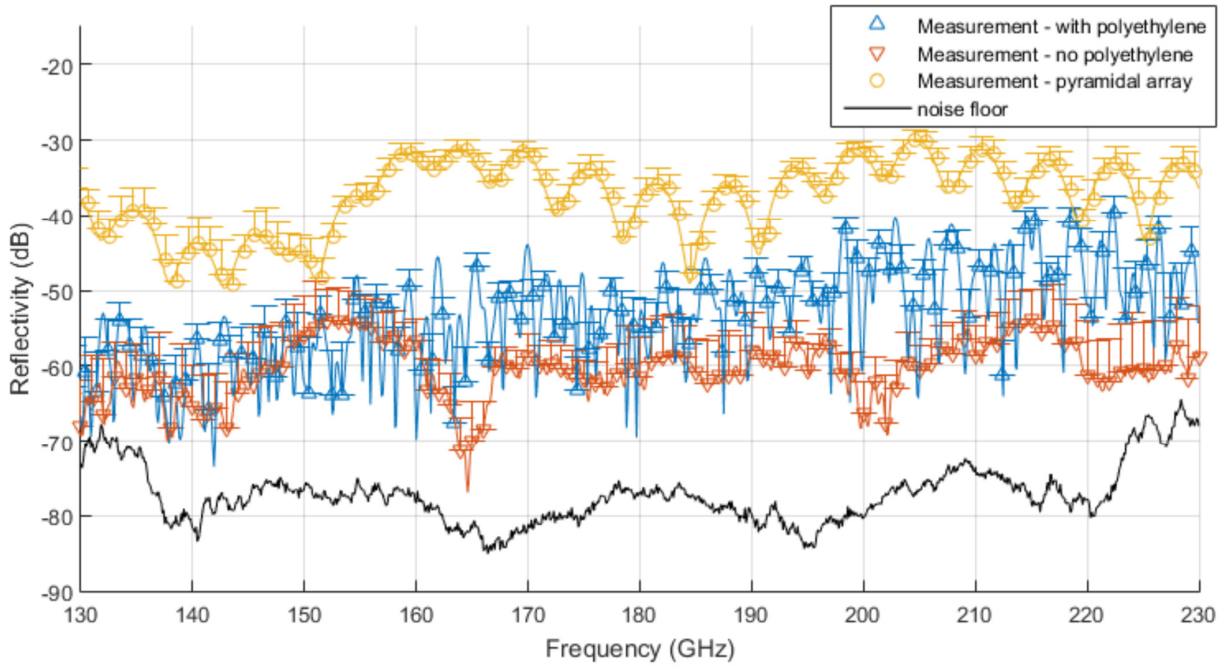


Fig. 6. Monostatic reflectivity results in G-band. The blue line shows the measured result with the polyethylene insulation. The red line shows the measured result with no polyethylene insulation. The yellow line shows the measured result for the pyramidal array blackbody. The black line shows the estimated measurement noise floor from the stationary measurement. The plotted errorbars have a magnitude of one standard error in the positive direction, the lower errorbars have been omitted as they mostly reach to 0 which is undefined in a log scale. Data lines are plotted at 0.1 GHz steps but for clarity errorbars and symbols are shown at 1 GHz intervals.

Gu, Dazhen; Houtz, Derek.
"G-band Reflectivity Results of Conical Blackbody for Radiometer Calibration."
Paper presented at ARFTG 90th Microwave Measurement Conference, Boulder, CO, United States. November 28, 2017 - December 1, 2017.

surface at the seam of the polyethylene sheet causing a relatively strong return signal not predicted in the idealized design model. Though the insulated target has poorer performance than the non-insulated case, the reflectivity magnitudes are all still well within the requirements for accurate radiometer calibration.

In Figure 6, we see that the reflectivity of the pyramidal array is considerably higher than for either of the conical blackbody configurations. This particular pyramidal array was designed to provide high-emissivity in channels at 150 GHz and around 183 GHz where we see dips in the reflectivity below -40 dB. Table 1 provides a summary of the results plotted in Figure 6.

Table 1. Summary of reflectivity results

	With polyethylene	No polyethylene	Pyramidal array
Maximum Reflectivity	-38.9 dB	-53.8 dB	-29.9 dB
Frequency of maximum	218.7 GHz	215 GHz	204.6 GHz
Mean G-band Reflectivity (130 GHz – 230 GHz)	-53.5 dB	-60.4 dB	-37.5 dB

The polyethylene layer reduces physical temperature gradients on the absorber surface by reducing radiative heat transfer. Uniform physical temperature is the other most crucial requirement for a high-performance blackbody along with high emissivity. The use of a thin closed-cell polyethylene foam insulation layer directly on the surface of the absorber was a novel approach to reducing surface temperature gradients, and we have demonstrated low reflectivity achievable with this approach.

We have also demonstrated the marginal and highly frequency-dependent performance of a traditional pyramidal array blackbody. This supports our claim that differences between the on-board calibration sources of various radiometer instruments can vary the resulting brightness temperature calibrations significantly. Introducing traceability back to a consistent standard would increase the long-term consistency and accuracy of radiometric remote sensing data.

IV. Conclusion

We have demonstrated low monostatic reflectivity in G-band for the smaller of the two NIST conical blackbodies. This band contains a number of critical remote sensing channels near 165 GHz, 183 GHz, and 229 GHz. Low monostatic reflectivity is critical to achieving high emissivity and allows us to relate physical temperature to microwave brightness temperature. In order to use the NIST conical blackbodies as a traceable standard for microwave brightness temperature, we must demonstrate performance and uncertainty equal to or better than that of remote-sensing instruments we intend to transfer this standard to. We have demonstrated the high emissivity achievable from the conical geometry and directly compared it to a traditional pyramidal array-type calibration source. The conical geometry also minimizes physical temperature gradients as compared to pyramidal array geometry. Previously, we had shown reflectivity results from 18 GHz to 110 GHz and, in this paper, we have more than doubled this frequency range by demonstrating its low-reflectivity performance between 130 GHz and 230 GHz. We intend to measure within the gap from 110 GHz to 130 GHz in the near future as this range contains a set of important channels in the oxygen absorption band near 118 GHz.

Acknowledgement

The authors acknowledge Richard Wylde and Thomas Keating Ltd. for manufacturing the blackbodies, and David Walker, formerly with NIST, for his long-time contributions and advice in the NIST remote-sensing project.

References

[1] C.Z. Zou and W. Wang, “Intersatellite calibration of AMSU-A observations for weather and climate applications,” J. Geophys. Res., vol. 116, no. D23113, 2011.

[2] W. J. Blackwell, L. Chidester, E. J. Kim, R. V. Leslie, C. H. Lyu and T. Mo, "NPP ATMS prelaunch performance assessment and Sensor Data Record validation," *2011 IEEE International Geoscience and Remote Sensing Symposium*, Vancouver, BC, 2011, pp. 32-34.

[3] D. A. Houtz, W. Emery, D. Gu, K. Jacob, A. Murk, D.K. Walker, R. Wylde, "Electromagnetic design and performance of a conical microwave blackbody target for radiometer calibration," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 8, pp. 4586-4596, Aug. 2017.

[4] D. Gu, D. Houtz, J. Randa and D. K. Walker, "Reflectivity Study of Microwave Blackbody Target," in *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 9, pp. 3443-3451, Sept. 2011.

Sensor Placement and Detection Coverage for Spectrum Sharing in the 3.5 GHz Band

Anirudha Sahoo, Mudumbai Ranganathan, Thao T. Nguyen, and Timothy A. Hall
National Institute of Standards and Technology
Gaithersburg, Maryland, U.S.A.
Email: {ans9, mranga, ttn1, tim.hall}@nist.gov

***Abstract*—The Federal Communications Commission rules for operation in the 3.5 GHz band require that an Environmental Sensing Capability (ESC) system detect the presence of a federal incumbent shipborne radar in order to protect it from harmful interference. Thus, ESC operators have to deploy ESC sensors along the coasts to comply with the rules. We formulate the ESC sensor deployment problem as a coverage problem where ESC sensors need to *cover* a predefined geometric area inside which radar may experience harmful interference. Using propagation models and radar parameters, we compute antenna lobe patterns for different beamwidths and detection thresholds of the ESC sensors. These patterns are then used to cover the geometric area such that both outage and excess coverage areas are minimized. We present a greedy algorithm and apply it to Dynamic Protection Areas currently being defined for the coasts of the contiguous United States. We evaluate its performance in terms of some key metrics important to the federal incumbent as well as commercial operators.**

I. Introduction

The Federal Communications Commission (FCC) has published rules [1] to allow the use of frequencies from 3550 MHz to 3700 MHz by commercial operators. However, this band, referred to as the Citizens Broadband Radio Service (CBRS) band, has to be shared by commercial operators with the incumbents. The incumbents have the highest priority, i.e., when an incumbent uses the band, CBRS devices (CBSDs) that cause harmful interference to the incumbent must vacate the spectrum. The CBSDs will be managed by a Spectrum Access System (SAS). Part of the CBRS band from 3550 MHz to 3650 MHz is currently being used by U.S. Navy radars. CBSDs deployed near the coast should not cause harmful interference to these incumbent shipborne radars, i.e., the interference to noise ratio (I/N) at the radar receiver should be below -6 dB [2]. The presence of this incumbent will be detected by an Environmental Sensing Capability (ESC) consisting of a number of strategically placed sensors along the coast, which will then inform the SAS about the presence of the incumbent.

The National Telecommunications and Information Administration (NTIA) is in the process of specifying Dynamic Protection Areas (DPAs). A DPA is a predefined protection area that may be activated to protect a federal incumbent radar or deactivated when the radar is outside the DPA [3]. The entire area of an activated DPA must be protected from aggregate interference from CBSDs. Hence, ESC sensors associated with each DPA are responsible for detecting the radar signals anywhere within the DPA. As per the draft version from the NTIA, there are 15 non-overlapping coastal DPAs covering the West Coast and 26 non-overlapping coastal DPAs covering the East Coast and Gulf Coast. DPAs around major Navy ports are closer to the coastline (depicted in red in Fig. 3) while the rest start approximately 10 km from the coastline (depicted in light blue in Fig. 3).

For each DPA, an ESC operator must decide the sites and operational parameters of sensors so that the incumbent shipborne radar is detected anywhere inside the DPA. In other words, for a given DPA, the set of one or more ESC sensors should provide *coverage* (in terms of detecting the radar) for that DPA. In this paper, we present a generalized approach and corresponding algorithm that determines ESC sensor locations, antenna orientations and detection thresholds, such that coverage to a geometric shape is achieved while the excess coverage area is minimized. Considering DPAs as a use case, we formulate performance metrics for our algorithm and present the results when applied to the entire coast of the contiguous United States (CONUS). To the best of our knowledge, most analyses in the literature have focused on covering the protection area of the incumbent, but there is no analogous study to minimize excess area of coverage outside the protection area, which is important to the commercial operators. Thus, our study addresses both incumbent protection and spectral utilization (by commercial operators) aspects of the CBRS band related to ESC sensor deployment.

II. Related Work

There is a rich literature on coverage of sensor networks. Sensor coverage requires that each location in the monitoring area of interest be covered or sensed by at least one sensor node. The sensor coverage problem can be classified as *area coverage* or *point coverage*. In area coverage, the goal is to cover a particular area of interest [4]–[6]. In point coverage, a set of points needs to be covered [7], [8]. A comprehensive survey of various coverage schemes is presented in [9]. However, our problem is quite different from the traditional sensor coverage problem studied in the aforementioned research. In the traditional sensor network coverage problem, typically each sensor is assumed to have fixed detection sensitivity, with an omnidirectional antenna. Hence, in most of those previous works, researchers assume the coverage area of each sensor is a circle of constant radius. In addition, the solutions typically require a multi-hop sensor network with suitable density in order to achieve certain optimization objectives (e.g., energy efficiency,

Hall, Timothy; Nguyen, Thao; Ranganathan, Mudumbai; Sahoo, Anirudha.
"Sensor Placement and Detection Coverage for Spectrum Sharing in the 3.5 GHz Band."
Paper presented at 29th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Bologna, Italy.
September 9, 2018 - September 12, 2018.

redundancy). Although in this paper we are also looking at an area coverage problem, the requirements and configuration issues are not the same. The ESC sensors, which will be deployed along the coasts, are required to cover an area out in the sea while limiting their coverage over land. Therefore, employing directional antennae pointing at a carefully computed azimuth angle towards sea is desirable to solve our problem. Furthermore, due to the security concerns of localization of the incumbent, a multi-hop sensor network is not applicable to our problem.

Simplifying assumptions were made in the first efforts to specify ESC sensor placement and detection criteria. An NTIA report used channel reciprocity to determine an ESC detection threshold [10] of (-64) dBm received radar peak power in a 1 MHz bandwidth. It proposed uniform ESC sensor spacing of about 50 km based on a geometric argument and the radio-horizon distance. The Wireless Innovation Forum (WINNF) Spectrum Sharing Committee (SSC) requirements [3] reference an ESC detection threshold of (-89) dBm/MHz from the NTIA Technical Memorandum 17-527 at which a coastline sensor must be able to detect shipborne radar. A technique for uniform placement of ESC sensors is presented in [11], using a linear coastline with a parallel line in the water to represent the required radar detection distance. It presents a distance calculation for redundant coverage, where every point between the two lines is covered by at least two sensors, as well as one for non-redundant coverage. In this paper, we derive a method for non-uniform placement and dynamic detection thresholds of ESC sensors.

The authors in [12] present an approach for optimal non-uniform sensor node placement. They use an abstract, piecewise linear representation of the coastline and of the isolation boundary. They use a sequential convex programming algorithm to solve for the minimum number of sensors needed for redundant as well as non-redundant coverage. The differences between this approach and ours are that we use an actual map of the coastline and DPA database, and then apply a greedy algorithm to solve it. Maps of the US coastlines and realistic CBSD deployments are used in [13] and [14] to compute aggregate interference, which are then used to determine non-uniform sensor locations and their detection thresholds. Sensor placement is formulated as a *set cover* problem and solved using a greedy approach for both redundant and non-redundant coverage. However, the above approach assumes omni-directional antennae and focuses on covering discrete points along the contour where radar starts to experience harmful interference while moving towards the coast. In this paper, we take into account directional antennae and attempt to cover the entire area within any given DPA.

We would like to point out that the solutions provided in [11]–[14] predate the concept of DPAs. They are either applicable to covering one large coastal area (e.g., entire east or west coast of the USA) [11], [12], or covering certain points along the boundary of one large protection area [13], [14], rather than a relatively smaller area like a DPA. Since those approaches deal with only one large area, they do not have to deal with any neighboring area. Hence, unlike our scheme, they do not need to consider false alarms due to coverage spilling into a neighboring DPA and, therefore, comparing their performance with that of our scheme is not appropriate.

III. Overview of Approach

A. Problem Formulation

In a CBRS system, ESC sensors will be deployed along the coast to detect the presence of shipborne radars anywhere inside a given DPA. We formulate this detection problem as a coverage problem in which the ESC sensors need to cover a *Required Coverage Area* (RCA), which can be any geometrical shape such as a DPA.

If we assume that the ESC sensors have omni-directional antennae and that the propagation loss from the radar to the ESC sensor is the same for a given distance regardless of the sensor location, then the sensor coverage geometry would be a circle. The circle radius depends on the detection threshold of the ESC sensor. A higher detection threshold leads to a smaller coverage radius and vice-versa.

Given a set of *candidate* ESC sensor site locations along the coast and an RCA, we can find the set of circles centered at any of the sensor sites such that the union of circles covers the entire RCA. However, there are two constraints to be satisfied: (1) the excess coverage area of the union of circles outside of the RCA is minimized and (2) the distance between centers of two consecutive circles has to be more than a specified distance. The first constraint is needed to minimize the occurrence of false alarms, i.e., ESC sensors detecting a radar outside of their associated RCA. The second constraint is needed to address the Operational Security (OPSEC) requirements of the federal incumbent [15].

However, since ESC sensors only need to detect radar out at sea, an omni-directional antenna is not necessary. Furthermore, an omni-directional antenna would incur more interference from CBSDs deployed on land to the ESC sensor. Therefore, in practice, ESC sensors will use sectorized antennae facing towards the sea. Using the Irregular Terrain Model (ITM) propagation model in area mode [16] and radar parameters, we can find a family of antenna coverage patterns at different detection thresholds for a given beamwidth of an ESC sensor antenna (see Fig. 2). Each antenna lobe corresponds to a given detection threshold. A point on a given lobe at a given angle (with respect to its boresight) represents the maximum distance at which the sensor can detect a radar at that detection threshold and angle.

The problem can now be defined as follows: given an RCA, a set of candidate ESC sensor sites, and a family of antenna lobe coverage patterns, find the subset of sites where ESC sensors should be placed, subject to a minimum distance constraint between adjacent ESC sensor sites. For each placement site, find the angles at which one or more member lobes should be used. When the deployment is complete, the entire RCA should be covered in such a way that the excess area, i.e., the difference between union of total area covered by all the lobes and the area of the RCA, is minimized.

Hall, Timothy; Nguyen, Thao; Ranganathan, Mudumbai; Sahoo, Anirudha.
"Sensor Placement and Detection Coverage for Spectrum Sharing in the 3.5 GHz Band."
Paper presented at 29th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Bologna, Italy.
September 9, 2018 - September 12, 2018.

Subroutine 1: $find\_max\_min\_circle$: Find the tightest circle that covers each point $p \in points\_to\_cover$, and return the circle with largest radius among these circles.

Input: $candidate\_sites$: Set of possible centers (candidate sites of ESC sensors).
$points\_to\_cover$: A set of points for which max_min circle is to be found.
Output: Circle (c, r): the largest circle (center c, radius r) among all the tight circles covering each point $p \in points\_to\_cover$.

1 **for** *each point* $p \in points\_to\_cover$ **do**
2 **for** *each* $c \in candidate\_sites$ **do**
3 d[c] := euclidean distance between p and c ;
4 $(center, radius) :=$ $(c, d[c])|\min_{\forall c \in candidate\_sites}(d[c])$ $min\_circle[p] := (center, radius)$;
5 $(c, r) := (min\_circle[p].center, min\_circle[p].radius) \mid \max_{\forall p \in points\_to\_cover}(min\_circle[p].radius)$;
6 return $Circle(c, r)$;

Subroutine 2: $min\_area\_circle\_cover\_greedy$: Greedy algorithm to find minimum area circle cover to a given set of protection points.

Input: $protection\_points$: Set of protection points to be covered.
$candidate\_sites$: Set of possible centers (candidate sites of ESC sensors).
$min\_distance$: Minimum distance permissible between two adjacent circle centers.
Output: $circle\_cover$: A set of circles that completely covers the points in $protection\_points$.

1 $circle\_cover = \emptyset$;
2 **while** $protection\_points$!= $\emptyset$ **do**
3 $Circle(center, radius) := find\_max\_min\_circle$ $(candidate\_sites, protection\_points)$;
4 $covered\_points$:= { p : $p \in protection\_points$ and p is inside $Circle(center, radius)$ } ;
5 $found := False$;
6 **for** *each* $Circle(c, r) \in circle\_cover$ **do**
7 **if** $c == center$ **then**
8 $circle\_cover$:= ($circle\_cover$ - $\{Circle(c, r)\}$) $\cup \{Circle(c, max(radius, r))\}$;
9 $found := True$;
10 **if** $found == False$ **then**
11 $circle\_cover := circle\_cover \cup$ $Circle(center, radius)$;
12 **for** *each center* $c \in (candidate\_sites - center)$ **do**
13 **if** *euclidean_distance*$(c, center) \leq min\_distance$ **then**
14 $candidate\_sites := candidate\_sites - c$;
15 $protection\_points$:= $protection\_points$ - $covered\_points$;
16 return $circle\_cover$

B. Approach

To simplify the problem, we discretize the RCA to a set of grid points, transforming the problem from covering an area to covering a set of points. Minimizing the cost of covering a set of points using a circle of radius r with a cost function $f(r) = r^\alpha$, $\alpha > 1$ (i.e., our circle-cover problem above) is shown to be NP-hard in [17]. Covering a set of points with lobe coverage patterns such that the excess area is minimized can also be shown to be NP-hard.

We use a two-step greedy approach to solve the problem. In the first step, we use a greedy method to cover the set of points with circles centered on a subset of candidate sensor site locations such that the excess area is minimized. In the second step, for each circle, we choose a lobe coverage pattern whose length along its major axis is larger than the radius of the circle by a factor greater than one. This ensures that a finite number of the same lobe coverage pattern can cover the entire circle with no outages. We then find the minimum number of the chosen lobes and corresponding orientation angles that tightly cover the points inside the circle.

C. Greedy Algorithm

For a given set of potential candidate ESC sensor sites and set of points to be covered, Subroutine 1 considers one point at a time and finds the tightest circle (centered at one of the candidate sensor sites) that covers that point (Line 4). It then returns the circle of maximum size (Line 5) out of all circles.

Subroutine 2 finds the set of circles that provides the minimum area cover to a set of points, such that the distance between the centers of any two consecutive circles is more than a predefined distance constraint. The algorithm finds the *max_min* circle for the given set of points using Subroutine 1 (Line 3). The centers that are within the distance constraint of the max_min circle are taken off of the possible center list (Line 14), and the points covered by the max_min circle are taken off the list of points to be covered (Line 15). This process is repeated until all the points are covered.

Our approach uses the minimum area circle cover as an *intermediate step* to provide minimum area coverage with antenna lobe patterns. Subroutine 3 takes a set of points to cover within a circle and an antenna lobe pattern to be used for coverage. The algorithm first finds the angle subtended by the two intersection points between the circle and the lobe at the center of the circle. Based on the subtended angle it then computes the incremental rotation angle to get a different orientation (angle with respect to horizontal direction) of the lobe. It first picks the lobe orientation (angle) that covers the maximum number of points (Line 16). This process is repeated for the rest of the points using the remaining lobe orientations (*while loop* in Line 10).

Algorithm 1 calls Subroutine 2 to get the greedy minimum area circle cover for the points (Line 1). For each circle, from the set of concentric lobes, it chooses the smallest lobe with radius larger than the radius of the circle by an *overlap_factor*. It then calls Subroutine 3 to find the angles (orientation) of the lobes, such that all the points are covered. This process is repeated for all the circles.

We illustrate our solution in Fig. 1. In the figure, an RCA is shown as the polygon ABCD (with some piece-wise linear sides in between). The points marked as "X" are potential ESC sensor locations. Subroutine 2 finds that two circles can cover the RCA (actually a discrete set of grid points in the RCA). For each circle, a lobe *overlap_factor* times the radius is chosen and the orientation of the lobes are computed using Subroutine 3. The larger circle is chosen first by Algorithm 1 and the common points within this circle and the RCA are

Hall, Timothy; Nguyen, Thao; Ranganathan, Mudumbai; Sahoo, Anirudha.
"Sensor Placement and Detection Coverage for Spectrum Sharing in the 3.5 GHz Band."
Paper presented at 29th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Bologna, Italy.
September 9, 2018 - September 12, 2018.

Subroutine 3: *find_antenna_overlay_for_sector* : Given a circle covering all points in *points_to_cover*, find the orientations of the lobe which tightly covers the points.

Input: *points_to_cover* : A set of grid points to be covered by lobe(s), which are within the circle $Circle(center, radius)$.
detection_coverage_lobe : Antenna lobe to be used for coverage of *points_to_cover* (of size $overlap\_factor * radius$ oriented in the horizontal direction).
Output: *angles* : A vector of angles giving the orientations of the lobe that covers the sector.

```
subtended_angle := find_subtended_angle (radius, detection_coverage_lobe) ;
npatterns := (int) (2 * π / subtendeded_angle) ;
delta_angle := 2 * π / npatterns ;
rotated_lobes := ∅ ;
for i = 1 to npatterns do
    angle := i* delta_angles ;
    lobe := rotate_and_translate(detection_coverage_lobe, center, angle) ;
    rotated_lobes := rotated_lobes ∪ (angle, lobe) ;
angles := ∅ ;
while points_to_cover != ∅ do
    max_points := 0 ;
    for each (angle, lobe) ∈ rotated_lobes do
        points_covered = find_cover(angle, lobe) ;
        if length(points_covered) > max_points then
            max_points := length(points_covered) ;
            (max_angle, max_lobe) := (angle, lobe) ;
            max_points_covered := points_covered ;
    angles := angles ∪ max_angle ;
    rotated_lobes := rotated_lobes - (max_angle, max_lobe) ;
    points_to_cover := points_to_cover - max_points_covered ;
return angles
```

Algorithm 1: *min_antenna_cover_greedy* : Find the antenna cover for all the points in *protection_points*.

Input: *protection_points*: Set of protection points to be covered.
candidate_sites: Set of candidate sites as centers.
min_distance: Minimum distance permissible between two adjacent circle centers.
detection_coverage: An array of non-intersecting concentric detection coverage lobes for the antenna. Each lobe has the same aperture angle (or beamwidth) and is oriented in the horizontal direction.
Output: $\{(center, lobe, \{angles\})\}$: A set containing ($center$, $lobe$, $\{angles\}$) for $lobe \in$ *detection_coverage* identifying the location, antenna lobe and azimuth angles of the lobes placed at the $center$ such that all the points in *protection_points* are covered.

```
cover := min_area_circle_cover_greedy(protection_points, candidate_sites,min_distance) ;
antenna_cover := ∅ ;
for each circle C ∈ cover do
    lobe := the smallest lobe ∈ detection_coverage such that lobe.radius ≥ overlap_factor * (C.radius) ;
    points_to_cover := { p | p ∈ protection_points and p is inside circle C } ;
    angles := find_antenna_overlay_for_sector(points_to_cover, C.center, C.radius, lobe) ;
    antenna_cover := antenna_cover ∪ {(C.center, lobe, angles)} ;
return antenna_cover
```

covered by four lobes (pink in color). Then the smaller circle is chosen which is covered by two lobes (blue in color).

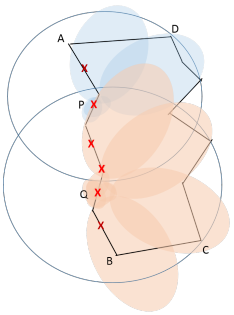


Fig. 1. Example of coverage of an RCA using our approach.

The time complexity of our algorithm is $O(n * m)$, where n is the number of discrete points to be covered in a RCA and m is the number of candidate sensor sites. It can be improved by using spatial data structures for a nearest neighbor search.

Our algorithm terminates when all the protected points are covered by antenna lobes. In the intermediate step, coverage is provided by circles. For every protection point, the Subroutine 1 will always find a circle that covers it. Hence, circle coverage provided by Subroutine 2 terminates. In Subroutine 3, lobe size is chosen to be larger than circle radius (by *overlap_factor*) and the angle of rotation of lobes is chosen such that the rotated lobes overlap with each other. Thus, it is guaranteed that all the protection points inside a circle will be covered by one or more lobes in *rotated_lobes*. Hence, Subroutine 3 also terminates.

In the final stage, we apply *simulated annealing* to our cover (not shown in the algorithm). To keep the search space reasonable, we apply random perturbations only to the orientation of the antenna lobes. The centers and the lobe sizes are obtained from the previous step. Simulated annealing attempts to drive a defined energy function to a minimum by randomly perturbing a given starting solution. A valid solution is one in which the antenna lobes cover the RCA. Area of the cover is used as the energy function, and the minimum energy solution in 1000 trials is picked. Finally, redundant lobes are removed, followed by the removal of sensors that have no lobe.

IV. Analysis Model

In this section, we describe the models and assumptions used in our analysis. Table I lists technical parameters used in this analysis.

The technical parameters for a federal incumbent radar transmitter, referred to as Shipborne Radar 1 in [2], are found in [2], [10]. The ESC sensor technical parameters are found in [14], [18]. The generalized mathematical model for the ESC sensor antenna gain pattern is calculated using the methodology in [19] as follows:

$$G_{ESC}(\theta) = G_{ESC\_peak} - \min\left[12\left(\frac{\theta}{\theta_{3dB}}\right)^2, A_H\right] \quad (1)$$

where $G_{ESC}(\theta)$ is the sensor antenna gain (dBi) at the off-axis angle θ, $-180^\circ \leq \theta \leq 180^\circ$, G_{ESC_peak} is

Hall, Timothy; Nguyen, Thao; Ranganathan, Mudumbai; Sahoo, Anirudha.
"Sensor Placement and Detection Coverage for Spectrum Sharing in the 3.5 GHz Band."
Paper presented at 29th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Bologna, Italy.
September 9, 2018 - September 12, 2018.

TABLE I
TECHNICAL PARAMETERS.

Shipborne Radar-1 Parameter	**Value**
Transmitted Power to Ant. (dBm)	90
Peak Antenna Gain (dBi)	32
Transmit/Receive Bandwidth (MHz)	1
Center Frequency (MHz)	3600
Antenna Height (m)	50
Insertion/Cable Losses (dB)	2
ESC Sensor Parameter	**Value**
Antenna Directivity/Patterns	3GPP
Peak Antenna Gain (dBi)	6.9
3-dB Beamwidth Ant. Gain (deg.)	60, 90, 120
Receive Bandwidth (MHz)	1
Center Frequency (MHz)	3600
Antenna Height (m)	25
Insertion/Cable Losses (dB)	2
ITM Input Parameter	**Value**
Polarization	1 (Vertical)
Dielectric constant	81 (Sea Water)
Conductivity	5 (Sea Water)
Surface Refractivity (N-units)	350 (Maritime, Over Sea)
Radio Climate	7 (Maritime, Over Sea)
Mode of Variability	3 (Broadcast)
Terrain Irregularity (m)	0 (Flat/Smooth Water)
Transmitter Siting Criteria	2 (Very Careful)
Receiver Siting Criteria	0 (Random)
Time/Location/Confidence Var. (%)	50/50/50

the ESC peak antenna gain (dBi), θ_{3dB} is the 3-dB beamwidth of the antenna (degree), and $A_H = 20$ dB is the maximum attenuation.

The path loss from the radar transmitter to the ESC sensor is computed using ITM propagation model [16]. The area prediction model is used to estimate the loss from empirical medians without details of the terrain profile between the radar and ESC sensor.

We define the detection coverage of an ESC sensor as a region within which the radar's peak signal level can be detected by the sensor. The area of the sensor detection coverage depends on the ESC detection threshold. For a given ESC detection threshold D_{th_esc} and a given angle θ that the radar subtends relative to the boresight of the ESC sensor antenna, the propagation loss from the radar transmitter to the ESC sensor is estimated as:

$$L(\theta) = P_{radar} + G_{peak\_radar} - L_{i\_radar} + G_{ESC}(\theta) - L_{i\_ESC} - B_{ESC/radar} - D_{th\_esc} \quad (2)$$

where $L(\theta)$ is the estimated propagation loss at a given angle θ (dB), P_{radar} is the transmit power of the radar (dBm), G_{peak_radar} is the peak antenna gain of the radar (dBi), L_{i_radar} is the radar transmitter insertion loss (dB), $G_{ESC}(\theta)$ is the ESC antenna gain in the direction of the radar (dBi), L_{i_ESC} is the ESC receiver insertion loss (dB), and $B_{ESC/radar}$ is the frequency dependent rejection (dB). The frequency dependent rejection is defined as $B_{ESC/radar} = 10\log_{10}(B_{ESC\_rx}/B_{radar\_tx})$, if $B_{ESC\_rx} < B_{radar\_tx}$; and $B_{ESC/radar} = 0$, otherwise. Note that B_{ESC_rx} and B_{radar_tx} are the bandwidths of the ESC receiver and the radar transmitter, respectively.

Once the propagation loss $L(\theta)$ at each angle θ is computed, the distance d corresponding to the propagation loss $L(\theta)$ is determined from a Propagation Loss vs. Distance graph (similar to Fig. B-1 in [11]) using the ITM area mode with appropriate parameter values. The point is then plotted as a polar point (d, θ). This procedure is repeated for different azimuth angles θ of the antenna, which results in a lobe coverage pattern for the detection threshold D_{th_esc}. A family of such antenna lobe coverage patterns is obtained by varying the detection threshold in the range of [-89, -50] dBm/MHz. Fig. 2 shows a family of lobe coverage patterns for antenna beamwidth of 60°.

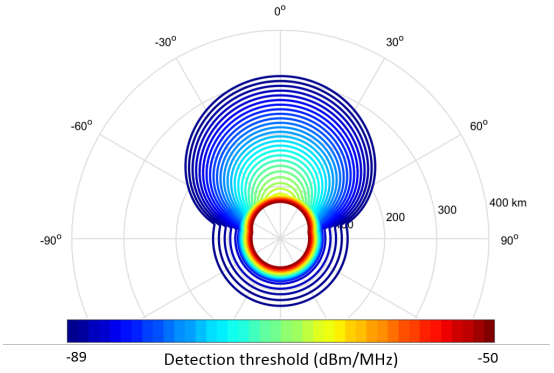


Fig. 2. Antenna coverage lobe patterns for θ_{3dB} of 60^0.

The draft DPA data from NTIA is expressed in latitude/longitude in World Geodetic System (WGS) 84 reference coordinate system. To convert them to northing/easting projected coordinates, we use the Hammer map projection technique to preserve areas. This conversion is required to carry out various geometric operations during DPA coverage analysis. The reference geographic center of the U.S. is chosen as (latitude, longitude) = (37.1669, -95.9669).

V. RESULTS

A. ESC Detection Coverage

We applied our proposed approach to find the detection coverage for all DPAs along the CONUS. Fig. 3 shows the final sensor locations (yellow pushpins) and their detection coverages (orange contours) computed using our method with *overlap_factor* set to 1.2.

For the West Coast, 15 sensors with sensitivities in the range of (-83 to -71) dBm/MHz and 18 antenna lobes were needed to cover 15 DPAs. Whereas, for the East and Gulf Coasts, 32 sensors with sensitivities in the range of (-89 to -75) dBm/MHz and 40 antenna lobes were required to cover 26 DPAs. These antenna lobes have beamwidths of 60° and 90°. For most DPAs, only a single sensor equipped with a single antenna lobe is needed to provide coverage. However, there are some exceptions for cases with large and/or irregular shape DPAs. As an artifact of the algorithm, which tries to minimize the excess area, some DPAs have multiple small lobes (with substantial overlapping areas among themselves) instead of having one large lobe.

Hall, Timothy; Nguyen, Thao; Ranganathan, Mudumbai; Sahoo, Anirudha.
"Sensor Placement and Detection Coverage for Spectrum Sharing in the 3.5 GHz Band."
Paper presented at 29th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Bologna, Italy.
September 9, 2018 - September 12, 2018.

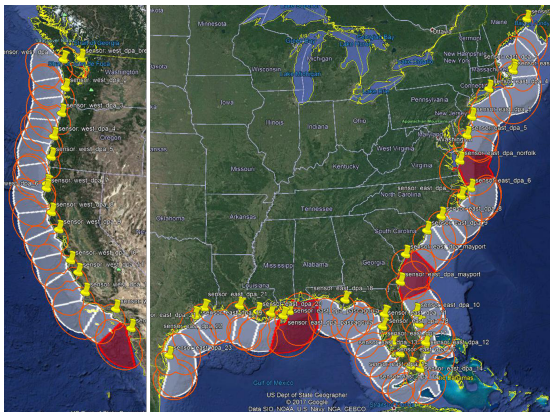


Fig. 3. Detection coverage of all coastal DPAs.

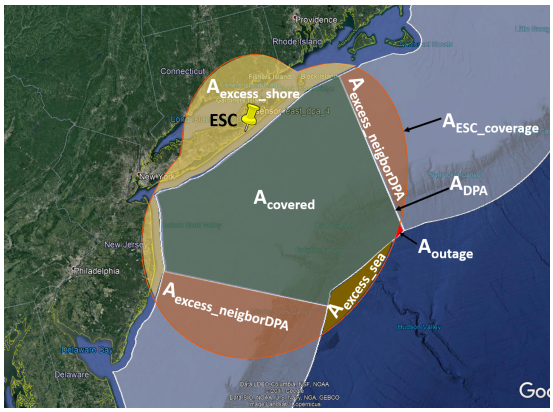


Fig. 4. Illustration of outage and excess areas.

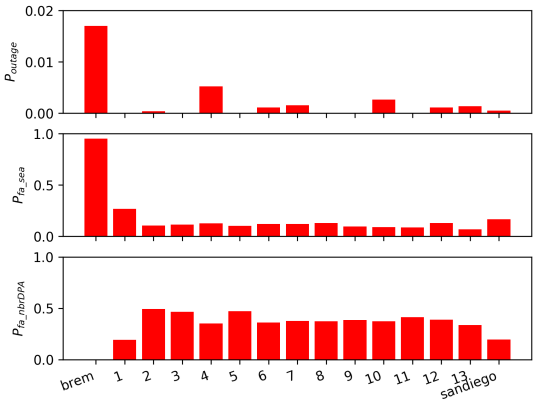


Fig. 5. Performance results of West Coast.

B. Performance Metrics

To formulate our performance metrics, let us define A_{DPA} as the total area of a DPA and A_{ESC} as the detection coverage area of ESC sensor(s) associated with the DPA. Furthermore, let the area that is inside a DPA as well as in its detection coverage area be defined as $A_{cov} = A_{DPA} \cap A_{ESC}$. The area, which is inside a DPA but is outside of its detection coverage area, is defined as $A_{outage} = A_{DPA} \cap \overline{A_{ESC}}$. Finally, the area that is outside of a DPA but is inside its detection coverage area is defined as $A_{excess} = \overline{A_{DPA}} \cap A_{ESC}$.

We further note that the excess area A_{excess} of a DPA has three components: a) excess area overlapping with its neighboring DPAs (A_{excess_nbrDPA}), b) excess area extending out to the sea (A_{excess_sea}), and c) excess area covering sea and land region along the shoreline (A_{excess_shore}). Thus, we have $A_{excess} = A_{excess\_nbrDPA} + A_{excess\_sea} + A_{excess\_shore}$.

The areas defined above are illustrated in Fig. 4. We now define the performance metrics used in our study.

1) Probability of Outage: For a given DPA, we define probability of outage as $P_{outage} = A_{outage}/A_{DPA}$. The probability of outage represents the probability of a shipborne radar not being detected when it is inside the DPA, assuming that its position inside the DPA is uniformly distributed. This value should be zero or significantly small to ensure that the DPA is fully monitored by the ESC sensor(s).

2) Probability of False Alarm: For a DPA we define two types of false alarms as follows.

a) False Alarm Out at Sea: This is the false alarm due to the excess coverage area further out at sea and is defined as $P_{fa\_sea} = A_{excess\_sea}/(A_{ESC} - A_{excess\_shore})$.

This metric captures the odds that an ESC sensor activates its associated DPA even though the radar is further out in the sea and outside of the DPA. P_{fa_sea} should be as low as possible to avoid unnecessary shutdown of CBSDs. We subtract the A_{excess_shore} from A_{ESC} in the denominator with the assumption that the shipborne radar is unlikely to operate in the A_{excess_shore}.

b) False Alarm from Neighboring DPAs: This false alarm is raised when a DPA is activated because its associated ESC sensor(s) detects signal from a shipborne radar present in its neighboring DPA. This is clearly an undesired event since the radar in the neighboring DPA should only be detected by the ESC sensor(s) in that neighboring DPA. The probability of this false alarm is defined as $P_{fa\_nbrDPA} = A_{excess\_nbrDPA}/(A_{ESC} - A_{excess\_shore})$.

C. Performance Results

Figs. 5 and 6 show performance results of our algorithm when applied to the DPAs along the West Coast and the combined East/Gulf Coasts, respectively.

The top subplot in each figure presents P_{outage} computed for each DPA. In most cases, P_{outage} is either zero or close to zero, except for the *brem* DPA (Bremerton, WA) in the West Coast which has a higher value. These small outages are artifacts of discretization of area of DPAs. They can be minimized further by having finer grid size. Nevertheless, the results indicate that the shipborne radar can be detected in any DPA with a very high probability.

The middle subplot in each figure depicts P_{fa_sea} for each DPA. On the West Coast, except for the *brem* DPA,

Hall, Timothy; Nguyen, Thao; Ranganathan, Mudumbai; Sahoo, Anirudha.
"Sensor Placement and Detection Coverage for Spectrum Sharing in the 3.5 GHz Band."
Paper presented at 29th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Bologna, Italy.
September 9, 2018 - September 12, 2018.

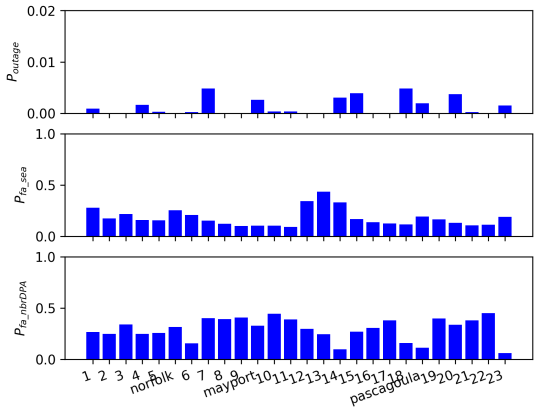


Fig. 6. Performance results of East and Gulf Coasts.

other DPAs have values in the range of (0.07 to 0.27). The *brem* DPA has an extremely narrow shape, for which an antenna lobe of even 60° beamwidth is too wide. This leads to an extremely large excess area into the sea, resulting in $P_{fa\_sea} = 0.95$. On the East and Gulf Coasts, P_{fa_sea} values vary in the range of (0.09 to 0.43). The large values of 0.34, 0.43, and 0.33 belong to DPAs 12, 13, and 14, respectively. These DPAs are close to Florida and have large sharp concave areas due to the islands in the Bahamas. This causes the lobes to cover substantial amount of areas outside of the concave parts of the DPAs. Using antennas with narrower beamwidths, e.g., 30°, will not considerably improve P_{fa_sea} performance, but it might cause OPSEC concern of geolocating incumbent activity [15].

The bottom subplot in each figure shows P_{fa_nbrDPA} for each DPA. P_{fa_nbrDPA} values for all DPAs are in the range of (0 to 0.5). Because of the geometric shapes of the DPAs and antenna lobes, improving P_{fa_nbrDPA} will worsen P_{fa_sea} for a given DPA. Weighting factors could be applied to these false alarms to achieve desired operational performance.

VI. Conclusion and Future Work

This paper presents an approach for an ESC operator to determine location and operational parameters of ESC sensors to detect the presence of federal incumbent shipborne radar. We formulate the problem as a generalized coverage problem where an ESC sensor covers a geometric shape (RCA) such that the excess area is minimized. We apply our algorithm to DPAs along the coasts of CONUS as a use case and present the performance results for each DPA.

We used the ITM area mode, without details of terrain information, to compute the antenna coverage patterns. Based on the final ESC sensor parameters, future work should consider using the ITM point-to-point mode [16] to evaluate the detection coverage areas and corresponding performance metrics. In addition, we used locations along the coast as candidate sites. However, ESC operators may prefer to use existing tower locations as candidate ESC sensor sites to minimize the cost of deployment. Hence, it would be worthwhile to study the performance of our algorithm for that scenario.

Acknowledgement

The authors acknowledge Anastase Nakassis of NIST for his insightful discussions which led to the refinement of our algorithm.

References

[1] "Citizens broadband radio service," 2 C.F.R. § 96, 2016.
[2] E. Drocella, J. Richards, R. Sole, F. Najmy, A. Lundy, and P. McKenna, "3.5 GHz Exclusion Zone Analyses and Methodology," National Telecommunications and Information Administration, Technical Report TR 15-517, Mar. 2016. [Online]. Available: http://www.its.bldrdoc.gov/publications/2805.aspx
[3] "Requirements for Commercial Operation in the U.S. 3550–3700 MHz Citizens Broadband Radio Service Band," Wireless Innovation Forum Document WINNF-15-S-0112, Version V1.3.0, Sep. 2017. [Online]. Available: https://workspace.winnforum.org/higherlogic/ws/public/document?document_id=4743&wg_abbrev=SSC
[4] J. Carle and D. Simplot-Ryl, "Energy-efficient area monitoring for sensor networks," *Computer*, vol. 37, no. 2, pp. 40–46, 2004.
[5] S. Slijepcevic and M. Potkonjak, "Power efficient organization of wireless sensor networks," in *Communications, 2001. ICC 2001. IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 472–476.
[6] D. Tian and N. D. Georganas, "A coverage-preserving node scheduling scheme for large wireless sensor networks," in *Proceedings of the 1st ACM international workshop on Wireless sensor networks and applications*. ACM, 2002, pp. 32–41.
[7] M. Cardei and D.-Z. Du, "Improving wireless sensor network lifetime through power aware organization," *Wireless Networks*, vol. 11, no. 3, pp. 333–340, 2005.
[8] K. Kar and S. Banerjee, "Node placement for connected coverage in sensor networks," in *WiOpt'03: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, 2003, pp. 2–pages.
[9] M. Cardei and J. Wu, "Energy-efficient coverage problems in wireless ad-hoc sensor networks," *Computer communications*, vol. 29, no. 4, pp. 413–420, 2006.
[10] F. H. Sanders, E. F. Drocella, and R. L. Sole, "Using On-Shore Detected Radar Signal Power for Interference Protection of Off-Shore Radar Receivers," National Telecommunications and Information Administration, Technical Report TR 16-521, Mar. 2016. [Online]. Available: http://www.its.bldrdoc.gov/publications/2828.aspx
[11] "Application of Google Inc. for Certification to Provide Spectrum Access System and Environmental Sensing Capability Services," GN Docket No. 15-319, Appendix B: Environmental Sensing Capability (ESC) Siting Considerations, 2016. [Online]. Available: https://ecfsapi.fcc.gov/file/60001851224.pdf
[12] S. Joshi and K. B. S. Manosha and M. Jokinen and T. Hänninen and Pekka Pirinen and H. Posti and M. Latva-aho, "ESC sensor nodes placement and location from moving incumbent protection in CBRS," *Proceedings of WInnComm 2016*, Mar. 2016.
[13] T. T. Nguyen, A. Sahoo, M. Souryal and T. A. Hall, "3.5 GHz Environmental Sensing Capability Sensitivity Requirements and Deployment," in *IEEE DySPAN '17*, March 2017, pp. 1–10.
[14] T. T. Nguyen and M. R. Souryal and A. Sahoo and T. A. Hall, "3.5 GHz Environmental Sensing Capability Detection Thresholds and Deployment," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 3, pp. 437–449, Sept 2017.
[15] "CBRS Operational Security," Wireless Innovation Forum Document WINNF-15-S-0071, Version V1.0.0, Jun. 2016. [Online]. Available: http://www.wirelessinnovation.org/assets/work_products/Specifications/winnf-15-s-0071-v1.0.0%20cbrs%20operational%20security.pdf
[16] "Irregular Terrain Model (ITM) (Longley-Rice) (20 MHz–20 GHz)." [Online]. Available: https://www.its.bldrdoc.gov/resources/radio-propagation-software/itm/itm.aspx
[17] H. Alt, E. M. Arkin, H. Brönnimann, J. Erickson, S. P. Fekete, C. Knauer, J. Lenchner, J. S. B. Mitchell, and K. Whittlesey, "Minimum-cost coverage of point sets by disks," in *Proceedings of the Twenty-second Annual Symposium on Computational Geometry*, ser. SCG '06. ACM, 2006, pp. 449–458.
[18] "3.5 GHz Radar Waveform Capture at Point Loma," National Institute of Standards and Technology (NIST), Technical Note NIST.TN.1954, May 2017. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.1954.pdf
[19] "Technical Specification Group Radio Access Network; Feasibility study for enhanced uplink for UTRA FDD," Technical Report 3GPP-TR25.896, Mar. 2004. [Online]. Available: http://www.qtc.jp/3GPP/Specs/25896-600.pdf

Hall, Timothy; Nguyen, Thao; Ranganathan, Mudumbai; Sahoo, Anirudha.
"Sensor Placement and Detection Coverage for Spectrum Sharing in the 3.5 GHz Band."
Paper presented at 29th Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Bologna, Italy.
September 9, 2018 - September 12, 2018.

Numerical Validation of a Boundary Element Method With $\boldsymbol{E}$ and $\frac{\partial \boldsymbol{E}}{\partial N}$ as the Boundary Unknowns

Johannes Markkanen
Department of Physics
University of Helsinki
P.O. Box 64, FI-00014, Finland
johannes.markkanen@helsinki.fi

Alex J. Yuffa, Joshua A. Gordon
National Institute of Standards and Technology
Boulder, Colorado 80305, USA
{alex.yuffa, josh.gordon}@nist.gov

***Abstract*—We recently developed a surface integral equation method where the electric field and its normal derivative are chosen as the boundary unknowns. After reviewing this formulation, we present preliminary numerical calculations that show good agreement with the known results. These calculations are encouraging and invite the further development of the numerical solution.**

I. Introduction

We have recently formulated a frequency domain surface integral equation method [1] that is applicable to penetrable closed surface scatterers. The method has several unique applications and advantages over the standard Stratton–Chu formulation as discussed in [1]. In our formulation, we choose the electric field (E-field) and its normal derivative as the boundary unknowns. This choice leads to 12 scalar unknowns on the surface of the scatterer; for each homogeneous region we have three scalar unknowns associated with the E-field and three scalar unknowns associated with its normal derivative. Similar to a typical surface integral equation formulation, our formulation is also based on the Green's theorem (Green's second identity). This formulation leads to *six* scalar equations, and thus it must be supplemented with *six* additional constraints in order to have the same number of equations as unknowns. Three of these constraints come from the well-known continuity condition of the E-field across an interface and the other three come from the recently derived continuity condition for the normal derivative of the E-field [1]–[3].

In this paper, we numerically solve the above discussed equations for spherical scatterers and compare the results to the well-known Mie series solution. We also comment on the choice of the basis functions in the Galerkin's method and its effects on numerical convergence.

II. Formulation Review

Consider a scatterer with permittivity $\overset{2}{\epsilon}$ and permeability $\overset{2}{\mu}$. The space surrounding the scatterer is assumed to be lossless with permittivity $\overset{1}{\epsilon}$ and permeability $\overset{1}{\mu}$, i.e., $\{\overset{1}{\epsilon}, \overset{1}{\mu}\} \in \mathbb{R}$. If we apply the Green's second identity to the scatterer and the surrounding space, then, after setting the observation point on the surface of the scatterer, we obtain:

$$\overset{\text{inc}}{\boldsymbol{E}}(\widetilde{S}) - \fint_{\Sigma} \left[\overset{1}{G} \frac{\partial \overset{1}{\boldsymbol{E}}}{\partial N} - \overset{1}{\boldsymbol{E}} \frac{\partial \overset{1}{G}}{\partial N} \right] \mathrm{d}S = \frac{1}{2} \overset{1}{\boldsymbol{E}}(\widetilde{S}), \tag{1a}$$

and

$$\fint_{\Sigma} \left[\overset{2}{G} \frac{\partial \overset{2}{\boldsymbol{E}}}{\partial N} - \overset{2}{\boldsymbol{E}} \frac{\partial \overset{2}{G}}{\partial N} \right] \mathrm{d}S = \frac{1}{2} \overset{2}{\boldsymbol{E}}(\widetilde{S}), \tag{1b}$$

where $\overset{\text{inc}}{\boldsymbol{E}}$ is the incident E-field, $\fint$ denotes the Cauchy principal value integral, Σ denotes the surface of the scatterer, $\frac{\partial}{\partial N}$ denotes the normal derivative, G is the free-space Green's function, and $\widetilde{S}$ is the observation point on Σ. In (1), the overset digit indicates if the quantity is associated with the scatterer or the surrounding space, e.g., $\overset{2}{\boldsymbol{E}}$ is the E-field just inside the scatterer. In the Gaussian unit system, the continuity condition for the E-field across an interface can be written as [1]:

$$\overset{2}{\boldsymbol{E}} = \overset{r}{\epsilon}{}^{-1} \left(\boldsymbol{N} \cdot \overset{1}{\boldsymbol{E}} \right) \boldsymbol{N} + \left(\boldsymbol{S}^{\alpha} \cdot \overset{1}{\boldsymbol{E}} \right) \boldsymbol{S}_{\alpha}, \tag{2a}$$

and the continuity condition for its normal derivative as [1]:

$$\begin{aligned} \frac{\partial \overset{2}{\boldsymbol{E}}}{\partial N} = \overset{r}{\mu} \Bigg(& \frac{\partial \overset{1}{\boldsymbol{E}}}{\partial N} - \nabla^{\alpha} \left[\left(\boldsymbol{N} \cdot \overset{1}{\boldsymbol{E}} \right) \boldsymbol{S}_{\alpha} - \left(\boldsymbol{S}_{\alpha} \cdot \overset{1}{\boldsymbol{E}} \right) \boldsymbol{N} \right] \Bigg) \\ & + \nabla^{\alpha} \left[\left(\boldsymbol{N} \cdot \overset{2}{\boldsymbol{E}} \right) \boldsymbol{S}_{\alpha} - \left(\boldsymbol{S}_{\alpha} \cdot \overset{2}{\boldsymbol{E}} \right) \boldsymbol{N} \right], \end{aligned} \tag{2b}$$

where $\overset{r}{\mu} = \overset{2}{\mu} / \overset{1}{\mu}$, $\overset{r}{\epsilon} = \overset{2}{\epsilon} / \overset{1}{\epsilon}$, $\boldsymbol{N}$ is the unit-normal pointing out of the scatterer, $\boldsymbol{S}_{\alpha}$ is the surface covariant basis [4], and ∇^{α} is the contravariant surface derivative [4]. Notice that (2) is written in the Einstein summation convention where the Greek indices range from 1 to 2. Substituting (2) into (1b) and using Gauss's theorem in two dimensions yields [1]:

$$\begin{aligned} \frac{1}{2} \overset{2}{\boldsymbol{E}}(\widetilde{S}) = & \fint_{\Sigma} \left[\overset{r}{\mu} \overset{2}{G} \frac{\partial \overset{1}{\boldsymbol{E}}}{\partial N} - \overset{1}{\boldsymbol{E}} \frac{\partial \overset{2}{G}}{\partial N} \right] \mathrm{d}S \\ & + \left(\overset{r}{\mu} - \overset{r}{\epsilon}{}^{-1} \right) \int_{\Sigma} \left(\boldsymbol{N} \cdot \overset{1}{\boldsymbol{E}} \right) \boldsymbol{\nabla} \overset{2}{G} \, \mathrm{d}S \\ & + (1 - \overset{r}{\mu}) \int_{\Sigma} \left(\overset{1}{\boldsymbol{E}} \cdot \boldsymbol{\nabla} \overset{2}{G} \right) \boldsymbol{N} \, \mathrm{d}S, \end{aligned} \tag{3a}$$

where

$$\overset{2}{\boldsymbol{E}} = \overset{1}{\boldsymbol{E}} + \left(\overset{r}{\epsilon}{}^{-1} - 1 \right) \left(\boldsymbol{N} \cdot \overset{1}{\boldsymbol{E}} \right) \boldsymbol{N}. \tag{3b}$$

This work was partially supported by U.S. government, not protected by U.S. copyright.

Gordon, Joshua; Markkanen, Johannes; Yuffa, Alexey.
"Numerical Validation of a Boundary Element Method With E and dE/dN as the Boundary Unknowns."
Paper presented at 2018 International Applied Computational Electromagnetics Society (ACES) Symposium, Denver, CO, United States. March 24, 2018 - March 29, 2018.

Equation (3) and (1a) form a set of *six* scalar integral equations with *six* scalar unknowns, namely, $\overset{1}{\boldsymbol{E}}$ and $\frac{\partial}{\partial N}\overset{1}{\boldsymbol{E}}$. This is the set of the integral equations that we numerically solve in the next section.

III. Numerical Calculations

We discretize the spherical scatterer with flat triangular elements and construct a basis for the E-field and its normal derivative. We use piecewise constant basis functions for each component associated with the triangle surfaces. Thus, the number of unknowns is six times the number of the triangular elements. Furthermore, we use Galerkin's method to discretize the equations. In other words, the test and basis functions are identical. It is worth noting that the basis functions *do not* enforce any continuity conditions for the E-field or its normal derivative along the surface. Hence, it is clear that we cannot obtain an optimal convergence rate. Moreover, we anticipate that the sharp wedges may also cause some difficulties. Finding a better set of basis functions is an interesting question for future research.

The integral equation set given by (3) and (1a) contains strongly singular integrals. The gradient of the Green's function has the strongest singularity and we decompose it into the normal and surface derivative parts. With the help of integration by parts, the latter one reduces to an integral over a triangle surface and a closed integral over the triangle's edges. We evaluate these integrals using the standard singularity extraction technique [5] in which the singular part is calculated analytically and the remaining part is calculated numerically.

To assess the method, we compare the radar cross section (RCS) of a sphere in free-space meshed by 940 flat triangular patches with the Mie series solution. Fig. 1 shows the RCS of a dielectric sphere with $\overset{1}{k}\rho = 1$, $\overset{2}{\epsilon} = 4$, and $\overset{2}{\mu} = 1$, where ρ is the radius of the sphere and Fig. 2 shows the RCS of a lossy sphere with $\overset{1}{k}\rho = 4$, $\overset{2}{\epsilon} = -2 + \mathrm{i}$, and $\overset{2}{\mu} = 1$. From the figures, we see that our solution agrees well with the Mie series solution in both the dielectric case and the lossy case. More specifically, the L^2-norm relative error of the far-field $\|\boldsymbol{E}\|^2$ integrated over a solid angle is 4.832×10^{-3} for Fig. 1 and 9.360×10^{-3} for Fig. 2.

IV. Conclusions

We numerically tested a recently formulated surface integral equation method where the electric field and its normal derivative are chosen as the boundary unknowns. The preliminary results presented here are in agreement with the Mie series solution for both dielectric and lossy spheres. Furthermore, the method seems to be viable for numerical computations and may be further improved if we employ basis functions that enforce the continuity conditions.

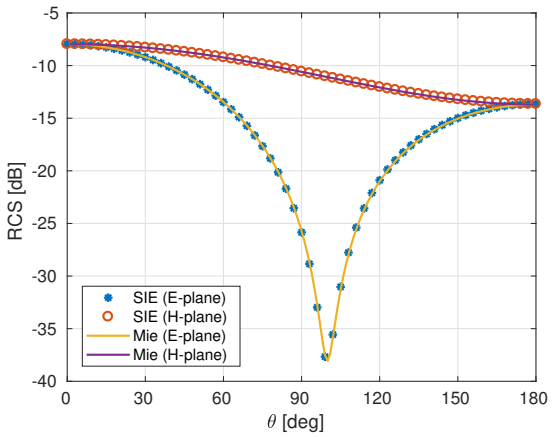


Fig. 1. (Color online) Comparison of the dielectric sphere's RCS as a function of the scattering angle θ computed via the surface integral equation (SIE) method with the Mie series solution.

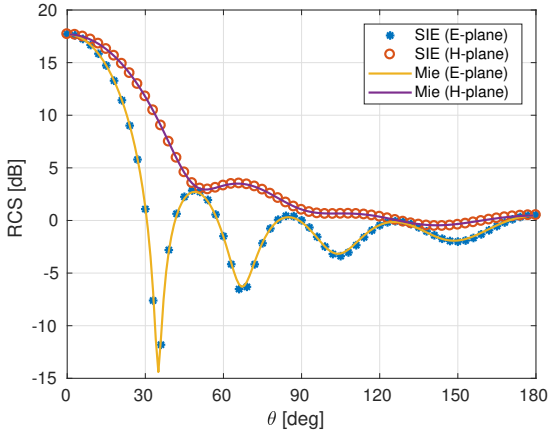


Fig. 2. (Color online) Comparison of the lossy sphere's RCS as a function of the scattering angle θ computed via the surface integral equation (SIE) method with the Mie series solution.

References

[1] A. J. Yuffa and J. Markkanen, "A 3D tensorial integral formulation of scattering containing intriguing relations," submitted for publication.

[2] J. DeSanto and A. Yuffa, "A new integral equation method for direct electromagnetic scattering in homogeneous media and its numerical confirmation," *Waves in Random and Complex Media*, vol. 16, no. 4, pp. 397–408, Nov. 2006. [Online]. Available: http://dx.doi.org/10.1080/17455030500486742

[3] J. A. DeSanto, "A new formulation of electromagnetic scattering from rough dielectric interfaces," *Journal of Electromagnetic Waves and Applications*, vol. 7, no. 10, pp. 1293–1306, Jan. 1993. [Online]. Available: http://dx.doi.org/10.1163/156939393X00480

[4] P. Grinfeld, *Introduction to Tensor Analysis and the Calculus of Moving Surfaces*. Springer, 2013.

[5] S. Järvenpää, M. Taskinen, and P. Ylä-Oijala, "Singularity subtraction technique for high-order polynomial vector basis functions on planar triangles," *IEEE Trans. Antennas Propag.*, vol. 54, no. 1, pp. 42–49, 2006.

Gordon, Joshua; Markkanen, Johannes; Yuffa, Alexey.
"Numerical Validation of a Boundary Element Method With E and dE/dN as the Boundary Unknowns."
Paper presented at 2018 International Applied Computational Electromagnetics Society (ACES) Symposium, Denver, CO, United States. March 24, 2018 - March 29, 2018.

Accurate, Precise and Traceable Laser Spectroscopy: Emerging Technology for the Study of Atmospheric Constituents

Adam J. Fleisher,*[1] David A. Long,[1] Joseph T. Hodges,[1] Gerd A. Wagner,[2] and David F. Plusquellic,[2]
[1]*National Institute of Standards & Technology, 100 Bureau Drive, Gaithersburg, MD 20899*
[2]*National Institute of Standards & Technology, 325 Broadway, Boulder, CO 80305*
*correspondence to: adam.fleisher@nist.gov

The National Academies of Science WORKSHOP ON THE FUTURE OF ATMOSPHERIC BOUNDARY LAYER OBSERVATIONS
October 24-26, 2017

1. Introduction

Accurate, precise, and traceable measurement science enables remote sensing over local, regional, continental, and planetary scales. Over the last 50 years, synergistic leaps in laboratory spectroscopy, optical engineering, theoretical physical chemistry, and computer science have enabled satellite, aircraft, and ground-based field campaigns with impressively low uncertainties. To further improve the accuracy of the reference data required by current and future remote sensing campaigns, as well as to explore fundamental molecular, atomic and collisional physics, new high-precision experimental techniques are required with traceability to the International System of Units (SI). Here we present an overview of ongoing NIST developments in laser instrumentation for the rapid acquisition of molecular spectra at the highest levels of accuracy and precision, and introduce extensions of these emerging laboratory technologies to applications in remote sensing and laser ranging.

2. Emerging laboratory technology – Rapid scanning and multiplexing

Frequency agile, rapid scanning (FARS) spectroscopy mitigates the largest sources of systematic uncertainty in the measurement of accurate spectroscopic parameters (e.g., transition frequencies and intensities, broadening and collisional parameters, temperature coefficients, etc.) by tuning the probe laser at kHz rates [1]. Compared to traditional laser scanning methods (e.g. temperature, current, mechanical grating), radio-frequency sideband tuning is more than 1 000 times faster, and thus "freezes" the systematic biases associated with drifts in sample conditions which general occur over ≥ 1 s. Rapid scanning integrated path, differential absorption light detection and ranging (IPDA LIDAR) of the Earth's planetary boundary layer (PBL) was used recently to measure

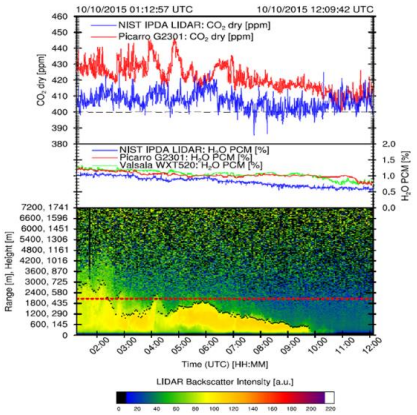


Figure 1. Top panel, CO_2 dry air concentrations measured by fast scanning IPDA LIDAR (blue) and a point sensor (red). Middle panel, corresponding H_2O concentrations. Bottom panel, LIDAR backscatter intensity, where red dots indicate the boundary layer height. [Fig. 13a; Wagner and Plusquellic, *Appl. Opt*. **55,** 6292-6310 (2016).]

Fleisher, Adam; Hodges, Joseph; Long, David; Plusquellic, David; Wagner, Gerd.
"Accurate, Precise and Traceable Laser Spectroscopy: Emerging Technology for the Study of Atmospheric Constituents."
Paper presented at National Academies of Science Workshop on the Future of Atmospheric Boundary Layer Observations, Warrenton, VA, United States. October 23, 2017 - October 26, 2017.

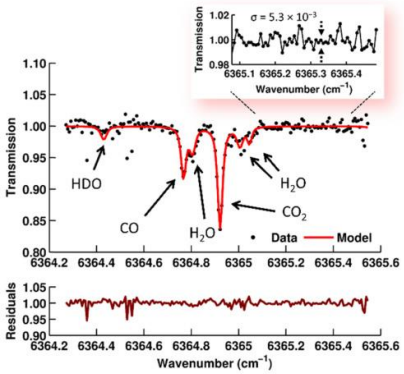


Figure 2. Cavity-enhanced dual-comb spectroscopy of synthetic air performed using electro-optic frequency combs. Constituent species including H_2O, HDO, CO, and CO_2, all acquired simultaneously and in as little as a few µs. [Fig. 4; Fleisher et al., *Opt. Express* **24,** 10424-10434 (2016).]

CO_2 and CH_4 dry air concentrations over an open path in proximity to the Rocky Mountain Flatirons in Boulder, Colorado (see Fig. 1) [2].

In addition to rapid scanning, we demonstrated multiplexed spectroscopy using electro-optic (EO) frequency combs and dual-comb spectroscopy (DCS). Multiplexed DCS is often described as Fourier transform spectroscopy without moving parts. Importantly, the optical generation of EO frequency combs provides unprecedented user control and frequency agility for multiplexed sensing, a high signal-to-noise ratio, and compatibility with a variety of enhancement cavities. EO frequency combs therefore enable the rapid spectroscopy of weak molecular transitions over effective path lengths of > 10 km from a compact and robust laboratory instrument. In Fig. 2, we apply cavity-enhanced DCS to the study of atmospheric constituents (included deuterated water) in a sample of synthetic air [3].

3. Enabling measurement science – Laser sensing of ambient radiocarbon (^{14}C)

Radiocarbon (^{14}C) is an important atmospheric tracer for CO_2 source apportionment [4]. In combination with atmospheric transport models, ^{14}C is widely used to identify carbon sources and sinks. However, ^{14}C analysis at the highest levels of precision is slowed by the necessary shipment of samples to off-site accelerator mass spectrometry (AMS) facilities. Recently we demonstrated the optical measurement of ^{14}C in CO_2 samples derived from the combustion of bioethanol [5]. With an uncertainty of 130 fmol/mol (130 parts-per-quadrillion, or ppq) in an acquisition time of 47 min, this emerging technology, based on the principles of linear absorption spectroscopy, will enable distributed, traceable, and *in situ* sensing of ^{14}C, and thus a new age of ^{14}C metrology.

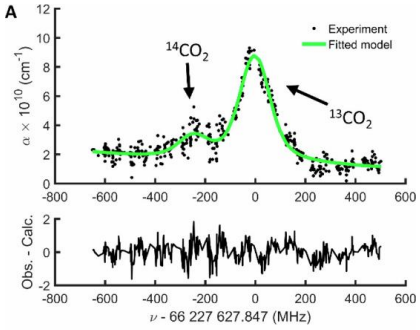


Figure 3. Cavity ring-down spectroscopy of CO_2 from the combustion of bioethanol. The peak at approximately −250 MHz is the signature ^{14}C absorption. The peak area is proportional to the mole fraction of ^{14}C. [Fig. 2; Fleisher et al., *J. Phys. Chem. Lett.* **8,** 4550-4556 (2017).]

4. Outlook

New methods in multiplexed spectroscopy have applications in remote sensing, including open-path sensing and ranging. Using EO frequency combs for IDPA LIDAR enables coherent averaging of backscattering in the time domain via ultrasensitive photon counting techniques (see Fig. 4) [6]. For IDPA LIDAR, the relatively narrow optical bandwidth of EO frequency combs is an advantage, as all the backscattered photons incident on the photon counting and multiplying detector carry information about integrated open-path absorption. In addition to DCS using EO frequency combs, dispersive spectrometers in combination with mid-infrared

mode-locked laser frequency combs also allow for rapid multiplexed spectroscopy over open atmospheric paths [7] and for the time-resolved spectroscopy of transient free radicals [8].

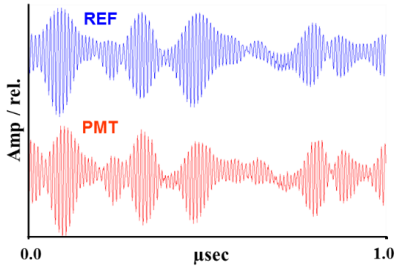


Figure 4. First demonstration of multiplexed IPDA LIDAR using electro-optic frequency combs. Averaged reference (blue) and coadded photon counting (red) interferograms are in excellent agreement following real-time phase corrections. The coadded photon counting interferogram contains absorption information which reveals the CO_2 dry air concentrations. [Plusquellic et al., *CLEO* (2017), paper AM1A.5.]

Distributed measurements of rare isotopologues, particularly the ^{14}C isotopologues of CO_2, CH_4, and carbonaceous aerosols, would enable the partitioning of carbon emissions into specific sources and sinks. Currently fossil fuel CO_2 emissions (CO_{2ff}) are routinely quantified via a ground-up approach, where economic data with uncertainties of approximately 5-10% are used to estimate CO_{2ff} [4]. A distributed network of *in situ* optical measurement of ^{14}C tracers could constrain atmospheric transport models, including key details on the PBL. As a worked example, a previous National Academies of Science report details how 10,000 measurements of ^{14}C per year could yield CO_{2ff} emissions uncertainties of 10-25% [9]. In lieu of a dedicated AMS facility, distributed optical sensors could potentially accomplish this ambitious measurement goal.

References

[1] G.-W. Truong, K. O. Douglass, S. E. Maxwell, R. D. van Zee, D. F. Plusquellic, J. T. Hodges, D. A. Long, "Frequency-agile, rapid scanning spectroscopy," *Nat. Photonics* **7,** 532-534 (2013).

[2] G. A. Wagner, D. F. Plusquellic, "Ground-based, integrated path differential absorption LIDAR measurement of CO_2, CH_4, and H_2O near 1.6 µm," *Appl. Opt.* **55,** 6292-6310 (2016).

[3] A. J. Fleisher, D. A. Long, Z. D. Reed, J. T. Hodges, D. F. Plusquellic, "Coherent cavity-enhanced dual-comb spectroscopy," *Opt. Express* **24,** 10424-10434 (2016).

[4] *Radiocarbon and Climate Change: Mechanisms, Applications and Laboratory Techniques*, E. A. G. Schuur, E. R. M. Druffel, eds. (Springer, 2016).

[5] A. J. Fleisher, D. A. Long, Q. Liu, L. Gameson, J. T. Hodges, "Optical measurement of radiocarbon below unity fraction modern by linear absorption spectroscopy," *J. Phys. Chem. Lett.* **8,** 4550-4556 (2017).

[6] D. F. Plusquellic, G. A. Wagner, A. J. Fleisher, D. A. Long, J. T. Hodges, "Multiheterodyne spectroscopy using multi-frequency combs," in *Conference on Lasers and Electro-Optics*, OSA Technical Digest (online) (Optical Society of America, 2017), paper AM1A.5.

[7] L. Nugent-Glandorf, F. R. Giorgetta, S. A. Diddams, "Open-air, broad-bandwidth trace gas sensing with a mid-infrared optical frequency comb," *Appl. Phys. B* **119,** 327-338 (2015).

[8] A. J. Fleisher, B. J. Bjork, T. Q. Biu, K. C. Cossel, M. Okumura, J. Ye, "Mid-infrared time-resolved frequency comb spectroscopy of transient free radicals," *J. Phys. Chem. Lett.* **5,** 2241-2246 (2014).

[9] *Verifying Greenhouse Gas Emissions: Methods to Support International Climate Agreements* (National Academies Press, 2010).

Fleisher, Adam; Hodges, Joseph; Long, David; Plusquellic, David; Wagner, Gerd.
"Accurate, Precise and Traceable Laser Spectroscopy: Emerging Technology for the Study of Atmospheric Constituents."
Paper presented at National Academies of Science Workshop on the Future of Atmospheric Boundary Layer Observations, Warrenton, VA, United States. October 23, 2017 - October 26, 2017.

Thermal Noise Metrology with Time-Based Synthesis

Dazhen Gu
National Institute of Standards and Technology, Boulder, CO USA
dazhen.gu@nist.gov

***Abstract*—We present a new measurement technique for quantifying noise temperature based on temperature and time. A single-pole, double-throw (SPDT) ultra-fast switch combines signals from outputs of two synchronized electromechanical (EM) switches. The input ports of EM switches are terminated by one unknown noise source, one known cold noise source, and two known noise references at ambient temperature. The magnitude of the combined noise signal can be adjusted by the duty cycle of the transistor-transistor logic (TTL) pulse that controls the fast switch. The measurement approaches the balanced mode by regulating the TTL duty cycle in order to minimize the difference between the combined noise signal and the ambient noise reference. In comparison to conventional total-power radiometers, this new instrumentation is more efficient and potentially provides a wider dynamic range.**

***Index Terms*—Balanced radiometer, feed-back loop, noise synthesis, switch, thermal noise.**

I. Introduction

Thermal electromagnetic noise originates from ergodic random processes and is one of the most important microwave parameters. Because of its ubiquitous presence in nature, thermal noise has become the subject of many practical applications. These include precision calibration of receivers, microwave sounding and imaging in earth science, non-invasive temperature measurement for industrial and medical applications, personnel and substance screening in security surveillance, and radiation detection of interstellar medium in radio astronomy.

Thermal noise measurements often present challenges due to the low signal strength. Radiometers have been developed to accurately measure the power of thermal noise signals. Filtering and amplification are almost always needed to bring the level of signal in a frequency band of interest up to the dynamic range of radiometers. However, the considerable amount of noise and gain instability introduced by components in a radiometer necessitate frequent calibration. Total-power radiometers are one of the most popular configurations for metrology applications. Other types of radiometers are also used, such as switched radiometers (Dicke type) and noise-injection radiometers among others. In this report, we present a new instrumentation using null balance between synthetic noise signal and references. Such a configuration, although bearing similarities with noise-injection radiometers, holds some specific traits and merits.

II. Background

A noise-injection radiometer is developed to address the temporal gain and noise-temperature (NT) fluctuation at the expense of radiometric resolution. It was first introduced by Goggins [1] and later Hardy demonstrated an S-band radiometer by improving the injection with modulated pulses [2]. The injected noise power level is constantly adjusted by gating the noise signal from a known source with pulses of variable width, until the injected noise, in combination with the noise signal under test, equals the power level of a reference noise signal; hence the radiometer approaches a null balance. It can be shown that the averaged noise injection power has a linear dependence on the pulse frequency or equivalently the pulse width, as long as the square-law detection holds in the radiometer backend. As a result, noise-temperature measurements essentially reduce to time measurements [2].

The instrumentation depicted above is successfully used in a variety of remote-sensing applications. However, there exists an underlying complication that impedes its utility in the metrology field. PIN diodes, and in fact any solid-state (SS) switches in a broader scope, are driven by pulse waveforms for fast switching. The leading edge voltage spike at both the rise and fall of the pulse waveform produces a spurious signal, called video leakage. This additional noise bleeds into the RF output ports of the switch when it is in active switching regardless of any RF signal present at the input ports of the switch. The video-leakage effects can be safely neglected when dealing with large signals. However, the extra noise introduced by pulse waveforms has to be accounted for, when the signal at the input ports of the switch is also small at thermal noise level.

Characterization of added noise by a SS switch deserves independent investigation in its own right. In principle, a SPDT SS switch is a 3-port device and its equivalent noise can be calibrated with traditional noise-temperature measurement methods. However, the calibration may be practically infeasible due to the dependence of added noise on many variables, such as the supply voltage, the TTL control voltage, and more importantly the switching frequency and the duty cycle.

III. Measurement Principle of New Technique

A block diagram of the proposed radiometer is shown in Fig. 1. Two EM switches are placed in the frontend and they operate synchronously at slow time periods of tens to hundreds of milliseconds. The switching speed of the EM switches is mainly limited by their own operating mechanism and the response time of the square-law detector in the radiometer backend. One of the EM switches (S1) is terminated by a device under test (DUT) with NT of T_x and an ambient noise reference with NT of T_a. The other switch (S2) is terminated by a cryogenic noise standard with NT of T_c and another

Gu, Dazhen.
"Thermal Noise Metrology with Time-Based Synthesis."
Paper presented at Conference on Precision Electromagnetic Measurements, Paris, France. July 8, 2018 - July 13, 2018.

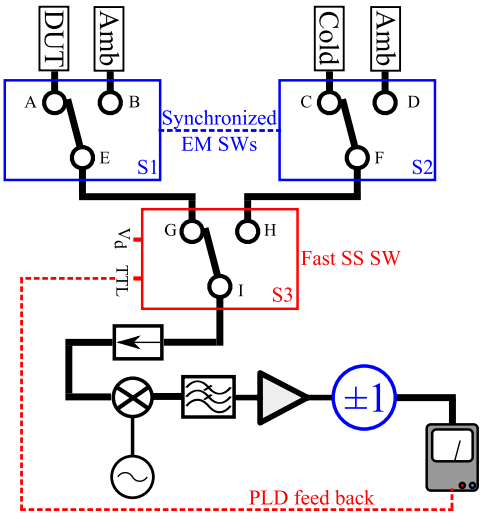


Fig. 1. Block diagram of a radiometer with a feed-back loop to adjust the magnitude of the synthesized noise signal balancing with ambient noise references. The component in the red box is a SS switch at a rapid switching speed and the ones in the blue boxes are two synchronized EM switches at much slower switching speed.

ambient reference with the same NT of T_a as the one on S1. The output of S1 and S2 are connected individually to the two input ports of a SS switch (S3). S3 operates at a very fast speed, which generates a synthesis of noise signals from port G and H measurable by the backend detector. When S1 toggles to port A and S2 toggles to port C, the NT at port I is related to the weighted average of T_x and T_c plus the added noise from S3 (ΔT_{S3}). When S1 toggles to port B and S2 toggles to port D, the NT at port I is roughly the sum of T_a and ΔT_{S3}. The NTs of synthesized noise signals at these two states can be expressed with approximation as

$$T_{\text{AC}} = \alpha T_x + (1-\alpha)T_c + \Delta T_{\text{S3}}, \tag{1a}$$

$$T_{\text{BD}} = T_a + \Delta T_{\text{S3}}, \tag{1b}$$

where α is the duty cycle of the pulse that drives S3.

The rest of the radiometer backend is fairly standard, consisting of isolators, down converters, filters, amplifiers, and a detector. The detector is also synchronized with the EM switches by either digital or analog methods. The resultant detection amounts to the difference between T_{AC} and T_{BD} and is fed back to adjust the TTL duty cycle of S3. The excessive noise (ΔT_{S3}) introduced by the SS switch always contributes equally to the synthesized signals and consequently the need for calibration of ΔT_{S3} is eliminated. Once a null balance is reached, we can easily infer T_x from (1a) and (1b) without knowing ΔT_{S3}.

The formulation in (1), especially (1a), does not include the mismatch and loss corrections. To account for these items, scattering (S-) parameter measurements of components and transmission paths are required. In practical implementation all the radiometric components are expected to be thermally stabilized at T_a, so that (1b) still holds true. After corrections are made to (1a), T_x can be determined from:

$$T_x = \frac{\left[\begin{array}{r} \alpha T_a(M^{\text{I}}_{\text{BEGI}} - M^{\text{I}}_{\text{AEGI}} + M^{\text{A}}_{\text{AEGI}}\eta_{\text{AEGI}}) \\ +(1-\alpha)(T_a - T_c)M^{\text{C}}_{\text{CFHI}}\eta_{\text{CFHI}} \\ +(1-\alpha)T_a(M^{\text{I}}_{\text{DFHI}} - M^{\text{I}}_{\text{CFHI}}) \end{array}\right]}{\alpha M^{\text{A}}_{\text{AEGI}}\eta_{\text{AEGI}}}, \tag{2}$$

where $M^{\diamond}_{\star\ast\circ\text{I}}$ refers to the mismatch factor at the reference point '$\diamond$' when the signal path '$\star\ast\circ$I' between the frontend ports and port I is established under pertinent switching conditions, and $\eta_{\star\ast\circ\text{I}}$ is the efficiency of the signal path '$\star\ast\circ$I'.

IV. Discussion and Conclusion

We proposed a greatly improved variant of a null-balanced, noise-injection radiometer suitable for metrology applications. A few noteworthy items are highlighted as follows:

1) With a specialized arrangement of components and proper synchronization, the added noise introduced by electronic switches, such as PIN diode switches and SS switches, are correctly addressed for the first time. This represents a major step forward for implementing noise-synthesis (or injection) techniques in microwave thermal-noise metrology. Its applications to other fields can evidently improve measurement accuracy.
2) The linearity requirement of backend detectors is relieved because of null-balance operations. This not only reduces cost by choosing economical components for instrumentation but also eliminates the measurement uncertainty due to detector nonlinearity and broadens the detection dynamic range.
3) In comparison to total-power radiometers used in most metrology labs, this new instrument can improve the measurement efficiency by at least 33% since the number of switching states is reduced from 3 to 2.
4) Time and frequency represent the most precise measurands available to metrologists. As long as the pulse duration is kept much longer than the switching time (~30 ns), the overall NT measurement uncertainty would still be dominated by S-parameter measurements.
5) Synchronization and feed-back functions can be realized by analog electronics to achieve fully automation at high speed. However, a digital implementation with simplified software developments may be adequate for validating a prototype system.

In summary, we presented a new instrumentation concept for thermal-noise metrology traceable to temperature and time. The radiometer consists of two synchronized EM switches and one SS switch. The arranged switch operation produces a synthesis of adjustable noise signals to match a known reference. This new approach allows cancellation of added noise from the SS switch and enables precision NT measurements.

References

[1] W. B. Goggins, "A Microwave Feedback Radiometer," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-3, no. 1, pp. 83-90, Jan. 1967.
[2] W. N. Hardy, K. W. Gray, and A. W. Love, "An S-Band Radiometer Design with High Absolute Precision," *IEEE Trans. Microw. Theory Techn.*, vol. 22, no. 4, pp. 382-390, Apr. 1974.

Gu, Dazhen.
"Thermal Noise Metrology with Time-Based Synthesis."
Paper presented at Conference on Precision Electromagnetic Measurements, Paris, France. July 8, 2018 - July 13, 2018.

A Self-Calibrated Transfer Standard for Microwave Calorimetry

Dazhen Gu[1], Xifeng Lu[1], Ben Jamroz[1], Dylan Williams[1], Bill Riddle[1], and Xiaohai Cui[2]
[1]National Institute of Standards and Technology, Boulder, CO USA
[2]National Institute of Metrology, Beijing, China
dazhen.gu@nist.gov

***Abstract*—We developed a new calibration technique for measuring the correction factor of a calorimeter with a vector network analyzer. Based on a wave-parameter formulation, we developed analytic formulas for the correction-factor (g) and effective-efficiency (η). This allows us to calibrate both the calorimeter and the thermistor mount in a single step. This greatly reduces the number of physical parameters involved in calibration processes while tracking correlated uncertainty.**

***Index Terms*—Correction factor, microcalorimeter, microwave power, transfer standard, vector wave parameter.**

I. Introduction

Power is one of the fundamental parameters in microwave metrology. A number of other microwave parameters are derived from precision measurements of either relative or absolute power. A thermistor mount serves as a popular metrology-grade tool and is recognized as a transfer standard for microwave power measurements. A thermistor mount is characterized by its effective efficiency that accounts for the difference between the DC substituted power and the microwave power delivered to it.

Calibrations of the thermistor-mount effective efficiency are often conducted in a calorimeter to achieve high precision. A thermal isolation section is integrated in the calorimeter to stabilize the thermal process for efficient measurements. A considerable amount of research has focused on determining the correction factor (g-factor) of calorimeters, which removes the loss contribution of the thermal isolation section to the power measurement while the calibration of a thermistor mount is undertaken. In this study, we present a new measurement technique using wave parameters to find the g-factor of calorimeters with a self-calibrated thermistor mount.

II. Calorimeter G-Factor

The g-factor is characterized by monitoring the thermopile voltage rise associated with the power dissipation in the thermal isolation section. There have been various methods developed for g-factor measurements, such as the line method, the thru method, and the short and offset-short method [1]. Regardless of what method we choose, the incident power (P_{inc}) at the reference plane of interest is almost always needed for g-factor measurements.

In order to obtain P_{inc}, a three-port coupler is often inserted between the calorimeter and the signal generator in the traditional experimental setup. The side arm of the coupler is terminated by a power sensor, usually a thermistor mount, and P_{inc} can be inferred from its reading. There are complications associated with this approach; 1) the equivalent source mismatch and the coupling coefficient of the coupler need to be measured separately, and 2) it requires a calibrated power sensor with known effective efficiency. Although complication 1 can be overcome with lengthy scattering (S-) parameter measurements, calibrated power sensors may not be available to begin with. In the following section, we show how to measure the g-factor without the need of calibrated sensors and couplers.

III. Theoretical Framework of New Technique

Since a vector network analyzer (VNA) consists of bi-directional couplers in its test ports, the use of a VNA to replace the combination of the signal generator, the coupler, and the power sensor is natural. Calibrated wave parameters can be obtained after calibrations to deduce P_{inc} needed for g-factor measurements. We formulate the short method in this report. Other methods can be handled in a similar manner.

A. Intrinsic G-Factor

We first introduce a reduced form of the g-factor and call it the intrinsic g-factor (g_c). For the foil-short method, we express this as

$$g_c = \frac{P_{\text{FS}}}{P_{\text{inc}}\left(1+|\mathbf{\Gamma}_{\text{FS}}|^2\right)} - \frac{1-|\mathbf{\Gamma}_{\text{FS}}|^2}{1+|\mathbf{\Gamma}_{\text{FS}}|^2}, \tag{1}$$

where P_{FS} is the measured power dissipation related to the loss of the thermal isolation section and the foil short, P_{inc} is the power incident on the foil short, and $\mathbf{\Gamma}_{\text{FS}}$ is the reflection coefficient of the flush short. The first term in (1) indicates the portion of the power dissipated in the thermal isolation section and the second term represents the part of the loss due to the flush short. The correction given by the second term is necessary due to the fact that only the thermal isolation section will be present in the final effective-efficiency measurement.

B. VNA S-Parameter and Power Calibrations

Prior to the g-factor measurements with the short method, we performed a one-port S-parameter calibration and power calibration at the reference plane behind the thermal isolation section where the foil short will be attached. A measurement diagram with network analysis is shown in Fig. 1.

Three or more calibration artifacts (e.g., short, open, and load) are sequentially connected at the reference plane to

Cui, Xiaohai; Gu, Dazhen; Jamroz, Benjamin; Lu, Xifeng; Riddle, Billy; Williams, Dylan.
"A Self-Calibrated Transfer Standard for Microwave Calorimetry."
Paper presented at 2018 Conference on Precision Electromagnetic Measurements, Paris, France. July 8, 2018 - July 13, 2018.

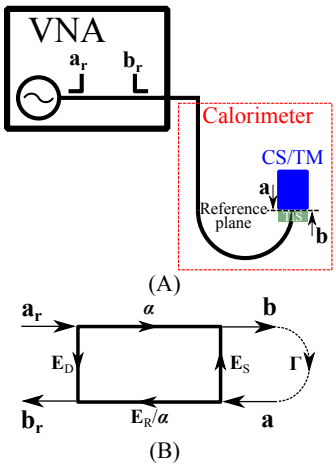


Fig. 1. (A) Illustration of measurement setup. Calibration standards (CS) and a thermistor mount (TM) are sequentially connected at the reference plane to perform the calibration. (B) Equivalent network for error-term analysis.

determine error coefficients $\mathbf{E}_D$, $\mathbf{E}_S$, and $\mathbf{E}_R$ of Fig. 1B. This would be adequate if we were only interested in measuring the reflection coefficient of a device under test (DUT) attached to the reference plane after the calibration. However, to know the incident power on the DUT, the knowledge of $\boldsymbol{\alpha}$ (or more precisely its magnitude $|\boldsymbol{\alpha}|$) is required.

We now connect a thermistor mount with unknown η to the reference plane and record the VNA raw readings $\mathbf{a}_r^{TM}$ and $\mathbf{b}_r^{TM}$. In wave-parameter representation, the power delivered to the thermistor mount is simply $|\mathbf{b}^{TM}|^2 - |\mathbf{a}^{TM}|^2$. The deliverable power can also be calculated from the DC substituted power scaled by the effective efficiency as

$$P = \frac{V_{off}^2 - V_{on}^2}{\eta R_0}, \tag{2}$$

where V_{on} and V_{off} represent the RF on-and off-voltage, respectively and R_0 is the thermistor mount resistance. As a result, the modulus of $\boldsymbol{\alpha}$ can be determined as follows:

$$|\boldsymbol{\alpha}| = \sqrt{\frac{V_{off}^2 - V_{on}^2}{\eta R_0 \left(\left| \frac{\mathbf{E}_R - \mathbf{E}_S\mathbf{E}_D}{\mathbf{E}_R}\mathbf{a}_r^{TM} + \frac{\mathbf{E}_S}{\mathbf{E}_R}\mathbf{b}_r^{TM} \right|^2 - \left| -\frac{\mathbf{E}_D}{\mathbf{E}_R}\mathbf{a}_r^{TM} + \frac{1}{\mathbf{E}_R}\mathbf{b}_r^{TM} \right|^2 \right)}}. \tag{3}$$

Note that η is required to complete the power calibration. To this end, we define scaled wave quantities $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$. So that their magnitude can be resolved directly without knowing η,

$$\begin{bmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \end{bmatrix} = \sqrt{\eta} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \frac{\sqrt{\eta}\boldsymbol{\alpha}}{\mathbf{E}_R} \begin{bmatrix} -\mathbf{E}_D & 1 \\ \mathbf{E}_R - \mathbf{E}_S\mathbf{E}_D & \mathbf{E}_S \end{bmatrix} \begin{bmatrix} \mathbf{a}_r \\ \mathbf{b}_r \end{bmatrix}. \tag{4}$$

In essence, the true wave quantities $\mathbf{a}$ and $\mathbf{b}$ differ merely by a factor of $\sqrt{\eta}$ from $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$. This manipulation enables us to reach a closed form of η evaluation in what follows.

C. Foil-Short Measurement

We now insert a copper foil as a flush short at the reference plane between the thermistor mount and the thermal isolation section and drive the calorimeter with the VNA set at the frequency of interest. The thermistor mount is biased by the power meter at the nominal DC voltage value. Aside from P_{FS} calculated from the thermopile voltage, all other variables in (1) can be computed from raw readings of wave quantities with completely solved error coefficients ($\mathbf{E}$'s) and partially solved $|\boldsymbol{\alpha}|$. They are expressed as

$$P_{FS} = e_{FS}/k_{FS} - e_{DC}/k_{DC}, \tag{5a}$$

$$\boldsymbol{\Gamma}_{FS} = \mathbf{a}^{FS}/\mathbf{b}^{FS} = \hat{\mathbf{a}}^{FS}/\hat{\mathbf{b}}^{FS}, \tag{5b}$$

$$P_{inc} = \left|\mathbf{b}^{FS}\right|^2 = \left|\hat{\mathbf{b}}^{FS}\right|^2/\eta. \tag{5c}$$

Here, e_{FS} and e_{DC} are stabilized thermopile voltages while the VNA is turned on and off respectively, and k_{FS} and k_{DC} are the proportionality factors of the thermopile. We now have explicitly acquired g_c as a function of η after substitution of (5)'s for variables in (1).

D. Effective-Efficiency Measurement

Next, we move on to the measurement of effective efficiency using a signal source with the standard approach. η is related to g_c as follows

$$\eta = \left(1 + g_c \frac{1 + |\boldsymbol{\Gamma}_{TM}|^2}{1 - |\boldsymbol{\Gamma}_{TM}|^2}\right) \frac{1 - \left(\frac{V_2}{V_1}\right)^2}{\frac{e_2/k_2}{e_1/k_1} - \left(\frac{V_2}{V_1}\right)^2}. \tag{6}$$

Thermal processes arrive at equilibrium while the signal source is off and then turned on. Under these two conditions, the power meter reads V_1 and V_2, the thermopile reaches e_1 and e_2, and its proportionality factors are k_1 and k_2. $\boldsymbol{\Gamma}_{TM}$ is the reflection coefficient of the thermistor mount. The term in the parenthesis corresponds to the conventional g-factor, differing from g_c by mismatch corrections of the thermistor mount.

The expansion of g_c in terms of η in (6) yields η in closed form as shown in (7), which completes the self calibration of the thermistor mount:

$$\eta = \frac{1 - \frac{1+\left|\hat{\mathbf{a}}^{TM}/\hat{\mathbf{b}}^{TM}\right|^2}{1-\left|\hat{\mathbf{a}}^{TM}/\hat{\mathbf{b}}^{TM}\right|^2} \cdot \frac{1-\left|\hat{\mathbf{a}}^{FS}/\hat{\mathbf{b}}^{FS}\right|^2}{1+\left|\hat{\mathbf{a}}^{FS}/\hat{\mathbf{b}}^{FS}\right|^2}}{\frac{\frac{e_2/k_2}{e_1/k_1} - \left(\frac{V_2}{V_1}\right)^2}{1-\left(\frac{V_2}{V_1}\right)^2} - \frac{1+\left|\hat{\mathbf{a}}^{TM}/\hat{\mathbf{b}}^{TM}\right|^2}{1-\left|\hat{\mathbf{a}}^{TM}/\hat{\mathbf{b}}^{TM}\right|^2} \cdot \frac{\frac{e_{FS}}{k_{FS}} - \frac{e_{DC}}{k_{DC}}}{\left(|\hat{\mathbf{a}}^{FS}|^2 + \left|\hat{\mathbf{b}}^{FS}\right|^2\right)}}. \tag{7}$$

With the η value at our disposal, we can accurately determine P_{inc} in (5c) and that in turn provides the g_c value in (1).

IV. Conclusion

We demonstrated a new method of measuring the effective efficiency of a thermistor mount in an uncalibrated calorimeter. By use of a partially calibrated VNA for g-factor measurements, both η and g_c can be acquired in closed form in terms of wave parameters and voltages. Experimental validation with correlated uncertainty will be reported at the conference.

References

[1] X. Cui and T. P. Crowley, "Comparison of Experimental Techniques for Evaluating the Correction Factor of a Rectangular Waveguide Microcalorimeter," *IEEE Trans. Instrum. Meas.*, vol. 60, no. 7, pp. 2690-2695, July 2011.

Cui, Xiaohai; Gu, Dazhen; Jamroz, Benjamin; Lu, Xifeng; Riddle, Billy; Williams, Dylan.
"A Self-Calibrated Transfer Standard for Microwave Calorimetry."
Paper presented at 2018 Conference on Precision Electromagnetic Measurements, Paris, France. July 8, 2018 - July 13, 2018.

Large Field of View Quantitative Phase Imaging of Induced Pluripotent Stem Cells and Optical Pathlength Reference Materials

Edward Kwee*, Alexander Peterson, Jeffrey Stinson, Michael Halter, Liya Yu, Michael Majurski, Joe Chalfoun, Peter Bajcsy, John Elliott
National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899

ABSTRACT

Induced pluripotent stem cells (iPSCs) are reprogrammed cells that can have heterogeneous biological potential. Quality assurance metrics of reprogrammed iPSCs will be critical to ensure reliable use in cell therapies and personalized diagnostic tests. We present a quantitative phase imaging (QPI) workflow which includes acquisition, processing, and stitching multiple adjacent image tiles across a large field of view (LFOV) of a culture vessel.

Low magnification image tiles (10x) were acquired with a Phasics SID4BIO camera on a Zeiss microscope. iPSC cultures were maintained using a custom stage incubator on an automated stage. We implement an image acquisition strategy that compensates for non-flat illumination wavefronts to enable imaging of an entire well plate, including the meniscus region normally obscured in Zernike phase contrast imaging. Polynomial fitting and background mode correction was implemented to enable comparability and stitching between multiple tiles. LFOV imaging of reference materials indicated that image acquisition and processing strategies did not affect quantitative phase measurements across the LFOV. Analysis of iPSC colony images demonstrated mass doubling time was significantly different than area doubling time.

These measurements were benchmarked with prototype microsphere beads and etched-glass gratings with specified spatial dimensions designed to be QPI reference materials with optical pathlength shifts suitable for cell microscopy.

This QPI workflow and the use of reference materials can provide non-destructive traceable imaging method for novel iPSC heterogeneity characterization.

Disclaimer: Commercial products are identified in this document in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the products identified are necessarily the best available for the purpose.

Keywords: Stem cells, quantitative phase imaging, reference materials

*edward.kwee@nist.gov; phone 1 301 975-2618;

1. INTRODUCTION

Robust regenerative medicine therapies require a source of stem and progenitor cells. Induced pluripotent stem cells (iPSCs) are a compelling stem cell population that can meet the needs of regenerative medicine. iPSCs are derived from somatic cells that are genetically reprogrammed back to a pluripotent state. The reprogramming process and maintenance of pluripotency during culture expansion requires a trained technician to distinguish heterogeneity of iPSCs and differentiated cells and select iPSCs based on cell morphology using phase contrast imaging [1, 2]. Traditional Zernike phase contrast is primarily used to characterize and identify desired iPSC colonies based on qualitative information about cellular features. Such characterization is subjective and non-transferable between instruments or laboratories. Additionally, Zernike phase contrast features are only maintained in cell cultures where the media surface is relatively flat. Due to the surface tension of cell culture media in well plates, the meniscus-free region is relatively small compared to the entire well, limiting the effective viewing area using Zernike phase contrast.

Bajcsy, Peter; Chalfoun, Joe; Elliott, John; Halter, Michael; Kwee, Edward; Majurski, Michael; Peterson, Alexander; Stinson, Jeffrey; Yu, Liya. "Large Field of View Quantitative Phase Imaging of Induced Pluripotent Stem Cells and Optical Pathlength Reference Materials." Paper presented at SPIE Photonics West BIOS: Quantitative Phase Imaging IV, San Francisco, CA, United States. January 27, 2018 - February 1, 2018.

Quantitative phase imaging (QPI) microscopy is an imaging modality that is different from conventional Zernike phase and can quantitatively determine the phase change in light through cells [3]. We demonstrate the use of QPI to characterize iPSC colony heterogeneity based on dynamic colony mass and area measurements. QPI measurements can also be made traceable to physical reference materials, enabling calibration and comparability between different instruments and laboratories.

2. METHODOLOGY

We developed a large field of view imaging (LFOV) method that performs tiled image acquisition across well plates and characterize iPSC heterogeneity across multiple colonies (Fig 1). In this study, we used a Phasics SID4BIO camera (Phasics S.A., Saint-Aubin, France) which uses quadriwave lateral shearing to perform QPI imaging [4]. QPI methods require physical flattening of the cell culture media to remove the meniscus effect [5]. We implement a method to correct for the non-flat illumination wavefront created by the meniscus and enable LFOV QPI imaging in regions of a well normally obscured in Zernike phase contrast images. We used reference materials to evaluate quantitative capabilities of the developed image processing and analysis workflow. Polymethymethacrylate (PMMA) beads suspended in mineral oil with known diameters and refractive index were evaluated as a stable reference material that can be used well plates. A custom fused silica phase grating with biologically relevant optical pathlength differences (OPD) that mimic those found with stem cells was also evaluated as a potential reference material with defined OPD. This QPI image acquisition workflow in conjunction with reference materials can overcome limitations of Zernike phase contrast and enable quantitative large field of view imaging of stem cells.

2.1 Quantitative Phase Imaging

Imaging was performed using a Axiovert 200M inverted microscope (Carl Zeiss Microscopy, Thornwood, New York) with a motorized stage and incubator system (Kairos Instruments, Pittsburgh, PA), maintaining the stage samples at 37 $^{\circ}$C and 5 % carbon dioxide. QPI was performed using a Phasics SID4BIO camera with a 485 nm LED transmitted light source (Thorlabs, Newton, NJ). Automated large field of view imaging was performed using a custom acquisition script for Micro-Manager [6].

Using a SID4BIO camera, a reference interferogram image was needed to construct the quantitative phase image [7]. To address this need, two wells in a well plate were used. The first well contained the sample of interest with media and the second well contained only media without sample. Interferogram image tiles were acquired on both wells in a positionally matched manner such that the meniscus on the sample tile were positionally matched with the meniscus on the blank wells. This acquisition method was chosen to compensate for the non-flat illumination created by the meniscus. The signal to noise ratio (SNR) was compared between the meniscus corrected and non-corrected QPI images.

Raw interferogram images were batch processed in MATLAB (MATLAB, Natick, MA) using a software development kit from Phasics to produce the QPI image. Resulting QPI images were segmented for foreground and background for reference beads or cells, as described in sections 2.3 and 2.4. Background was corrected using a third order polynomial fit. For each tile, the mode of the background was subtracted across the tile to enable tile-to-tile comparability. The resulting images were stitched using the Microscopy Image Stitching Tool (MIST) [8].

Bajcsy, Peter; Chalfoun, Joe; Elliott, John; Halter, Michael; Kwee, Edward; Majurski, Michael; Peterson, Alexander; Stinson, Jeffrey; Yu, Liya. "Large Field of View Quantitative Phase Imaging of Induced Pluripotent Stem Cells and Optical Pathlength Reference Materials." Paper presented at SPIE Photonics West BIOS: Quantitative Phase Imaging IV, San Francisco, CA, United States. January 27, 2018 - February 1, 2018.

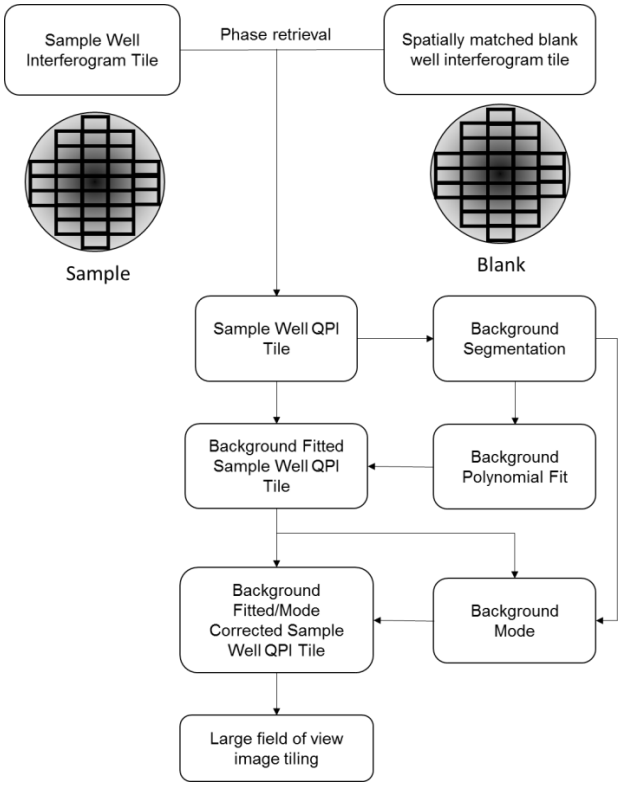


Figure 1: Image acquisition and processing workflow to acquire large field of view tiled QPI image. Diagram of spatially matched acquisition of meniscus in sample and reference sample shown.

2.2 Reference Beads

Polymethylmethacrylate (PMMA) (Cospheric, Santa Barbara, CA) beads were suspended in BioUltra (MilliporeSignma, St. Louis, MO) mineral oil and placed in a 24 well plate. Tiled interferogram and Zernike phase contrast images were taken across a 6x6 tile region (6.2 mm x 4.7 mm) of one well. Reference interferogram images were taken in a well containing only mineral oil. Bead images were processed with and without meniscus correction. Beads were segmented using the active contour method, processed, and stitched with MIST as described in 2.1. Beads were characterized for refractive index and diameter with respect to reference specifications [9, 10].

2.3 Phase Grating

Fused silica phase gratings were fabricated with a period of 40 micrometers (µm), 20 µm feature width, and 100 nanometer (nm) etch depth using reactive ion etching. Samples were imaged using QPI. Resulting images were analyzed in MATLAB. Bilevel waveform pulse analysis was used to characterize optical pathlength differences of the grating when compared to fabrication specifications.

2.4 Cell Culture

ND2.0 iPSCs were plated on Matrigel (Corning, Corning, NY) using E8 Flex culture media (Life Technologies, Carlsbad, CA) in six well plates under three different culture conditions: clump passaged cells plated out of cryopreservation thaw, single cell passaged cells plated out of cryopreservation thaw, and single cell passaged cells at passage three. Cells were cultured in an incubator at 37 $^{\circ}$C and 5 % carbon dioxide for 24 hours to allow the cells to settle and adhere. After 24 hours, cells were imaged using quantitative phase imaging every hour for two days. Images were processed and stitched as described in subsection 2.1. Image analysis was performed on non-fitted, background

Bajcsy, Peter; Chalfoun, Joe; Elliott, John; Halter, Michael; Kwee, Edward; Majurski, Michael; Peterson, Alexander; Stinson, Jeffrey; Yu, Liya. "Large Field of View Quantitative Phase Imaging of Induced Pluripotent Stem Cells and Optical Pathlength Reference Materials." Paper presented at SPIE Photonics West BIOS: Quantitative Phase Imaging IV, San Francisco, CA, United States. January 27, 2018 - February 1, 2018.

mode corrected images. Colonies were segmented at every time point using the empirical gradient threshold method [11]. Manual tracking was performed to identify clonal colonies that did not merge over the course of the time lapse. Clonal colony time-lapse images were processed in MATLAB to determine the area and mass growth rates of colonies. Optical pathlength difference at each pixel was converted to picograms mass using the constant 0.18 $\mu m^3/pg$ [12, 13]. Paired *t*-test was used to compare mass and area doubling times. Analysis of variance (ANOVA) and Tukey comparison tests were used to compare different culture conditions.

3. RESULTS

3.1 Reference beads

Large field of view images of reference beads were acquired (Fig 2). Zernike phase imaging demonstrated the extent of the meniscus in the well (Fig 1A). Tiles that were not corrected for the meniscus (Fig 2B) had a SNR of 2.52 ± 1.11 (mean ± standard deviation, *n*=36 tiles). Tiles that were corrected for the meniscus (Fig 2B) had an improved SNR of 13.8 ± 6.53 (*n*=36) and allowed imaging beads in the meniscus region of the well. Optical pathlength difference measures of the beads reproduced the bead refractive index specifications and dimensions according to reported specifications in both the meniscus-free and meniscus regions of the well (Fig 2D) [9, 10].

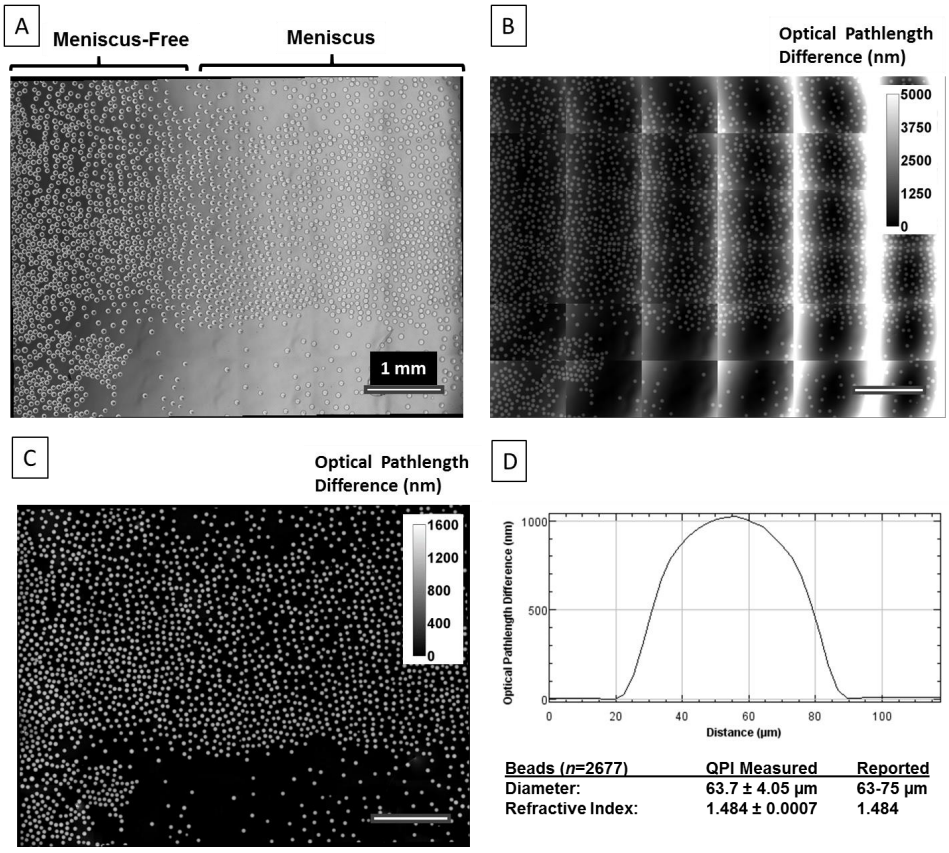


Figure 2: Reference bead characterization across a 6x6 tile region (6.2 mm x 4.7 mm) of a 24 well plate: A) Stitched Zernike phase contrast image of PMMA beads in mineral oil viewing both the meniscus and meniscus-free regions. B) Stitched QPI image of PMMA beads using a single reference image. C) Stitched QPI image using location matched reference images from a blank well to compensate for the cell culture media meniscus. D) Representative line scan of bead from (C). Measured diameter and calculated refractive image of beads from (C).

Bajcsy, Peter; Chalfoun, Joe; Elliott, John; Halter, Michael; Kwee, Edward; Majurski, Michael; Peterson, Alexander; Stinson, Jeffrey; Yu, Liya. "Large Field of View Quantitative Phase Imaging of Induced Pluripotent Stem Cells and Optical Pathlength Reference Materials." Paper presented at SPIE Photonics West BIOS: Quantitative Phase Imaging IV, San Francisco, CA, United States. January 27, 2018 - February 1, 2018.

3.2 Phase grating

QPI images were taken of the phase grating. QPI measurements reproduced optical pathlength difference with 5.6 % error and feature width within 3 % error of fabrication specifications across 27 periods of the grating (Fig 3).

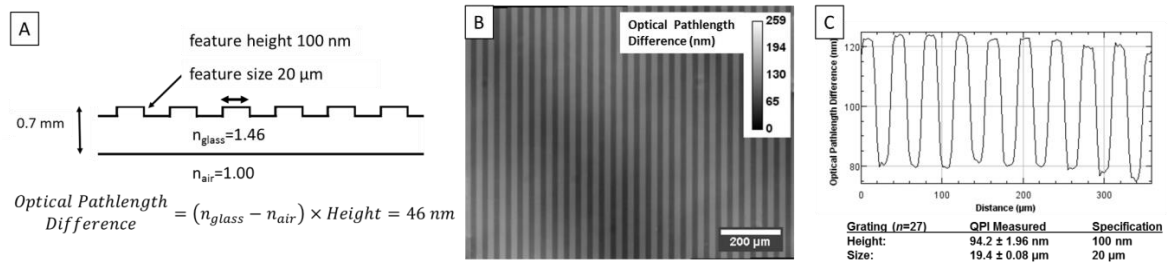


Figure 3: Phase grating characterization: A) Schematic design of phase grating and equation for OPD. B) 10x field of view of phase grating using QPI. C) Line scan of grating. Peak to valley heights and size of spacing were from equation in (A).

3.3 iPSC imaging

LFOV QPI images of iPSC colonies were acquired, processed, and stitched (Fig 4). A portion of the well that was imaged included the meniscus (Fig 4A). Using the developed image processing workflow, the tiled QPI image compensated for the meniscus effect. 70 colonies were identified across all culture conditions. Colonies had a mean mass doubling time of 24.7 hours and a mean area doubling time of 40.6 hours.

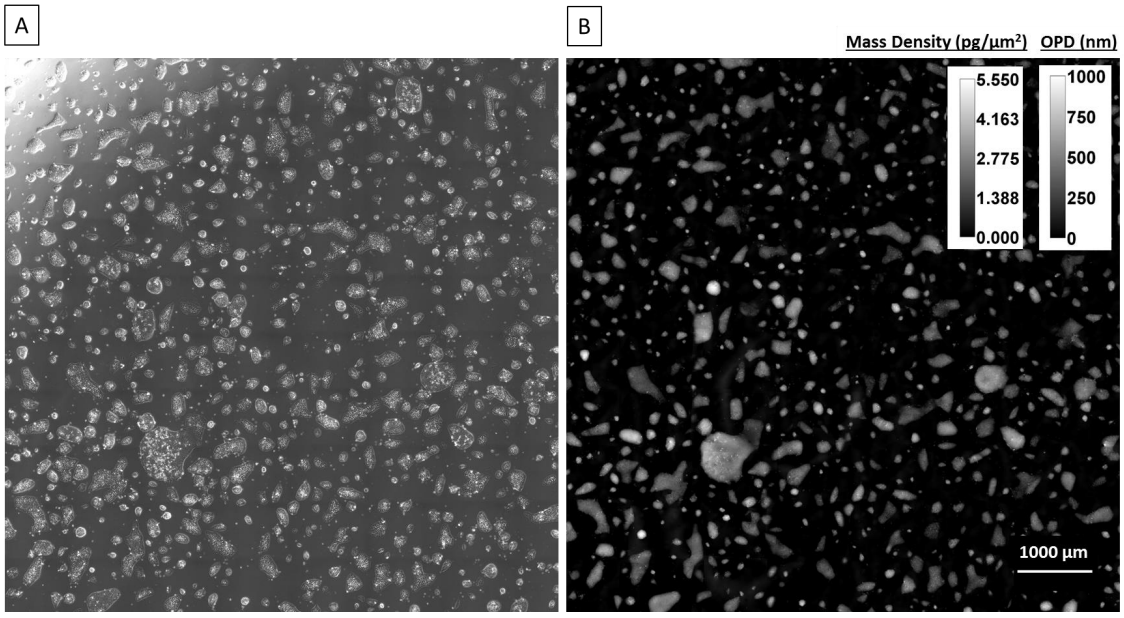


Figure 4: QPI large field of view iPSC colonies of 8x10 tiles (7.4 mm x 7.4 mm). A) Example stitched image of well using Zernike phase contrast image with meniscus. B) Background corrected QPI image with intensity scale expressed in mass density and optical pathlength difference.

Bajcsy, Peter; Chalfoun, Joe; Elliott, John; Halter, Michael; Kwee, Edward; Majurski, Michael; Peterson, Alexander; Stinson, Jeffrey; Yu, Liya. "Large Field of View Quantitative Phase Imaging of Induced Pluripotent Stem Cells and Optical Pathlength Reference Materials." Paper presented at SPIE Photonics West BIOS: Quantitative Phase Imaging IV, San Francisco, CA, United States. January 27, 2018 - February 1, 2018.

Comparing growth rates, mass growth rates were found to be significantly different than area growth rates (Fig 5D). Mass doubling times for iPSCs cultured as single cells were significantly different than clump passaged cells (Fig 5F). No difference in growth rate was found after three passages in the single cell passaged cells. Initial colony mass did not correlate with mass growth rate. Initial colony mass was not significantly associated with mass doubling time.

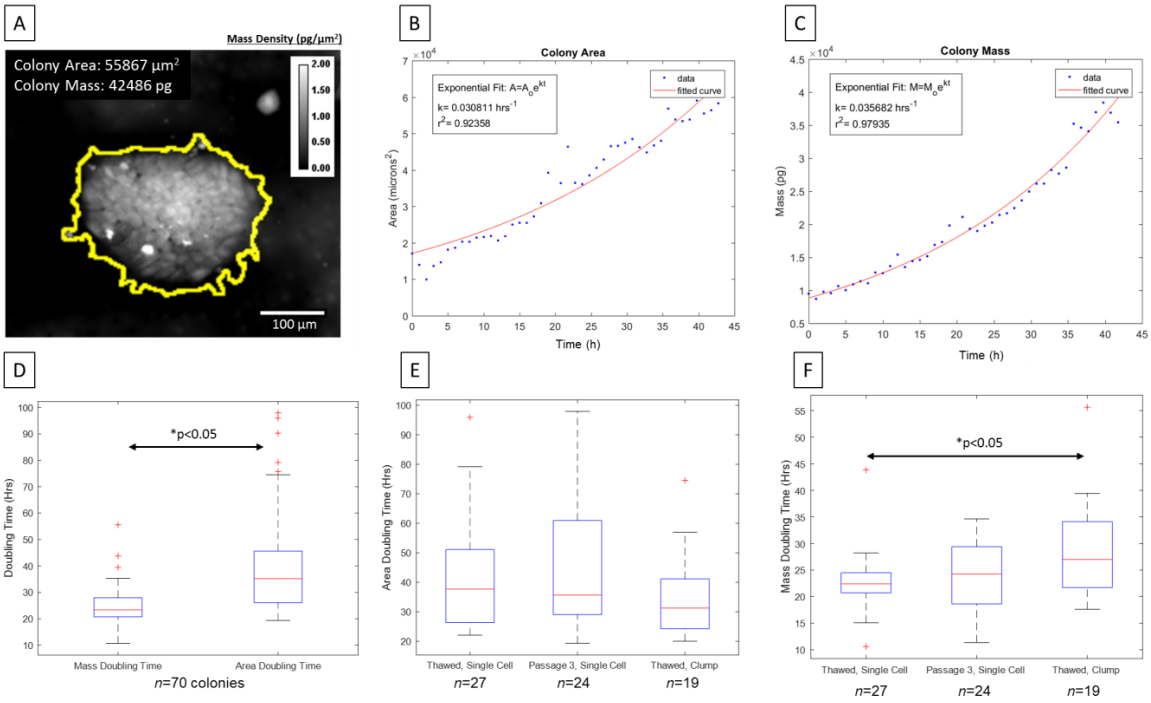


Figure 5: QPI colony analysis. QPI image of representative colony in (A). Based on cell segmentation (yellow), colony area and mass were calculated for each time point. Exponential fits for area (B) and mass (C) was determined for n=70 colonies. Box plots for doubling time based on mass and area for all colonies (D), area doubling time based on culture condition (E), and mass doubling time based on culture condition (F). Box plots in (D),(E), and (F) show the median (red), interquartile range (blue), 1.5 times the interquartile range (dotted), and outliers (red plus).

4. DISCUSSION

We developed an imaging workflow that enables large field of view QPI imaging of cell cultures in well plates with comparability enabled by reference materials. This method overcomes limitations of Zernike phase contrast as a reproducible, quantitative method. This QPI method also enabled imaging in the meniscus to perform imaging of multiple colonies in a well plate to characterize colony heterogeneity. Optical pathlength difference measurements reproduced reference specifications of PMMA beads in mineral oil. The combination of PMMA beads and mineral oil can serve as a stable reference that is not susceptible to evaporative effects and enable long term calibration and traceability of QPI measurements. While useful, the optical pathlength difference created by PMMA beads in mineral oil, approximately 1000 nm, which are higher than the OPD created by biological cells, which are between 50 nm to 250 nm. Additionally, beads may not be the appropriate OPD reference material for all QPI instruments and applications. we found that a custom, fabricated phase grating could be used to produce optical pathlength shifts closer to biological cells. This grating can be used for non-well plate applications and can be placed directly on a QPI instrument without requiring any immersion liquid. A phase grating format also presents the opportunity for an orthogonal measurement of the optical pathlength difference by measuring the diffraction pattern on the back focal plane of a microscope, which is currently

Bajcsy, Peter; Chalfoun, Joe; Elliott, John; Halter, Michael; Kwee, Edward; Majurski, Michael; Peterson, Alexander; Stinson, Jeffrey; Yu, Liya. "Large Field of View Quantitative Phase Imaging of Induced Pluripotent Stem Cells and Optical Pathlength Reference Materials." Paper presented at SPIE Photonics West BIOS: Quantitative Phase Imaging IV, San Francisco, CA, United States. January 27, 2018 - February 1, 2018.

under investigation. Phase gratings with larger etch depths are under evaluation to characterize QPI sensitivity and dynamic range across the biological range of optical pathlength differences induced by cells. Depending on the QPI instrument and application, these materials can serve as a stable reference to calibrate and enable comparability between QPI instruments and experiments.

QPI imaging of iPSC colonies enabled measurement of growth rates by both area and dry mass. Mass growth rates were found to be significantly different than area growth rates. For 2D microscopy techniques like Zernike phase contrast and brightfield, area is the primary measure of cell growth and confluency. QPI provides different metabolic information of the mass growth rate of cells that is not captured by area measures. This could be due to the capability for QPI mass measurements to capture mass changes that occur during the resting or interphase periods of the cell cycle. These periods are not captured by changes in cell area when cells are not actively dividing. Additionally, the contraction and expansion of cell area over the course of the time lapse makes area a poor measure of exponential growth.

Mass growth rate was able to distinguish differences between iPSCs cryopreserved as single cells from those cryopreserved as clump passaged cells. Proliferative ability and self-renewal are critical characteristics of iPSCs. Colonies with the highest mass growth rates may represent the clones with the greatest regenerative capabilities [14]. Further work is needed to confirm the pluripotency state of the cells. The use of QPI to distinguish pluripotency state between stem cells and differentiated cells is under investigation. We expect differences in nucleus condensation and mesenchymal-epithelial transitions as measured by QPI can characterize cell pluripotency state. Beyond culture conditions described here, QPI can enable characterization and identification of other critical manufacturing processes in the culture of iPSCs. This can include assessment of iPSC response to reprogramming methodologies, cell culture media formulations, and differentiation protocols. The quantitative capability and traceability to reference materials of QPI can advance the study and manufacturing of iPSCs as a viable cell therapy for patients.

Disclaimer: Commercial products are identified in this document in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the products identified are necessarily the best available for the purpose.

REFERENCES

[1] E. M. Chan, S. Ratanasirintrawoot, I. H. Park *et al.*, "Live cell imaging distinguishes bona fide human iPS cells from partially reprogrammed cells," Nat Biotechnol, 27(11), 1033-7 (2009).

[2] H. Masaki, T. Ishikawa, S. Takahashi *et al.*, "Heterogeneity of pluripotent marker gene expression in colonies generated in human iPS cell induction culture," Stem Cell Research, 1(2), 105-115 (2008).

[3] M. Mir, B. Bhaduri, R. Wang *et al.*, [Chapter 3 - Quantitative Phase Imaging] Elsevier, (2012).

[4] P. Bon, G. Maucort, B. Wattellier *et al.*, "Quadriwave lateral shearing interferometry for quantitative phase microscopy of living cells," Opt Express, 17(15), 13080-94 (2009).

[5] J. Marrison, L. Räty, P. Marriott *et al.*, "Ptychography – a label free, high-contrast imaging technique for live cells using quantitative phase information," Scientific Reports, 3, 2369 (2013).

[6] A. D. Edelstein, M. A. Tsuchida, N. Amodaj *et al.*, "Advanced methods of microscope control using μManager software," Journal of biological methods, 1(2), e10 (2014).

[7] P. Bon, G. Maucort, B. Wattellier *et al.*, "Quadriwave lateral shearing interferometry for quantitative phase microscopy of living cells," Optics Express, 17(15), 13080-13094 (2009).

[8] J. Chalfoun, M. Majurski, T. Blattner *et al.*, "MIST: Accurate and Scalable Microscopy Image Stitching Tool with Stage Modeling and Error Minimization," Sci Rep, 7(1), 4988 (2017).

[9] A. W. Peterson, M. Halter, A. L. Plant *et al.*, "Surface plasmon resonance microscopy: Achieving a quantitative optical response," Review of Scientific Instruments, 87(9), 093703 (2016).

[10] Z. Hu, and D. C. Ripple, "The Use of Index-Matched Beads in Optical Particle Counters," J Res Natl Inst Stand Technol, 119, 674-82 (2014).

[11] J. Chalfoun, M. Majurski, A. Peskin *et al.*, "Empirical gradient threshold technique for automated segmentation across image modalities and cell lines," J Microsc, 260(1), 86-99 (2015).

Bajcsy, Peter; Chalfoun, Joe; Elliott, John; Halter, Michael; Kwee, Edward; Majurski, Michael; Peterson, Alexander; Stinson, Jeffrey; Yu, Liya. "Large Field of View Quantitative Phase Imaging of Induced Pluripotent Stem Cells and Optical Pathlength Reference Materials." Paper presented at SPIE Photonics West BIOS: Quantitative Phase Imaging IV, San Francisco, CA, United States. January 27, 2018 - February 1, 2018.

[12] R. Barer, and S. Joseph, “Refractometry of Living Cells,” Part I. Basic Principles, s3-95(32), 399-423 (1954).
[13] T. A. Zangle, and M. A. Teitell, “Live-cell mass profiling: an emerging approach in quantitative biophysics,” Nat Methods, 11(12), 1221-8 (2014).
[14] S. Ruiz, A. D. Panopoulos, A. Herrerías *et al.*, “A high proliferation rate is required for cell reprogramming and maintenance of human embryonic stem cell identity,” Current biology : CB, 21(1), 45-52 (2011).

Bajcsy, Peter; Chalfoun, Joe; Elliott, John; Halter, Michael; Kwee, Edward; Majurski, Michael; Peterson, Alexander; Stinson, Jeffrey; Yu, Liya.
"Large Field of View Quantitative Phase Imaging of Induced Pluripotent Stem Cells and Optical Pathlength Reference Materials."
Paper presented at SPIE Photonics West BIOS: Quantitative Phase Imaging IV, San Francisco, CA, United States. January 27, 2018 - February 1, 2018.

HFERP - A New Multivariate Encryption Scheme

Yasuhiko Ikematsu[3], Ray Perlner[2], Daniel Smith-Tone[1,2], Tsuyoshi Takagi[3], and Jeremy Vates[1]

[1]Department of Mathematics, University of Louisville, Louisville, Kentucky, USA
[2]National Institute of Standards and Technology, Gaithersburg, Maryland, USA
[3]Kyushu University, Institute of Mathematics for Industry, Fukuoka, Japan

y-ikematsu@imi.kyushu-u.ac.jp, ray.perlner@nist.gov, daniel.smith@nist.gov, takagi@imi.kyushu-u.ac.jp, jeremy.vates@louisville.edu

Abstract. In 2016, Yasuda et al. presented a new multivariate encryption technique based on the Square and Rainbow primitives and utilizing the plus modifier that they called Square Rainbow Plus (SRP). The scheme achieved a smaller blow-up factor between the plaintext space and ciphertext space than most recent multivariate encryption proposals, but proved to be too aggressive and was completely broken by Perlner et al. in 2017. The scheme suffered from the same MinRank weakness that has allowed effective attacks on several notable big field multivariate schemes: Hidden Field Equations (HFE), multi-HFE, HFE-, for example. We propose a related new encryption scheme retaining the desirable traits of SRP and patching its weaknesses. We call the scheme HFE Rainbow Plus (HFERP) because it utilizes a similar construction as SRP with an HFE primitive replacing the Square polynomial. The effect of this substitution is to increase the Q-rank of the pubic key to such a degree that the MinRank attack is impossible. HFERP still retains the relatively small blow-up factor between the plaintext space and ciphertext space, and is thus a candidate for secure multivariate encryption without an essential doubling in size between plaintext and ciphertext.

Key words: Multivariate Cryptography, HFE, encryption, MinRank, Q-rank

1 Introduction

Ever since the discovery of polynomial time algorithms for factoring and computing discrete logarithms on a quantum computer by Peter Shor [1], creating schemes that resist such developments has fallen upon the shoulders of today's

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

cryptographers. In recent years, quantum computing has made significant advances leading some experts to make more confident predictions that the post-quantum world will soon be upon us, see, for example, [2].

There has also been an explosive development in public key technologies relying on mathematics for which there is no known significant computational advantage quantum computers possess. In particular, multivariate public key cryptography (MPKC) produced numerous schemes for public key encryption and digital signatures in the late 1990s. These schemes further fueled the development of computational algebraic geometry, and seem to have inspired the advancement of some of the symbolic algebra techniques we now apply to all areas of post-quantum cryptography, that is, cryptography designed with quantum computers in mind.

With the development of such techniques, many multivariate schemes have been cryptanalyzed and broken. Specifically, multivariate encryption seems to be challenging. The purpose of this article is to confront this challenge, advancing a new multivariate encryption scheme Hidden Field Equations Rainbow Plus (HFERP), based on Square Rainbow Plus (SRP), see [3], developed to eradicate the deficiencies of its predecessor.

1.1 Recent History

While there may be many trustworthy candidates for multivariate signatures, such as Unbalanced Oil and Vinegar (UOV) [4], Rainbow [5], and Gui [6], developing multivariate schemes for encryption has been a bit of a struggle. While some older ideas have have been reborn with better parameter sets due to the advancement of the science, such as applying HFE-, see [7], to encryption, most of the surviving multivariate encryption schemes are relatively young.

In the last few years, there have been a few new proposals for multivariate encryption, mostly following the idea that it is easier to hide the structure of an injective mapping into a large codomain than to hide the structure of a bijection, as is needed for any encryption mapping into a codomain of the same size as the domain. The ABC Simple Matrix encryption scheme of [8, 9] and ZHFE, see [10] are examples of this idea. Most of these encryption ideas, both new and old, have inspired recent surprising cryptanalyses that affect parameter selection or outright break the scheme, see [11–15], for example.

Such a tale describes the life of SRP, see [3], the design of which aimed to be very efficient and holds a comparably small blow up factor between the plaintext and ciphertext sizes. The scheme also claimed security against attacks efficient against the Square and Rainbow schemes by combining them into one. Unfortunately, SRP is also the victim of a new cryptanalysis, see [16]. The attack exploits the low Q-rank of the Square map, a vulnerability inherited by the public key. A modified MinRank attack was able to pull apart the Square polynomials from the Rainbow and Plus polynomials in the public key.

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

1.2 Our Contribution

We present a new composite scheme in the manner of SRP by replacing the weaker Square layer with an HFE polynomial of higher Q-rank and finding the correct balance in the sizes of the HFE, Rainbow and Plus layers for efficiency and security. We call our scheme HFERP. We further establish the complexity of the relevant attack models: the algebraic attack, the MinRank attack, and the invariant attack.

1.3 Organization

The paper is organized as follows. In the next section, we present isomorphisms of polynomials and describe the structure of HFE and SRP. The subsequent section reviews the Q-rank of ideals in polynomial rings and discusses invariant properties of Q-rank and min-Q-rank. In section 4, we review more carefully the previous cryptanalyses of HFE and SRP. We then present HFERP in the next section. Section 6 discusses the complexity of all known relevant attacks on HFERP. Our choice of parameters to optimize security and performance along with experimental results are then presented in the following section. Finally, we conclude discussing why a similar approach to SRP seems to produce such a different technology in HFERP.

2 Big Field Schemes

HFE and SRP are members of a family of cryptosystems known as "big field" schemes. This term is based on the system exploiting the vector space structure of a degree n extension of $\mathbb{K}$ over a finite field $\mathbb{F}_q$. Using core maps within the extension field allows us to take advantage of Frobenius automorphisms $x \mapsto x^q$ for any function of the form $f(x) = x^{q^i+q^j}$, noting that $\phi^{-1} \circ f \circ \phi$ is a vector-valued quadratic function over $\mathbb{F}_q$ where $\phi : \mathbb{F}_q^n \to \mathbb{K}$ is an $\mathbb{F}_q$-vector space isomorphism. By observing that any vector-valued quadratic function on $\mathbb{F}_q^n$ is isomorphic to a sum of such monomials, it is clear that any quadratic function f over $\mathbb{K}$ can be represented as a vector-valued function, F, over $\mathbb{F}_q$.

This equivalence allows us to construct cryptosystems in conjunction with the following concept, the isomorphisms of polynomials.

Definition 1 *Two vector-valued multivariate polynomials F and G are said to be* isomorphic *if there exist two affine maps T, U such that $G = T \circ F \circ U$.*

The equivalence and isomorphism marry in a method commonly referred to as the butterfly construction. Given a vector space isomorphism $\phi : \mathbb{F}_q^n \to \mathbb{K}$ and an efficiently invertible map $f : \mathbb{K} \to \mathbb{K}$, we compose two affine transformations $T, U : \mathbb{F}_q^n \to \mathbb{F}_q^n$ in order to obscure our choice of basis for the input and output. This construction generates a vector-valued map $P = T \circ \phi^{-1} \circ f \circ \phi \circ U = T \circ F \circ U$,

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

where $F = \phi^{-1} \circ f \circ \phi$.

$$\begin{array}{ccccccc} & & \mathbb{K} & \xrightarrow{f} & \mathbb{K} & & \\ & & \uparrow{\scriptstyle \phi} & & \downarrow{\scriptstyle \phi^{-1}} & & \\ \mathbb{F}_q^n & \xrightarrow{U} & \mathbb{F}_q^n & \xrightarrow{F} & \mathbb{F}_q^n & \xrightarrow{T} & \mathbb{F}_q^n \end{array}$$

2.1 HFE

The Hidden Field Equation Scheme was first introduced by Patarin, see [17], as an improvement on the well known C^* construction of [18]. Patarin's contribution was to use a general polynomial with degree bound D in place of the central monomial map of C^*.

Explicitly, one chooses a quadratic map $f : \mathbb{K} \to \mathbb{K}$ of the form:

$$f(x) = \sum_{\substack{i \leq j \\ q^i + q^j \leq D}} \alpha_{i,j} x^{q^i + q^j} + \sum_{\substack{i \\ q^i \leq D}} \beta_i x^{q^i} + \gamma, \tag{1}$$

where the coefficients $\alpha_{i,j}, \beta_i, \gamma \in \mathbb{K}$ and the degree bound D is sufficiently low for efficient inversion using the Berlekamp algorithm, see [19] .

The public key is computed as $P = T \circ F \circ U$, where $F = \phi^{-1} \circ f \circ \phi$. Inversion is accomplished by taking a ciphertext $y = P(x)$, computing $v = T^{-1}(y)$, solving $\phi(v) = f(u)$ for u via the Berlekamp algorithm and then recovering $x = U^{-1}(\phi^{-1}(u))$.

2.2 Rainbow

The Rainbow scheme is a generalization of Patarin's UOV, see [4]. The key idea, introduced by Ding, see [5], was constructing multiple layers of UOV.

Let $\mathbb{F}$ be a finite field with a degree n extension $\mathbb{F}^n$. Let $\mathcal{V} = \{1, 2, \ldots, n\}$. For a chosen u, let $v_1, \ldots, v_u$ be integers such that $0 < v_1 < \cdots < v_u = n$ and let $\mathcal{V}_l = \{1, \ldots, v_l\}$ for each $l \in \{1, \ldots, u\}$. Note that $|\mathcal{V}_i| = v_i$.

Let $o_i = v_{i+1} - v_i$ for each $i \in \{1, \ldots, u-1\}$ and $\mathcal{O}_i = S_{i+1} - S_i$ for each $i \in \{1, \ldots, u-1\}$. Define P_l to be the space generated by the span of polynomials of the following form:

$$f(x_1, \ldots, x_n) = \sum_{i \in \mathcal{O}_l, j \in \mathcal{V}_l} \alpha_{i,j} x_i x_j + \sum_{i,j \in \mathcal{V}_l} \beta_{i,j} x_i x_j + \sum_{i \in \mathcal{V}_l} \gamma_i x_i + \eta$$

One can refer to the previous constructions using the following terminology: $\mathcal{O}$ is the collection of oil variables, $\mathcal{V}$ is the collection of vinegar variables, and a polynomial $f \in P_l$ is an l-th layer Oil and Vinegar polynomial.

The Rainbow map $F : \mathbb{F}^n \to \mathbb{F}^{n-v_1}$ is defined as (with $x_1, \ldots, x_n$ being referred to as $\bar{x}$ for convenience)

$$F(\bar{x}) = (\tilde{F}_1(\bar{x}), \ldots, \tilde{F}_{u-1}(\bar{x})) = (F_1(\bar{(x)}, \ldots, F_{n-v_1}(\bar{x})$$

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

where each $\tilde{F}_i$ consists of o_i randomly chosen quadratic polynomials from P_i. F is a Rainbow polynomial map with $u-1$ layers. The public key is generated in the usual fashion by applying two affine transformations, T and U, where $T : \mathbb{F}^{n-v_1} \to \mathbb{F}^{n-v_1}$ and $U : \mathbb{F}^n \to \mathbb{F}^n$: $T \circ F \circ U$

2.3 SRP

In Section 5, we present in detail the construction of our proposed scheme, HFERP. For reference, we will include the Square Map definition as well as method of inversion presented in the original SRP paper, see [3].

Instead of using the HFE core map described in section 5, SRP uses the Squaring map where the Square component is defined as $\mathcal{F}_S : \mathbb{F}_q^{n'} \to \mathbb{F}_q^d$ (where $q^d + 1$ is divisible by 4) and it is the result of the following composition:

$$\mathbb{F}_q^{n'} \xrightarrow{\pi_d} \mathbb{F}_q^d \xrightarrow{\phi} \mathbb{K} \xrightarrow{X \mapsto X^2} \mathbb{K} \xrightarrow{\phi^{-1}} \mathbb{F}_q^d$$

Upon inversion step 3, the user would compute

$$R_{1,2} = \pm X^{(q^d+1)/4}$$

and use it to find $\mathbf{y} = (y_1^{(i)}, \ldots, y_d^{(i)}) = \phi^{-1}(R_i) \in \mathbb{F}_q^d$. The choice of the Square map was made because of the speed of inversion it provided when compared to any other quadratic maps. Unfortunately, due to this choice, SRP was quickly broken in [16] by isolating the squaring public polynomials and exploiting its low Q-rank.

3 Q-Rank

The min-Q-rank of the public key is a critical quantity when analizing the security of big field schemes within multivariate cryptography. For clarification, the definition is as follows:

Definition 2 *The Q-rank of any quadratic map $f(\overline{x})$ on $\mathbb{F}_q^n$ is the rank of the quadratic form $\phi^{-1} \circ f \circ \phi$ in $\mathbb{K}[X_0, \ldots, X_{n-1}]$ via the identification $X_i = \phi(\overline{x})^{q^i}$.*

Usually, the definition of the rank of a quadratic form is given as the minimum number of variables required to express an equivalent quadratic form due to quadratic form equivalences corresponding to matrix congruence. Note that congruent matrices have the same rank. This same quantity is equal to the rank of the matrix representations of the quadratic form, even in characteristic 2, where the quadratics x^{2q^i} are additive, but not linear for $q > 2$.

Q-rank is invariant under one-sided isomorphisms $f \mapsto f \circ U$, but is not invariant under isomorphisms of polynomials in general. The quantity that is often meant by the term Q-rank, but more properly called min-Q-rank, is the minimum Q-rank among all nonzero linear images of f. This min-Q-rank is invariant under isomorphisms of polynomials and is the quantity relevant for cryptanalysis.

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

4 Previous Cryptanalysis of Relevant Schemes

SRP was a designed as a concatenation of two known multivariate schemes and a scheme modifier. The first component was Square, see [20], which can be seen as a degenerate version of HFE. The second component was oil-and-vinegar (OV) or, more generally, Rainbow, see [21, 5]. The final component was the plus modifier, first proposed in [22]. The algebraic properties of these schemes were intended to complement their weaknesses when used in conjunction. This patchwork design requires, however, a careful consideration of the relevant cryptanalyses within all of these families.

The original oil-and-vinegar (OV) scheme, proposed in [21], was completely broken in [23] by what we call the invariant method. Specifically, the balanced OV scheme contains an equal number of oil variables, variables which only occur linearly in the central map, and vinegar variables, which occur quadratically. Thus, the differential of any central polynomial has the shape

$$Df_i = \begin{bmatrix} a_{1,1} & \cdots & a_{1,v} & a_{1,v+1} & \cdots & a_{1,2v} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{1,v} & \cdots & a_{v,v} & a_{v,v+1} & \cdots & a_{v,2v} \\ a_{1,v+1} & \cdots & a_{v,v+1} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ a_{1,2v} & \cdots & a_{v,2v} & 0 & \cdots & 0 \end{bmatrix},$$

under an appropriate basis of $\mathbb{F}^{2v} = V \oplus O$, where V is the subspace spanned by the vinegar variables and O is the subspace spanned by the oil variables.

The invariant attack proceeds by computing the differential of random linear combinations of the public polynomials until two full rank differentials, Df_1 and Df_2, are produced. Then O is left invariant by $Df_1^{-1}Df_2$ and is thus easily recovered. A similar technique has been used in conjunction with rank attacks to assault schemes with a similar structure whenever $\dim(V) \leq \dim(O)$, see, in particular, [11, 24, 13].

HFE and some of its modifications have been the target of effective cryptanalyses utilizing the low Q-rank property of the central map. Each of these cryptanalyses can be described as a big field MinRank attack, recovering a low rank quadratic form over the extension $\mathbb{E}$ from which an isomorphism relating the public key to an equivalent private key can be derived.

The earliest iteration of this technique is the well-known Kipnis-Shamir (KS) attack of [25], also known by the name MinRank, due to the close relationship between the attack and the MinRank problem in algebraic complexity theory, see [26]. The KS-attack recovers a private key for HFE by exploiting the fact that the low Q-rank of the central map is a property preserved by isomorphisms. Considering an odd characteristic instance of HFE. We may write the homogeneous

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

quadratic part of the central map as

$$\begin{bmatrix} x & x^q & \cdots & x^{q^{n-1}} \end{bmatrix} \begin{bmatrix} \alpha_{0,0} & \alpha'_{0,1} & \cdots & \alpha'_{0,d-1} & 0 \cdots 0 \\ \alpha'_{0,1} & \alpha_{1,1} & \cdots & \alpha'_{1,d-1} & 0 \cdots 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \ddots \vdots \\ \alpha'_{0,d-1} & \alpha'_{1,d-1} & \cdots & \alpha_{d-1,d-1} & 0 \cdots 0 \\ 0 & 0 & \cdots & 0 & 0 \cdots 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \ddots \vdots \\ 0 & 0 & \cdots & 0 & 0 \cdots 0 \end{bmatrix} \begin{bmatrix} x \\ x^q \\ \vdots \\ x^{q^{n-1}} \end{bmatrix},$$

where $\alpha'_{i,j} = \frac{1}{2}\alpha_{i,j}$ and $d = \lceil \log_q(D) \rceil$. The KS-attack first interpolates an univariate representation of the public key over $\mathbb{E}$. This representation of the public key is isomorphic to the central map of Q-rank bounded by the ceiling of the logarithm of the degree bound. Thus, there is a linear map T^{-1} which when composed with the public key has Q-rank d, and so there is a low rank matrix that is an $\mathbb{E}$-linear combination of the Frobenius powers of G. This turns recovery of the transformation T into the solution of a MinRank problem over $\mathbb{E}$.

Another version of this attack, utilizing the same property, is the key recovery attack of [27]. The authors prove the existence of an $\mathbb{E}$-linear combination of the *public* key with low rank over $\mathbb{E}$. Setting the unknown coefficients of this linear combination as variables, they construct the ideal $I \subseteq R = \mathbb{F}[T]$ of minors of this sum of the appropriate dimension such that $V(I) \cap \mathbb{E}^{\dim(R)}$ consists of exactly such linear coefficients. Thus a Gröbner basis needs to be computed over $\mathbb{F}$ and the variety computed over $\mathbb{E}$. This modeling of the KS-attack is called minors modeling and dramatically improves the efficiency of the KS-attack in many circumstances.

The KS-attack with either KS modeling or with minors modeling has also been used to break other HFE descendants. In [27], the minors modeling approach is used to break multi-HFE. In [15], the KS-attack is extended to provide key recovery for HFE-. In [14], both the KS modeling and minors modeling versions of the KS-attack are used to undermine the security of ZHFE.

The MinRank methodology is also employed in [16], where an effective key recovery attack on SRP is presented. It was shown that the low Q-rank of Square is exposed by the SRP construction. Specifically, the Q-rank of the square map $f(x) = x^2$ is one over an odd characteristic field. Since this low Q-rank map is in the span of the public polynomials, there is an $\mathbb{E}$-linear combination of the public polynomials of rank one! Thus the ideal generated by the two-by-two minors is resolved at degree two and the complexity of the attack is $\mathcal{O}(\binom{m+1}{2}^{\omega})$, where $2 \leq \omega \leq 3$ is the linear algebra constant. The attack is applied practically, breaking the 80-bit parameters in about 8 minutes.

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

5 HFERP

In this section, we present a significant modification of SRP that we call HFERP. The key observation is that by replacing the Square map with a higher Q-rank instance of HFE, one can make the MinRank attack inefficient while maintaining efficient inversion. For simplicity of the exposition, we present the scheme with a single layer UOV component, noting that it is trivial to replace UOV with a multi-layer Rainbow via the same construction.

Choose a finite field $\mathbb{F}_q$ and let $\mathbb{E}$ be a degree d extension field over $\mathbb{F}_q$. Let $\phi : \mathbb{F}_q^d \to \mathbb{E}$ be an $\mathbb{F}_q$-vector space isomorphism. Also, let $o, r, s,$ and l be non-negative integers.

Key Generation Let $n = d + o - l$, $n' = d + o$ and $m = d + o + r + s$. The central map of HFERP is the concatenation of an HFE core map, $\mathcal{F}_{HFE}$, an UOV (or alternatively, Rainbow) section, $\mathcal{F}_R$, and the plus modifier, $\mathcal{F}_P$. Formal definitions of the maps are provided below:

- The HFE component is defined as $\mathcal{F}_{HFE} : \mathbb{F}_q^{n'} \to \mathbb{F}_q^d$ and is the result of the following composition:

$$\mathbb{F}_q^{n'} \xrightarrow{\pi_d} \mathbb{F}_q^d \xrightarrow{\phi} \mathbb{E} \xrightarrow{f} \mathbb{E} \xrightarrow{\phi^{-1}} \mathbb{F}_q^d$$

  where f is the HFE core map described in (1) and $\pi_d : \mathbb{F}_q^{d+o} \to \mathbb{F}_q^d$ is the projection onto the first d coordinates.
- The UOV (or alternatively, Rainbow) component is defined as

$$\mathcal{F}_R = (g^{(1)}, \ldots, g^{(o+r)}) : \mathbb{F}_q^{n'} \to \mathbb{F}_q^{o+r}$$

  following the normal construction of the UOV signature scheme where $\mathcal{V} = \{1, \ldots, d\}$ and $\mathcal{O} = \{d + 1, \ldots, d + o\}$. For every $k \in \{1, \ldots, o + r\}$, the quadratic polynomial $g^{(k)}$ is of the following form:

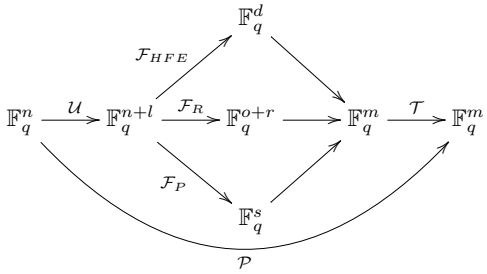
$$g^{(k)}(x_1, \ldots, x_{n'}) = \sum_{i \in \mathcal{O}, j \in \mathcal{V}} \alpha^{(k)} x_i x_j + \sum_{i,j \in \mathcal{V}, i \leq j} \beta_{i,j}^{(k)} x_i x_j + \sum_{i \in \mathcal{V} \cup \mathcal{O}} \gamma_i^{(k)} x_i + \eta^{(k)}$$

  where $\alpha^{(k)}$, $\beta_{i,j}^{(k)}$, $\gamma_i^{(k)}$, and $\eta^{(k)}$ are chosen at random from $\mathbb{F}_q$.
- The Plus modification is defined as $\mathcal{F}_P = (h^{(1)}, \ldots, h^{(s)}) : \mathbb{F}_q^{n'} \to \mathbb{F}_q^s$ which consists of s randomly generated quadratic polynomials.

An affine embedding $\mathcal{U} : \mathbb{F}_q^n \to \mathbb{F}_q^{n'}$ of full rank and an affine isomorphism $\mathcal{T} : \mathbb{F}_q^m \to \mathbb{F}_q^m$ are chosen for the butterfly construction as is common in big field schemes. The public key is given by $\mathcal{P} = \mathcal{T} \circ \mathcal{F} \circ \mathcal{U} : \mathbb{F}_q^n \to \mathbb{F}_q^m$, where $\mathcal{F} = \mathcal{F}_{HFE} \| \mathcal{F}_R \| \mathcal{F}_P$ ($\|$ being the concatination function), and the private key

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

is represented by the following figure:



Encryption Given a message $M \in \mathbb{F}_q^n$, the ciphertext is computed as $C = \mathcal{P}(M) \in \mathbb{F}_q^m$.
Decryption Given a ciphertext $C = (c_1, \ldots, c_m) \in \mathbb{F}_q^m$, the decryption process is the following:

1. Compute $\mathbf{x} = (x_1, \ldots, x_m) = \mathcal{T}^{-1}(C)$.
2. Compute $\mathbf{X} = \phi(x_1, \ldots, x_d) \in \mathbb{E}$.
3. Use the Berlekamp algorithm to compute the inverse of the HFE polynomials to recover $\mathbf{y} = (y_1, \ldots, y_d)$.
4. Given the vinegar values $y_1, \ldots, y_d$, solve the system of $o+r$ linear equations in the $n' - d = o$ variables $u_{d+1}, \ldots, u_{n'}$ given by

$$g^{(k)}(y_1, \ldots, y_d, u_{d+1}, \ldots, u_{n'}) = x_{d+k}$$

   for $k = 1, \ldots, o + r$. The solution is denoted $(y_{d+1}, \ldots, y_{n'})$.
5. Compute the plaintext $M \in \mathbb{F}_q^n$ by finding the preimage of $(y_1, \ldots, y_{n'})$ under the affine embedding $\mathcal{U}$.

6 Complexity of Attack

In this section we derive tight complexity estimates or proofs of resistance for the principal relevant attacks on HFERP. These attacks include the direct algebraic attack, the MinRank attack, the small field MinRank and dual rank attacks, and the invariant attack.

6.1 Algebraic Attack

The algebraic attack attempts to invert the public key at a ciphertext directly via the calculation of a Gröbner basis. It is commonly believed that the closeness of the solving degree of a polynomial system, the degree at which the Gröbner basis is resolved, and the degree of regularity, the degree at which a non-trivial syzygy producing a degree fall first occurs, is a generic property. Thus the lower bound on the complexity of the algebraic attack that the degree of regularity provides is likely a tight bound, and is consequently a critical quantity for analyzing the security of the scheme.

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

Theorem 1 *The degree of regularity of the public key of HFERP is bounded by*

$$d_{reg} \leq \begin{cases} \frac{(q-1)\lceil \log_q(D) \rceil}{2} + 2 \text{ if } q \text{ is odd or } \lceil \log_q(D) \rceil \text{ is even,} \\ \frac{(q-1)\left(\lceil \log_q(D) \rceil + 1\right)}{2} + 1 \text{ otherwise.} \end{cases}$$

Proof. There is a linear function of the public key separating the HFE polynomials $\mathcal{H}$ from the non-HFE polynomials $\mathcal{N}$. Trivially, the d_{reg} is bounded by the degree of regularity of the system $\mathcal{H}$, which, via [28, Theorem 4.2], produces the above bound.

One must note that the above bound is not what is needed to ensure security. Instead we require a lower bound. Extensive experimentation shows that for very small q, the above estimate is tight. We have, however, a further complication. In general, adding more polynomials to an ideal may decrease its degree of regularity. To address this issue we have conducted small scale experiments showing that the degree of regularity and solving degree behave similarly to those of random systems, see Section 7.

Conjecture 1 *Under the assumption that the degree of regularity is at least* $\lceil \log_q(D) \rceil + 2$ *for small odd* q *and sufficiently large* n*, the complexity of the algebraic attack is given by*

$$Comp._{alg} = \mathcal{O}\left(\binom{n + d_{reg}}{d_{reg}}^2 \binom{n}{2} \right) = \mathcal{O}\left(n^{2\lceil \log_q(D) \rceil + 6} \right).$$

6.2 MinRank Attack

The min-rank attack proposed in [16] is so successful due to the Q-rank of the squaring map within SRP being equal to one. By changing the square map component to an HFE core map, we are able to thwart such an attack on HFERP. This subsection walks through the attack proposed in [16] , with HFERP in mind, and proves that the min-Q-rank of HFERP differs from SRP.

Note that, similar to SRP, the public key of HFERP has an analogous scheme without embedding as long as $\pi_d \circ \mathcal{U}$ is of full rank, which it is defined to be in this scheme. Let $\pi'_d : \mathbb{F}_q^n \to \mathbb{F}_q^d$ be the projection onto the first d coordinates and find a projection $\rho : \mathbb{F}_q^{n+l} \to \mathbb{F}_q^n$ such that $\mathcal{U}' = \rho \circ \mathcal{U}$ has full rank and $\pi'_d \circ \mathcal{U}' = \pi_d \circ \mathcal{U}$. Let $\mathcal{F}^* : \mathbb{E} \to \mathbb{E}$ represent the chosen high Q-rank HFE core map so that $\mathcal{F}_{HFE} = \phi^{-1} \circ \mathcal{F}^* \circ \phi \circ \pi_d$. Then identify the Rainbow and random components as $\mathcal{F}'_R : \mathcal{F}_R \circ \mathcal{U} \circ {\mathcal{U}'}^{-1}$ and $\mathcal{F}'_P : \mathcal{F}_P \circ \mathcal{U} \circ {\mathcal{U}'}^{-1}$ respectively. Thus, one can see that

$$\mathcal{T} \circ \begin{bmatrix} \phi \circ \mathcal{F}^* \circ \phi^{-1} \circ \pi_d \\ \mathcal{F}_R \\ \mathcal{F}_P \end{bmatrix} \circ \mathcal{U} = \mathcal{T} \circ \begin{bmatrix} \phi \circ \mathcal{F}^* \circ \phi^{-1} \circ \pi'_d \\ \mathcal{F}'_R \\ \mathcal{F}'_P \end{bmatrix} \circ \mathcal{U}'.$$

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

Notice that the attack on SRP was not just a min-rank attack on the public key of SRP, but on a linear combination of public forms of SRP that had low Q-rank over the degree d extension used by the squaring component. This method allowed the attack to ignore the fact that the public key of an instance of SRP was expected to be of high rank. Thus, to demonstrate that HFERP resists such an attack, we briefly outline the method of deriving the linear combination of public forms from [16] for HFERP and prove that the min-Q-Rank of the result is sufficiently high to resist such an attack.

Let α be a primitive element of the degree d extension $\mathbb{E}$ of $\mathbb{F}_q$. Fix a vector space isomorphism $\phi : \mathbb{F}_q^d \to \mathbb{E}$ defined by $\phi(\bar{x}) = \sum_{i=0}^{d-1} x_i \alpha^i$. Then, fix a one dimensional representation $\Phi : \mathbb{E} \to \mathbb{A}$ defined by $a \overset{\Phi}{\mapsto} (a, a^q, \ldots, a^{q^{d-1}})$. Next, define $\mathcal{M}_d : \mathbb{F}_q^d \to \mathbb{A}$ by $\mathcal{M}_d = \Phi \circ \phi$. It was demonstrated you can look at this map through the following matrix representation

$$\mathbf{M}_d = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ \alpha & \alpha^q & \ldots & \alpha^{q^{d-1}} \\ \alpha^2 & \alpha^{2q} & \ldots & \alpha^{2q^{d-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^{d-1} & \alpha^{(d-1)q} & \ldots & \alpha^{(d-1)q^{d-1}} \end{bmatrix} \in \mathcal{M}_{d\times d}(\mathbb{E})$$

This matrix allows the passage from $\mathbb{F}_q^d$ and $\mathbb{A}$ easily by right multiplication with $\mathbf{M}_d$ or $\mathbf{M}_d^{-1}$. Next are a few more definitions necessary to be able to look at a matrix representation of the public key:

$$\widetilde{\mathbf{M}}_d = \begin{bmatrix} \mathbf{M}_d & 0 \\ 0 & \mathbf{I}_{o+r+s} \end{bmatrix} \in \mathcal{M}_{m\times m}(\mathbb{E})$$

$$\widehat{\mathbf{M}}_d = \begin{bmatrix} \mathbf{M}_d \\ \mathbf{0}_{o\times d} \end{bmatrix} \in \mathcal{M}_{(d+o)\times d}(\mathbb{E})$$

Finally, define $\mathbf{F}^{*i}$ be the matrix representation of the quadratic form over $\mathbb{A}$ of the i^{th} Frobenius power of the chosen HFE core map. Now we have all the necessary notation to view the public key as a matrix equation.

Denote the m-dimensional vector of $(d+o)\times(d+o)$ symmetric matrices associated by the private key as follows:

$$(\mathbf{F}_{(HFE,0)}, \ldots, \mathbf{F}_{(HFE,d-1)}, \mathbf{F}_{(R,0)}, \ldots, \mathbf{F}_{(R,o+r-1)}\mathbf{F}_{(P,0)}, \ldots, \mathbf{F}_{(P,s-1)}). \quad (2)$$

Note that the function corresponding to the application of each coordinate of a vector of the quadratic forms followed by the application of a linear map represented by a matrix is denoted as a right product of the vector and a matrix representation of the linear map.

Next, observe

$$(\mathbf{F}_{(HFE,0)}, \ldots, \mathbf{F}_{(HFE,d-1)})\mathbf{M}_d = (\widehat{\mathbf{M}}_d\mathbf{F}^{*0}\widehat{\mathbf{M}}_d^\top, \ldots, \widehat{\mathbf{M}}_d\mathbf{F}^{*(d-1)}\widehat{\mathbf{M}}_d^\top),$$

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

which yields

$$(\bar{x}\mathbf{F}_{(HFE,0)}\bar{x}^\top,\ldots,\bar{x}\mathbf{F}_{(HFE,d-1)}\bar{x}^\top)\mathbf{M}_d = \\ (\bar{x}\widehat{\mathbf{M}}_d\mathbf{F}^{*0}\widehat{\mathbf{M}}_d^\top\bar{x}^\top,\ldots,\bar{x}\widehat{\mathbf{M}}_d\mathbf{F}^{*(d-1)}\widehat{\mathbf{M}}_d^\top\bar{x}^\top),$$

as a function of $\bar{x}$. This gives the following equation:

$$(\mathbf{F}_{(HFE,0)},\ldots,\mathbf{F}_{(HFE,d-1)},\mathbf{F}_{(R,0)},\ldots,\mathbf{F}_{(P,s-1)})\widetilde{\mathbf{M}}_d = \\ (\widehat{\mathbf{M}}_d\mathbf{F}^{*0}\widehat{\mathbf{M}}_d^\top,\ldots,\widehat{\mathbf{M}}_d\mathbf{F}^{*(d-1)}\widehat{\mathbf{M}}_d^\top,\mathbf{F}_{(R,0)},\ldots,\mathbf{F}_{(P,s-1)}) \tag{3}$$

Now, look to the relation between the public key and its corresponding private key central maps:

$$(\mathbf{P}_0,\ldots,\mathbf{P}_{m-1})\mathbf{T}^{-1} = (\mathbf{U}\mathbf{F}_{(HFE,0)}\mathbf{U}^\top,\ldots,\mathbf{U}\mathbf{F}_{(P,s-1)}\mathbf{U}^\top). \tag{4}$$

By combining equations 3 and 4, we have the following:

$$(\mathbf{P}_0,\ldots,\mathbf{P}_{m-1})\mathbf{T}^{-1}\widetilde{\mathbf{M}}_d = \\ (\mathbf{U}\widehat{\mathbf{M}}_d\mathbf{F}^{*0}\widehat{\mathbf{M}}_d^\top\mathbf{U}^\top,\ldots,\mathbf{U}\widehat{\mathbf{M}}_d\mathbf{F}^{*(d-1)}\widehat{\mathbf{M}}_d^\top\mathbf{U}^\top,\mathbf{U}\mathbf{F}_{(R,0)}\mathbf{U}^\top,\ldots,\mathbf{U}\mathbf{F}_{(P,s-1)}\mathbf{U}^\top)$$

As in [16], let $\widehat{\mathbf{T}} = \mathbf{T}^{-1}\widetilde{\mathbf{M}}_d = [t_{i,j}] \in \mathcal{M}_{m\times m}(\mathbb{E})$ and $\mathbf{W} = \mathbf{U}\widehat{\mathbf{M}}_d$. This identification produces

$$\sum_{i=0}^{m-1} t_{i,0}\mathbf{P}_i = \mathbf{W}\mathbf{F}^{*0}\mathbf{W}^\top. \tag{5}$$

Since the rank of $\mathbf{F}^{*i}$ is equal to the Q-rank of the quadratic form of the HFE core map for all i, the rank of this $\mathbb{E}$-linear combination of the public matrices is bounded by the minimum of the rank of $\mathbf{U}\widehat{\mathbf{M}}_d$ and the rank of $\mathbf{F}^{*0}$, *id est* the Q-rank of our HFE core map. This statement forms the following theorem:

Theorem 2 *The min-Q-rank of the public key P of HFERP(q,d,o,r,s,l) is given by:*

$$\textit{min-Q-rank(P)} \leq \textit{min}\{\textit{Rank}(\mathbf{U}\widehat{\mathbf{M}}_d), \textit{Rank}(\mathbf{F}^{*0})\}$$

Proof. The proof in [16] describes the parameters in which the min-Q-rank(P) can be equal to zero. So, we move forward with the assumption that $\mathbf{U}\widehat{\mathbf{M}}_d \neq 0$, which occurs with high probability when $d > l$. In (5) we have a linear combination of the public key equations equal to the following:

$$\mathbf{W}\mathbf{F}^{*0}\mathbf{W}^\top = \mathbf{U}\widehat{\mathbf{M}}_d\mathbf{F}^{*0}\widehat{\mathbf{M}}_d^\top\mathbf{U}^\top. \tag{6}$$

This proves our result.

It should be noted that $\mathbf{U}$, $\widehat{\mathbf{M}}_d$, and $\mathbf{F}^{*0}$ are chosen by the user. They can easily be chosen in such a way such that

$$\text{min-Q-rank(P)} = \min\{\textit{Rank}(\mathbf{U}\widehat{\mathbf{M}}_d), \textit{Rank}(\mathbf{F}^{*0})\}.$$

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

This would also occur with high probability if $\mathbf{U}$, $\widehat{\mathbf{M}}_d$, and $\mathbf{F}^{*0}$ were randomly generated. Directly from [15], we also have the following complexity for the MinRank attack on HFERP:

Corollary 1 *The complexity of the MinRank attack with minors modeling on HFERP is given by*

$$Comp._{Minors} = \mathcal{O}\left(\binom{m + \lfloor \log_q(D) \rfloor}{\lceil \log_q(D) \rceil}^2 \binom{m}{2}\right) = \mathcal{O}\left(m^{2\lceil \log_q(D) \rceil + 2}\right).$$

6.3 Base-Field Rank and Invariant Attacks

Variants of several attacks applicable to other versions of the Rainbow cryptosystem are applicable to HFERP. These include the linear-algebra-search version of MinRank [29], the HighRank attack [29] and the UOV invariant attack [4].

The MinRank attack works by randomly choosing one or more vectors $\mathbf{w}_j$ in the plaintext space and solving for a linear combination $t_i \in \mathbb{F}$ of the plaintext equations satisfying:

$$\sum_{i=1}^{m} t_i D f_i(\mathbf{w}_j) = 0$$

The attack succeeds when $\mathbf{w}_j$ is in the kernel of a low rank linear combination of differentials of the public polynomials. In the case of HFERP, the HFE component equations form a d-dimensional subspace of the public equations having rank d over $\mathbb{F}$. Note that the attacker can remove up to $d-1$ equations while preserving at least a one dimensional subspace of low rank maps. Thus, the attack can succeed with a one dimensional solution space for t_i and only a single $\mathbf{w}_j$ as long as $m \leq n + d$.

If $m > n + d$, the adversary may still use a single vector $\mathbf{w}_j$ to constrain the t_i's rather than attempting to find two vectors in the kernel of the HFE equations. In this case, the attacker must search through an $m - n - d + 1$ dimensional space of spurious solutions to find the useful 1 dimensional space of t_is. This method is still less expensive than searching for two vectors in the kernel of the HFE equations when $m < n + 2d$.

It should be further noted that, since the differentials of the oil maps will map any vector in the kernel of the HFE equations to the d-dimensional HFE input space, we expect an $o_1 + r_1 - d$ dimensional subspace of the oil equations to also have such a vector in the kernel of their differentials, see Figure 1. Thus, when $m < n + \max(d, o_1 + r_1)$, vectors in the HFE kernel can be recognized, because they are in the kernel of an unusually large subspace of the public equations, and when $2d < n$ the linear combinations of the public equations from the HFE and oil spaces can be recognized due to their low rank.

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

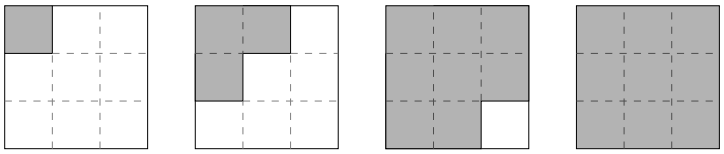


Fig. 1. The shape of the matrix representations of the central maps of HFERP. The shaded regions represent possibly nonzero values while unshaded areas have coefficients of zero.

Thus the complexity of MinRank (for plausible choices of m) is

$$Comp._{MinRank} = \begin{cases} \mathcal{O}\left(q^d m^\omega\right) & m < n + \max(d, o_1 + r_1) \\ \mathcal{O}\left(q^{d+m-n-\max(d,o_1+r_1)} n^\omega\right) & m \geq n + \max(d, o_1 + r_1) \\ & m < n + d + \max(d, o_1 + r_1) \\ & n > 2d \\ \mathcal{O}\left(q^{m-n} n^\omega\right) & m \geq n + \max(d, o_1 + r_1) \\ & m < n + 2d \\ & n \leq 2d \\ \mathcal{O}\left(q^{2d} m^\omega\right) & m < 2n + \max(d, o_1 + r_1 - d) \\ & \text{No better attack.} \end{cases}$$

In the HighRank attack, the attacker randomly selects linear combinations of the public polynomials with the hope of selecting a polynomial with significantly less than full rank. This attack takes advantage of the $d + o_1 + r_1$-dimensional subspace of the public polynomials generated by the HFE maps and either the Rainbow-1 maps of Figure 1 or for UOV of the d-dimensional HFE subspace. The complexity of the attack is then:

$$Comp._{HighRank} = \mathcal{O}\left(q^{m-d-o_1-r_1} n^\omega\right).$$

It should also be noted that linear combinations of HFE and Rainbow-1 polynomials form an $m - s$ dimensional subspace of the public polynomials, that act linearly on the $o_2 - l$ dimensional preimage under $\mathcal{U}$ of the oil subspace. This bounds their rank to be at most $2d$. Noting that the probability that a random square matrix has corank a is approximately q^{-a^2}, we see that, the high rank attack can be straightforwardly applied if $2d < n - \sqrt{m - d - o_1 - r_1}$.

Additionally, the HighRank attack can be combined with the oil and vinegar invariant attack to distinguish linear combinations of the HFE and Rainbow

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

maps from other linear combinations of the public maps. Here, a pair of maps from the HFE and Rainbow subspace can be identified by restricting their differentials to a subspace of the plaintext space in which both maps are full rank, and checking to see if $(Dp_1)^{-1}Dp_2$ has a large invariant subspace (which will be the intersection of the preimage of the oil subspace under $\mathcal{U}$ and the subspace used to restrict the differentials). This allows the high rank attack to be applied with similar complexity as long as $2d < n - \sqrt{\frac{m-d-o_1-r_1}{2}}$: Applying the attack will involve testing no more than $\left(q^{\frac{m-d-o_1-r_1}{2}}\right)^2 = q^{m-d-o_1-r_1}$ pairs of rank $n-2d$ maps, and therefore this step will not dominate the complexity of the approximately $q^{m-d-o_1-r_1}$ rank computations involved in the HighRank step.

If $2d \geq \zeta$, where $\zeta_1 = n - \sqrt{\frac{m-d-o_1-r_1}{2}}$, the complexity of HighRank is given by:

$$Comp._{HighRank} = \begin{cases} Comp._{HighRank} = \mathcal{O}\left(q^{m-d}n^{\omega}\right) & 2d \geq \zeta_1 \\ Comp._{HighRank} = \mathcal{O}\left(q^{m-d-o_1-r_1}n^{\omega}\right) & 2d < \zeta_1. \end{cases}$$

Finally, when $2d \geq n - \sqrt{\frac{m-d-o_1-r_1}{2}}$, as in the UOV attack, the previous steps must be combined with a projection, aimed at removing enough vinegar variables that the restriction of the differentials of linear combinations of HFE and Rainbow maps to the projected plaintext space is less than full rank. This yields a complexity for hybrid HighRank/UOV invariant type attacks of:

$$Comp._{UOV} = \begin{cases} \mathcal{O}\left(q^{m-d-o_1-r_1}n^{\omega}\right) & n > \zeta_2 \\ \mathcal{O}\left(q^{m-d-o_1-r_1+\sqrt{\frac{m-d-o_1-r_1}{2}}+2d-n}(o_1+o_2-l)^4\right) & n \leq \zeta_2. \end{cases}$$

where $\zeta_2 = 2d + \sqrt{\frac{m-d-o_1-r_1}{2}}$. This attack may also be applied to the Rainbow-2 maps of Figure 1 in which case the complexity is:

$$Comp._{UOV2} = \begin{cases} \mathcal{O}\left(q^s n^{\omega}\right) & n > 2d + 2o_1 + \sqrt{\frac{s}{2}} \\ \mathcal{O}\left(q^{s+\sqrt{\frac{s}{2}}+2d+2o_1-n}(o_2-l)^4\right) & n \leq 2d + 2o_1 + \sqrt{\frac{s}{2}}. \end{cases}$$

7 Parameter Selection and Experimental Results

We propose single-layer parameters (A) and (B) for 80-bit security and multi-layer parameters (C) and (D) for 128-bit security :

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

(A) $(q = 3, d = 42, o = 21, r = 15, s = 17, l = 0, D = 3^7 + 1)$

(B) $(q = 3, d = 63, o = 21, r = 11, s = 10, l = 0, D = 3^7 + 1)$

(C) $(q = 3, d = 85, o_1 = o_2 = 70, r_1 = r_2 = 89, s = 61, l = 0, D = 3^7 + 1)$

(D) $(q = 3, d = 60, o_1 = o_2 = 40, r_1 = r_2 = 23, s = 40, l = 0, D = 3^9 + 1)$

Then we have the following values for (n, m): $(63, 95)$ for (A), $(84, 105)$ for (B), $(225, 464)$ for (C), and $(140, 226)$ for (D). The security level for suggested parameters is estimated by all the attack in §6. Here, we assume that the degree of regularity for direct attack is 10 by Conjecture 1 for (A),(B), and (C) while it is 12 for (D).

To draw a direct comparison with HFE, note that to achieve the same security level as HFERP, an HFE scheme requires m equations, and hence $n = m$ variables. Therefore secure HFE public keys are far larger while offering slower decryption due to the use of the Berlekamp algorithm in a far larger field.

We ran a series of experiments with Magma, see [30], on a 2.6 GHz Intel® Xeon$^{\mathrm{R}}$ CPU[1]. These are not optimized implementations.

	(A)	(B)	(C)	(D)
Key Generation	0.299 s	0.572 s	20.498 s	3.43 s
Encryption	0.001 s	0.001 s	0.006 s	0.001 s
Decryption	3.977 s	8.671 s	49.182 s	124.27 s
Secret Key Size	19.8KB	31.7KB	1344.0KB	226.0KB
Public Key Size	48.2KB	93.6KB	2905.7KB	552.3KB

Table 1. Experimental results for HFERP.

We also investigated the growth of the first fall degree (d_{reg}) as well as the solving degree with five experiments performed at each of eight different parameters sets. We directly compared these data with randomly generated systems, see Table 2.

For comparison, we include the semi-regular degree for systems of m equations in n variables. This quantity was calculated by computing the first non-positive coefficient in the series

$$S_{n,m}(t) = \frac{(1 - t^q)^n (1 - t^2)^m}{(1 - t)^n (1 - t^{2q})^m}.$$

[1] Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

Noting that the degree of regularity of the zero-dimensional ideal is the same as the first fall degree of the ideal generated by the homogeneous components of the generators of highest degree. We derive the above formula as the fusion of the techniques in [31] and [32].

It is clear that the degree of regularity of the small scale instances of HFERP grows in relation to that of random schemes. By the data in the tables, we can estimate that the degree of regularity for direct attack on (A) and (B) is greater than 9 at least.

Table 2. Direct attack experiment data for various values of d, o, r, s. (s.r.d. stands for semi- regular degree)

			HFERP		Random		
(q, d, o, r, s, l, D)	n	m	d_{reg}	sol. deg	d_{reg}	sol. deg	s.r.d.
$(3, 8, 4, 3, 3, 0, 2188)$	12	18	$4, 4, 4, 4, 4$	$4, 4, 4, 4, 4$	$4, 4, 4, 4, 4$	$4, 4, 4, 4, 4$	4
$(3, 10, 5, 4, 3, 0, 2188)$	15	22	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	5
$(3, 12, 6, 5, 4, 0, 2188)$	18	27	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	5
$(3, 14, 7, 5, 5, 0, 2188)$	21	31	$6, 5, 5, 5, 5$	$6, 6, 6, 6, 6$	$5, 5, 5, 5, 5$	$6, 6, 6, 6, 6$	6

Table 2.A. Direct Attack, $d = 2o, d + o \fallingdotseq 2(r + s), o = 4, 5, 6, 7$

			HFERP		Random		
(q, d, o, r, s, l, D)	n	m	d_{reg}	sol. deg	d_{reg}	sol. deg	s.r.d.
$(3, 9, 3, 2, 2, 0, 2188)$	12	16	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	5
$(3, 12, 4, 2, 2, 0, 2188)$	16	20	$5, 6, 6, 5, 5,$	$5, 6, 6, 6, 5$	$6, 5, 6, 6, 5$	$6, 6, 6, 6, 6$	6
$(3, 15, 5, 3, 3, 0, 2188)$	20	26	$6, 5, 5, 5, 5$	$6, 6, 6, 6, 6$	$5, 5, 5, 6, 5$	$6, 6, 6, 6, 6$	6
$(3, 18, 6, 3, 3, 0, 2188)$	24	30	$5, 5, 5, 5, 5$	$7, 7, 7, 7, 7$	$5, 5, 5, 5, 7$	$7, 7, 7, 7, 7$	7

Table 2.B. Direct Attack, $d = 3o, r + s \fallingdotseq o, o = 3, 4, 5, 6$

			HFERP		Random		
(d, o, r, s, l, D)	n	m	d_{reg}	sol. deg	d_{reg}	sol. deg	s.r.d.
$(3, (3, 3), (4, 4), 2, 0, 2188)$	9	19	$3, 3, 3, 3, 3$	$3, 3, 2, 3, 2$	$3, 3, 3, 3, 3$	$2, 3, 3, 2, 2$	3
$(7, (6, 6), (7, 7), 5, 0, 2188)$	19	38	$4, 4, 4, 4, 4$	$4, 4, 4, 4, 4$	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	5
$(10, (8, 8), (11, 11), 7, 0, 2188)$	26	55	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	5
$(14, (11, 11), (14, 14), 10, 0, 2188)$	36	74	5		5		6

Table 2.C. Direct Attack,
$d \fallingdotseq 3.4a, o \fallingdotseq (2.8a, 2.8a), r \fallingdotseq (3.56a, 3.56a), s \fallingdotseq 2.44a, a = 1, 2, 3, 4$

			HFERP		Random		
(d, o, r, s, l, D)	n	m	d_{reg}	sol. deg	d_{reg}	sol. deg	s.r.d.
$(5, (3, 3), (2, 2), 3, 0, 3^9 + 1)$	11	18	$4, 4, 4, 4, 4$	$4, 4, 4, 4, 4$	$4, 4, 4, 4, 4$	$4, 4, 4, 3, 4$	4
$(7, (5, 5), (3, 3), 5, 0, 3^9 + 1)$	17	28	$4, 4, 4, 4, 4$	$4, 4, 4, 4, 4$	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	5
$(10, (6, 6), (4, 4), 6, 0, 3^9 + 1)$	22	36	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	$5, 5, 5, 5, 5$	$6, 6, 6, 6, 6$	6
$(12, (8, 8), (5, 5), 8, 0, 3^9 + 1)$	28	46	$5, 5$	$6, 6$	$5, 5$	6	6

Table 2.D. Direct Attack,
$d \fallingdotseq 2.4a, o \fallingdotseq (1.6a, 1.6a), r \fallingdotseq (0.92a, 0.92a), s \fallingdotseq 1.6a, a = 2, 3, 4, 5$

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

8 Conclusion

SRP was an ambitious encryption scheme attempting to combine the efficiency of the inversion of Square with the security of Rainbow to achieve security with a small blow-up factor between the plaintext and ciphertext. Unfortunately, this technique was a bit too ambitious.

Interestingly, the idea of replacing Square with a more general and higher Q-rank HFE primitive seems to solve this problem. Even more interestingly, the resulting scheme, HFERP, though in principle assailable via essentially every major cryptanalytic technique available in multivariate cryptography, appears to be out of range of these myriad attacks.

The parameter ℓ in SRP was introduced for efficiency, attempting to reduce the public key size while maintaining the algebraic structure of the scheme. We have found that this quantity adds nothing to security and have set it equal to zero for our suggested parameters. An interesting possible future problem is to determine whether ℓ can be securely set to a value larger than zero and thereby reduce public key size. For now, we err on the side of caution, and conservatively use all of the entropy we can get.

9 Acknowledgments

The first and fourth authors were supported by JST CREST (Grant Number JPMJCR14D6).

References

1. Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J. Sci. Stat. Comp. **26, 1484** (1997)
2. Mosca, M.: Cybersecurity in a quantum world: will we be ready? Workshop on Cybersecurity in a Post-Quantum World, Invited Presentation (2015) https://csrc.nist.gov/csrc/media/events/workshop-on-cybersecurity-in-a-post-quantum-world/documents/presentations/session8-mosca-michele.pdf.
3. Yasuda, T., Sakurai, K. In: A Multivariate Encryption Scheme with Rainbow. Springer International Publishing, Cham (2016) 236–251
4. Kipnis, A., Patarin, J., Goubin, L.: Unbalanced oil and vinegar signature schemes. EUROCRYPT 1999. LNCS **1592** (1999) 206–222
5. Ding, J., Schmidt, D.: Rainbow, a new multivariable polynomial signature scheme. ACNS 2005, LNCS **3531** (2005) 164–175
6. Petzoldt, A., Chen, M.S., Yang, B.Y., Tao, C., Ding, J. In: Design Principles for HFEv- Based Multivariate Signature Schemes. Springer Berlin Heidelberg, Berlin, Heidelberg (2015) 311–334
7. Patarin, J.: Hidden Field Equations (HFE) and Isomorphisms of Polynomials: two new Families of Asymmetric Algorithms. Eurocrypt '96, Springer **1070** (1996) 33–48
8. Tao, C., Diene, A., Tang, S., Ding, J.: Simple matrix scheme for encryption. In Gaborit, P., ed.: PQCrypto. Volume 7932 of Lecture Notes in Computer Science., Springer (2013) 231–242

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

9. Ding, J., Petzoldt, A., Wang, L.: The cubic simple matrix encryption scheme. [33] 76–87
10. Porras, J., Baena, J., Ding, J.: ZHFE, A new multivariate public key encryption scheme. [33] 229–245
11. Moody, D., Perlner, R.A., Smith-Tone, D.: An asymptotically optimal structural attack on the ABC multivariate encryption scheme. [33] 180–196
12. Moody, D., Perlner, R.A., Smith-Tone, D.: Key recovery attack on the cubic abc simple matrix multivariate encryption scheme. In: Selected Areas in Cryptography – SAC 2016: 23rd International Conference, Revised Selected Papers, LNCS, Springer (2017)
13. Moody, D., Perlner, R.A., Smith-Tone, D.: Improved attacks for characteristic-2 parameters of the cubic ABC simple matrix encryption scheme. [34] 255–271
14. Cabarcas, D., Smith-Tone, D., Verbel, J.A.: Key recovery attack for ZHFE. [34] 289–308
15. Vates, J., Smith-Tone, D.: Key recovery attack for all parameters of HFE-. [34] 272–288
16. Perlner, R.A., Petzoldt, A., Smith-Tone, D. In: Total Break of the SRP Encryption Scheme. Springer, In press. (2017)
17. Patarin, J.: Hidden Fields Equations (HFE) and Isomorphisms of Polynomials (IP): Two New Families of Asymmetric Algorithms. In: EUROCRYPT. (1996) 33–48
18. Matsumoto, T., Imai, H.: Public quadratic polynomial-tuples for efficient signature verification and message-encryption. Eurocrypt '88, Springer **330** (1988) 419–545
19. Berlekamp, E.R.: Factoring polynomials over large finite fields. Mathematics of Computation **24** (1970) pp. 713–735
20. Clough, C., Baena, J., Ding, J., Yang, B.Y., Chen, M.S.: Square, a New Multivariate Encryption Scheme. In Fischlin, M., ed.: CT-RSA. Volume 5473 of Lecture Notes in Computer Science., Springer (2009) 252–264
21. Patarin, J.: The oil and vinegar algorithm for signatures. Presented at the Dagsthul Workshop on Cryptography (1997)
22. Patarin, J., Goubin, L., Courtois, N.: C^*_{-+} and HM: Variations Around Two Schemes of T. Matsumoto and H. Imai. In Ohta, K., Pei, D., eds.: ASIACRYPT. Volume 1514 of Lecture Notes in Computer Science., Springer (1998) 35–49
23. Shamir, A., Kipnis, A.: Cryptanalysis of the oil & vinegar signature scheme. CRYPTO 1998. LNCS **1462** (1998) 257–266
24. Moody, D., Perlner, R.A., Smith-Tone, D.: Key recovery attack on the cubic ABC simple matrix multivariate encryption scheme. In Avanzi, R., Heys, H.M., eds.: Selected Areas in Cryptography - SAC 2016 - 23rd International Conference, St. John's, NL, Canada, August 10-12, 2016, Revised Selected Papers. Volume 10532 of Lecture Notes in Computer Science., Springer (2016) 543–558
25. Kipnis, A., Shamir, A.: Cryptanalysis of the HFE public key cryptosystem by relinearization. Advances in Cryptology - CRYPTO 1999, Springer **1666** (1999) 788
26. Faugère, J., Din, M.S.E., Spaenlehauer, P.: Computing loci of rank defects of linear matrices using gröbner bases and applications to cryptology. In Koepf, W., ed.: Symbolic and Algebraic Computation, International Symposium, ISSAC 2010, Munich, Germany, July 25-28, 2010, Proceedings, ACM (2010) 257–264
27. Bettale, L., Faugère, J., Perret, L.: Cryptanalysis of HFE, multi-HFE and variants for odd and even characteristic. Des. Codes Cryptography **69** (2013) 1–52

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

28. Ding, J., Hodges, T.J.: Inverting HFE systems is quasi-polynomial for all fields. In Rogaway, P., ed.: Advances in Cryptology - CRYPTO 2011 - 31st Annual Cryptology Conference, Santa Barbara, CA, USA, August 14-18, 2011. Proceedings. Volume 6841 of Lecture Notes in Computer Science., Springer (2011) 724–742
29. Goubin, L., Courtois, N.: Cryptanalysis of the ttm cryptosystem. In Okamoto, T., ed.: ASIACRYPT. Volume 1976 of Lecture Notes in Computer Science., Springer (2000) 44–57
30. Bosma, W., Cannon, J., Playoust, C.: The Magma algebra system. I. The user language. J. Symbolic Comput. **24** (1997) 235–265 Computational algebra and number theory (London, 1993).
31. Yang, B., Chen, J.: Theoretical analysis of XL over small fields. In Wang, H., Pieprzyk, J., Varadharajan, V., eds.: Information Security and Privacy: 9th Australasian Conference, ACISP 2004, Sydney, Australia, July 13-15, 2004. Proceedings. Volume 3108 of Lecture Notes in Computer Science., Springer (2004) 277–288
32. Bardet, M., Faugre, J., Salvy, B., Yang, B.: Asymptotic behaviour of the degree of regularity of semi-regular polynomial systems. In MEGA '05, 2005. Eighth International Symposium On Effective Methods In Algebraic Geometry (2005)
33. Mosca, M., ed.: Post-Quantum Cryptography - 6th International Workshop, PQCrypto 2014, Waterloo, ON, Canada, October 1-3, 2014. Proceedings. Volume 8772 of Lecture Notes in Computer Science., Springer (2014)
34. Lange, T., Takagi, T., eds.: Post-Quantum Cryptography - 8th International Workshop, PQCrypto 2017, Utrecht, The Netherlands, June 26-28, 2017, Proceedings. Volume 10346 of Lecture Notes in Computer Science., Springer (2017)

A Toy Example

The purpose of the following toy example is to help the reader understand the process of generating a public key for an instance of HFERP as well as an example of encryption and decryption. The parameters used are by no means secure and are soley for instructional purposes.

Parameters of this toy example are as follows: $q = 7,\ d = o = r = 2,\ s = 1,$ and $l = 0$. Then, construct $\mathbb{E}$ a degree 2 extension field over $\mathbb{F}_7$. The chosen HFE core map is $f = \xi^{12}X^{14} + \xi^6 X^8 + \xi^{29}X^2$ where $\xi \in \mathbb{E}$. Let $\mathcal{T}$ and $\mathcal{U}$ be the following affine maps:

$$\mathcal{T} = \begin{bmatrix} 2\,1\,2\,4\,5\,0\,3 \\ 1\,1\,3\,3\,4\,4\,4 \\ 4\,2\,1\,3\,1\,0\,6 \\ 0\,1\,0\,1\,5\,5\,5 \\ 5\,5\,3\,6\,4\,2\,4 \\ 2\,5\,1\,6\,5\,6\,0 \\ 1\,1\,2\,2\,6\,4\,3 \end{bmatrix}, \mathcal{U} = \begin{bmatrix} 4\,6\,6\,4 \\ 3\,2\,0\,2 \\ 1\,1\,6\,5 \\ 3\,6\,6\,6 \end{bmatrix}$$

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

With the parameters described above, $\mathcal{F}$ can be represented as the follwing matrices over $\mathbb{F}_7$

$$F_1 = \begin{bmatrix} 0&1&0&0 \\ 4&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \end{bmatrix}, F_2 = \begin{bmatrix} 0&3&0&0 \\ 1&6&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \end{bmatrix}, F_3 = \begin{bmatrix} 3&1&6&1 \\ 3&1&4&5 \\ 3&4&0&0 \\ 3&2&0&0 \end{bmatrix},$$

$$F_4 = \begin{bmatrix} 5&1&0&3 \\ 0&5&0&3 \\ 0&4&0&0 \\ 6&1&0&0 \end{bmatrix}, F_5 = \begin{bmatrix} 6&0&3&4 \\ 6&2&4&2 \\ 6&3&0&0 \\ 0&3&0&0 \end{bmatrix}, F_6 = \begin{bmatrix} 4&4&1&1 \\ 3&0&0&3 \\ 3&6&0&0 \\ 1&2&0&0 \end{bmatrix}, F_7 = \begin{bmatrix} 6&3&2&3 \\ 4&4&0&6 \\ 2&3&1&3 \\ 6&4&0&6 \end{bmatrix}$$

P_1 and P_2 represent the HFE component, $P_3 \to P_6$ represent the rainbow component, and P_7 represents the plus component. With the public key generated by $\mathcal{P} = \mathcal{T} \circ \mathcal{F} \circ \mathcal{U}$, its matrix form over $\mathbb{F}_7$ is:

$$P_1 = \begin{bmatrix} 1&1&2&5 \\ 1&2&3&2 \\ 3&2&4&4 \\ 3&3&0&3 \end{bmatrix}, P_2 = \begin{bmatrix} 0&2&0&6 \\ 4&5&2&0 \\ 6&3&3&4 \\ 3&1&2&2 \end{bmatrix}, P_3 = \begin{bmatrix} 2&3&1&4 \\ 4&5&4&5 \\ 3&5&5&1 \\ 5&1&0&6 \end{bmatrix},$$

$$P_4 = \begin{bmatrix} 0&6&0&2 \\ 1&3&0&2 \\ 5&1&5&1 \\ 5&3&0&5 \end{bmatrix}, P_5 = \begin{bmatrix} 4&3&2&3 \\ 6&5&2&4 \\ 4&3&1&5 \\ 5&2&4&5 \end{bmatrix}, P_6 = \begin{bmatrix} 1&4&2&2 \\ 3&3&6&2 \\ 5&4&0&0 \\ 3&5&5&4 \end{bmatrix}, P_7 = \begin{bmatrix} 1&3&6&0 \\ 0&3&4&0 \\ 1&2&4&2 \\ 2&1&6&4 \end{bmatrix}$$

Given the following plaintext, $(2, 6, 1, 5)$, the resulting ciphertext is $(0, 0, 1, 3, 0, 4, 0)$.

Decryption: Given a ciphertext $(0, 0, 1, 3, 0, 4, 0)$, the following process is how you would obtain its corresponding plaintext.
Part of the secrect key:

$$\mathcal{T}^{-1} = \begin{bmatrix} 1&6&4&2&2&2&5 \\ 5&4&4&6&0&5&2 \\ 5&3&5&2&3&2&4 \\ 5&6&5&5&2&1&1 \\ 2&5&4&2&1&5&2 \\ 2&5&6&6&3&5&5 \\ 1&2&5&4&4&0&5 \end{bmatrix}, \mathcal{U}^{-1} = \begin{bmatrix} 4&5&2&1 \\ 3&1&3&1 \\ 4&1&2&0 \\ 5&6&1&1 \end{bmatrix}$$

Feed the ciphertext through $\mathcal{T}^{-1}$ to get

$$(0, 6, 2, 6, 0, 4, 6) \tag{7}$$

The first $d = 2$ elements are the corresponsing HFE outputs. Take these elements and adjust the HFE core map as follows:

$$f := f - 0\xi^{1-1} - 6\xi^{2-1} = \xi^{12}X^{14} + \xi^{6}X^{8} + \xi^{29}X^{2} + \xi$$

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

Perform the Berlekamp algorithm to find the preimage of f. In doing so in this toy example, you get $(0, 6)$. Next, construct the vector:

$$\overline{u} = [0, 6, u_1, u_2]\,.$$

Construct equations of the form $\overline{u}F_1\overline{u}^\top = x_i$ where x_i refers to the i^{th} element of (7), for $i \in \{3, 4, 5, 6\}$. This will result with the following equations:

$$\begin{bmatrix} 6u_1 + 1 \\ 3u_1 + 3u_2 + 5 \\ 2u_2 + 2 \\ u_1 + 2u_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 6 \\ 0 \\ 4 \end{bmatrix}$$

Solving this system of equations gives us $u_1 = 6$ and $u_2 = 6$. Thus,

$$\overline{u} = [0, 6, 6, 6]\,.$$

Finally, feed this through $\mathcal{U}^{-1}$ to get the plaintext, $[2, 6, 1, 5]$.

Ikematsu, Yashuhiko; Perlner, Ray; Smith-Tone, Daniel; Takagi, Tsuyoshi; Vates, Jeremy.
"HFERP - A New Multivariate Encryption Scheme."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

Managed Blockchain Based Cryptocurrencies with Consensus Enforced Rules and Transparency

Peter Mell
National Institute of Standards and Technology
Gaithersburg, Maryland 20899
Email: peter.mell@nist.gov

***Abstract*—Blockchain based cryptocurrencies are usually unmanaged, distributed, consensus-based systems in which no single entity has control. Managed cryptocurrencies can be implemented using private blockchains but are fundamentally different as the owners have complete control to do arbitrary activity without transparency (since they control the mining). In this work we explore a hybrid approach where a managed cryptocurrency is maintained through distributed consensus based methods. The currency administrator can perform ongoing management functions while the consensus methods enforce the rules of the cryptocurrency and provide transparency for all management actions. This enables the introduction of money management features common in fiat currencies but where the managing entity cannot perform arbitrary actions and transparency is enforced. We thus eliminate the need for users to trust the currency administrator but also to enable the administrator to manage the cryptocurrency. We demonstrate how to implement our approach through modest modifications to the implicit Bitcoin specification, however, our approach can be applied to most any blockchain based cryptocurrency using a variety of consensus methods.**

***Index Terms*—cryptocurrency, blockchain, managed, trust**

I. Introduction

Blockchain based cryptocurrencies are usually unmanaged, distributed, consensus-based systems in which no single entity has control [1]. They use open consensus based approaches that allow anyone to participate in maintaining the blockchain, even retaining their anonymity. Such systems remove the need for a third party in financial transactions and eliminate the double spending problem (where the same digital cash is spent multiple times) [2]. This lack of a need for a trusted third party is supposed to result in reduced transaction fees over non-cryptocurrency based systems (e.g., credit cards), enabling efficient micropayments [3]. Recently however, limitations with some cryptocurrencies on transaction throughput has caused transaction fees to be high. Lastly, such systems generally provide a level of anonymity where individuals are not linked to accounts and where it is trivial for an individual to produce and use new accounts. Examples of such systems include Bitcoin [4], Ethereum [5], Bitcoin Cash [6], Litecoin [7], Cardano [8], NEM [9], Dash [10][1].

In this work, we consider how to bring many of the advantages of such open consensus based cryptocurrencies to the area of managed cryptocurrencies[2]. We refer to a currency as 'managed' if there exists an owner that can exert control over the currency. Managed currencies include electronic representations of fiat currencies as well as virtual world and in-game currencies. In the cryptocurrency realm, they are often referred to as 'permissioned blockchains' (examples include Multichain [12] and Ripple). With managed currencies, the identity of individuals is often, but not necessarily, linked to the accounts (e.g., as when someone opens a bank checking account). Furthermore, the managing entity usually reserves the right to control the money supply (i.e., they can print money). And law enforcement related functions may include freezing or confiscating assets. Managed cryptocurrencies can be implemented with private blockchains using tools such as Multichain. However, in such implementations the owners have complete control to perform arbitrary activity without transparency. This is because the owners authorize (and thus control) the servers maintaining the blockchain.

In our research we explore a hybrid approach where we merge strengths of open consensus based cryptocurrencies with features often found in managed currencies. In doing so we design not a particular cryptocurrency, but instead a flexible architecture that allows for different implementations. From the open consensus approach we leverage the ability of the mining community to enforce the rules of the currency and to enforce transparency, where all transactions are publicly viewable. In this way the managing entity of the cryptocurrency cannot perform arbitrary actions, but only those explicitly allowed in the cryptocurrency design and all such management actions are publicly recorded in the blockchain. From the managed currencies, we leverage concepts such as the ability of the currency administrator to create funds, tie user identity to accounts, freeze/confiscate funds (e.g., due to illegal activity), and set the block awards for miners. This last feature indirectly enables the currency administrator to control the electricity consumption of the consensus mechanism (since fewer miners will participate if the rewards are lower). Energy consumption has often been cited as a major problem with consensus 'proof-of-work' systems; in 2014 Bitcoin mining consumed as much electricity as Ireland [13].

Since our approach is an architecture, the creator of any

[1]Any mention of commercial products is for information only; it does not imply recommendation or endorsement. The blockchain based cryptocurrencies listed are the ones with the largest market capitalization in descending order as of 2017-12-29 according to [11].

[2]Note that managed cryptocurrencies also use consensus methods but they are not open to public participation.

Mell, Peter.
"Managed Blockchain Based Cryptocurrencies with Consensus Enforced Rules and Transparency."
Paper presented at The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, NY, United States. August 1, 2018 - August 3, 2018.

particular managed cryptocurrency instance can choose which features to include or exclude. Our architecture is flexible such that it can be used to implement open consensus environments like Bitcoin as well as closed controlled environments achievable with systems like Multichain. However, our approach is not intended for that purpose. Our area of interest is where the architecture is used to create hybrid approaches that combine the strengths (and weaknesses) of both. Note that we are not advocating any particular approach in this work and our goal is not to propose the creation of any specific cryptocurrency. Rather, we explore here the technological foundations that can enable the merging of the managed cryptocurrency idea with an open consensus based architecture and explore the resultant strengths and weaknesses.

To enable management of the currency, we propose using a genesis transaction. All blockchains have a genesis block which is the first block, but this genesis transaction is a first transaction from which all subsequent transactions are authorized. The genesis transaction authorizes a special root account that has the currency manager role and that will be controlled by the currency administrator (the entity issuing the cryptocurrency). Our tagging of accounts with roles is key to our architecture. Accounts with the currency manager role can configure the currency to have different properties through defining policy (e.g., adjusting the roles implemented and mining rewards). Also, these accounts can issue transactions to create other accounts with different roles, in a hierarchical fashion with accounts closer to the root being more authoritative. The possible roles include currency manager, central banker, law enforcement, user, and account manager. The central bankers can create and delete funds. Law enforcement can freeze account and confiscate funds (e.g., for fraudulently gained funds being sent to terrorist organizations [14])[3]. Users can perform monetary transactions without the need for a trusted third party. And account managers can create user accounts (and may be required to link them to physical identities).

We demonstrate how to implement our approach through modest modifications to the implicit Bitcoin specification. We chose Bitcoin because it is was the first blockchain based cryptocurrency and is the most used. However, our approach can be applied to most any blockchain based cryptocurrency (including smart contract approaches such as Ethereum). We modify Bitcoin as little as possible to facilitate implementation of our specification; all of our features were implemented through small changes to the Bitcoin transaction format. Currency managers can issue policy in such a way that the changes are reversible or permanent. Permanent changes restrict the currency manager's future actions (since they cannot be undone). Such changes are important as they can provide users confidence in the system through knowledge that the currency administrator will abide by a set of self-established rules. Added to this, the architecture requires that all management actions be transparent to the users.

Key to this approach are our solutions for maintaining a balace of power. The consensus based methods must ensure that the currency administrator (who owns the root currency manager node) abides by the stated rules of the cryptocurrency and enforces transparency of all management actions. However, the participants in the consensus methods should not be able to take control away from the currency administrator nor exclude any management transactions from entering the blockchain.

In summary, open consensus based unmanaged cryptocurrencies provide significant new benefits over previous electronic cash efforts. They eliminate the need for trusted third parties by eliminating the double spending problem, remove the need for a dedicated and centralized infrastructure, and allow for the possibility of very low transaction fees thus enabling inexpensive micro-transactions [4]. However, this model is unsuitable for managed cryptocurrencies because it is completely controlled by whomever joins the cryptocurrency network to maintain the blockchain (an open and anonymous group). Previous efforts to support managed cryptocurrencies have used permission-based blockchains where the administrators can control all access to the blockchain, ability of users to issue transactions, and ability of miners to maintain the blockchain. This is a powerful and efficient paradigm for many use cases. However, the user base must have complete trust in the currency administrator. In our work, we are attempting to eliminate the need for users to trust the currency administrator but also to enable the administrator to manage the cryptocurrency. At the same time, we are attempting to incorporate the many benefits achieved by unmanaged cryptocurrencies while mitigating the weaknesses (especially in the area of power consumption in maintaining the blockchain).

The main deliverable this paper is a novel architecture for maintaining a managed cryptocurrency through distributed consensus based approaches (eliminating the need for users to trust the currency administrator), as well as an evaluation of the resultant benefits and weaknesses. It also provides technical bit-level details on how to modify the Bitcoin specification in order to implement the approach. In future work, we will provide such an implementation and perform empirical studies. We expect the necessary code changes to be relatively straightforward given our modest changes to the specification, but this cannot be claimed until a prototype implementation has been developed.

II. Related Work

To our knowledge, this is the only work combining the idea of a managed cryptocurrency with the open consensus model used by unmanaged currencies. The work most similar to ours is Multichain. It provides a platform for creating and deploying 'private' blockchains within or between organizations. It is designed to provide the following features [12]:

[3]Note that in most consensus based cryptocurrencies, restoration of funds is impossible without forking the currency.

[4]Bitcoin has high transaction fees due to limits on transaction throughput, but this is a technical problem not necessarily present in other cryptocurrencies.

Mell, Peter.
"Managed Blockchain Based Cryptocurrencies with Consensus Enforced Rules and Transparency."
Paper presented at The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, NY, United States. August 1, 2018 - August 3, 2018.

1) 'to ensure that the blockchain's activity is only visible to chosen participants'
2) 'to introduce controls over which transactions are permitted'
3) 'to enable mining to take place securely without proof of work and its associated costs'

Instances of Multichain have an administrator or group of administrators that define the ongoing policy of the system. They have complete control in defining who can view the blockchain, who can put transactions on the blockchain, and who can maintain the blockchain (those mining new blocks). This last feature enables them to maintain the blockchain at very little cost since the computationally expensive proof-of-work consensus methods of Bitcoin can be dispensed with. This is replaced with a flexible round robin approach where the miners mostly take turns publishing the new blocks and generally do not receive any reward for doing so (since the work is trivial).

While a powerful approach for organization-run blockchains, Multichain cannot be used to satisfy our stated objectives since the administrators have complete control. There is no mechanism to implement a balance of power where the administrators can manage the currency in an ongoing fashion but where the maintainers of the blockchain can ensure that the administrators follow the stated rules of the cryptocurrency.

Country specific managed cryptocurrencies exist or are in the process of being deployed, not all of them being blockchain based, and the degree to which they are 'managed' varies greatly. Dubai has launched its own cryptocurrency called emCash [15]. Singapore has announced experimentation with one [16] and Estonia has announced thier 'estcoin' [17]. The company Monetas [18] offers a product to enable countries to issue their own digital currencies; it is being actively used by several countries. Senegal is piloting a digital currency called eCFA using the Monetas platform that, if successful, will be used by Cote d'Ivoire, Benin, Burkina Faso, Mali, Niger, Togo and Lusophone Guinea Bissau [19]. Tunisia has done the same using the Monetas platform [20]. The Russian Central Bank has publicly pushed for a national cryptocurrency [21]. Venezuela has announced that it will launch an oil-backed cryptocurrency [22]. And lastly, the Bank for International Settlements released a report noting that countries may need to replace cash with national cryptocurrencies [16].

In the area of unmanaged cryptocurrencies, there exist hundreds of them. Bitcoin was the first to use blockchains and was introduced in 2008 [4]. There exist many forks and variants of Bitcoin, mostly optimizing certain features but often introducing novel and revolutionary architectural changes. We review here the blockchain based cryptocurrencies with the largest market capitalization, as of 2017-12-29. Ethereum was the first production product to enable executable programs (called smart contracts) to be put on a cryptocurrency blockchain [5]. Ripple [23] provides a solution for banks to send payments globally. Bitcoin Cash [6] is a fork of Bitcoin with a much larger block size limit. This enables many more transactions per block thereby increasing throughput and driving down transaction fees. Litecoin [7] is almost identical to Bitcoin but with several differences: smaller block publication time, larger maximum number of coins, and a change in hashing algorithm. Cardano [8] is based on [24] describing a 'provably secure proof-of-stake blockchain protocol'. NEM [9] incorporates a reputation system, proof-of-importance, and multisignature accounts. Dash [10] is 'privacy-centric' with a two-tiered administration network and an ability for users to instantly send coin.

III. Managed Cryptocurrency Architecture

All blockchains contain a 'genesis block'. This is the first block on the blockchain and it has no pointer to a previous block (being the first one). All users of the blockchain must agree on this first block for a consistent view of the blockchain to exist. We propose the addition of a 'genesis transaction'[5]. This is the first transaction in the blockchain and it defines an account that has the currency manager role (and is owned by the currency administrator). In our system, only accounts with roles can issue transactions and only accounts with the currency manager role can create other accounts with roles (with one important exception, discussed later). Thus, the genesis transaction is the transaction that enables all other transactions.

The initial account is the root of a hierarchical tree of nodes, where each node represents an account labeled with a set of roles[6]. The root node not only has the currency manager (M) role[7], but it has all other available roles: central banker (C), law enforcement (L), user (U), and account manager (A). We label the roles of an account by concatenating all applicable labels. Thus, the root node has the role set 'MCLUA'.

When a node with the M role creates a new account (more precisely, it labels some unlabeled account created by some user), it bestows on that account a, not necessarily proper, subset of its roles. Thus, the cardinality of the set labels for nodes monotonically decreases as one traverses higher in the hierarchy tree. One exception to this monotonicity rule is that nodes with the M label may also modify the role sets of nodes higher in the tree (provided they are on the path from the target node to the root), restricted again to the set of roles possessed.

Nodes with the A role may also create and delete accounts, but such created accounts may only have the U role. The currency administrator then can delegate user account management to third party organizations by giving them the A role.

The different roles provide different accounts different capabilities:

[5]This is related to the "asset genesis" metadata transaction idea [12] but is more powerful as it controls all transactions on the blockchain.

[6]We use the terms node and account interchangeably depending upon the desired perspective (node in a tree versus account owned by a user)

[7]The M role is distinct from the currency administrator. Many accounts may have the M role but there exists a single entity which is the currency administrator.

Mell, Peter.
"Managed Blockchain Based Cryptocurrencies with Consensus Enforced Rules and Transparency."
Paper presented at The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, NY, United States. August 1, 2018 - August 3, 2018.

- The U role enables an account to receive and spend coins. An account for which the U role has been removed has its funds frozen.
- The A role enables an node to create accounts with the U role (and only the U role). It may also remove the U label for its descendants.
- The C role enables the creation of new coins (apart from the block mining rewards).
- The L role enables an account to forcibly move funds between accounts, to remove the U label, and to restore a previously removed U label. However, these actions can only be performed against nodes with the same or greater distance from the root.

The currency administrator, who will own the root M labelled node, may require that A nodes verify users' identities prior to providing an account. In this case, the architecture enables a system where the 'know your customer' (KYC) laws might be satisfied. Individual transacting parties would not know each other's identities but some account authorizing entity would have a record for each account with the U role. Fulfilling KYC laws is a general problem for cryptocurrencies [25].

Figure 1 shows an example account hierarchy where we label nodes with their roles (e.g., a MUA node has the M, U, and A roles). The initial node created by the genesis transaction is at the bottom. Each node is labeled with its set of roles. Each UA node represents an organization authorized to manage user accounts. The MUA nodes authorize the UA nodes and can undo any undesired action taken by the UA nodes, since they are on the path from all UA nodes to the root. This action could be taken if there is negligence on the part of a UA node in creating U nodes or if a UA node's credentials are stolen. Note that there are two MUA nodes, one on top of the other. The topmost node will be used to create and delete UA nodes, the bottom one will be used to fix the system in the event that the topmost node's credentials are stolen. This is also the reason why there are two MCLUA nodes, one on top of the other. The root node ideally is never used again after creating the MCLUA node above it. This helps prevent the root node's credentials from being stolen. In general, actions should be performed by nodes higher up in the tree that have the least privilege possible since the use of a node puts it in a more vulnerable position. The credentials of nodes not used can be secured simply by converting them to physical form and locking them in a safe (which we recommend doing with the initial node's credentials). This hierarchical node and role structure then enables the currency administrator to create a defense in depth security model. Accounts lower in the hierarchy have greater power and their credentials should be locked securely and rarely used.

A last capability not yet discussed is that accounts with M roles can issue policy that alters the cryptocurrency specification. In the event of policy conflicts between different M nodes, the nodes closer to the root are more authoritative. For M nodes the same distance from the root, those labeled with the M role in earlier blocks are more authoritative. In the event of a tie, the node labeled with the M role first within the same block wins.

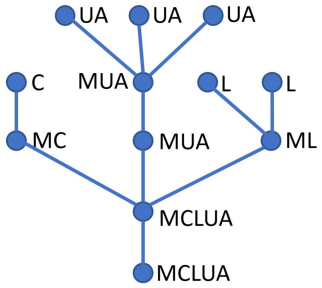


Fig. 1. Example Managed Cryptocurrency Hierarchy.

The policy deployed by the M nodes define the cryptocurrency. It is this policy that makes our approach an architecture. The policy can be set such that the cryptocurrency acts in an entirely unmanaged mode like the many popular open consensus cryptocurrencies in use today. The policy can also be set to allow the currency administrator full control as with the administrators in Multichain. More interesting to our research though is when the policy combines both open consensus and managed currency features. The policy enables each of the roles to be enabled or disabled and grants/limits the power of each role. Policy also can affect the mining community. A policy transaction can set a particular block reward or define a minimum transaction fee. Controlling these will affect the size of the mining community. For a proof-of-work based consensus mechanism such as Bitcoin, this will then indirectly control the amount of electricity used to manage the cryptocurrency (trading off power consumed against robustness of the mining pool against attack). This approach can enable an energy efficient proof-of-work consensus system where the currency administrator balances overall mining power desired vs. energy consumed. The exact capabilities available with policy are covered in section V-C.

IV. Bitcoin Specification Overview

There does not exist an official Bitcoin specification. The original Bitcoin paper [4] contained the primary architectural details but the specification is defined by the applications that maintain it on the network. That said, there exists a Bitcoin reference client 'bitcoind' and related protocol documentation [26]. From this was created a useful developers reference [27]. An in depth research analysis of Bitcoin is available in [28].

In this section we briefly review the features of the Bitcoin specification that will be of use for our modified specification. Figure 2 shows the layout of a Bitcoin transaction (copied from [27], see this for details). The vin[] sections describe the inputs to a Bitcoin transaction (the particular coins to be spent). The hash and n values specify particular coins from the output of some other Bitcoin transaction. The scriptSig is a script to provide cryptographic evidence that the owner of the coins approves of the coins being spent. It is a response

Mell, Peter.
"Managed Blockchain Based Cryptocurrencies with Consensus Enforced Rules and Transparency."
Paper presented at The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, NY, United States. August 1, 2018 - August 3, 2018.

Field name		Type (Size)	Description
nVersion		int (4 bytes)	Transaction format version (currently 1).
#vin		VarInt (1-9 bytes)	Number of transaction input entries in *vin*.
vin[]	hash	uint256 (32 bytes)	Double-SHA256 hash of a past transaction.
	n	uint (4 bytes)	Index of a transaction output within the transaction specified by *hash*.
	scriptSigLen	VarInt (1-9 bytes)	Length of *scriptSig* field in bytes.
	scriptSig	CScript (Variable)	Script to satisfy spending condition of the transaction output ($hash$,n).
	nSequence	uint (4 bytes)	Transaction input sequence number.
#vout		VarInt (1-9 bytes)	Number of transaction output entries in *vout*.
vout[]	nValue	int64_t (8 bytes)	Amount of 10^{-8} BTC.
	scriptPubkeyLen	VarInt (1-9 bytes)	Length of *scriptPubkey* field in bytes.
	scriptPubkey	CScript (Variable)	Script specifying conditions under which the transaction output can be claimed.
nLockTime		unsigned int (4 bytes)	Timestamp past which transactions can be replaced before inclusion in block.

Fig. 2. Bitcoin Transaction Format for Sending Bitcoin (BTC), copied from [27].

script that meets the conditions of the challenge script in the transaction containing the coins that are to be spent (see the vout[] scriptPubKey field below). These conditions are usually met by proving ownership of the private key associated with the coins.

The vout[] sections describe the outputs to a Bitcoin transaction (groupings of coins along with who owns each group). Ownership is specified within each scriptPubkey which is a script defining how the coins can be spent (usually specifying a public key). To satisfy the scripPubkey challenge script and spend the coins at some future time, the owner will need to generate a scriptSig response script in some vin[] field for some transaction in which they prove ownership of the private key associated with the specified public key. This is the Pay-to-Pubkey (P2PK) Bitcoin transaction type for moving coins between accounts (see section 4.3.1 of [27] for a detailed explanation).

Figure 3 shows how a vin[] field in a new transaction can reference a specific vout[] field in a previous transaction (copied from [27], see this for details). The vin[] hash value specifies the transaction and the n value specifies the specific vout[] field. The scriptSig in the vin[] of the new transaction then satisfies the scriptPubkey from the vout[] field specified from a previous transaction so that the coins can be spent (i.e., proving that the owner of the coins wants them spent).

V. Technical Design Using Bitcoin Specification Modifications

This section provides the technical specification for our managed cryptocurrency architecture described in section III. Our approach is to implement our architecture using only modest changes to the Bitcoin specification, changing the regular Bitcoin transaction format. Section IV provided the necessary background on the Bitcoin specification. Interested readers should also consult the de facto Bitcoin specifications [26] and [27] to better understand these changes in the context of the larger blockchain system.

To implement our architecture's functionality, we repurpose the regular Bitcoin transaction. The format remains the same as the Bitcoin transaction shown previously in figure 2 with a few exceptions. Our primary change is to leverage and revamp the vout[] nValue field in order to implement account roles and cryptocurrency policy. Another major change is to require in a transaction the inclusion of vin[] fields that provide the necessary roles for a transaction to be valid.

Our first modification was to change the transaction format version, nVersion, to 1944[8]. Transaction format version 1 is used by the regular Bitcoin transactions and is disallowed by our architecture.

The vin[] field operates similarly as before. In Bitcoin, a vin[] field specifies a set of coins from a particular transaction already posted on the blockchain. The vin[] field then provides the evidence that the owner of those coins wants to spend them by providing a vin[] scriptSig field that satisfies the vout[] scriptPubkey field of the coins to be spent. In our design, the vin[] field works the same way for coin transfers.

However, the vin[] field can also be used to bring roles into a transaction to authorize activities that require roles (which is most any activity in our architecture, depending upon the specific policy enacted). Functionally, it is like we are 'spending' a role to use it to authorize some action given the usual use of a vin[] field (but roles can be 'spent' an infinite number of times and are not transferred like coin). A vin[] field

[8]This is the year big band leader Glenn Miller died while flying to France to encourage allied troops.

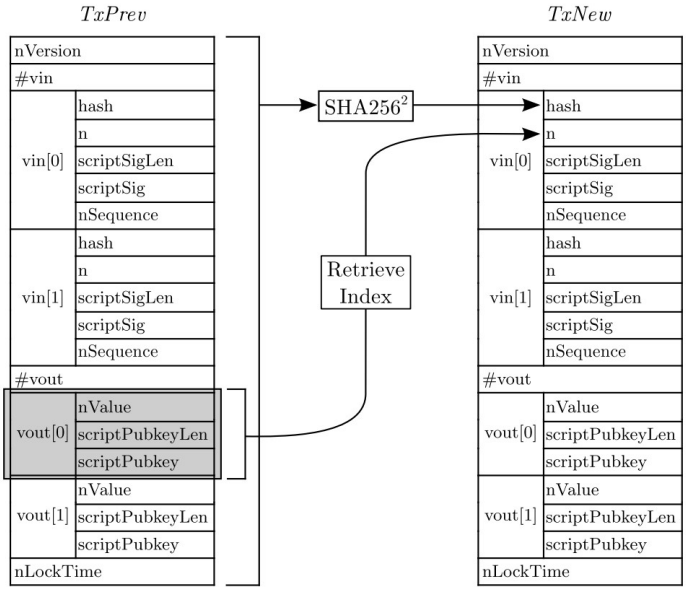


Fig. 3. Bitcoin vin[] Reference to a Previous Transaction (copied from [27]).

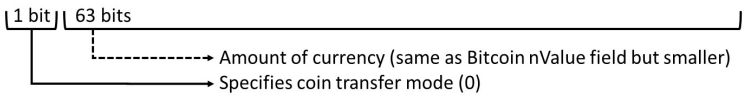


Fig. 4. 64 bit nValue Field Format for the Coin Transfer Mode

can specify a former transaction where an account was given a role. The vin[] scriptSig field then provides evidence that the owner of that account wants to use their role in this transaction (the scriptSig field must satisfy the scriptPubkey field of the transaction where the account was given the role). Thus, each vin[] field can bring a particular role from a particular account into a transaction in order to meet the role requirements for that transaction.

The vout[] field was also reinterpreted. The nValue field now specifies the mode in which its encompassing vout[] field will operate. There are three modes: coin transfer mode, role change mode, and policy change mode. Coin transfer mode moves coin between accounts similarly to a normal Bitcoin transaction. However, we restrict the transaction types that can be used in order to ensure that coins are linked to accounts. Role change mode enables accounts with the M, A, and L roles to modify the role labels of other accounts. Policy change mode enables accounts with the M role to enact and/or modify cryptocurrency policy (to essentially define the ongoing rules for the cryptocurrency). If the first bit of an nValue field is a 0, the encompassing vout[] field is in coin transfer mode. If the first two bits of an nValue field are '10', the encompassing vout[] field is in role change mode. And a nValue field beginning with '11' specifies policy change mode.

Also within the vout[] field, we restrict the scriptPubkey field to only use the Pay-to-Pubkey (P2PK) transaction type. P2PK associates coins with a specific public key (an account in our architecture). If set up to do so, this enables cryptocurrencies implemented from our architecture to link accounts to account owners. This linkage can take place when an account with the A role grants the U role to another account (thereby authorizing it for coin transfers). In this case, the authorizing entity checks the user's identity using out-of-band traditional methods (e.g., passports, drivers licenses, and identity cards).

A. Coin Transfer Mode

If an nValue field has its first bit set to 0, the encompassing vout[] field is in coin transfer mode and is used to move coin between accounts. Since the first bit was used to specify this, the remaining 63 bits specify the amount of coin to be transferred (in Bitcoin all 64 bits are used). Figure 4 shows the changes to the nValue field for the transfer of coin (those nValue fields beginning with 0). Note that for all figures showing the revised nValue format (including this one), solid lines originate from bits that define the action to be taken while dotted lines originate from parameter values.

Anytime a transaction has one or more vout[] fields in coin transfer mode, the original accounts owning the coins and the destination accounts for the coins must all have the U role. This is accomplished by including in the transaction vin[]

fields that bring in the U roles for the accounts either sending or receiving coin.

Lastly, coinbase transactions (the first transaction of each block where the miner sends itself the reward coins) are handled the same as with Bitcoin. However, the vout[] nValue field will start with a 0 bit, putting it in coin transfer mode. Also, the miner must include a vin[] field after the normal coinbase transaction vin[] field in which the miner provides the U role for the account to which the coins are destined.

B. Role Change Mode

If an nValue field has its first two bits set to '10', then the encompassing vout[] field is used to change the roles for a set of accounts. The third bit represents whether or not the vout[] field is removing or adding roles. 0 indicates that roles are being removed and a 1 represents that they are being added. The subsequent bits are flags referring to the different roles. Bits 4, 5, 6, 7, and 8 map to roles M, C, L, U, and A respectively. The remaining 56 bits are undefined. This may be wasteful of space but role change transactions will be relatively rare and we are trying to change the Bitcoin specification as little as possible. Figure 5 shows these changes to the nValue field.

The vout[] scriptPubkeyLen and scriptPubkey fields specify the public key for the account that has these roles. The roles granted by the transaction can then be used in future transactions by the future transaction providing a vin[] scriptSig field that satisfies the vout[] field of the transaction granting the roles. Essentially, an owner of an account uses their private key in some future transaction to prove ownership of a public key documented in a past transaction where the roles were granted. Note that cryptocurrency participants, specifically the miners, will have to make sure that the roles being accessed by a transaction haven't been previously removed from the relevant accounts (roles can be removed by accounts with the M, L, or A roles). This check is similar to miners in Bitcoin checking to make sure that particular coins haven't already been spent.

Every transaction requires one or more roles in order to be valid. Each role has different rules that must be satisfied for the applicable transaction to be valid:

1) M Role Processing: Any addition or removal of roles requires the M role to be provided in one or more of the vin[] datastructures (with two exceptions, see the A and L roles). Each role change vout[] datastructure must be 'covered' by a vin[] scriptSig field where the address specified is located between the root and the node affected in the node hierarchy. Also, the 'covering' address (referenced by the vin[] scriptSig field) must have the role that is to be added or removed in the 'covered' vout[] datastructure.

2) C Role Processing: The inclusion of a vin[] datastructure that has a scriptSig field that satisfies an account having the C role means that the transaction may create coins. There is no need then for other vin[] datastructures. The vout[] datastructures provide coins to the designated addresses.

3) L Role Processing: The inclusion of a vin[] datastructure that has a scriptSig field that satisfies an account having the L role means that the other vin[] fields do NOT need the scriptSigLen or scriptSig fields (for bringing coin into the transaction). Coins may be transferred without the permission of the owners with the inclusion of the L role in the transaction. Also, having the inclusion of the L role enables vout[] datastructures that remove the U role from other accounts. Also, the U role may be added back to accounts for which it was previously revoked. However, these abilities only apply to nodes in the hierarchy that are at a greater distance from the root than the vin[] specified node with the L role (this is to enable the currency administrator to limit this power by creating L role accounts at differing distances from the root).

4) U Role Processing: Any movement of funds requires the U role for the original owner of the coins (specified in the vin[] fields). The recipients of any coins (specified in the vout[] field) must also have the U role.

5) A Role Processing: The inclusion of a vin[] datastructure that has a scriptSig field that satisfies an account having the A role means that the vout[] fields may add role U to accounts. Doing so adds them as descendants in the hierarchical account tree. Accounts with the A role may likewise remove the U role from any descendant. If an A node removes one of its descendants U roles, another A node may add the U role to that node. In this case, the affected node becomes a descendant of the A node adding the U role. Note that if a node with the L role removes the U role from a node, it is put on a special list of frozen nodes and only another node with the L role may remove the affected node from the list.

C. Policy Change Mode

If an nValue field has its first two bits set to '11', then the encompassing vout[] field is in policy change mode, used to create or modify cryptocurrency policy. Note that a vout[] field in policy change mode is only allowed in a transaction if at least one of the vin[] fields provides the M role (since only currency managers can modify policy).

The third bit of the nValue field defines the permanence of the policy (0 is not permanent and 1 is permanent). If an account issues permanent policy, it may not change it in the future. However, M accounts with greater priority, as described in section III, can still trump the issued policy. If the initial root node issues permanent policy, it cannot be changed for the life of the cryptocurrency. This enables the issuance of a static instance of our cryptcurrency architecture. Some features may be made permanent while others are left open for change. It may not be immediately clear why an issuer of a currency would make anything permanent, because it reduces their flexibility. However, by making certain features permanent it provides guarantees to the users. The currency administrator is then constrained to operate within the published rules of the cryptocurrency even though they still manage it. This idea of permanence is important in order to limit the currency administrator from having absolute rule (which is the case

Mell, Peter.
"Managed Blockchain Based Cryptocurrencies with Consensus Enforced Rules and Transparency."
Paper presented at The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, NY, United States. August 1, 2018 - August 3, 2018.

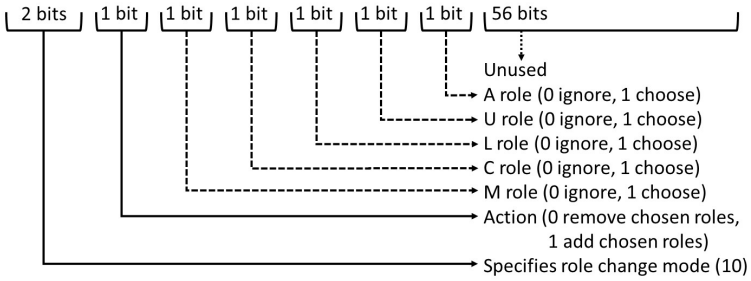


Fig. 5. 64 bit nValue Field Format for the Role Change Mode

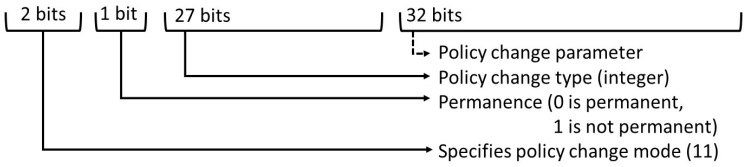


Fig. 6. 64 bit nValue Field Format for the Policy Change Mode

in many of the private blockchain managed cryptocurrencies, such as with Multichain [12]).

After the first three bits of an nValue field are set (to 110 for not permanent or 111 for permanent), the remaining 61 bits specify the policy setting to be made. There is just one policy change made per nValue field, and just one nValue field per vout[] datastructure. However, a single transaction may have many vout[] datastructures.

The next 27 bits specify an integer representing the policy change type while the last 32 bits are used to hold the policy change parameter. The structure of the nValue field in the policy change mode is shown in figure 6.

For the policy change mode, there are currently 14 policy change types with associated parameters, shown in table I. For the binary parameters, 0 means disable and 1 means enable. Binary parameters default to 1 (these policies are enabled by default when the cryptocurrency is initiated).

Policy change types 0 to 5 enable or disable the various roles in available in the architecture (discussed in section III). Type 5 enables or disables the L role from moving coins (disabling would limit the L role to freezing accounts). Type 6 sets a limit for how much coin the set of C roles may create within any particular block. Type 7 sets the block reward mode (0 is the automated approach used by the base cryptocurrency system, Bitcoin in our case, while 1 enables a mode where a currency manager explicitly sets rewards). Type 8 and 9 are for the manual mode and enable setting the block reward and setting a minimum block reward. The purpose of the type 9 is to allow a currency manager to permanently set a minimum while still having the flexibility to adjust the current reward with type 8. Types 10 and 11 are for the self-adjusting mode and enable setting the decay rate for block rewards as well as setting a maximum decay rate. Again, the latter is intended to be used in a mode where it is set permanently. Type 12 sets a transaction fee minimum.

Types 13-15 are important for setting security policy (discussed in detail in section VI). Type 13 sets how often management transactions must appear in a consecutive sequence of blocks (0 disables this feature). For example, a setting of 5 indicates that a certain number of management transactions must appear within every subsequent grouping of 5 blocks. Type 14 specifies the minimum on how many management transactions must appear in that grouping of blocks. A management transaction is one that requires the M role to be present in one of the vin[] fields (see section V-B1). If the currency administrator doesn't have enough management transactions that they wish to put on the blockchain to meet the minimum, then they may issue one or more no operation (no-op) policy change mode transactions of type 15 using one of their M nodes. These do nothing but meet the requirement. A last nuance of this mechanism is that at least one of the management transactions must be a policy change mode transaction. This is to ensure that the currency administrator can always change policy (as the miners might just include non-policy management transactions to meet the minimum requirement).

VI. Security Models

A key aspect of our architecture is to ensure that a balance of power is maintained. Users of the system, including currency managers, should be able to issue any valid transaction onto the blockchain (pursuant to the current policy settings). Miners should be able to enforce policy restrictions and provide transparency for all transactions added to the blockchain.

There are two security models that can be used to enforce this balance of power. Each model slightly favors one party,

Mell, Peter.
"Managed Blockchain Based Cryptocurrencies with Consensus Enforced Rules and Transparency."
Paper presented at The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, NY, United States. August 1, 2018 - August 3, 2018.

TABLE I
CRYPTOCURRENCY POLICY SETTINGS

Policy Change Type	Description	Parameter
0	Enable or disable the M role globally	0 or 1
1	Enable or disable the C role globally	0 or 1
2	Enable or disable the L role globally	0 or 1
3	Enable or disable the U role globally	0 or 1
4	Enable or disable the A role globally	0 or 1
5	Enable or disable the L roles from moving coins	0 or 1
6	C role coin creation limit per block (0 means no limit)	Integer
7	Set block reward mode (0 means manual, 1 means self-adjusting)	0 or 1
8	For manual mode, set block reward	Integer
9	For manual mode, set minimum block reward	Integer
10	For self-adjusting, set geometric decay rate	Float between 0 and 1
11	For self-adjusting, set maximum decay rate	Float between 0 and 1
12	Set transaction fee minimum (0 means no minimum)	Integer
13	Periodicity of management transaction inclusion in blocks	Integer
14	Minimum number of management transactions per period	Integer
15	No operation (used to prove the currency administrator is active)	0

currency managers or miners, although both achieve a reasonable balance (dependent upon the use case).

A. Independent Mining Model

In the independent mining model, the currency administrator permanently disables the requirement to include management transactions periodically (thus the blockchain is not dependent on receiving management transactions). This can be done by having the initial node permanently set the policy change type 13 to 0. In this mode the currency administrator cannot take over maintenance of the blockchain (since mining is unrestricted as with Bitcoin). However, if at least 51 % of the miners collude to 'revolt' against the currency managers, they can prevent future management transactions from entering the blockchain (as well as issuing the well known set of 51 % attacks present with most blockchains [29]). The way this attack works is that the miners controlling 51 % of the computational power simply work on a chain with only their own blocks, excluding the blocks produced by others. Over time, their chain will be longer since they own the majority of the computational power and the other miners will follow their chain (fruitlessly trying to append blocks in a competition they will never win)

B. Dependent Mining Model

Even though the 51 % attack possibility exists in Bitcoin and most other cryptocurrencies, the risk may be too great for some issuers of cryptocurrency; in such a case, the currency administrator can use our dependent mining model. In this case the blockchain is dependent on receiving management transactions. With this approach, the currency administrator using an M node sets policy change types 13 and 14. This forces the miners to include a certain number of management transaction per a certain number of blocks. We advise setting this liberally (type 13 large and type 14 small) since the expectation is that 51 % of the miners will not revolt. If a revolt occurs and miner only include the minimum necessary, then these policy values can be changed to force the miners to allow for more management transactions.

If the miners completely revolt and violate policy, the 'compliant' miners will reject their blocks. This would fork the blockchain into a compliant chain and a non-compliant chain. This is the same thing that would happen with any cryptocurrency if a group of miners begin producing blocks that do not satisfy the specification requirements.

An important aspect of this second model is that it gives more power to the currency administrator than the first model. This can be seen as a positive feature or a weakness depending upon the use case and perspective. With the second model, the currency managers accounts can refuse to submit management transactions, which will eventually cause block creation to halt (issuing management transactions would immediately restart production). This may not be considered a significant threat as the currency administrator initiated the blockchain and inherently will want it to continue operating (this argument is somewhat analogous to the one explaining why Bitcoin in practice is resistant to a 51 % attack even though theoretically it is vulnerable [29]: the miners have a huge stake in the system and won't want it to fail). This could even be considered a feature as owners of a blockchain could eventually deprecate it and move the data to a new blockchain with enhanced technical capabilities. Note that using such an option would be extremely visible and necessarily be rare as it would require all of the users' cryptocurrency software to be updated and reconfigured.

C. Node Software Security

We should note that in all cryptocurrency systems, the authors of the software used by the participating nodes (especially the mining nodes) have significant power. Our architecture is no exception. However, here there is also a balance of power. The currency administrator will likely be a maintainer of the software used by nodes to maintain the blockchain. Hypothetically, they could use this to violate established permanent policy and/or take control of the blockchain from

Mell, Peter.
"Managed Blockchain Based Cryptocurrencies with Consensus Enforced Rules and Transparency."
Paper presented at The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, NY, United States. August 1, 2018 - August 3, 2018.

the miners through the creation and publication of 'malicious' software. However, this can only occur if the majority of miners adopt the malicious software. Even if this did happen (e.g., through miners blindly adopting an update), the miners could simply roll back to a previous non-malicious version to restore the proper function of the architecture.

If miners author the node software, they publish 'malicious' software, and the majority of miners adopt it, the miners could revolt against the currency administrator. However, this is identical to a 51 % attack as described above. The result would be a forking of the blockchain, creating compliant and non-compliant chains. The compliant chain would continue to implement our architecture with a reduce set of compliant miners.

VII. Conclusion

We provide a novel cryptocurrency architecture which is a hybrid approach where a managed cryptocurrency is maintained through distributed open consensus based methods. Key to this architecture is the idea of a genesis transaction upon which all other transactions are based and which enables the establishment of a hierarchy of accounts with differing roles. It is these roles that enabled us to introduce features from fiat currencies into a cryptocurrency: law enforcement, central banking, and account management. Another novel feature is that the architecture allows the cryptocurrency policy to be maintained dynamically by the currency administrator, but certain policy settings can be made permanent in order to facilitate confidence in the stability of the system. This is especially important for the relationship between the currency administrator and an independent community of miners. The currency administrator can control block rewards, which indirectly enables the currency administrator to adjust the power consumption of blockchain maintenance. However, the currency administrator can enact permanent policy to guarantee the miners a certain level of reward. This is important not only to the miners but it prevents the currency administrator from lowering the block reward to nothing and then taking over the mining (and thus completely controlling the blockchain as with many permissioned blockchain systems). Our policy system thus enables a cryptocurrency to be set up that has a balance of power where the currency administrator can perform management functions but where a group of independent miners enforce policy and provide transparency through recording all administrative activity on the blockchain. However, the possibility still exists that the currency administrator or miners could violate policy and attempt to take control of the system. To mitigate this, we provide two security policies that can enforce the balance of power (each with a small bias one direction or the other). Lastly, we showed that our architecture can be implemented through modest changes to the Bitcoin specification. We note though that our approach is not tied to Bitcoin and can be implement on differing cryptocurrency platforms.

References

[1] A. Baliga, "Understanding blockchain consensus models," Tech. rep., Persistent Systems Ltd, Tech. Rep., 2017.

[2] M. Swan, *Blockchain: Blueprint for a new economy*. "O'Reilly Media, Inc.", 2015.

[3] A. Narayanan, J. Bonneau, E. Felten, A. Miller, and S. Goldfeder, *Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction*. Princeton University Press, 2016.

[4] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[5] G. Wood, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum Project Yellow Paper*, vol. 151, 2014.

[6] "Bitcoincash," accessed: 2017-12-29. [Online]. Available: https://www.bitcoincash.org/

[7] "Litecoin," accessed: 2017-06-16. [Online]. Available: https://litecoin.org/

[8] "Why we are building cardano," accessed: 2017-12-29. [Online]. Available: https://whycardano.com/

[9] "Nem the smart asset blockchain," accessed: 2017-12-29. [Online]. Available: https://nem.io/

[10] E. Duffield and D. Diaz, "Dash: A privacy-centric crypto-currency," 2014.

[11] "Cryptocurrency market capitalizations," accessed: 2017-12-29. [Online]. Available: https://coinmarketcap.com/

[12] G. Greenspan, "Multichain private blockchainwhite paper," 2015.

[13] K. J. O'Dwyer and D. Malone, "Bitcoin mining and its energy footprint," 2014.

[14] T. Lee, "Feds charge new york woman with sending bitcoins to support isis," Dec. 2017. [Online]. Available: https://arstechnica.com/tech-policy/2017/12/feds-charge-new-york-woman-with-sending-bitcoins-to-support-isis/

[15] J. Buck, "Dubai will issue first ever state cryptocurrency," Oct. 2017. [Online]. Available: https://cointelegraph.com/news/dubai-will-issue-first-ever-state-cryptocurrency

[16] E. Cheng, "Fedcoin? central banks may need 'digital alternative to cash,' global financial watchdog says," Sep. 2017. [Online]. Available: https://www.cnbc.com/2017/09/18/central-banks-may-need-a-digital-alternative-to-cash-bis-says.html

[17] K. Korjus, "Were planning to launch estcoinŁ Łand thats only thestart," Dec. 2017. [Online]. Available: https://medium.com/e-residency-blog/were-planning-to-launch-estcoin-and-that-s-only-the-start-310aba7f3790

[18] "Monetas," accessed: 2017-12-29. [Online]. Available: https://monetas.net/

[19] L. Chutel, "West africa now has its own digital currency," Dec. 2016. [Online]. Available: https://qz.com/872876/fintech-senegal-is-launched-the-ecfa-digital-currency

[20] E. Smart, "Could a national cryptocurrency like fedcoin save the establishment from economic self-destruction?" Digital Currency Executive, Feb. 2016.

[21] K. Helms, "Russia's central bank pushes for national cryptocurrency," Oct. 2017. [Online]. Available: https://news.bitcoin.com/russias-central-bank-pushes-for-national-cryptocurrency

[22] D. B. Alexandra Ulmer, "Enter the 'petro': Venezuela to launch oil-backed cryptocurrency," Reuters, Dec. 2017.

[23] "Ripple solutions guide," accessed: 2017-12-29. [Online]. Available: https://ripple.com/files/ripple_solutions_guide.pdf

[24] A. Kiayias, A. Russell, B. David, and R. Oliynykov, "Ouroboros: A provably secure proof-of-stake blockchain protocol," in *Annual International Cryptology Conference*. Springer, 2017, pp. 357–388.

[25] M. Staples, S. Chen, S. Falamaki, A. Ponomarev, P. Rimba, A. Tran, I. Weber, X. Xu, and J. Zhu, "Risks and opportunities for systems using blockchain and smart contracts. data61," 2017.

[26] "bitcoinwiki protocol documentation," accessed: 2017-12-29. [Online]. Available: https://en.bitcoin.it/wiki/Protocol_documentation

[27] Okupski, "Bitcoin developer reference," 2014. [Online]. Available: http://enetium.com/resources/Bitcoin.pdf

[28] J. Bonneau, A. Miller, J. Clark, A. Narayanan, J. A. Kroll, and E. W. Felten, "Sok: Research perspectives and challenges for bitcoin and cryptocurrencies," in *Security and Privacy (SP), 2015 IEEE Symposium on*. IEEE, 2015, pp. 104–121.

[29] J. Yli-Huumo, D. Ko, S. Choi, S. Park, and K. Smolander, "Where is current research on blockchain technology? a systematic review," *PloS one*, vol. 11, no. 10, p. e0163477, 2016.

Mell, Peter.
"Managed Blockchain Based Cryptocurrencies with Consensus Enforced Rules and Transparency."
Paper presented at The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, NY, United States. August 1, 2018 - August 3, 2018.

Improved Cryptanalysis of HFEv- via Projection

Jintai Ding[1], Ray Perlner[2], Albrecht Petzoldt[2], and Daniel Smith-Tone[2,3]

[1]Department of Mathematical Sciences, University of Cincinnati,
Cincinnati, Ohio, USA
[2]National Institute of Standards and Technology,
Gaithersburg, Maryland, USA
[3]Department of Mathematics, University of Louisville,
Louisville, Kentucky, USA

jintai.ding@uc.edu, ray.perlner@nist.gov,
albrecht.petzoldt@gmail.com, daniel.smith@nist.gov✉

Abstract. The Hidden Field Equations with vinegar and minus modifiers (HFEv-) signature scheme is one of the most studied multivariate schemes and one of the major candidates for the upcoming standardization of post-quantum digital signature schemes. In this paper, we propose three new attack strategies against HFEv-, each of them using the idea of projection. Especially our third attack is very effective and is, for some parameter sets, the most efficient known attack against HFEv-. Furthermore, our attack requires much less memory than direct and rank attacks. By our work, we therefore give new insights in the security of the HFEv- signature scheme and restrictions for the parameter choice of a possible future standardized HFEv- instance.

Key words: Multivariate Cryptography, HFEv-, MinRank, Gröbner Basis, Projection

1 Introduction

Multivariate cryptography is one of the main candidates for establishing cryptosystems which resist attacks with quantum computers (so called post-quantum cryptosystems). Especially in the area of digital signatures, there exists a large number of practical multivariate schemes such as Unbalanced Oil and Vinegar (UOV) [1] and Rainbow [2].

Another well known multivariate signature scheme is the HFEv- signature scheme, which was first proposed by Patarin, Courtois and Goubin in [3]. Most notably about this scheme are its very short signatures, which are currently the shortest signatures of all existing schemes (both classical and post-quantum).

In this paper we propose three new attacks against the HFEv- signature scheme, each of them using the idea of projection. This means that each of our attacks reduces the number of variables in the system by guessing, either before or after the attack itself.

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

The most interesting results hereby are provided by a distinguishing based attack, which is related to the hybrid approach of the direct attack [4]. The goal of our attack is to remove the vinegar modifier. This allows the attacker to follow up with any key recovery or signature forgery attack applicable to an HFE-instance with the same degree bound and the same number of removed equations as the original HFEv- instance. The attack is very effective and outperforms, for selected parameter sets, all other attacks against HFEv-. Furthermore, the memory requirements of our attack are far less than those of direct and MinRank attacks.

The rest of the paper is organized as follows. In Section 2, we give a short overview of multivariate cryptography and introduce the HFEv- cryptosystem, while Section 3 reviews the previous cryptanalysis of this scheme. Section 4 describes our first two attacks, which combine the MinRank attack with the idea of projection. In Section 5, we present then our distinguishing based attack, whose complexity is analyzed in Section 6. Finally, Section 7 discusses ideas for future work.

2 Hidden Field Equations

2.1 Multivariate cryptography

The basic objects of multivariate cryptography are systems of multivariate quadratic polynomials over a finite field $\mathbb{F}$. The security of multivariate schemes is based on the *MQ Problem* of solving such a system. The MQ Problem is proven to be NP-Hard even for quadratic polynomials over the field GF(2) [5] and believed to be hard on average (both for classical and quantum computers).

To build a multivariate public key cryptosystem (MPKC), one starts with an easily invertible quadratic map $\mathcal{F} : \mathbb{F}^n \to \mathbb{F}^m$ (*central map*). To hide the structure of $\mathcal{F}$ in the public key, we compose it with two invertible affine (or linear) maps $\mathcal{T} : \mathbb{F}^m \to \mathbb{F}^m$ and $\mathcal{U} : \mathbb{F}^n \to \mathbb{F}^n$. The *public key* of the scheme is therefore given by $\mathcal{P} = \mathcal{T} \circ \mathcal{F} \circ \mathcal{U} : \mathbb{F}^n \to \mathbb{F}^m$. The relation between the easily invertible central map $\mathcal{F}$ and the public key $\mathcal{P}$ is referred to as a morphism of polynomials.

The *private key* consists of the three maps $\mathcal{T}, \mathcal{F}$ and $\mathcal{U}$ and therefore allows to invert the public key. To generate a signature for a document (hash value) $\mathbf{h} \in \mathbb{F}^m$, one computes recursively $\mathbf{x} = \mathcal{T}^{-1}(\mathbf{h}) \in \mathbb{F}^m$, $\mathbf{y} = \mathcal{F}^{-1}(\mathbf{x}) \in \mathbb{F}^n$ and $\mathbf{z} = \mathcal{U}^{-1}(\mathbf{y}) \in \mathbb{F}^n$. To check the authenticity of a signature $\mathbf{z} \in \mathbb{F}^n$, one simply computes $\mathbf{h}' = \mathcal{P}(\mathbf{z}) \in \mathbb{F}^m$. If the result is equal to $\mathbf{h}$, the signature is accepted, otherwise rejected. This process is illustrated in Figure 1.

2.2 HFE Variants

The HFE encryption scheme was proposed by J. Patarin in [6]. The scheme belongs to the *BigField* family of multivariate schemes, which means that it uses

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

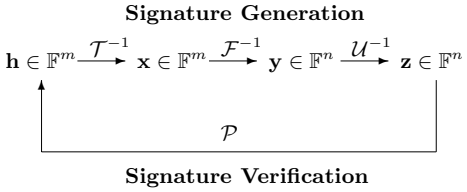


Fig. 1. Signature Generation and Verification for Multivariate Signature Schemes

a degree n extension field $\mathbb{E}$ of $\mathbb{F}$ as well as an isomorphism $\phi : \mathbb{F}^n \to \mathbb{E}$. The central map is a univariate polynomial map over $\mathbb{E}$ of the form

$$\mathcal{F}(X) = \sum_{0\le i,j}^{q^i+q^j\le D} \alpha_{ij} X^{q^i+q^j} + \sum_{i=0}^{q^i\le D} \beta_i X^{q^i} + \gamma.$$

Due to the special structure of $\mathcal{F}$, the map $\bar{\mathcal{F}} = \phi^{-1} \circ \mathcal{F} \circ \phi$ is a quadratic map over the vector space $\mathbb{F}^n$. In order to hide the structure of $\mathcal{F}$ in the public key, $\bar{\mathcal{F}}$ is composed with two affine maps $\mathcal{T}$ and $\mathcal{U}$, i.e. $\mathcal{P} = \mathcal{T} \circ \bar{\mathcal{F}} \circ \mathcal{U}$.

After the basic scheme was broken by direct [7] and rank attacks [8], several versions of HFE for digital signatures have been proposed. Basically, these schemes use two different techniques: the minus and the vinegar modification. For the HFEv- signature scheme [3], the central map $\mathcal{F}$ has the form

$$\mathcal{F}(X, \overline{x}_V) = \sum_{0\le i,j}^{q^i+q^j\le D} \alpha_{ij} X^{q^i+q^j} + \sum_{i=0}^{q^i\le D} \beta_i(x_{n+1}, \ldots, x_{n+v}) X^{q^i} + \gamma(x_{n+1}, \ldots, x_{n+v}),$$

where β_i and γ are linear and quadratic maps in the vinegar variables $\overline{x}_V = (x_{n+1}, \ldots, x_{n+v})$ respectively. Defining $\psi : \mathbb{F}^{n+v} \to \mathbb{E} \times \mathbb{F}^v$ by $\psi = \phi \times id_v$, the public key has the form

$$\mathcal{P} = \mathcal{T} \circ \phi^{-1} \circ \mathcal{F} \circ \psi \circ \mathcal{U} : \mathbb{F}^{n+v} \to \mathbb{F}^{n-a}$$

with two affine maps $\mathcal{T} : \mathbb{F}^n \to \mathbb{F}^{n-a}$ and $\mathcal{U} : \mathbb{F}^{n+v} \to \mathbb{F}^{n+v}$, and is a multivariate quadratic map with coefficients and variables over $\mathbb{F}$.

Signature Generation: To generate a signature $\mathbf{z}$ for a document d, one uses a hash function $\mathcal{H} : \{0,1\}^\star \to \mathbb{F}^{n-a}$ to compute a hash value $\mathbf{h} = \mathcal{H}(d) \in \mathbb{F}^{n-a}$ and performs the following four steps

1. Compute a preimage $\mathbf{x} \in \mathbb{F}^n$ of $\mathbf{h}$ under the affine map $\mathcal{T}$ and set $X = \phi(\mathbf{x}) \in \mathbb{E}$.
2. Choose random values for the vinegar variables $x_{n+1}, \ldots, x_{n+v}$ and substitute them into the central map to obtain the parametrized map $\mathcal{F}_V$.
3. Solve the univariate polynomial equation $\mathcal{F}_V(Y) = X$ over the extension field $\mathbb{E}$ by Berlekamp's algorithm.
4. Compute the signature $\mathbf{z} = \mathcal{U}^{-1}(\phi^{-1}(Y)||x_{n+1}|| \ldots ||x_{n+v}) \in \mathbb{F}^{n+v}$.

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

Signature Verification: To check the authenticity of a signature $\mathbf{z} \in \mathbb{F}^{n+v}$, the verifier computes $\mathbf{h} = \mathcal{H}(d)$ and $\mathbf{h}' = \mathcal{P}(\mathbf{z})$. If $\mathbf{h}' = \mathbf{h}$ holds, the signature is accepted, otherwise rejected.

3 Previous Cryptanalysis

3.1 Direct Algebraic Attack

The direct algebraic attack is the most straightforward way to attack a multivariate cryptosystem such as HFEv-. In this attack, one considers the public equation $\mathcal{P}(\mathbf{z}) = \mathbf{h}$ as an instance of the MQ-Problem. In the case of HFEv-, this public system is slightly underdetermined. In order to make the solution space zero dimensional, one therefore fixes $a + v$ variables in order to get a determined system before applying an algorithm like XL [9] or a Gröbner basis method such as F_4 or F_5 [10, 11]. In some cases one gets better results by guessing additional variables, even if this requires running the Gröbner basis algorithm several times (hybrid approach [4]).

The complexity of a direct attack using the hybrid approach against a system of m quadratic equations in n variables can be estimated as

$$Comp_{direct} = \min_k \quad q^k \cdot 3 \cdot \binom{n-k+d_{\text{reg}}}{d_{\text{reg}}}^2 \cdot \binom{n-k}{2}\Bigg($$

where d_{reg} is the so called *degree of regularity* of the multivariate system. Note that this formula gives only a rough estimate and lower bound of the complexity of a direct attack, since it assumes that the linear systems appearing during the attack are very sparse systems. It is not clear if this assumption holds and if the used Wiedemann algorithm can work with the assumed complexity.

Experiments have shown that the public systems of HFE and its variants can be solved significantly faster than random systems [7, 12]. This phenomenon was studied by Ding et al. in a series of papers [13–15]. In [15] it was shown that the degree of regularity of solving an HFEv- system is upper bounded by

$$d_{\text{reg, HFEv}-} \leq \begin{cases} \frac{(q-1)\cdot(r+a+v-1)}{2} + 2 & q \text{ even and } r+a \text{ odd} \\ \frac{(q-1)\cdot(r+a+v)}{2} + 2 & \text{otherwise} \end{cases}. \tag{1}$$

3.2 MinRank

The historically most effective attack on the HFE family of cryptosystems is the MinRank attack which exploits the algebraic consequence of a low degree bound D. This low degree bound leads to the fact that the central map has a low Q-rank.

Definition 1 *The Q-rank of a multivariate quadratic map $\mathcal{F} : \mathbb{F}^n \to \mathbb{F}^n$ over the finite field $\mathbb{F}$ with q elements is the rank of the quadratic form $\mathcal{Q}$ on $\mathbb{E}[X_1, \ldots, X_n]$ defined by $Q(X_1, \ldots, X_n) = \phi \circ \mathcal{F} \circ \phi^{-1}(X)$, under the identification $X_1 = X, X_2 = X^q, \ldots, X_n = X^{q^{n-1}}$.*

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

Q-rank is invariant under one-sided isomorphisms of polynomials of the form $\mathcal{G} = \mathcal{I} \circ \mathcal{F} \circ \mathcal{U}$, where $\mathcal{I}$ is the identity transformation. Q-rank is not, however, invariant under isomorphisms of polynomials in general. The min-Q-Rank of a quadratic map $\mathcal{F}$ is the minimum Q-rank of any quadratic map in the isomorphism class of $\mathcal{F}$. This quantity is invariant under isomorphisms of polynomials, and is the relevant quantity for cryptanalysis. For historical reasons, language is often abused and the term Q-rank is used in place of min-Q-rank.

As an example, consider an odd characteristic instance of HFE. We may write the homogeneous quadratic part of F as

$$\begin{bmatrix} X & X^q & \cdots & X^{q^{n-1}} \end{bmatrix} \begin{bmatrix} \alpha_{1,1} & \alpha'_{1,2} & \cdots & \alpha'_{1,d} & 0 & \cdots & 0 \\ \alpha'_{1,2} & \alpha_{2,2} & \cdots & \alpha'_{2,d} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \alpha'_{1,d} & \alpha'_{2,d} & \cdots & \alpha_{d,d} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} X \\ X^q \\ \vdots \\ X^{q^{n-1}} \end{bmatrix},$$

where $\alpha'_{i,j} = \frac{1}{2}\alpha_{i,j}$ and $d = \lceil \log_q(D) \rceil$. Clearly, this quadratic form over the ring $\mathbb{E}[X_1, \ldots, X_n]$ has rank d, and thus the HFE central map has Q-rank d.

The first iteration of the MinRank attack in the *BigField* setting is the Kipnis-Shamir (KS) attack of [8]. Via polynomial interpolation, the public key can be expressed as a quadratic polynomial $\mathcal{G}$ over the degree n extension field $\mathbb{E}$. By construction there is an $\mathbb{F}$-linear map $\mathcal{T}^{-1}$ such that $\mathcal{T}^{-1} \circ \mathcal{G}$ has rank d, thus there is a rank d matrix that is an $\mathbb{E}$-linear combination of the Frobenius powers of $\mathcal{G}$. This turns recovery of the transformation $\mathcal{T}$ into the solution of a MinRank problem over $\mathbb{E}$.

A significant improvement to this method for HFE is the key recovery attack of Bettale et al. [16]. The first significant observation made was that an $\mathbb{E}$-linear combination of the *public* polynomials has low rank as a quadratic form over $\mathbb{E}$. By constructing a formal linear combination of the public polynomials with variable coefficients, one can collect the polynomials representing $(d+1) \times (d+1)$ minors of this linear combination, which must be zero by the Q-rank bound. The advantage this technique offers is that the coefficients of the polynomial are in $\mathbb{F}$; thus, the Gröbner basis calculation can be performed over $\mathbb{F}$, while the variety is computed over $\mathbb{E}$. This *minors modeling* method is significantly more efficient than the KS-attack when the number of equations is similar to the number of variables. (In contrast, for schemes such as Zhuang-zi Hidden Field Equations (ZHFE), see [17], it seems that the KS modeling is more efficient, probably due to the large number of variables in the Gröbner basis calculation, see [18].) To make the ideal zero-dimensional, we fix one variable; thus ,the complexity of the KS-attack with minors modeling is asymptotically $\mathcal{O}(n^{(\lceil log_q(D) \rceil)\omega})$, where $2 \leq \omega \leq 3$ is the linear algebra constant.

The MinRank approach can also be effective in attacking HFE-. The key observation in [19] is that not only does the removal of an equation increase the

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

Q-rank by merely one, there is also a basis in which it increases the degree only by a factor of q. Thus HFE- schemes with large base fields are vulnerable to the minors modeling method of [16], even when multiple equations are removed. The complexity of the KS-attack with minors modeling for HFE- is asymptoticaly $\mathcal{O}(n^{(\lceil log_q(D)\rceil+a)\omega})$, where a is the number of equations removed and $2 < \omega \leq 3$ is the linear algebra constant.

4 Variants of MinRank with Projection

As first explicitly noted in [15], the Q-rank of the central map is increased by v with the introduction of v vinegar variables and therefore the min-Q-rank of HFEv- is $\lceil log_q(D)\rceil+a+v$. We now discuss techniques for turning this observation into a key recovery attack. From this point on, let r denote $\lceil log_q(D)\rceil$, that is, the Q-rank of the HFE component of the central map.

4.1 MinRank then Projection

The simplest way to attempt an attack utilizing the low Q-rank of the central map of HFEv- is to directly apply a MinRank attack and then try to discover the vinegar subspace by considering the solution as a quadratic form. To this end, consider the surjective $\mathbb{E}$-algebra representation $\Phi : \mathbb{E} \to \mathbb{A}$ defined by $\Phi(X) = (X, X^q, \ldots, X^{q^{n-1}})$. We may map directly from an n-dimensional vector space over $\mathbb{F}$ to $\mathbb{A}$ via right multiplication by the matrix

$$\mathbf{M}_n = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \theta & \theta^q & \cdots & \theta^{q^{n-1}} \\ \theta^2 & \theta^{2q} & \cdots & \theta^{2q^{n-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \theta^{n-1} & \theta^{(n-1)q} & \cdots & \theta^{(n-1)q^{n-1}} \end{bmatrix},$$

with the choice of a primitive element $\theta \in \mathbb{E}$ (i.e. $\mathbb{E} = \mathbb{F}(\theta)$). Right multiplication by $\mathbf{M}_n$ corresponds to the linear map $\Phi \circ \phi$, where the choice of isomorphism ϕ is determined by the choice of primitive element θ.

We may incorporate the vinegar variables into the picture by simply appending them to $\mathbb{A}$. Specifically, define the map $\widetilde{M}_n : \mathbb{F}^{n+v} \to \mathbb{A} \times \mathbb{F}^v$ by right multiplication by the matrix

$$\widetilde{\mathbf{M}}_n = \begin{bmatrix} \mathbf{M}_n & \mathbf{0}_{n\times v} \\ \mathbf{0}_{v\times n} & I_v \end{bmatrix},$$

where I_v is the identity matrix. We may then represent any HFEv- map as a single $(n+v)\times(n+v)$ matrix with coefficients in $\mathbb{E}$. Note specifically that any function bilinear with respect to the vinegar variable x_n and the HFE variables $x_0, \ldots, x_{n-1}$ can be encoded in row and/or column n of the quadratic form

$$\mathbf{x}\mathbf{Q}\mathbf{x}^\top = \mathbf{x}\widetilde{\mathbf{M}}_n\mathbf{R}\widetilde{\mathbf{M}}_n^\top\mathbf{x}^\top,$$

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

where $\mathbf{R} \in \mathcal{M}_{(n+v)\times(n+v)}(\mathbb{E})$.

Let $\mathbf{F}$ be defined by $\mathbf{x}\widetilde{\mathbf{M}}_n\mathbf{F}\widetilde{\mathbf{M}}_n^\top\mathbf{x}^\top = \mathcal{F}(x)$ where $\mathcal{F}$ is the central map of HFEv-. We will say that $\mathbf{F}$ is the matrix representation of $\mathcal{F}$ over $\mathbb{A}\times\mathbb{F}^v$. Let $\mathbf{F}^{*i}$ be the matrix representation of the ith Frobenius power of $\mathcal{F}$ over $\mathbb{A}\times\mathbb{F}_v$. Then we have, for example the following shape for $\mathbf{F}^{*0}$:

$$\begin{bmatrix}
\begin{pmatrix} \alpha_{0,0} & \cdots & \alpha_{0,d-1} \\ \vdots & \ddots & \vdots \\ \alpha_{0,d-1} & \cdots & \alpha_{d-1,d-1} \end{pmatrix} & 0\cdots 0 & \begin{pmatrix} \beta_{0,n} & \cdots & \beta_{0,n+v-1} \\ \vdots & \ddots & \vdots \\ \beta_{d-1,n} & \cdots & \beta_{d-1,n+v-1} \end{pmatrix} \\
0\cdots 0 & 0\cdots 0 & 0\cdots 0 \\
\begin{pmatrix} \beta_{0,n} & \cdots & \beta_{d-1,n} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} & 0\cdots 0 & \begin{pmatrix} \beta_{n,n} & \cdots & \beta_{n,n+v-1} \\ \vdots & \ddots & \vdots \\ \beta_{0,n+v-1} & \cdots & \beta_{d-1,n+v-1} \end{pmatrix}
\end{bmatrix}.$$

Here we see that $\operatorname{rank}(\mathbf{F}^{*0}) = r + v$. The structure of $\mathbf{F}^{*1}$ is similar with the upper left HFE block consisting of $\alpha_{i,j}$ shifted down and to the right and raised to the power of q, and the symmetric blocks of mixing monomials shifted down and to the right with a more complicated function applied to the $\beta_{i,j}$ coefficients to respect the Frobenius map.

Now let $\mathbf{U}, \mathbf{T}$ and $\mathbf{P}_i$ be the matrix representations of the affine isomorphisms $\mathcal{U}$ and $\mathcal{T}$ and the public quadratic forms $\mathcal{P}_i$, respectively. Then we derive the relation

$$(\mathbf{P}_1,\ldots,\mathbf{P}_n)\mathbf{T}^{-1}\mathbf{M}_n = (\mathbf{U}\widetilde{\mathbf{M}}_n\mathbf{F}^{*0}\widetilde{\mathbf{M}}_n^\top\mathbf{U}^\top,\ldots,\mathbf{U}\widetilde{\mathbf{M}}_n\mathbf{F}^{*(n-1)}\widetilde{\mathbf{M}}_n^\top\mathbf{U}^\top).$$

Thus $\mathbf{U}\widetilde{\mathbf{M}}_n\mathbf{F}^{*0}\widetilde{\mathbf{M}}_n^\top\mathbf{U}^\top$ is an $\mathbb{E}$-linear combination of the public quadratic forms. Since $\mathbf{U}\widetilde{\mathbf{M}}_n$ is invertible, the rank of this linear combination is the rank of $\mathbf{F}^{*0}$, which is $r + v$.

Following the analysis of [19, Theorem 2], we see that the effect of the minus modifier on the matrix representation of $\mathcal{F}$ over $\mathbb{A}\times\mathbb{F}^v$ is to add to it constant multiples of itself with a cyclic shift of the rows and columns down and to the right within the HFE block. Thus for HFEv-, $\mathbf{F}^{*0}$ has the shape given in Figure 2. The rank of this quadratic form is $r + a + v$.

The solution of the MinRank instance provides an equivalent transformation $\mathcal{T}'$ to the output transformation $\mathcal{T}$ (up to the choice of extension to full rank) and a matrix $\mathbf{L}$ representing the low Q-rank quadratic form $\mathbf{U}'\widetilde{\mathbf{M}}_n\widehat{\mathbf{F}}^{*0}\widetilde{\mathbf{M}}_n^\top\mathbf{U}'^\top$ over $\mathbb{A}\times\mathbb{F}^v$, where $\mathcal{P} = \mathcal{T}'\circ\phi^{-1}\circ\widehat{\mathcal{F}}\circ\phi\circ\mathcal{U}'$ for an equivalent private key $(\mathcal{T}', \widehat{\mathcal{F}}, \mathcal{U}')$. Now that the correct output transformation is recovered, it remains to recover the vinegar subspace of the map $\mathcal{L}$ defined by $\mathbf{L} = \mathbf{U}'\widetilde{\mathbf{M}}_n\widehat{\mathbf{F}}^{*0}\widetilde{\mathbf{M}}_n^\top\mathbf{U}'^\top$.

First, note that the kernel of $\mathbf{L}$ as a linear map is orthogonal to the vinegar subspace, so we may simplify the analysis by projecting onto the orthogonal complement of a codimension one subspace of the kernel. Let $\widehat{\mathcal{L}}$ denote the composition of $\mathcal{L}$ with this projection. The strategy now is to compose codimension

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

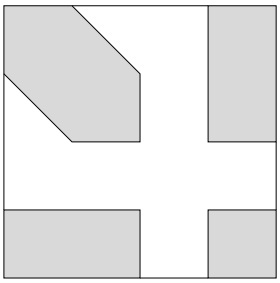


Fig. 2. The shape of the central map of HFEv- composed with the minus projection over $\mathbb{A} \times \mathbb{F}^v$. The shaded areas represent possibly nonzero entries.

one projection mappings π with the transformation $\widehat{\mathcal{L}}$ to filter out the vinegar variables. It suffices to choose projections whose kernels are orthogonal to $\ker(\widehat{\mathbf{L}})$.

If there is a nontrivial intersection between the kernel of π and the vinegar subspace, the rank of the matrix representation of $\widehat{\mathcal{L}} \circ \pi$, $\mathbf{\Pi}\widehat{\mathbf{L}}\mathbf{\Pi}^\top$, will be reduced. In contrast, if this intersection is empty, the rank of $\mathbf{\Pi}\widehat{\mathbf{L}}\mathbf{\Pi}^\top$ should remain the same. To see this, note that by an argument symmetric to that of [19, Lemma 1] we may equivalently define $\widehat{\mathcal{L}} \circ \pi$ by

$$\widehat{\mathcal{L}} \circ \pi = \mathcal{U}^{-1} \circ [(\phi \circ \pi_1 \circ \phi^{-1} \circ \mathcal{S}_1) \times \pi_2] \circ \mathcal{S}_2,$$

where $\mathcal{S}_1 : \mathbb{F}^n \to \mathbb{F}^n$ is nonsingular, $\mathcal{S}_2 : \mathbb{F}^{n+v} \to \mathbb{F}^n \times \mathbb{F}^v$ is an isomorphism, $\pi_1 : \mathbb{E} \to \mathbb{E}$ has degree at most q^{n-r-a} (since the intersection of the image of $\widehat{\mathcal{L}} \circ \pi$ and the HFE subspace is at least $(r+a)$-dimensional) and $\pi_2 : \mathbb{F}^v \to \mathbb{F}^v$ is linear. Since the degree bound of the central HFE quadratic form is q^{r+a}, the highest monomial degree in the composition of π_2 with this map is bounded by q^{n-1}, thus the polynomials $\pi_1, \pi_1^q, \ldots, \pi_1^{q^{r+a}}$ are linearly independent.

The probability that the linear form defining $\ker(\pi)$ which is orthogonal to the kernel of $\widehat{\mathbf{L}}$ lies in the vinegar subspace is $q^{-(r+a+1)}$. Once such a vector is recovered, this step is repeated on the orthogonal complement of the discovered vectors until a basis for the vinegar subspace is found. Thus the complexity of this method when fixing one variable to make the ideal zero dimensional is

$$Comp_{MP} = \mathcal{O}\left(\binom{n+r+v}{r+a+v}^2 \binom{n-a}{2} (r+a+v+1)^3 q^{r+a+1}\right).$$

4.2 Projection then MinRank

Another approach using MinRank is a "project-then-MinRank" approach. In this strategy, one randomly projects the plaintext space onto a codimension k subspace and then applies the MinRank attack. Since the projection π cannot increase the Q-rank of the central map, the Q-rank is at most $r + a + v$.

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

We may choose $k = n - r - a - v$, and expect that the rank of $\mathcal{P} \circ \pi$ is still $r + a + v$, due to the fact that the HFE component is still of full rank, as noted in the previous section. If, however, there is a nontrivial intersection between the kernel of π and the vinegar subspace, the rank of this quadratic form will be less than $r + a + v$. The probability this occurs is $q^{k-n} = q^{-(r+a+v)}$.

Generalizing, we may project further in an attempt to eliminate possibly more vinegar variables and reduce the rank further. The minors system of a MinRank attack at rank r is fully determined if the square of r less than the number of variables bounds the number of public equations; thus, if the image of π is of dimension at least the sum of $\sqrt{n-a}$ and r, the minors system is still fully determined. Therefore, consider eliminating c vinegar variables. This requires k to be at least $n - a - r + c - \sqrt{n-a}$. The probability that there is a c-dimensional intersection between the kernel of π and the vinegar subspace is then $q^{c(k-n)-\binom{c}{2}} \geq q^{\binom{c+1}{2}-cr-ca-c\sqrt{n-a}}$.

Once at least one vinegar variable is found, the new basis can be utilized to filter out the remaining vinegar variables as in the previous method. The complexity of the this method with one variable fixed is

$$Comp_{PM} = \mathcal{O}\quad q^{c(r+a+\sqrt{n-a})-\binom{c+1}{2}}\binom{n+r+v-c}{r+a+v-c}^2\left(r\binom{n-a}{2}\right)\left($$

5 The Distinguishing Based attack

In this section we present our distinguishing based attack against the HFEv- signature scheme. We restrict to the case of $\mathbb{F} = \mathrm{GF}(2)$. The idea of the attack is closely related to the direct attacks with projection (also known as the hybrid approach). We define

$$\mathcal{V} = \left\{ \sum_{i=n+}^{n+v} \left(\lambda_i \mathcal{U}_i | \lambda_i \in \{0,1\} \right\} \right.,$$

where $\mathcal{U}_i$ denotes the i-th component of the affine transformation $\mathcal{U} : \mathbb{F}^{n+v} \to \mathbb{F}^{n+v}$. Therefore, $\mathcal{V}$ is the space spanned by the affine representations of the vinegar variables $x_{n+1}, \ldots, x_{n+v}$. Our attack is based on the following two observations.

- Consider the two HFEv- public keys $\mathcal{P}_1 = \mathrm{HFEv}{-}(n, D, a, v_1)$ and $\mathcal{P}_2 = \mathrm{HFEv}{-}(n, D, a, v_2)$. Before applying a Gröbner basis algorithm to the systems, we fix $a+v_1$ variables in $\mathcal{P}_1$ and $a+v_2$ variables in $\mathcal{P}_2$ to get determined systems. As shown in Table 1 and Figure 3, direct attacks against these systems behave differently. In particular, we can distinguish between determined instances of the two systems $\mathcal{P}_1$ and $\mathcal{P}_2$ by looking at the step degrees of the F_4 algorithm. This remains possible even when adding (not too many) additional linear equations to the systems $\mathcal{P}_1$ and $\mathcal{P}_2$ (thus guessing some of the variables) before applying a Gröbner basis method (hybrid approach).

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

v	HFEv-$(26, 17, 1, v)$	HFEv-$(33, 9, 3, v)$
0	2,3,4,3,4	2,3,4,4,4
1	2,3,4,4,4	2,3,4,5,4
2	2,3,4,5,4	2,3,4,5,5
3	2,3,4,5,5	2,3,4,5,5,5,5,5,6
4	2,3,4,5,5,5,5,5	2,3,4,5,6,6
5	2,3,4,5,6	
random system	2,3,4,5,6	2,3,4,5,6,6

Table 1. Step degrees of the F_4 algorithm against determined HFEv- systems for different values of v

– Let us consider the special case where $v_2 = v_1 - 1$ holds. By adding one linear equation $\ell \in \mathcal{V}$ to $\mathcal{P}_1$, we remove the influence of one of the vinegar variables from the system $\mathcal{P}_1$. A direct attack against the so obtained system $\mathcal{P}_1'$ therefore behaves in exactly the same way as a direct attack against the system $\mathcal{P}_2$ (see Table 2).

5.1 The Distinguisher

Based on the two above observations, we can now construct a distinguisher as follows. We start with an HFEv- public key $\mathcal{P} = \text{HFEv}-(n, D, a, v)$. $\mathcal{P}$ consists of $n-a$ quadratic equations in $n+v$ variables over the field GF(2). After adding the field equations $\{x_i^2 - x_i : i = 1, \ldots, n+v\}$, we append k randomly chosen linear equations $\ell_1, \ldots, \ell_k$ to the system. Therefore, our new system $\mathcal{P}'$ consists of

– the $n-a$ quadratic HFEv- equations from $\mathcal{P}$
– $n+v$ field equations $x_i^2 - x_i = 0$ $(i = 1, \ldots, n+v)$
– the k linear equations $\ell_1, \ldots, \ell_k$.

Altogether, the system $\mathcal{P}'$ consists of $2n-a+v+k$ equations in $n+v$ variables.

After having constructed the system $\mathcal{P}'$, we solve it via a Gröbner basis algorithm. Due to Observation 2, the behaviour of this algorithm should depend on the fact whether one of the linear equations ℓ_i added to the system (or a linear combination of the ℓ_i) is an element of the vinegar space $\mathcal{V}$. In fact, we can observe a difference in the step degrees of the algorithm (see Example 1 below).

Formally written, we can use our technique to distinguish between the two cases

$$\left\{ \sum_{i=1}^{k} \left(\lambda_i \ell_i \mid \lambda_i \in \{0,1\} \right\} \right. \cap \mathcal{V} = \emptyset \text{ and}$$
$$\left\{ \sum_{i=1}^{k} \left(\lambda_i \ell_i \mid \lambda_i \in \{0,1\} \right\} \right. \cap \mathcal{V} \neq \emptyset. \qquad (2)$$

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

However, in most cases that $\left\{\sum_{i=1}^{k} \lambda_i \ell_i \mid \lambda_i \in \{0,1\}\right\} \cap \mathcal{V} \neq \emptyset$, the intersection contains only a single equation $\tilde{\ell}$.

Remark: We have to note here that the number k of linear equations added to the system $\mathcal{P}$ is upper bounded by a value $\bar{k}(n, D, a, v)$. When adding more than $\bar{k}$ linear equations to the system, a distinction between the two cases of (2) is no longer possible.

Example 1: We consider HFEv- systems with $(n, D, a) = (33, 9, 3)$ and varying values of $v \in \{0, \ldots, 4\}$. The resulting HFEv- public keys are systems of $n - a = 30$ quadratic equations in $n + v$ variables. After appending the field equations $\{x_i^2 - x_i = 0\}$ to the systems, we added randomly chosen linear equations to reduce the effective number of variables in our systems. Figure 3 shows the degree of regularity of a direct attack using F_4 against the (projected) systems. For comparison, the figure also contains data for a random system of the same size.

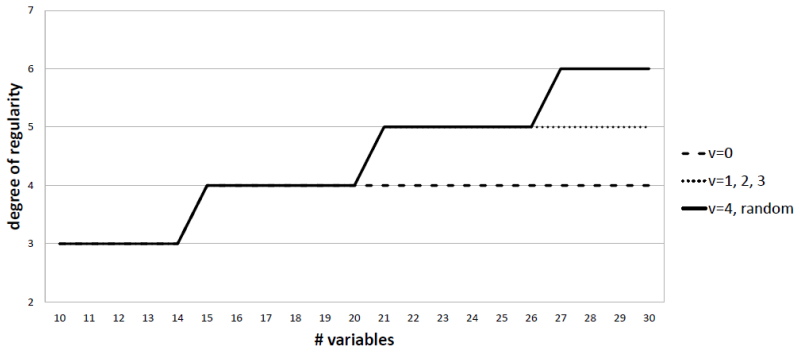


Fig. 3. Direct attack against (projected) HFEv- systems with $(n, D, a) = (33, 9, 3)$ and varying values of v

As Figure 3 shows, there exists, for every parameter set (n, D, a, v) a number $\bar{k}$ such that

1) When adding less than $\bar{k}$ linear equations to the system, the degree of regularity of a direct attack against the projected system is the same as that of a direct attack against the unprojected system.
2) When adding $k \geq \bar{k}$ linear equations, the system behaves exactly like a random system of the same size.

Let us now look at our distinguisher. For this, we skip the parameter set $(n, D, a, v) = (33, 9, 3, 0)$ since, in this case, $\mathcal{V} = \emptyset$ holds. However, as Table 2

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

shows, we can, for each of the values $v \in \{1, \ldots, 4\}$, disitnguish between the two cases of (2).

			step degrees of F_4	
v	$\bar{k}$	$n - \bar{k}$	for $\mathcal{L} \cap \mathcal{V} = \emptyset$	for $\mathcal{L} \cap \mathcal{V} = \{\tilde{\ell}\}$
4	3	27	1,2,3,4,5,6	1,2,3,4,5,5,5
3	4	26	1,2,3,4,5,5,5	1,2,3,4,5,5
2	4	26	1,2,3,4,5,5	1,2,3,4,5,4
1	9	21	1,2,3,4,5	1,2,3,4,4,4

Table 2. Distinguisher Experiments on HFEv-$(33, 9, 3, v)$ systems for different values of v

For abbreviation, we use in the table $\mathcal{L} := \left\{\sum_{i=1}^{k} \lambda_i \ell_i \mid \lambda_i \in \{0, 1\}\right\}$. Note that the evolution of the step degrees for HFEv-(33,9,3,4) is the same as for a random system of the same size.

5.2 The Attack

Based on the distinguisher presented in the previous section, we can construct an attack against HFEv- as follows. By performing the distinguishing experiment with a large number of systems $\mathcal{P}'$ (containing different linear equations), we can find a set of k linear equations $\ell_1, \ldots, \ell_k$ such that $\left\{\sum_{i=1}^{k} \lambda_i \ell_i \mid \lambda_i \in \{0, 1\}\right\} \cap \mathcal{V} = \{\tilde{\ell}_1\}$. Using this, we can determine the exact form of $\tilde{\ell}_1$ as follows. Note that there exist coefficients $\alpha_i \in \{0, 1\}$ $(i = 1, \ldots k)$ such that

$$\tilde{\ell}_1 = \sum_{i=1}^{k} \alpha_i \cdot \ell_i.$$

In order to determine the exact form of this linear combination, we remove one of the linear equations (say ℓ_1) from the system $\mathcal{P}'$ and add another randomly chosen linear equation. If we still can observe a difference in the behaviour of a direct attack compared to a random choice of linear equations, we know that the coefficient α_1 must be 0. Otherwise, the coefficient α_1 must be 1, and we have to add ℓ_1 back to the system.

We repeat this step for $i = 2, \ldots, k$ to determine the values of all the coefficients α_i $(i = 1, \ldots, k)$. This will give us the exact form of the linear equation $\tilde{\ell}_1 \in \mathcal{V}$. We denote this technique as "remove-and-add" strategy.

Having found $\tilde{\ell}_1$, we add it to the original HFEv-(n, D, a, v) system. The resulting system will behave exactly like an HFEv-$(n, D, a, v - 1)$ system, and we can again use our distinguisher and repeat the above procedure to find a second linear equation $\tilde{\ell}_2 \in \mathcal{V}$. Note that this will be much easier than finding $\tilde{\ell}_1$ (see next section).

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

After having found v linear independent equations $\tilde{\ell}_1, \ldots \tilde{\ell}_v \in \mathcal{V}$ and adding them to the HFEv- system, the resulting system will behave exactly like an HFE-(n,D,a) system (i.e. we have no vinegar variables any more). We can then use any attack against HFE- (e.g. the key recovery attack of Vates et al. [19] or a direct attack) to break the scheme. We analyze the complexity of our distinguisher and this attack in the next section.

Let us briefly return to Example 1. When we start with the system $\mathcal{P}$ =HFEv-(33,9,3,4), we can use our distinguisher to find a set $\{\ell_1, \ldots, \ell_k\}$ of linear equations such that $\left\{\sum_{i=1}^{k} \lambda_i \ell_i \mid \lambda_i \in \{0,1\}\right\} \cap \mathcal{V} = \{\tilde{\ell}_1\}$. After having recovered the exact form of $\tilde{\ell}$, we can append it to the system $\mathcal{P}$, which will then behave exactly like an HFEv-(33,9,3,3) system. Let us denote this new system by $\mathcal{P}^{(1)}$. We can then use the distinguisher on $\mathcal{P}^{(1)}$ to obtain a second linear equation $\tilde{\ell}_2 \in \mathcal{V}$. Adding $\tilde{\ell}_2$ to the system $\mathcal{P}^{(1)}$ leads to a system $\mathcal{P}^{(2)}$ behaving exactly like a HFEv-(33,9,3,2) system. By continuing this process, we finally obtain the system $\mathcal{P}^{(4)}$ corresponding to an HFEv- (33,9,3,0) system. We can then break this scheme by using any attack on HFE-.

Algorithm 1 Our distinguishing based attack

Input: HFEv-(n, D, a, v) public key $\mathcal{P}$

Output: equivalent HFE-(n, D, a) public key $\tilde{\mathcal{P}}$

1: Append $\bar{k}$ randomly chosen linear equations $\ell_1, \ldots, \ell_{\bar{k}}$ in the variables $x_1, \ldots, x_{n+v}$ (as well as the field equations $x_i^2 - x_i = 0$) to the system $\mathcal{P}$ and solve it by F_4.
2: Repeat this step until the F_4-step degrees differ from the standard case. This means that we have found a set of linear equations $\ell_1, \ldots, \ell_k$ such that $\left\{\sum_{i=1}^{k} \lambda_i \ell_i \mid \lambda_i \in \{0,1\}\right\} \cap \mathcal{V} = \{\tilde{\ell}_1\}$
3: Determine the exact form of $\tilde{\ell}$ by the above described "remove-and-add" strategy.
4: Append the linear equation $\tilde{\ell}$ to the system $\mathcal{P}$. The resulting system $\mathcal{P}'$ will behave exactly like an HFEv-(n,D,a,v-1) public key.
5: Repeat the above steps until having found v linear independent equations $\tilde{\ell}_1, \ldots, \tilde{\ell}_k \in \mathcal{V}$.
6: **return** $\tilde{\mathcal{P}} = (\mathcal{P}, \tilde{\ell}_1, \ldots, \tilde{\ell}_v)$

6 Complexity Analysis

In the first step of our attack, we have to find one linear equation $\tilde{\ell} \in \mathcal{V}$ by using our distinguisher and a following application of the "remove-and-add" strategy described in the previous section. Therefore, the complexity of this first step of our attack is determined by three factors:

1. The number of times we have to run the distinguisher in order to find a set of linear equations $\ell_1, \ldots, \ell_k$ such that $\left\{\sum_{i=1}^{k} \lambda_i \ell_i \mid \lambda_i \in \{0,1\}\right\} \cap \mathcal{V} = \{\tilde{\ell}\}$,
2. The cost of one run of the distinguisher and

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

3. The cost of recovering the exact form of $\tilde{\ell}$.

The first number is determined by

– The probability that a randomly chosen linear equation in $n+v$ variables is contained in the space $\mathcal{V}$ spanned by the linear representation of the vinegar variables $\mathcal{U}_{n+1}, \ldots, \mathcal{U}_{n+v}$ A randomly chosen linear equation $\bar{\ell}$ in $n+v$ variables can be seen as a linear combination of the components of $\mathcal{U}$, i.e.

$$\bar{\ell} = \sum_{i=1}^{n+v} \lambda_i \cdot \mathcal{U}_i. \tag{3}$$

The reason for this is that $\mathcal{U}$ is an invertible map from $\mathbb{F}^{n+v}$ to itself, which means that the components of $\mathcal{U}$ form a basis of this space. There are 2^{n+v} choices for the parameters λ_i $(i = 1, \ldots, n+v)$. On the other hand, every element $\tilde{\ell}$ of the space $\mathcal{V}$ spanned by the linear transformations of the vinegar variables $v_1, \ldots, v_v$ can be written in the form

$$\tilde{\ell} = \sum_{i=n+1}^{n+v} \lambda_i \cdot \mathcal{U}_i.$$

The probability that a randomly chosen linear equation $\bar{\ell}$ lies in $\mathcal{V}$ is therefore given by

$$\text{prob}(\bar{\ell} \in \mathcal{V}) = 2^{-n}. \tag{4}$$

The reason for this is that all the coefficients λ_i $(i = 1, \ldots, n)$ in the representation (3) of $\bar{\ell}$ must be zero.

– The number of linear equations (and linear combinations thereof) added to the public key. When adding k linear equations $\ell_1, \ldots, \ell_k$ to the public key, we do not have to consider only the k equations $\ell_1, \ldots, \ell_k$ itself, but also all linear combinations of the form

$$\ell = \sum_{i=1}^{k} \lambda_i \cdot \ell_i.$$

The total number of linear equations we have to consider is therefore not k, but 2^k.

Therefore, when adding k linear equations $\ell_1, \ldots, \ell_k$ to the public key, the probability of finding one linear equation $\tilde{\ell} \in \mathcal{V}$, is given by

$$\text{prob} = 1 - (1 - 2^{-n})^{2^k} \approx 2^{k-n}.$$

In order to find one linear equation $\tilde{\ell} \in \mathcal{V}$, we therefore have to run our distinguisher about 2^{n-k} times.

A single run of our distinguisher corresponds to one run of the F_4 algorithm. The cost of this can be estimated as

$$Comp_{F_4} = 3 \cdot \binom{h'}{d_{\text{reg}}}^2 \cdot \binom{n'}{2},$$

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

where n' is the number of variables in the quadratic system and d_{reg} is the so called degree of regularity.

Note that this formula assumes that the linear systems appearing during the attack are solved using a sparse Wiedemann solver. Furthermore we use the fact that the system is defined over the field GF(2), which reduces the number of terms in the high degree polynomials.

With regard to the number n' of variables we find that the linear equations added to the public key are "absorbed" at a very early step of the F_4 algorithm, i.e. they are used to reduce the number of variables in the system. This fact is illustrated in Table 3. In the table, we consider two random systems, both containing 25 quadratic equations. However, while the equations of system A are polynomials in 25 variables, the polynomials of system B contain 35 variables. On the other hand, the system B additionally contains 10 linear equations.

	25 equations, 25 variables			25 quadr. + 10 lin. equations, 35 variables		
step	degree	matrix size	time (s)	degree	matrix size	time (s)
				1	10 × 36	0.0
				1	20 × 36	0.0
1	2	25 × 326	0.0	2	330 × 631	0.0
2	3	652 × 2626	0.02	3	650 × 2626	0.02
3	4	7894 × 14 498	1.27	4	7864 × 15 568	1.34
4	5	52 488 × 52 956	79.86	5	52 197 × 52 665	80.26
5	6	248 705 × 245 506	179.34	6	248 273 × 108 524	182.24

Table 3. Experiments with random systems

As the table shows, both systems behave very similarly. Starting at step 2 (degree 3), there is no significant difference between the matrix sizes or the running times of the single steps between the two systems.

We can therefore conclude that the quadratic systems we consider in our distinguishing based attack ($n - a$ quadratic equations + k linear equations in $n+v$ variables) behave just like systems of $n-a$ quadratic equations in $n+v-k$ variables.

The cost of recovering the exact form of $\tilde{\ell}$ is negligible in comparison to finding linear equations $\ell_1, \ldots, \ell_k$ such that $\left\{\sum_{i=1}^{k} \lambda_i \ell_i \mid \lambda_i \in \{0,1\}\right\} \cap \mathcal{V} = \{\tilde{\ell}\}$. Remember that $\tilde{\ell}$ can be written as a linear combination of $\ell_1, \ldots, \ell_k$, i.e. $\tilde{\ell} = \sum_{i=1}^{k} \lambda_i \cdot \ell_i$.

As described in the previous section, we remove for this one linear equation ℓ_i from the system $\mathcal{P}'$. By adding a randomly chosen linear equation, we obtain a system $\mathcal{P}''$ of the same dimensions. We apply the F_4 algorithm against the two systems $\mathcal{P}'$ and $\mathcal{P}''$. If we observe a difference in the behavior of the algorithm, we know that the coefficient λ_i in the above linear combination is 1. Otherwise we have $\lambda_i = 0$. By running this test for all $i \in \{1, \ldots, k\}$, we can determine

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

all the coefficients λ_i and therefore recover $\tilde{\ell}$. In order to recover $\tilde{\ell}$, we therefore need $2 \cdot k$ runs of the F_4 algorithm, which is far less than the 2^{n-k} F_4-runs above. Therefore, we do not have to consider this step in our complexity analysis.

Altogether, we can estimate the complexity of this first step of our attack by

$$Comp_{\text{Dist; classical}} = 2^{n-k} \cdot 3 \cdot \binom{n+v-k}{d_{\text{reg}}}^2 \cdot \binom{n+v-k}{2}. \quad (5)$$

In the presence of quantum computers, we can speed up the searching step of this attack using Grover's algorithm. Thus we get

$$Comp_{\text{Dist; quantum}} = 2^{(n-k)/2} \cdot 3 \cdot \binom{n+v-k}{d_{\text{reg}}}^2 \cdot \binom{n+v-k}{2}.$$

Note that this assumption of the complexity is very optimistic, since it assumes a perfect "square-root" speed up by Grover's algorithm. Since quantum algorithms must be reversible, it is not clear if this is possible.

As equation (5) shows, the complexity decreases when we increase the number k of linear equations added to the public key. However, as already mentioned in the previous section, our distinguisher fails when k is too large. We denote the maximal value of k for which our distinguisher works by $\bar{k}(n, D, a, v)$.

In order to remove all the vinegar variables from the system $\mathcal{P}$, we have to repeat the above process v times. However, with decreasing v we find (see Table 2)

1) the number $\bar{k}$ of linear equations that we can add to the public system increases, reducing the number of F_4-runs.
2) the degree of regularity of the systems generated by our distinguisher decreases, reducing the complexity of a single F_4-run.

Therefore, the following steps of our attack will be much faster than the first step. This means, that we can estimate the complexity of the whole attack as in formula (5).

However, in order to estimate the complexity of our attack against an HFEv-(n, D, a, v) scheme in practice, we still have to answer the following two questions.

- What is the maximal number $\bar{k}$ of linear equations we can add to the public key such that our distinguisher works?
- What is the degree of regularity of the systems generated by our distinguisher?

In order to answer these questions, we once more consider Example 1 (see previous section).

First, let us consider the second question. As a comparison of Table 2 and Figure 3 shows, the degree of regularity of solving the systems generated by our distinguisher corresponds exactly to the degree of regularity of an unprojected HFEv- system with parameters (n, D, a, v). As stated in [20], we can estimate this value as

$$d_{\text{reg}} = \left\lfloor \frac{r + a + v + 7}{3} \right\rfloor. \quad (6)$$

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

where $r = \lfloor \log_q(D-1) \rfloor + 1$.

To answer the first question, let us take a closer look at the behavior of the hybrid approach against random systems (see Figure 3). We start with a random system of 30 quadratic equations in 30 variables over GF(2). After appending the field equations $x_i^2 - x_i = 0$ $(i = 1, \ldots, 30)$, we add $k \in \{0, \ldots, 20\}$ linear equations to the system. Table 4 shows for which values of k we reach given values of regularity.

d_{reg}	$\#k$ of added linear equations
3	for $k \geq 16$
4	for $10 \leq k \leq 15$
5	for $4 \leq k \leq 9$
6	for $k \leq 3$

Table 4. Degree of regularity of projected random systems with 30 equations

Let us define $\hat{k}(d)$ to be the maximal number of linear equations we can add to the random system, such that the degree of regularity of a direct attack against the system is greater or equal to d, i.e $\hat{k}(6) = 3$, $\hat{k}(5) = 9$ and $\hat{k}(4) = 15$.

By comparing these numbers with the values of $\bar{k}$ listed in Table 2, we find

$$\hat{k}(d^\star) \leq \bar{k} \leq \hat{k}(d^\star) + 1,$$

where $d^\star$ is the degree of regularity of a direct attack against an HFEv-(n, D, a, v) scheme (see equation (6)).

In order to estimate the complexity of our attack against an HFEv-(n, D, a, v) scheme, we therefore proceed as follows.

1. We compute the degree of regularity of the unprojected HFEv-(n, D, a, v) system (see equation (6)). Denote the result by $d^\star$.
2. We estimate the maximal number $\bar{k}$ of linear equations we can add to the public HFEv- system by $\hat{k}(d^\star)$. This value can be obtained as follows.
   The degree of regularity of a random system of $m = n - a$ quadratic equations in n' variables over GF(2) can be estimated as the smallest index d for which the coefficient of X^d in

$$\frac{1}{1-X} \cdot \left(\frac{1-X^2}{1-X}\right)^{n'} \cdot \left(\frac{1-X^2}{1-X^4}\right)^{m}$$

   is non-positive [21].
   We can use this equation to determine the values of $\hat{k}(d^\star)$.

By substituting the so obtained values of $\bar{k}$ and $d^\star$ into formula (5), we therefore get a close estimation of the complexity of our distinguishing based attack against an HFEv-(n, D, a, v) system.

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

Remark: The above procedure allows us to get an estimation of the complexity of our distinguishing based attack against a given HFEv- scheme. However, it seems to be a very hard task to find a closed formula for this complexity.

Example 2: Consider an HFEv- system over GF(2) with $(n, D, a, v) = (91, 5, 3, 2)$. We obtain $r = \lfloor \log_2(D-1) \rfloor + 1 = 3$. The degree of regularity of a direct attack against the HFEv- system (with field equations) is given by

$$d_{\text{reg}} = \left\lfloor \frac{3+3+2+7}{3} \right\rfloor = 5.$$

Therefore, we get

$$Comp_{\text{direct}} = 3 \cdot \binom{88}{5}^2 \cdot \binom{88}{2} \approx 2^{63.9}.$$

After adding $k = 68$ randomly chosen linear equations to the system, the step degrees of the $F4$ algorithm look like $1; 1, 2, 3, 4$. When one of the linear equations was chosen from the vinegar space $\mathcal{V}$, we obtain $1; 1, 2, 3, 3$.

Therefore, we can estimate the complexity of our distinguisher by

$$Comp_{\text{Distinguisher}} = 2^{23} \cdot \binom{25}{4}^2 \cdot \binom{25}{2} \approx 2^{60.1},$$

which is nearly 16 times faster than a direct attack.

The "MinRank-then-project" approach has a complexity estimated by

$$Comp_{\text{MP}} = 3 \cdot \binom{96}{8}^2 \binom{88}{2} \approx 2^{87.4},$$

while the complexity of the "project-then-MinRank" approach has complexity

$$Comp_{\text{PM}} = 2^{14} \cdot 3 \cdot \binom{95}{7}^2 \binom{88}{2} \approx 2^{92.6}.$$

Therefore, for the above parameter set, the distinguishing based attack is the most efficient classical attack against HFEv-.

With regard to the memory consumption, we get

$$\text{Memory}_{\text{direct}} = \binom{88}{5}^2 \approx 2^{50.4},$$

$$\text{Memory}_{\text{MP}} = \binom{96}{8}^2 \approx 2^{73.9},$$

$$\text{Memory}_{\text{PM}} = \binom{95}{7}^2 \approx 2^{66.7},$$

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

$$\text{Memory}_{\text{Distinguisher}} = \binom{25}{4}^2 \approx 2^{27.3}.$$

As these data show, the distinguishing based attack requires much less memory than the direct and the MinRank attack. Since attacks against large instances of multivariate schemes often fail due to memory restrictions, the small memory consumption is a huge benefit of this attack.

Remark: The comparably low complexity of our attack in Example 2 is caused by the small number of vinegar variables in the system. Due to this, the distinguisher works also for relatively small numbers of variables, which enables us to add a large number of linear equations to the system. This again reduces the number of distinguisher runs and therefore the complexity of the attack. (In the case of the example, we found that the distinguisher works for only 25 variables in the system, due to which we had to run our distinguisher only 2^{23} times.)

When the number v of vinegar variables increases, we can not distinguish between the two cases at 25 variables any more. We have to reduce the number of linear equations added to the system and therefore have to run the distinguisher much more often (and for larger systems). Therefore, for larger values of v, the complexity of our attack increases.

For the parameter sets usually used in HFEv- like schemes (and suggested for the National Institute of Standards and Technology (NIST) call for proposals), the direct attack is usually more efficient than our attack. However, in terms of memory consumption, our attack is still much better.

7 Possible Future Work

In this section we shortly describe a strategy to reduce the complexity of our attack. However, since we have neither enough space nor time to present our idea completely, we leave a detailed analysis as future work.

In the distinguishing step of our attack, we solve a large number of multivariate systems using a direct attack. These systems are obtained by adding k linear equations to a multivariate quadratic system $\mathcal{P}$ of m equations in $n+v$ variables (or equivalently projecting the system to a $n+v-k$ dimensional subspace). In Section 5, these projections were chosen at random.

The main idea to reduce the complexity of this step is now to select the projection in a slightly nonrandom fashion. In particular, we consider a projection in two steps.We apply a projection $\tilde{\pi}$ of corank $k+1$ to the system $\mathcal{P}$ and derive from this a set of corank k projections $\{\pi_i\}$. In this case, we can treat the image of $\tilde{\pi}$ in the plaintext space as being generated by the variables $x_1, \ldots, x_{n+v-k-1}$, while the image of each of the projections π_i is generated by the same variables plus one additional variable x_{n+v-k}, which defines a 1-dimensional subspace of the kernel of $\tilde{\pi}$, which will vary depending on the choice of π_i.

During the computation of a Gröbner basis of $\mathcal{P}(\pi_i) = (f_1(\pi_i), \ldots, f_m(\pi_i))$, the F_4 algorithm looks for polynomials p_j of degree $d-2$ such that $\sum p_j \cdot f_j(\pi_i) = q$, where q is a polynomial of degree at most $d-1$.

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

Our strategy will be to first solve for all p_j in the variables $x_1, \ldots, x_{n+v-k-1}$, such that

$$\sum p_j f_j(\pi_i) = q \pmod{x_{n+v-k}}.$$

As the above equation is equivalent to $\sum p_j f_i(\tilde{\pi}) = q$, this computation can be reused for multiple different choices of π_i. By doing so, we therefore can reduce the effort of computing the Gröbner basis needed during the application of our distinguisher.

However, in order to find the exact amount of saving, much more work is required. We therefore leave an exact analysis of the above mentioned idea as future work.

Another topic for future work is a precise complexity analysis of our attack. The complexity analysis presented in Section 6 is based much on heuristics and experiments. In particular, formula (5) contains the parameters $\bar{k}$ and $d^{\star}_{\text{reg}}$, which (so far) could only be determined experimentally. It therefore would be desireable to develop a formula which computes the complexity of our attack for given HFEv- parameters n, D, a and v.

Acknowledgements

We thank the anonymous reviewers of PQCrypto for their valuable comments which helped to improve this paper. In particular we want to thank the shepherd for her help in creating the final version of this paper.

References

1. Kipnis, A., Patarin, J., Goubin, L.: Unbalanced oil and vinegar signature schemes. EUROCRYPT 1999. LNCS **1592** (1999) 206–222
2. Ding, J., Schmidt, D.: Rainbow, a new multivariable polynomial signature scheme. ACNS 2005, LNCS **3531** (2005) 164–175
3. Patarin, J., Courtois, N., Goubin, L.: Quartz, 128-bit long digital signatures. In Naccache, D., ed.: CT-RSA. Volume 2020 of Lecture Notes in Computer Science., Springer (2001) 282–297
4. Bettale, L., Faugère, J.C., Perret, L.: Hybrid approach for solving multivariate systems over finite fields. Journal of Mathematical Cryptology **3** (2009) 177–197
5. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman & Co., New York, NY, USA (1979)
6. Patarin, J.: Hidden Fields Equations (HFE) and Isomorphisms of Polynomials (IP): Two New Families of Asymmetric Algorithms. In: EUROCRYPT. (1996) 33–48
7. Faugere, J.C.: Algebraic cryptanalysis of hidden field equations (HFE) using grobner bases. CRYPTO 2003, LNCS **2729** (2003) 44–60
8. Kipnis, A., Shamir, A.: Cryptanalysis of the HFE public key cryptosystem by relinearization. Advances in Cryptology - CRYPTO 1999, Springer **1666** (1999) 788

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

9. Courtois, N., Klimov, A., Patarin, J., A.Shamir: Efficient algorithms for solving overdefined systems of multivariate polynomial equations. EUROCRYPT 2000, LNCS **1807** (2000) 392–407
10. Faugere, J.C.: A new efficient algorithm for computing grobner bases (f4). Journal of Pure and Applied Algebra **139** (1999) 61–88
11. Faugere, J.C.: A new efficient algorithm for computing grobner bases without reduction to zero (f5). ISSAC 2002, ACM Press (2002) 75–83
12. Mohamed, M.S.E., Ding, J., Buchmann, J.: Towards algebraic cryptanalysis of hfe challenge 2. In: ISA. Volume 200 of Communications in Computer and Information Science., Springer (2011) 123–131
13. Ding, J., Hodges, T.J.: Inverting HFE systems is quasi-polynomial for all fields. In Rogaway, P., ed.: Advances in Cryptology - CRYPTO 2011 - 31st Annual Cryptology Conference, Santa Barbara, CA, USA, August 14-18, 2011. Proceedings. Volume 6841 of Lecture Notes in Computer Science., Springer (2011) 724–742
14. Ding, J., Kleinjung, T.: Degree of regularity for HFE-. IACR Cryptology ePrint Archive **2011** (2011) 570
15. Ding, J., Yang, B.Y.: Degree of regularity for hfev and hfev-. In Gaborit, P., ed.: PQCrypto. Volume 7932 of Lecture Notes in Computer Science., Springer (2013) 52–66
16. Bettale, L., Faugère, J., Perret, L.: Cryptanalysis of HFE, multi-HFE and variants for odd and even characteristic. Des. Codes Cryptography **69** (2013) 1–52
17. Porras, J., Baena, J., Ding, J.: ZHFE, A new multivariate public key encryption scheme. In Mosca, M., ed.: Post-Quantum Cryptography - 6th International Workshop, PQCrypto 2014, Waterloo, ON, Canada, October 1-3, 2014. Proceedings. Volume 8772 of Lecture Notes in Computer Science., Springer (2014) 229–245
18. Cabarcas, D., Smith-Tone, D., Verbel, J.A.: Key recovery attack for ZHFE. [22] 289–308
19. Vates, J., Smith-Tone, D.: Key recovery attack for all parameters of HFE-. [22] 272–288
20. Petzoldt, A.: On the complexity of the hybrid approach on hfev-. Cryptology ePrint Archive, Report 2017/1135 (2017) `https://eprint.iacr.org/2017/1135`.
21. Yang, B., Chen, J.: Theoretical analysis of XL over small fields. In Wang, H., Pieprzyk, J., Varadharajan, V., eds.: Information Security and Privacy: 9th Australasian Conference, ACISP 2004, Sydney, Australia, July 13-15, 2004. Proceedings. Volume 3108 of Lecture Notes in Computer Science., Springer (2004) 277–288
22. Lange, T., Takagi, T., eds.: Post-Quantum Cryptography - 8th International Workshop, PQCrypto 2017, Utrecht, The Netherlands, June 26-28, 2017, Proceedings. Volume 10346 of Lecture Notes in Computer Science., Springer (2017)

Ding, Jintai; Perlner, Ray; Petzoldt, Albrecht; Smith-Tone, Daniel.
"Improved Cryptanalysis of HFEv- via Projection."
Paper presented at PQCrypto 2018: The Ninth International Conference on Post-Quantum Cryptography, Fort Lauderdale, FL, United States.
April 9, 2018 - April 11, 2018.

A System for Centralized ABAC Policy Administration and Local ABAC Policy Decision and Enforcement in Host Systems using Access Control Lists*

David Ferraiolo
National Institute of Standards and Technology
Gaithersburg, Maryland 20899
USA
dferraiolo@nist.gov

Serban Gavrila
National Institute of Standards and Technology
Gaithersburg, Maryland 20899
USA
serban.gavrila@nist.gov

Gopi Katwala
National Institute of Standards and Technology
Gaithersburg, Maryland 20899
USA
gopi.katwala@nist.gov

ABSTRACT

We describe a method that centrally manages Attribute-Based Access Control (ABAC) policies and locally computes and enforces decisions regarding those policies for protection of resource repositories in host systems using their native Access Control List (ACL) mechanisms. The method is founded on the expression of an ABAC policy that conforms to the access control rules of an enterprise and leverages the ABAC policy expression by introducing representations of local host repositories into the ABAC policy expression as objects or object attributes. Repositories may be comprised of individual files, directories, or other resources that require protection. The method further maintains a correspondence between the ABAC representations and repositories in local host systems. The method also leverages an ability to conduct policy analytics in such a way as to formulate ACLs for those representations in accordance with the ABAC policy and create ACLs on repositories using the ACLs of their corresponding representations. As the ABAC policy configuration changes, the method updates the ACLs on affected representations and automatically updates corresponding ACLs on local repositories. Operationally, users attempt to access resources in local host systems, and the ABAC policy is enforced in those systems in terms of their native ACLs.

*The U.S. Government has filed a patent application of certain aspects of the subject matter disclosed in this paper.

Products may be identified in this document, but identification does not imply recommendation or endorsement by NIST, nor that the products identified are necessarily the best available for the purpose.

ABAC'18, March 19–21, 2018, Tempe, AZ, USA
ACM 978-1-4503-5633-6/18/03
https://doi.org/10.1145/3180457.3180460

ACM Reference Format:
David Ferraiolo, Serban Gavrile and Gopi Katwala. 2018. A System for Centralized ABAC Policy Administration and Local ABAC Policy Decision and Enforcement in Host Systems using Access Control Lists. *In Proceedings of 3rd Workshop on Attribute Based Access Control (ABAC'18)*. ACM, Tempe, AZ, USA, 8 pages, March.
DOI: 10.1145/3180457.3180460

KEYWORDS

ABAC; Attribute Based Access Control; ACLs; Access Control Lists; NGAC; Next Generation Access Control; Architecture; Authorization; Policy Machine

1. INTRODUCTION AND BACKGROUND

We describe a method for the enforcement of Attribute Based Access Control (ABAC) policies in host systems using their Access Control Lists (ACLs). The method centrally manages ABAC policies using a Policy and Attribute Administrative Point and a database for storing attributes and policy information in what we refer to as the Minimum ABAC implementation.

The Minimum ABAC implementation also includes a Policy Analytics Engine, that computes Access Control Lists (ACLs) in terms of local host repositories that are represented as ABAC objects or object attributes. As a consequence of the method, ABAC policies are enforced over user access requests to resource repositories in local host systems in terms of their ACLs.

An ACL is a simple mechanism that dates back to the early 1970s and remains in widespread use to protect resource repositories of varying types (e.g., files and directories). Each resource repository is associated with an ACL that stores the users and their approved access rights for the repository. The list is checked by the access control system to determine if access is granted or denied. Lists need not be excessively long if groups of users with common access rights to the repository (rather than individual members) are attached.

The principal advantages of ACL mechanisms are that they are extremely efficient when computing access decisions and

Ferraiolo, David; Gavrila, Serban; Katwala, Gopi.
"A System for Centralized ABAC Policy Administration and Local ABAC Policy Decision and Enforcement in Host Systems using Access Control Lists."
Paper presented at 8th ACM Conference on Data and Applications Security and Privacy (CODASPY 2018), Tempe, AZ, United States. March 21, 2018 - March 21, 2018.

help simplify the review of users' access rights to a repository. Another advantage is that they allow access to a repository to be easily revoked by simply deleting an ACL entry, or deleting a user or group membership of an ACL entry. However, because ACLs make it difficult to determine the access rights users have to repositories, ACLs are cumbersome when managing access capabilities of users.

Attribute-Based Access Control (ABAC) represents the latest milestone in the evolution of authorization approaches [1]. ABAC is an access control method wherein user requests to perform operations on resources are granted or denied based on the assigned attributes of the user, the assigned attributes of the resource, and a set of policies that are specified in terms of those attributes.

A key ABAC advantage is how easily it manages policy. When a user enters on duty or when a user's job function, authority, affiliations, or any other user characteristic changes, an administrator simply assigns/reassigns the user to the appropriate attributes, and the user automatically gains appropriate access capabilities to system resources. Similarly, when a resource is created or different accesses to a resource are required, appropriate object attribute assignments are created/deleted, automatically enabling policy-preserving user access rights to the resource.

ABAC implementations typically include four layers of functional decomposition working together to bring about policy-preserving access control: Enforcement, Decision, Access Control Data, and Administration. Among these components is a Policy Enforcement Point (PEP) that traps access requests and enforces policy. To determine whether to grant or deny access, the PEP submits the request to a Policy Decision Point (PDP). The PDP computes and returns a decision to the PEP based on policy and attribute information stored in one or more databases (access control data). Information is managed in the policy and attribute store through an ABAC system administrative API.

There are currently two standards [2] that address these ABAC features: Extensible Access Control Markup Language (XACML) [3] and Next Generation Access Control (NGAC) [4, 5, 6]. For both standards, there exist open-source and commercial-compliant implementations. While these implementations deliver ABAC's administrative advantages, not all ABAC implementations enable efficient policy analytics—the ability to answer key questions regarding the access state.

In many ways, the method offers the best of both ABAC and ACLs. By leveraging an ABAC implementation, the method provides a means of access control policy support that goes beyond what is feasible through direct management of ACLs. For instance, enforced policies may combine privileges of sub-policies (e.g., discretionary access control and role-based access control) and may consider denials for expressing privilege exceptions to sub-policies. By expressing policies in terms of combinations of user attributes, the method requires the creation and management of fewer attributes than the number of otherwise required groups. By conducting policy enforcement and decision-making using ACLs, the method provides enhanced performance in granting or denying user access requests beyond what is possible using an ABAC system alone. This enhanced performance is not only desirable in applications that manually access system resources on an individual basis, but is absolutely essential in such environments as big data processing and supercomputing, that require batch processing. Although ACLs preclude the ability to answer important policy review questions, the method allows the full breadth of policy analytics that is permissible to ABAC in general including the identification of the access capabilities of a user. Perhaps most appealing, the method achieves enforcement of ABAC policies in local host systems with minimal or no required changes to those systems beyond implementation of agent software.

2. MINIMUM ABAC REQUIREMENTS

The method of central management of ABAC policies and local enforcement of those ABAC policies through ACLs on local host repositories is dependent on an efficient means of conducting policy analytics in an ABAC system. Determining what group of users can access a resource with an access right (e.g., read or write) is especially crucial. The open source implementation, Harmonia 1.6, is an NGAC reference implementation on GitHub that exemplifies an ABAC implementation with the capability to efficiently conduct policy analytics [7]. Annex A of [6] describes, in detail, a linear-time algorithm to calculate the access rights a user has to objects representing protected resources based on the work published in [8]. The algorithm can also be easily adapted to make various other key policy determinations such as identifying the access rights a group of users have to protected resources.

Although a PEP and PDP are normally included in an ABAC implementation, the method does not depend on these components since the functions of enforcing ABAC policy and computing decisions are achieved by the local host's ACL mechanism. What is required of the ABAC system, in addition to conducting policy analytics, is the ability to administer and store policies and attributes. We generally refer to this administrative component as the Policy and Attribute Administrative Point, although XACML does not prescribe a means of managing its attributes.

3. METHOD

The method for centralized ABAC policy management and local host enforcement of ABAC policies using native host ACLs includes the following steps:

- Expressing an ABAC policy that defines privileges, or privileges and prohibitions, of users using a policy and attribute administration point for configuring the authorization data of a centralized ABAC system that, in part, defines policies in terms of objects and object attributes;
- Introducing representations of resource repositories needing protection in local systems that protect their data using ACLs into the ABAC policy expression as objects or object attributes;
- Establishing a one-to-one correspondence between the representations of resource repositories and actual resource repositories;
- Formulating ACLs for representations in accordance with the ABAC policy by determining the group of users that can exercise the access right for each access right (e.g., read, write) relevant for a representation r;
- Creating a group on the local system with user members for each determined group and hosting the resource repository corresponding to representation r, using agent software;

Ferraiolo, David; Gavrila, Serban; Katwala, Gopi.
"A System for Centralized ABAC Policy Administration and Local ABAC Policy Decision and Enforcement in Host Systems using Access Control Lists."
Paper presented at 8th ACM Conference on Data and Applications Security and Privacy (CODASPY 2018), Tempe, AZ, United States. March 21, 2018 - March 21, 2018.

- Creating a user account, if one does not yet exist, for each user member of each created group on the local system hosting the resource repository corresponding to representation r, using agent software;
- Creating an ACL on the resource repository corresponding to representation r, using the ACL formulated for representation r and agent software;
- When required, altering the expression of ABAC policy using the policy and attribute administration point of the centralized ABAC system and mandating the update of ACLs of each representation affected by the alteration;
- Updating group memberships, user accounts, and ACLs on local systems with resource repositories that correspond to affected representations using agent software.

Although the method could be implemented in different ways, Figure 1 illustrates a preferred approach that includes a Control Center surrounded on three sides by an Administrator, a Minimum ABAC implementation, and a local Host System.

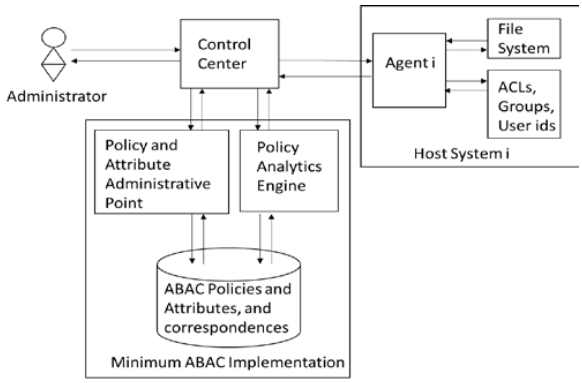


Figure 1: A preferred architecture of method

Administrators express ABAC policies, introduce representations of local repositories into the policy expression, and instruct the creation of ACLs for repositories as administrative commands using the administrative API of the ABAC Control Center. The Control Center provides status and resulting information in reply to administrative commands. The Minimum ABAC Implementation consists of a Policy and Attribute Administrative Point, a Policy Analytics Engine, and a database for storing ABAC Policies and Attributes. In addition, the method may store correspondences between repositories and representations in the database.

The Control Center, through the Policy and Attribute Administrative Point, creates and manages ABAC policies and attributes that are stored in computer memory and/or on disk referred to here as the database. The Control Center issues commands to the Policy and Attribute Administrative Point for managing attributes and policies. The Policy and Attribute Administrative Point implements administrative routines that, when executed, create and delete information stored in the database. These administrative routines may pertain to viewing or reading database information, which would be returned to the Control Center.

The Host System normally implements a File System comprised of repositories of files and directories and normally maintains an access control system with data comprising ACLs, groups, and user identities. In addition to these native components, the method implements Agent software on the Host System with administrative privileges for identifying and viewing repositories and creating, deleting, and updating groups, user identities, and ACLs for repositories. The main function of Agent software is to translate centralized Control Center administrative commands to native host administrative commands. Although the commands issued to Agent software by the Control Center may be uniform across a variety of Host Systems, Agent software on Host Systems are specific to the ACL, group, and user semantics of a host and, in this case, Host i. Agent software response to the Control Center may be uniform across Host Systems. Agent commands to the File System and commands to the host access control system are host-specific. Similarly, status and data returned to the Agent from the File System and access control system status information returned to the Agent are also host-specific.

The Control Center, through Agent software identifies repositories requiring protection in the File System, creates a representation of each such repository as either an object or an object attribute in the ABAC Policy using the Policy and Attribute Administrative Point, and creates a correspondence between the representation and repository in the database.

The Control Center, through the Policy Analytics Engine, computes ACLs with required groups for representations in accordance with ABAC Policies and Attributes stored in the database and subsequently creates ACLs for corresponding repositories, creates groups, and, if necessary, creates user identities in the host access control system using host agent software. To complete this function, the Control Center passes a representation (an object or object attribute) to the Policy Analytics Engine, which then issues read commands to the database resulting in the returns of requested ABAC policy and attribute data. Once the Policy Analytics Engine computes an ACL with required groups, that information is passed back to the Control Center.

The Control Center, through the Policy and Attribute Administrative Point, may update ABAC policies and/or attributes stored in the database. In such cases, the Control Center instructs the Policy Analytics Engine to re-compute ACLs and Groups for affected representations. Using Agent software, ACLs are updated for corresponding repositories, groups, and, if necessary, creates/deletes user identities in the host access control system.

The Policy and Attribute Administrative Point and Policy Analytics Engine could be built as modules of the Control Center on the same machine. The database could be hosted on that machine, or these components could reside as independent network components. Although portrayed as a single store, the attributes and policies may physically reside in different stores. In the case that the method provides ABAC support to a single host system, the Control Center, the entire Minimum ABAC Implementation, and the Agent could reside on that host system.

4. ILLUSTRATIVE EXAMPLE

Figure 2 illustrates an example directory structure of a file system on a host system owned by a Bank to serve as a running use case to highlight the features of the method. The structure includes a root directory ("Products") with two subdirectories ("loans" and "accounts"), each with subdirectories (e.g., loans 2 and accounts 1) for storing and organizing loan and account

products as files and with respect to the branches of the bank. For instance, loans 2 maintains loan files belonging to branch 2.

Although ACL features for protecting resource repositories can vary from system to system and different terminology is sometimes used to express the same feature, we identify semantics common to most if not all ACL mechanisms.

- ACLs on directories are treated differently than ACLs on files.
- Read on a directory implies the right to list children of the directory.
- Write on a directory implies the right to create/delete children of the directory.
- Read and write on a file implies the same right.
- ACLs on a directory or file can inherit or block ACLs of parent directories.

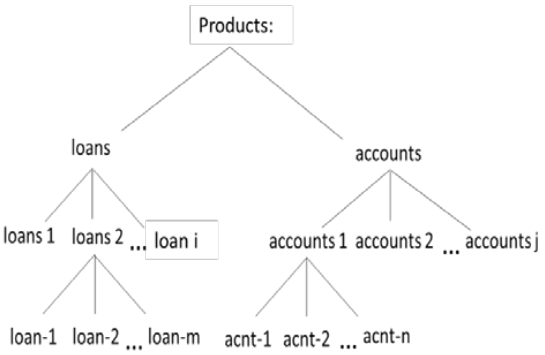


Figure 2: Example directory structure

In addition to the directory structure illustrated in Figure 2, we assume these ACL semantics for the purposes of our illustrative example.

4.1 ABAC Policy Expression

The method begins with the creation of an ABAC policy using the Policy and Attribute Administrative Point of an ABAC implementation. Figure 3 is an illustration of an example bank policy in terms of NGAC policy elements and relations wherein users (e.g., u1, u2) and user attributes (e.g., Teller, Branch1) are shown on the left side of the graph, and object attributes (e.g., Accounts 1 and Loans) and objects (e.g., **loan-1**, o2) are on the right side. The arrows denote assignments and imply a containment relation (e.g., loan-1 is contained in loans 2, Loans, Br2 Products, Products, and RBAC). The policy takes into consideration two sub-policies referred to by NGAC as policy classes: RBAC and BranchAccess.

Access rights to perform operations are acquired through associations. The dashed lines illustrate association relations. By *ua---ars---oa,* we denote an association where *ua* is a user attribute, *ars* is a set of resource and/or administrative access rights, and *oa* is an object attribute[1]. The ars depicted in Figure 3, pertain to both resource access rights and administrative access rights. The r and w are read and write, resource access rights, and c-ooa and d-ooa are administrative access rights for "creating an object in object attribute" and "deleting an object in object attribute." The meaning of an association *ua---ars---oa* is that the users contained in *ua* can execute the access rights in *ars* on the policy elements referenced by *oa*. The set of policy elements referenced by *oa* is dependent on (and meaningful to) the access rights in *ars*. For instance, the association Loan Officer---{r, w, c-ooa, d-ooa}---Loans pertains to capabilities to read and write objects (representing files) contained in Loans (i.e, o2 and loan-1) and create and delete object assignments (a type of relation) in Loans, Loans 1, and Loans 2.

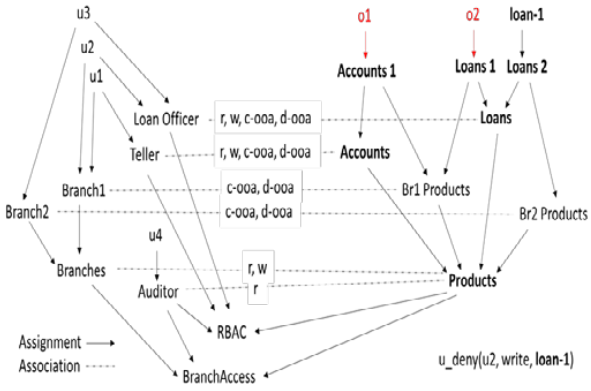


Figure 3: Example policy configuration

Collectively, associations and assignments indirectly specify privileges with respect to policy classes of the form (*u*, *ar*, *e*), with the meaning that user *u* is permitted (or has a capability) to execute the access right *ar* on element *e*, where *e* can represent an object attribute or object.

NGAC includes an algorithm for determining privileges with respect to one or more policy classes and associations. Specifically, (*u*, *ar*, *e*) is a privilege if and only if, for each policy class *pc* in which *e* is contained, the following is true:

1. The user *u* is contained by the user attribute of an association;
2. The element *e* is contained by the attribute *at* of that association;
3. The attribute *at* of that association is contained by the policy class *pc*; and
4. The access right *ar* is a member of the access right set of that association.

Table 1 lists the derived privileges for the policy configuration depicted in Figure 3.

Table 1. List of derived privileges for Figure 2

(u1, r, o1), (u1, w, o1), (u2, r, o2), (u2, w, o2), (u2, r, loan-1), (u2, w, loan-1), (u3, r, o2), (u3, w, o2), (u3, r, loan-1), (u3, w, loan-1), (u1, c-ooa, Accounts 1), (u1, d-ooa, Accounts 1), (u2, c-ooa, Loans 1), (u2, d-ooa, Loans 1), (u3, c-ooa, Loans 2), (u3, d-ooa, Loans 2), (u4, r, o1), (u4, r, o2), (u4, r, o3), (u4, r, loan-1)

[1] For the purposes of this paper, we specify an association using a simpler notation than formally specified in the NGAC standard.

Ferraiolo, David; Gavrila, Serban; Katwala, Gopi.
"A System for Centralized ABAC Policy Administration and Local ABAC Policy Decision and Enforcement in Host Systems using Access Control Lists."
Paper presented at 8th ACM Conference on Data and Applications Security and Privacy (CODASPY 2018), Tempe, AZ, United States. March 21, 2018 - March 21, 2018.

In addition to assignments and associations, NGAC includes prohibitions or deny relations. In general, deny relations specify privilege exceptions. Among these prohibitions is a user-based deny, denote by, u_deny(*u*, *ars*, *pe*), where *u* is a user, *ars* is an access right set, and *pe* is a policy element used as a reference for itself and the policy elements contained by the policy element. The meaning is that user *u* cannot execute access rights in *ars* on policy elements in *pe*. User-deny relations can be created by an administrator. An administrator, for example, might impose a condition wherein no user is able to alter their own loan file, even if the user is assigned to Loan Officer with capabilities to read/write all Loans. The u-deny relation depicted in Figure 3, prohibits u2 from writing to **loan-1**. This privilege exception is reflected in Table 1 using red font.

A natural language description of the policy expressed by Figure 3 is as follows:

- Tellers can read and write accounts objects in all branches.
- Tellers can create and delete accounts objects in the branches for which they are assigned.
- Loan Officers can read and write loans objects in all branches.
- User u3 (a Loan Officer) cannot write to Loan-1.
- Loan Officers can create and delete loans objects in the branches to which they are assigned.
- An Auditor can read account and loan products in all branches.

4.2 Creating ACLs for Representations

The method leverages an ABAC policy expression by introducing representations of host repositories as either an object attribute in the case of a directory or an object in the case of a file. The method further maintains a correspondence between the ABAC representations of the repository and the actual repository in host systems. In Figure 3, **Accounts 1**, **Loans 1**, **Loans 2, Accounts, Loans, Products,** and **loan-1** are in bold to indicate that they represent host system repositories in the directory structure depicted in Figure 2.

Figure 4 illustrates an establishment of a correspondence between **Loans 2** in the ABAC configuration and loans 2 in the directory structure of the local host file system.

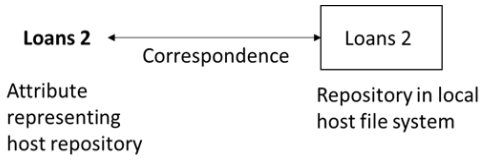


Figure 4: Correspondence between the representation of Loans 2 in the ABAC system and loans 2 in the local host File System.

Once a representation has been established, the method conducts a policy review in such a way as to formulate an ACL for the representation in accordance with the ABAC policy. A central aspect of the policy review involves determining the group of users who can perform specific operations (e.g., read and write) on the representation or on an object contained in the representation. Since the meaning of an ACL differs for directories and files, the logic of the Policy Analytics Engine may make a distinction between representations of files, directories containing files, and directories that do not contain files. For the purposes of this paper we assume a Policy Analytics Engine that makes such a distinction. In describing its logic, we use the notion of a "Custom" ACL to indicate the blocking of ACL privilege inheritance of parent directories.

Let us consider **loan-1**, a representation of the file loan-1. To read **loan-1** a user needs to be assigned to Loan Officer or Auditor. The group of users that meet this criterion are u2, u3, and u4. To write **loan-1** a user needs to be assigned to Loan Officer. The group of users that meet this criterion are u2 and u3. However, in accordance with the overall policy, u2 is denied the ability to write to **loan-1**, and, as such, user u2 is not included in the group for writing. Any convention can be used for naming groups. In our example, we will use gr1 for the group that can read and gr2 for the group that can write to **loan-1** in deriving an ACL for **loan-1**:

loan-1: Custom
gr1, r; gr2, w -- where gr1=u2, u3, u4, and gr2=u3

The ACL is designated as "Custom" to indicate that it does not inherit access rights from its parent directory (**loans 2**). In the case of a representation of a directory containing files, the logic creates a custom ACL for the directory and an ACL for inheritance by the files (the directory's children). While establishing correspondence with a directory repository that contains files, the logic also creates an arbitrary-unique object and assigns that object to the repositories representation if no object is currently assigned to the representation. The red object to object-attribute assignments in Figure 3 illustrates such an assignment. To read an object in **Loans 2** under the policy of Figure 3, a user needs to be assigned to Loan Officer or Auditor. We will refer to the group of users that meet this criterion as gr3. To write to an object in **Loans 2**, a user needs to be assigned to Loan Officer. We refer to that group of users as gr4. Now, let us consider the groups that can list and create/delete the children of **Loans 2**.

In general, a user needs to have permissions to list children for all directories along the path to a file for which they have read access. In the case of a representation of a directory of any type, this group would correspond to the users with read access to an object contained in the representation. In the case of **Loans 2,** that is gr3.

Now, let us consider the group of users that can create/delete children. This group of users would correspond to the users that can create/delete objects in **Loans 2**. In accordance with the policy, these users would be required to be assigned to both Loan Officer and Branch 2, namely u3. Given that read on a directory implies list and write on a directory implies create/delete children, we can derive the follow ACL for **Loans 2**.

Loans 2: Custom
file (inherit) – gr3, r; gr4, w
directory – gr3, r (list); gr5, w (create/delete children)
-- where gr3=u2, u3, u4; gr4=u2, u3; and gr5=u3

Because file level permissions apply to children (files) of the directory, ACL file inheritance is specified. Again, due to its designation as "Custom," this ACL file inheritance is blocked for **loan-1,** enabling the preservation of u2's denial to write to loan-1. Using the same approach used for **Loans 2**, an ACL can be created for **Loans 1** and **Accounts 1** that also contain files:

Ferraiolo, David; Gavrila, Serban; Katwala, Gopi.
"A System for Centralized ABAC Policy Administration and Local ABAC Policy Decision and Enforcement in Host Systems using Access Control Lists."
Paper presented at 8th ACM Conference on Data and Applications Security and Privacy (CODASPY 2018), Tempe, AZ, United States. March 21, 2018 - March 21, 2018.

Loans 1: Custom
file (inherit) – gr6, r; gr7, w
directory – gr6, r (list); gr8, w (create/delete children)
-- where gr6=u2, u3, u4; gr7=u2, u3; gr8=u2

Accounts 1: Custom
file (inherit) – gr9, r; gr10, w
directory – gr9, r (list); gr11, w (create/delete children)
-- where gr9=u1, u4; gr10=u1; gr11=u1

Now, let us consider representations of directory repositories that do not contain files. For these representations, a read (list) ACL is required. Given a user needs to have permissions to list children for all directories along the path to a file for which they have read access, the Policy Analytic Engine could simply identify the users who can read an object contained in the representation. Applying this approach to **Loans**, **Accounts**, and **Products**, we formulate their ACLs:

Loans: Custom
directory – gr12, r (list) --
where gr12=u2, u3, u4

Accounts: Custom
directory – gr13, r (list) --
where gr13=u1, u4

Products: Custom
directory – gr14, r (list) --
where gr14=u1, u2, u3, u4

4.3 Creating Host Access Control Data

The method further creates corresponding group(s) as well as user account(s) and an ACL on the local host repositories using the computed group and the ACL of the corresponding representation. Figure 5 depicts the creation of such access control information on a local host system regarding **loan-1**.

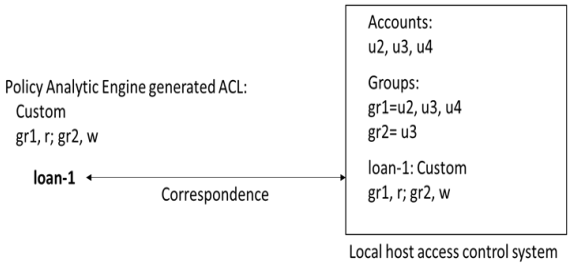


Figure 5: Creation of accounts, groups, and ACLs in local host access control system corresponding to loan-1.

Subsequently to creation of access control information pertaining to **loans-1** the method could create access control information pertaining to **Loans 2**, as shown in Figure 6.

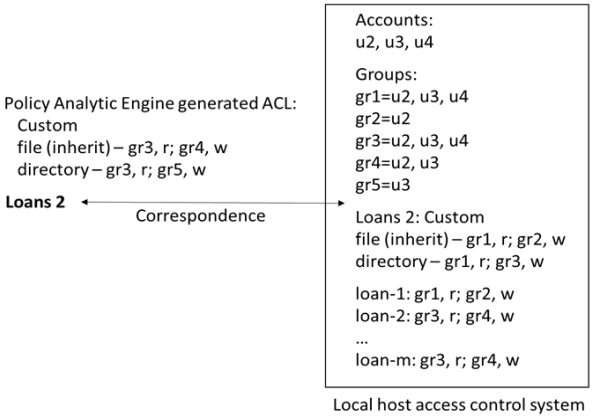


Figure 6: Creation of accounts, groups, and ACLs in local host access control system corresponding to Loans 2.

4.4 Updating Host Access Control Information

As the ABAC policy changes, the method updates appropriate accounts, groups, and ACLs pertaining to affected representations and automatically updates ACLs on corresponding local repositories. Consider the update of the ABAC policy of Figure 3 as indicated in Figure 7.

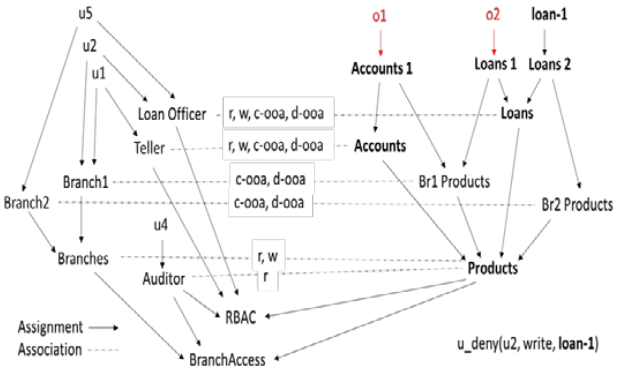


Figure 7: Updating ABAC policy.

Under the updated policy, user u3 has been deleted and replaced by user u5, a new Loan Officer in Branch 2. **Loans 2** is affected by this policy change, and consequently, the logic automatically updates the access control data of the local host access control system as illustrated in Figure 8.

Ferraiolo, David; Gavrila, Serban; Katwala, Gopi.
"A System for Centralized ABAC Policy Administration and Local ABAC Policy Decision and Enforcement in Host Systems using Access Control Lists."
Paper presented at 8th ACM Conference on Data and Applications Security and Privacy (CODASPY 2018), Tempe, AZ, United States. March 21, 2018 - March 21, 2018.

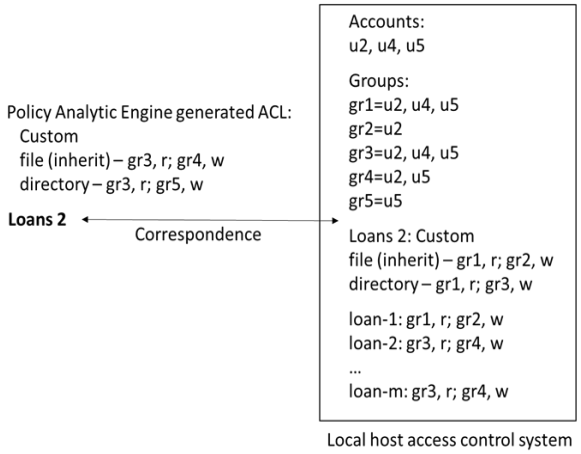


Figure 8: Local changes to the user accounts, groups, and ACLs in correspondence to the updated ABAC policy of Figure 4

5. SYSTEM OPERATION

Operationally, administrators express ABAC policies, introduce representations of local repositories into the policy expression, and instruct the creation of ACLs for repositories through the administrative API of the ABAC Control Center.

Host users attempt to access repositories in local host systems as they normally would, and ABAC policies are enforced in those systems in terms of host ACLs managed by the method. Although the examples used to describe the method pertain to a single local host, the method allows for the centralized management of ACLs in multiple hosts, each within an independent administrative domain as shown in Figure 9.

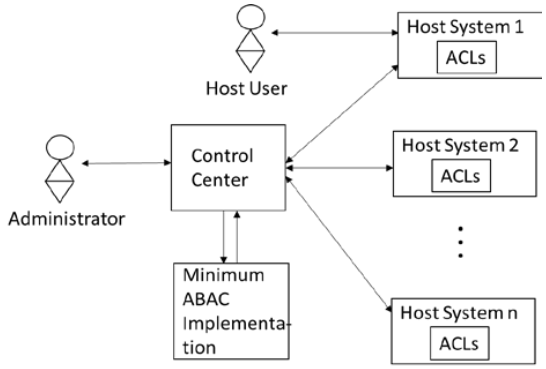


Figure 98: Centralized ABAC policy management and local decision-making and enforcement across multiple security domains

Because of this use of ACLs, access decisions are computed and policy is enforced with an efficiency far superior to an ABAC system that includes PEP and PDP components.

6. RELATED WORK

The method described in this paper is not the only system used for the centralized management of ACLs. In fact, an entire class of products exist, referred to as Enterprise Security Management Systems (ESMSs), which are used for centralized management of authorizations for resources resident in host systems and distributed throughout the enterprise. A common abstraction used by these systems is that of roles and RBAC in general [9, 10]. For instance, roles stored and managed in a directory are used to formulate groups used on ACLs or create ACLs in accordance with role membership and permissions directly associated with roles. The Role Control Center (RCC) [11] is a robust implementation that makes use of much of the entire RBAC abstraction. RCC supports an ESMS model with general role hierarchies, static separation of duty constraints, and an advanced permission review facility (as defined in NIST's proposed RBAC standard [12]). The RCC server is responsible for mapping selected subgraphs of the role graph (called views) to user accounts and groups on heterogeneous hosts as well as for mapping abstract objects and role permissions to actual objects and permission structures (e.g., ACLs) on those hosts. For these tasks, RCC, like our method, uses agent software running on each host to create/delete groups and user accounts, populate the groups with user accounts, and set up ACLs according to commands received from the RCC server. Consequently, RBAC policies are enforced using host ACL mechanisms.

Although there are architectural similarities with RCC and other ESMS products, the method described in this paper is the first to achieve enforcement of ABAC policies using host ACL mechanisms. The enforced policies are based on combinations of user attributes (including but not limited to roles) and object attributes. The ACLs that enforce policy are arrived at not through one-to-one mapping of roles to groups or role permissions to ACLs, but through policy analytics. In particular, the method is based on the determination of a group of users that can access an object or an object in an object attribute with an access right (e.g., read or write) where the source of the group may pertain to a multitude of user attributes.

7. CONCLUSION AND FUTURE WORK

The method described in this paper enables centralized management of ABAC policies for resources repositories distributed throughout an enterprise using host ACLs. It includes a centralized Control Center surrounded by an Administrator, a Minimum ABAC implementation, and Local Host Systems. Administrators express ABAC policies, introduce representations of local repositories into the policy expression, and instruct the creation of ACLs for repositories as administrative commands using the administrative API of the Control Center. The Minimum ABAC Implementation consists of a Policy and Attribute Administrative Point, Policy Analytics Engine, and database for storing ABAC Policies and Attributes. The Control Center maps the authorization data to the various host system ACL mechanisms using the Policy Analytics Engine, through agent software implemented on the host systems. The Control Center, through the Policy and Attribute Administrative Point, may update ABAC policies and/or attributes stored in the database. In such cases, the Control Center instructs the Policy Analytics Engine to re-compute ACLs for affected representations. Using Agent software, ACLs are updated for corresponding repositories in their host access control systems. Operationally, users attempt to access resources in local host systems, and the ABAC policy is enforced in those systems in terms of their ACLs.

All components of the Minimum ABAC implementation are

Ferraiolo, David; Gavrila, Serban; Katwala, Gopi.
"A System for Centralized ABAC Policy Administration and Local ABAC Policy Decision and Enforcement in Host Systems using Access Control Lists."
Paper presented at 8th ACM Conference on Data and Applications Security and Privacy (CODASPY 2018), Tempe, AZ, United States. March 21, 2018 - March 21, 2018.

available as open source. As such, given ABAC policy decision and enforcement are conducted using native host ACL mechanisms, the only components necessary for implementation of the method are the Control Center and host-agent software.

To date we have conducted a variety of experiments to demonstrate the viability of the method, to include development of agent software for the Windows operating system. We plan on development of agent software for other operating environments along with the development of a Control Center component. In addition, we are using a subset of the components available by Harmonia 1.6 to meet the requirements of the Minimum ABAC implementation.

8. REFERENCES

[1] V.C. Hu, D. Ferraiolo, R. Kuhn, A. Schnitzer, K. Sandlin, R. Miller, and K. Scarfone, *Guide to Attribute Based Access Control (ABAC) Definition and Considerations*, National Institute of Standards and Technology (NIST) Special Publication (SP) 800-162, January 2014. http://dx.doi.org/10.6028/NIST.SP.800-162

[2] D. F. Ferraiolo, R. Chandramouli, V. Hu, and R. Kuhn, National Institute of Standards and Technology DRAFT (NIST) SP-800-178, A Comparison of Attribute Based Access Control (ABAC) Standards for Data Services, October 2016. http://dx.doi.org/10.6028/NIST.SP.800-178

[3] The eXtensible Access Control Markup Language (XACML), Version 3.0, OASIS Standard, January 22, 2013. http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.pdf

[4] NICITS 499 – Information technology – Next Generation Access Control - Functional Architecture (NGAC-FA), INCITS 499-2013, American National Standard for Information Technology, American National Standards Institute, March 2013.

[5] INCITS 526 – Information technology – Next Generation Access Control – Generic Operations and Data Structures, INCITS 526-2016, American National Standard for Information Technology, American National Standards Institute, 2016.

[6] NICITS 525 – Information technology – Next Generation Access Control - Implementation Requirements, Protocols and API Definitions (NGAC-IRPADS), in initial public review (December 1, 2017 to January 30, 2018).

[7] NIST Policy Machine Versions 1.5 and 1.6 - Harmonia [Website].

[8] Peter Mell, James Shook, Richard Harang and Serban Gavrila, *Linear Time Algorithms to Restrict Insider Access using Multi-Policy Access Control Systems*, Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, vol. 8, num. 1, March 2017, pp. 4-25, URL: http://isyou.info/jowua/papers/jowua-v8n1-1.pdf.

[9] Enterprise Security Architecture using IBM Tivoli Security Solutions (2002) – IBM Corporation

[10] Enterprise Security Station – User Guide (Windows GUI) – BMC Software Inc., 2002.

[11] Ferraiolo D, Chandramouli R, Ahn GJ, Gavrila SI (2003) The role control center: features and case studies. Proc. of the 8th ACM symposium on access control models and technologies, Como, Italy, pp12–20, June.

[12] D. Ferraiolo, R. Sandhu, S. Gavrila, R. Kuhn, R. Chandramouli, Proposed NIST standard for role-based access control, ACM Transactions on Information and Systems Security 4 (3) (2001).

Ferraiolo, David; Gavrila, Serban; Katwala, Gopi.
"A System for Centralized ABAC Policy Administration and Local ABAC Policy Decision and Enforcement in Host Systems using Access Control Lists."
Paper presented at 8th ACM Conference on Data and Applications Security and Privacy (CODASPY 2018), Tempe, AZ, United States. March 21, 2018 - March 21, 2018.

Quantifying Information Exposure in Internet Routing

Peter Mell
National Institute of Standards and Technology
Gaithersburg MD, USA
peter.mell@nist.gov

Assane Gueye
Electrical Engineering
University of Maryland, College Park
University Alioune Diop of Bambey,
assane1.gueye@uadb.edu.sn

Christopher Schanzle
National Institute of Standards and Technology
Gaithersburg MD, USA
christopher.schanzle@nist.gov

***Abstract*—Data sent over the Internet can be monitored and manipulated by intermediate entities in the data path from the source to the destination. For unencrypted communications (and some encrypted communications with known weaknesses), eavesdropping and man-in-the-middle attacks are possible. For encrypted communication, the identification of the communicating endpoints is still revealed. In addition, encrypted communications may be stored until such time as newly discovered weaknesses in the encryption algorithm or advances in computer hardware render them readable by attackers.**

In this work, we use public data to evaluate both advertised and observed routes through the Internet and measure the extent to which communications between pairs of countries are exposed to other countries. We use both physical router geolocation as well as the country of registration of the companies owning each router. We find a high level of information exposure; even physically adjacent countries use routes that involve many other countries. We also found that countries that are well 'connected' tend to be more exposed. Our analysis indicates that there exists a *tradeoff* between robustness and information exposure in the current Internet.

***Index Terms*—Measurement, Privacy, Internet**

I. Introduction

Data sent over the Internet can be monitored and manipulated by intermediate entities in the path from the source to the destination. For unencrypted communications (and some encrypted communications with known weaknesses), eavesdropping and man-in-the-middle attacks are possible. For encrypted communication, the identification of the communicating endpoints is still revealed. This is important for certain security sensitive communications (e.g., communication between military commands and units). In addition, encrypted communications may be stored until such time as newly discovered weaknesses in the encryption algorithm or advances in computer hardware render them readable by attackers. This kind of attack is especially dangerous as quantum computers, that can break widely used public key encryption, become a reality.

This work is an attempt to quantify this *global information exposure* in the Internet by measuring the extent to which communications between pairs of countries are exposed to other countries. We focus on the routers relaying the packets in Internet communications and use two publicly available datasets to evaluate both advertised and observed routes through the Internet.

With the first dataset, we focus on the physical geolocation of the routers. Every router resides within a unique national boundary and is required to operate according to the laws of that nation. Thus, the data traversing the nation may be exposed to government eavesdropping or control. This dataset was essentially traceroute data from a worldwide collection of monitors (probing sites). We determined the country of residence of each router by geolocating it Internet Protocol (IP) address and used that to convert the router paths to country paths. This gives us a security model in which each router is mapped to their country of residence. Whether or not countries use due process and/or provide transparency for such data access does not affect our results.

With the second dataset, we focus on the legal ownership of the routers. Every router is part of an autonomous system (AS) and every AS is owned by a company that has a country of registration. In this approach, we map each router to the country in which their AS is registered. This models companies abiding by the laws of their country of registration and providing access to the routers that they own. This may be required in some countries or optional in others for assets outside of the physical boundary of the country. These distinctions are irrelevant for our research as we are evaluating worst case data exposure. This dataset was from BGP router tables where we converted advertised routes to country paths through mapping ASs to their country of registration.

Using these two datasets we performed several experiments. In the first experiment we evaluated how well the data from a set of monitors (BGP routing tables or traceroute probing sites) in a country 'generalized' to other sites in the country. This experiment undergirds the utility of the other measurements. We then measure the number of countries 'involved' in communication between pairs of countries with respect to the distance between the paired countries. We next randomly choose increasing sets of untrusted countries to be 'excluded' from communication exchanges. We saw how well pairs of countries could avoid that their communications transit through the excluded countries. Lastly, we perform graph centrality analyses (closeness, degree, eigenvalue, and load) on the graphs generated from using the country paths from

Gueye, Assane; Mell, Peter; Schanzle, Christopher.
"Quantifying Information Exposure in Internet Routing."
Paper presented at The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, NY, United States. August 1, 2018 - August 3, 2018.

both datasets.

Our generalization results showed that it is possible to use a small number of monitoring sites within a country, x, in order to represent the country to country network traffic of the rest of the sites in x to a high degree of coverage.

With the 'involved' country experiments, the geolocation and registration data produced very similiar results. We discovered that adjacent countries (on our country communication graph of the Internet) still have a high number of involved countries between them (those that can view their network traffic). Even more surprising, the number of involved countries actually increases as the country graph distance decreases. In general the number of involved countries was slightly higher for the registration data compared to the geolocation data. Lastly, we found that countries were extremely close together (usually just 1 or 2 hops for most countries).

With the 'excluded' country experiments, the results from the geolocation and registration data differed significantly. We show that it is insufficient to use just the geolocation data to determine information exposure, as has been done in previous work. We show that for a small number of excluded countries (e.g., less than 10), in general there is a high chance that a country can send data to another country where all known routes avoid the excluded countries. However, this likelihood decreases extremely rapidly with more than 10 excluded countries.

With the country graph centrality experiments, we compute the average number of involved countries (between a given country to all the other possible destinations) with respect to the centrality (closeness, degree, eigenvalue, and load) of that country. We show that countries with high centrality values (i.e., well connected) tend to have higher information exposure. This has been observed with all centrality metrics and with both datasets. This observation is consistent with the findings of the 'involved' country experiments, where we discovered that adjacent countries still have a high number of involved countries between them. Indeed, when a country has high (say degree) centrality, it has many direct neighbors. Since the number of involved countries is high for each neighbor, the average is consequently high. This seems to suggest that in the current Internet, there is a tradeoff between the "connectivity" of a country and its degree of information exposure. Indeed, a country that is well connected has many alternate paths to each destination (which is desirable for the robustness of routing). However, the diversity of paths also implies that many countries (some potentially adversarial) might be traversed by the communications to a given destination. We are not aware of any other study revealing this robustness-exposure tradeoff.

II. Data Description

We obtained our data from the public datasets provided by the Center for Applied Internet Data Analysis (CAIDA) [1], covering the years 2015 and 2016. We collected both Border Gateway Protocol (BGP) routing tables to view advertised AS routes through the Internet as well as traceroute type data from a worldwide set of monitors. We converted both AS and router paths into country paths (as described below). A challenge is that our datasources reveal Internet paths that are both advertised (registration data) and used (geolocation data), however, neither dataset reveals how often these paths are used. In addition, there likely exist additional routes not revealed from our data sources. Our experimental results thus are a lower bound on the extent to which information exposure is taking place at the country level. This said, it is reasonable to assume that our discovered routes cover the primary pathways through the Internet.

A. Geolocated Router Path Data

Our first dataset, which we call the 'geolocated' data, consisted of actual paths discovered through active scamper probing [2] (similiar to traceroute) by a worldwide set of CAIDA Archipelago (Ark) monitors [3]. We collected all daily traces from January 01, 2015 to December 31, 2016 (a total of 123 121 files totaling 2.3 TB). After pre-processing and duplicate removal, we ended up with more than 3.1 billion distinct probe traces for each year. We then used the MaxMind[1] service [4] to geolocate each router within a particular country and we converted the router paths into country paths. While geolocation data of routers can be inaccurate, previous work has found that it is more accurate at a country level of abstraction [5].

B. Autonomous System Path Data

Our second dataset, which we call the 'registration' data, consisted of Border Gateway Protocol (BGP) routing tables from a worldwide set of routers. This provided advertised routes between autonomous systems (ASs). We obtained the data using the BGPStream tools [6] to collect data from the University of Oregon Routeviews Project [7] from Jan 01, 2015 to Dec 31, 2016 (a total of 150 GB of raw data). After pre-processing and duplicate removal, we ended up with more than 2.5 billion path traces for each year. Using other CAIDA provided data, we mapped ASs to their countries of registration thereby converting the AS paths to country paths.

III. Data Generalization Experiment

Both datasets yield country paths through the Internet originating from specific locations or 'monitors' (a router that provided its BGP tables or a scamper probing site). Any particular country will have zero or more monitors. The monitors tend to be distributed throughout a country as there is little motivation to monitor the same location multiple times. Thus, we assume that the monitors have somewhat of a random distribution but acknowledge that this is not strictly true.

To undergird the results of our information exposure experiments, it is necessary to show that the set of monitors within particular countries provide sufficient data to represent the entire country (that they 'generalize'). More specifically, the set of country paths yielded by the monitors should closely

[1]Any mention of commercial entities or products is for information only; it does not imply recommendation or endorsement by NIST.

Gueye, Assane; Mell, Peter; Schanzle, Christopher.
"Quantifying Information Exposure in Internet Routing."
Paper presented at The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, NY, United States. August 1, 2018 - August 3, 2018.

approximate the set of country paths that would be revealed in the hypothetical case of having monitors in every location within the country.

We approximately test this by comparing the paths revealed by each monitor with the paths revealed by all other monitors within a particular country. Let M represent the set of monitors in a country. Let $R(W)$ represent the set of country paths revealed by the monitors in set W. Let $x \in M$ and $Y = M \backslash x$. For each x, we compute a ratio $|R(\{x\}) \cap R(Y)|/|R(\{x\})|$ which we refer to as the 'generalization ratio'. We then plot the mean of all such computed ratios for each country against the number of monitors in that country.

For the registration data, the mean generalization ratio increases very quickly to at least .7 with 20 monitors and around .9 with 60 monitors. One country had 3303 monitors and another 13 912 monitors resulting in generalization ratios of .99. For the geolocation data, the mean generalization ratio also increases very quickly to at around .8 with just 5 monitors. One country had 44 monitors with a generalization ratio of around .96. These results show that with a sufficient number of monitors within a country, our data generalizes to represent the country paths used by a vast majority of the countries.

IV. Information Exposure Experiments and Analysis

All experiments were performed on both datasets (geolocation and registration) for both 2015 and 2016. Due to space constraints, we limit ourselves to summarizing our findings from the data. Some example results are provided, anonymizing the countries as Bespin and Hoth.

A. Number of Countries Involved in Pairwise Communication

Our first experiment is to measure how many countries are involved in country to country Internet communications. The set of 'involved' countries, $I(x, y)$, for a pair of communicating countries x and y consists of all countries on any recorded country path from x to y (excluding both x and y). The involved countries represent the minimal set of countries that could observe or modify some fraction of the communications from x to y. Note that even with encrypted traffic, this measurement matters for certain high sensitivity communications (e.g., military commands) as the communicating endpoints are revealed.

1) Experiment: For our experiment, we calculated $I(x, y)$ for all paths between all pairs of countries. We did this for both the registration and geolocation data. We evaluated each source country x individually, creating a figure to analyze each country (an example is shown in figure 1). For each x we plotted each target country y using a figure with an x-axis representing the mean country distance among all paths and the y-axis representing the total number of involved countries using all paths. We then performed the same analysis but used the minimum country distance for the x-axis.

2) Discussion: We find that the number of involved countries can be high. This is true for both datasets. One large wealthy nation has up to 96 involved countries between it and another country. Surprisingly, we find this even for countries that are adjacent. This indicates that even when countries have close proximity (either geographically or logically), the routing structure of the Internet will use many non-direct paths. While likely necessary and appropriate for dynamic load balancing, it has a huge effect in increasing the worst case information exposure between countries on the Internet.

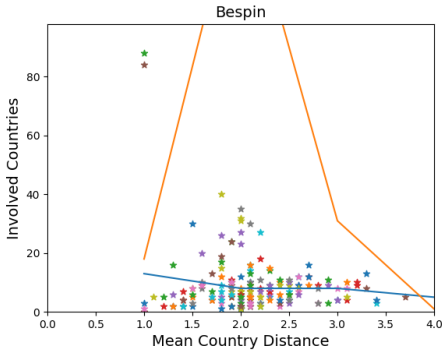


Fig. 1. Bespin Mean Involved Country Exposure from Geolocation data for 2016

Furthermore, we find a relationship between the number of involved countries and the distance between the countries (i.e., number of intervening countries). The number of involved countries generally decreases as the distance increases. This result seems counter intuitive and has a great impact on the evaluation of the privacy of communications between pairs of countries.

B. Excluding Countries from Communications

In this experiment we evaluate how easily particular countries can communicate to all other countries without their communications traversing some target set of countries.

1) Experiment: We execute 20k trials per country. We test excluded country list sizes ranging from 0 to 190 using a step of 10. For each size, we ran 500 trials; for each trial we choose a random destination country and a random set of countries to be on the excluded list for that trial. For each country we created a figure showing the mean ratio of paths containing no excluded countries over all paths (including data to all the other countries in a single figure). We also plotted the ratio for all paths containing excluded countries as well as when the paths contained a mixture (some having excluded countries and some not). An example figure is shown in figure 2.

2) Discussion: For all countries, as the number of countries on the excluded list increases, the ratio of paths containing no excluded countries decreases rapidly (a roughly geometric or exponential decay depending upon the country). For one wealthy large country, an excluded list of size 10 using the geolocation data yields a 57% chance of no paths having an excluded country (choosing some other country at random as the destination). However, increasing the list to 20 reduces the chance to just around 37%; at 50 it is 10%. These numbers

Gueye, Assane; Mell, Peter; Schanzle, Christopher.
"Quantifying Information Exposure in Internet Routing."
Paper presented at The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, NY, United States. August 1, 2018 - August 3, 2018.

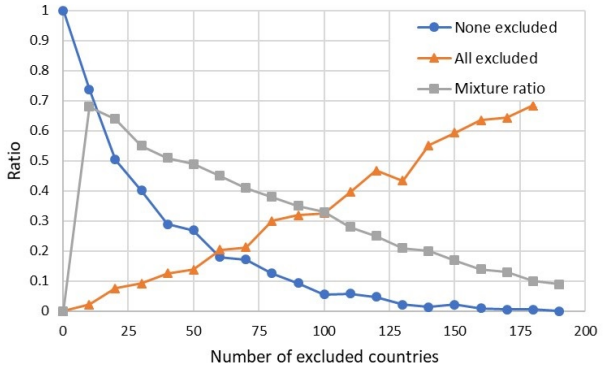


Fig. 2. Hoth Excluded Country Exposure from the Geolocation Data for 2016

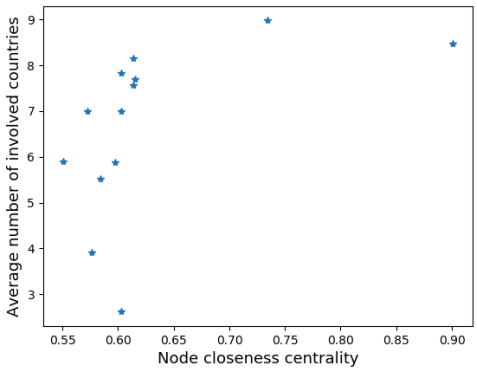


Fig. 3. Node Closeness Centrality for 2016 Geolocation Data

vary dramatically given specific scenarios with set country pairs and should not be used to evaluate the overall state of privacy for Internet communications.

Performing the same example evaluation using the registration data changes the results to around 9%, 6%, and 2% respectively. This shows that evaluating information exposure using just the geolocation data is insufficient as the registration data has significantly different result values (although the similarly shaped functions). To our knowledge, all past research in this area has focused solely on using the geolocation data (or they used the router registration BGP data as a substitute for router geolocation, which we show in section V to be flawed).

C. Comparison to Country-Country Communication Graph

We also analyzed the overall communication graphs using centrality metrics. To our knowledge, these are novel results as we focus on centrality measurements for the Internet as a whole from the perspective of country to country communications.

1) Experiment: In this experiment we take the following types of centrality metrics: closeness, degree, eigenvalue, and load. Degree centrality indicates the number of nodes connected to a given node. Closeness centrality measures the mean distance from a vertex to other vertices. Eigenvalue centrality is a measure of the structural importance of nodes, proportional to the structural importances of their connected neighborhood. The load centrality of a node is the fraction of all shortest paths that pass through that node.

The experiment works as follows. First, we generate a country-level communication graph by: (1) merging all routers (or ASs) within a same country into one node and removing all loops; (2) considering links between routers (or ASs) in different countries as links between the nodes representing the countries (all multi-edges are removed). Second, for each country (which represents a node in the graph) and for each centrality metric, we compute the average number of involved countries to all possible destinations from that country. We finally scatter-plot the average number of involved countries as a function of the value of the centrality metrics. An example result is provided in figure 3 (the rest are omitted due to space limitations.

2) Discussion: We find that the general trend is that as the centrality values of the nodes increase, the average number of involved countries increases. In other terms, as a country is more connected, its information exposure increases in the sense that there are more countries that might be involved in a communication between that country and a random destination in the world.

The observations above are consistent with what we have found with our 'involved' countries experiment. For recall, we observed that adjacent countries still have a high number of involved countries between them. As involved countries are determined using routes discovered in the dataset, this is equivalent to saying that even for pairs of countries between which there is a direct communication path, there exist many alternate and independent paths that go through at least one other country. Hence, if a country has many neighbors (in the communication graph), there will exist many such alternate paths (going through third countries), and the average number of involved countries will be large. This could also be explained by the dis-assortativity [8] property of the Internet, which states that "nodes with high centrality metric tend to be connected together". A consequence of this is that those nodes will tend to form together a very well-connected subgraph with many independent paths. On the other hand, as the distance between the two countries increases, there will be less of those independent paths, resulting to a lower average of involved countries.

Having many independent paths is by itself a well-desired property for the robustness of the Internet. Our study shows that there is an unintended cost to this robustness: *the exposure of information*. Since the Internet routing is based on "best effort", any of the alternate paths can be used for a given communication between two countries. However, with more alternate paths between two countries, there is a greater risk that their communications transit via an un-trusted country, hence increasing their information exposure. To our knowl-

Gueye, Assane; Mell, Peter; Schanzle, Christopher.
"Quantifying Information Exposure in Internet Routing."
Paper presented at The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, NY, United States. August 1, 2018 - August 3, 2018.

edge, this paper is the first study that reports on this tradeoff between robustness and information exposure.

V. Related Work

[9] is most similar to our work in that they evaluate routes to determine their information exposure at a country level of abstraction. It uses traceroute and due to limitations of their measurement infrastructure, they limit their analysis to five countries. They focus on measurements of regular users accessing the Alexa Top 100 websites and provide related analytics on specific countries. In contrast, our work is focused on information exposure measurements for the movement of high sensitivity data of interest to foreign nation-states. Instead of being limited to just a few countries, our approach using publicly available datasets can be applied to all countries. Of perhaps greater importance is that our work did not just evaluate router geolocation dataset as in [9], it includes router country of registration data as well.

[10] and [11] use BGP data for an analysis of nation-state routing. However, they use BGP data to approximate the country of residence of each router on a path. At the AS level of granularity, the country location of the specific routers used in a path cannot be accurately determined [9]. To easily see this, consider the Internet backbone provider Level 3. They own routers throughout the world, on every continent, and yet their country of registration is the United States [1]. Using the Level 3 Internet backbone, communications paths can physically traverse the entire globe and yet appear using the BGP data to stay within the United States. We use the more accurate geolocated traceroute data for this purpose and uniquely use the BGP data to show the countries that have influence over the companies owning the routers (regardless of their physical location). Both perspectives are important and our work is the first to take both into account.

[12] and [13] evaluate how groups of countries could collude to partition up the Internet into isolated chunks in order to prevent pairs of countries from communicating. The two papers use geolocated router paths converted to country paths as we do but focus on using them to measure the possibility of active attacks as opposed to measuring information exposure.

VI. Conclusion

We have quantified the information exposure in the global Internet with respect to countries having access to (or even modifying) data in transit between other countries. Our experiments covered all countries and we presented here the results for two representative countries. We have found that the level of exposure is significant. Even for communicating countries that are physically adjacent, many paths involving other countries are used. Physical proximity does not guarantee private communications. Our study has also shown that there is an apparent tradeoff between robustness and information exposure in the global Internet.

Our results motivate enhanced security with respect to international communications that may be of interest to foreign entities. We assume that strong encryption for such communications is already being implemented. Enhancements may be made in the area of hiding the communicating endpoints for communications in which this is relevant (e.g., communicating intelligence centers or military commands). The use of proxies and anonymous routing services can be assistive (although network throughput can then suffer and such services have had vulnerabilities). A remaining danger is that strongly encrypted traffic will be stored and then decrypted once quantum computers are accessible to nation states or until used cryptographic algorithms have been broken. Our findings on information exposure then promotes the changeover to quantum resistant encryption.

Lastly, our work has shown that it is possible to model the communications between countries to determine the information exposure. Our work is thus a template for how others can perform this calculation for operational use. Using our approach, allied countries can evaluate their Internet communications and determine a lower bound on which other countries have the access to eavesdrop on their traffic. If that set of countries contains only allies, the risk of information exposure is diminished (but not eliminated). If not, additional security measures should be considered for highly sensitive data.

Acknowledgment

This work was partially accomplished under NIST Cooperative Agreement No.70NANB16H024 with the University of Maryland. The authors would like to thank CAIDA and the University of Oregon for providing the data.

References

[1] "Center for applied internet data analysis data," Dec. 2017. [Online]. Available: http://www.caida.org/data
[2] "Scamper internet topology and performance probing tool," Dec. 2017. [Online]. Available: https://www.caida.org/tools/measurement/scamper
[3] "Archipelago (ark) measurement infrastructure," Dec. 2017. [Online]. Available: http://www.caida.org/projects/ark
[4] "Maxmind geolocation service," Dec. 2017. [Online]. Available: https://www.maxmind.com/en/home
[5] B. Huffaker, M. Fomenkov, and K. Claffy, "Geocompare: a comparison of public and commercial geolocation databases," *Proc. NMMC*, pp. 1–12, 2011.
[6] "Bgpstream toolset," Dec. 2017. [Online]. Available: https://bgpstream.caida.org
[7] "University of oregon routeviews project," Dec. 2017. [Online]. Available: http://www.routeviews.org
[8] M. E. Newman, "Assortative mixing in networks," *Physical review letters*, vol. 89, no. 20, p. 208701, 2002.
[9] A. Edmundson, R. Ensafi, N. Feamster, and J. Rexford, "Characterizing and avoiding routing detours through surveillance states," *arXiv preprint arXiv:1605.07685*, 2016.
[10] J. Karlin, S. Forrest, and J. Rexford, "Nation-state routing: Censorship, wiretapping, and bgp," *arXiv preprint arXiv:0903.3218*, 2009.
[11] A. Shah and C. Papadopoulos, "Characterizing international bgp detours," Technical Report CS-15-104, Colorado State University, Tech. Rep., 2015.
[12] P. Mell, R. Harang, and A. Gueye, "The resilience of the internet to colluding country induced connectivity disruptions," in *Proc. of the Workshop on Security of Emerging Networking Technologies*, 2015.
[13] ——, "Measuring limits on the ability of colluding countries to partition the internet," *International Journal of Computer Science: Theory and Application*, vol. 3, no. 3, 2015.

Gueye, Assane; Mell, Peter; Schanzle, Christopher.
"Quantifying Information Exposure in Internet Routing."
Paper presented at The 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications, New York, NY, United States. August 1, 2018 - August 3, 2018.

Mass Measurements of Focal Adhesions in Single Cells Using High Resolution Surface Plasmon Resonance Microscopy

Alexander W. Peterson*, Michael Halter, Alessandro Tona, Anne L. Plant, and John T. Elliott
Biosystems and Biomaterials Division, National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, MD 20899

ABSTRACT

Surface plasmon resonance microscopy (SPRM) is a powerful label-free imaging technique with spatial resolution approaching the optical diffraction limit. The high sensitivity of SPRM to small changes in index of refraction at an interface allows imaging of dynamic protein structures within a cell. Visualization of subcellular features, such as focal adhesions (FAs), can be performed on live cells using a high numerical aperture objective lens with a digital light projector to precisely position the incident angle of the excitation light. Within the cell-substrate region of the SPRM image, punctate regions of high contrast are putatively identified as the cellular FAs. Optical parameter analysis is achieved by application of the Fresnel model to the SPRM data and resulting refractive index measurements are used to calculate protein density and mass. FAs are known to be regions of high protein density that reside at the cell-substratum interface. Comparing SPRM with fluorescence images of antibody stained for vinculin, a component in FAs, reveals similar measurements of FA size. In addition, a positive correlation between FA size and protein density is revealed by SPRM. Comparing SPRM images for two cell types reveals a distinct difference in the protein density and mass of their respective FAs. Application of SPRM to quantify mass can greatly aid monitoring basic processes that control FA mass and growth and contribute to accurate models that describe cell-extracellular interactions.

Keywords: Surface plasmon, SPRM, focal adhesions, mass, density, cells, proteins

*alexander.peterson@nist.gov; phone 1 301 975-5665;

1. INTRODUCTION

Focal adhesions (FAs) are specialized multi-component protein complexes that permit communication between the interior of the cell and the extracellular matrix via integrin receptors and the actin cytoskeleton [1]. FAs contain many known proteins [2] and are involved in mechanical and chemical signaling. FA signaling is involved in a variety of cellular functions such as cell growth, morphogenesis, and cancer metastasis [3, 4]. Fluorescence microscopy is a primary tool used to quantitatively study FA morphology and dynamics [5]. This either requires immunofluorescent labelling of FAs or FAs with attached fluorescent labels. Therefore, only specifically labelled FA components can be visualized. Total internal reflection fluorescence microscopy (TIRFM) is one method used to create high resolution FA images [6]. The TIRFM evanescent wave is used to selectively probe fluorescence near the cell-substrate interface, allowing access to FA protein nanoarchitecture which resides well within 150 nm of the surface [7].

A technique such as surface plasmon resonance microscopy (SPRM) has the potential to be a useful orthogonal technique to that of fluorescence microscopy. It is a label-free surface sensitive imaging technique that uses conventional microscopy objectives to provide a mass and density measurement for FAs. This type of measurement fundamentally integrates all the protein assembly processes occurring in the FA growth process. Surface plasmon resonance (SPR) essentially measures the refractive index of a material at a thin metal surface [8]. The resonance minimum of SPR is sensitive to material near the surface and has the sensitivity to detect changes in surface protein binding [9]. SPR imaging is an approach to SPR that provides the ability to monitor spatial changes in reflectivity at an angle that is close to the resonance minimum [10]. These reflectivity values of the SPR image can be converted into index of refraction values by using the Fresnel model. SRPM is an extension of SPR imaging through a high numerical objective [11]. It provides the advantage of high magnification and compatibility with other microscopic imaging techniques. However, achieving the necessary spatial resolution needed to visualize subcellular features has been elusive until recently.

Elliott, John; Halter, Michael; Peterson, Alexander; Plant, Anne; Tona, Alessandro.
"Mass Measurements of Focal Adhesions in Single Cells Using High Resolution Surface Plasmon Resonance Microscopy."
Paper presented at SPIE Photonics West BIOS: Plasmonics in Biology and Medicine XV, San Francisco, CA, United States. January 27, 2018 - February 1, 2018.

We have developed an SPRM system that is ideally suited to measuring subcellular components such as focal adhesions [12]. Essentially, an incident arc of light, shaped by a digital light projector, illuminates a gold coated coverslip through a high NA microscope objective at a selected excitation angle and captures the reflected image on a CCD camera. By limiting the angle of excitation light, the SPR signal to noise ratio is enhanced and this allows near-diffraction limited lateral resolution with 150 nm penetration depth above the substrate. This spatial resolution enables visualization of subcellular organelles, such as cellular focal adhesions. Obtaining SPRM images through a high NA objective requires us to correct for optical aberrations prior to using the Fresnel model which provides optical parameters such as index of refraction [13]. Here we interpret cellular focal adhesions as an optical layer and measure the index of refraction, which we convert into a protein density. The differences in focal adhesion properties between two different cell lines were examined with the SPRM system.

2. METHODOLOGY

Disclaimer: Certain commercial equipment, instruments, or materials are identified here in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

2.1 SPR Microscopy

The details of the apparatus are described previously [12, 13]. Briefly, we performed SPR on an inverted microscope (Olympus IX-70, Center Valley, PA) with a high numerical aperture objective lens (100×, 1.65 NA, Olympus) by launching an arc of 590 nm incident light using a digital light projector at the SPR imaging angle through the objective and collecting the reflected light image onto a CCD camera.

2.2 Substrate Preparation

The details of substrate preparation have been described previously [12]. Essentially, specialized coverslips (18 mm diameter, n = 1.78, Olympus) were coated with ≈1 nm chromium and ≈45 nm gold. The gold coated coverslip was immersed in a 0.5 mmol/L hexadecanethiol solution in ethanol for 12 h to generate a self-assembled monolayer. The coverslip was then inserted into a sterile solution of 25 µg/mL bovine plasma fibronectin (Sigma, St. Louis, MO) in Ca^{2+}- and Mg^{2+}- free Dulbecco's phosphate buffered saline (DPBS; Invitrogen, Carlsbad, CA) for 1 h.

2.3 Cell Culture

The rat aortic vascular smooth muscle cell line, A10 (ATCC, Manassas, VA) was maintained in Dulbecco's Modified Eagles Medium with 25 mM HEPES (DMEM; Mediatech, Herndon, VA) supplemented with nonessential amino acids, glutamine, penicillin (100 units/mL), streptomycin (100 µg/mL), 10 % (v/v) fetal bovine serum (FBS) (Invitrogen); the human lung carcinoma cell line A549 (ATCC) was maintained in RPMI medium (Invitrogen) supplemented with glutamine, penicillin (100 units/mL), streptomycin (100 µg/mL), 10 % (v/v) FBS (Invitrogen); Both cell lines were maintained in a humidified 5 % (v/v) CO_2 balanced-air atmosphere at 37 °C. Cells were harvested with 0.25 % (w/v) trypsin-EDTA (Invitrogen), and seeded in growth medium onto the fibronectin coated substrates at a density of 1000 cells/cm^2. After 72 h incubation, cells on the substrates were washed with warm Hanks Balanced Salt Solution (HBSS; ICN Biomedicals, Costa Mesa, CA), fixed in 1 % (w/v) paraformaldehyde (EMS, Hatfield, PA) in Dulbecco's phosphate buffered saline (DPBS; Invitrogen) for 30 min at room temperature, quenched in 0.25 % (w/v) NH_4Cl in DPBS (15 min) and rinsed with DPBS. The substrates were then overlaid with a fluidic chamber made out of polydimethylsiloxane (PDMS) and kept in DPBS during all microscopy measurements.

2.4 SPRM Image Collection and Processing

A SPRM image is created with a p- and s-polarized image taken by rotating the linear polarizer 90° while using the arc-shaped incident illumination at an angle near the SPR minimum. The p-polarized image is divided by the s-polarized image to create a reflectivity image. The p/s intensities for each pixel are then divided by the apodization correction factor, which is a function of the calculated incident angle [13]. The result is an image with normalized and corrected reflectivity units. The images are further modified to convert the reflectivity units into Δ-reflectivity (ΔR) by using $\Delta R = R_1 - R_0$ where R_1 is the normalized reflectivity unit of the sample and R_0 is the average reflectivity of the SPR image

Elliott, John; Halter, Michael; Peterson, Alexander; Plant, Anne; Tona, Alessandro.
"Mass Measurements of Focal Adhesions in Single Cells Using High Resolution Surface Plasmon Resonance Microscopy."
Paper presented at SPIE Photonics West BIOS: Plasmonics in Biology and Medicine XV, San Francisco, CA, United States. January 27, 2018 - February 1, 2018.

background in phosphate buffered saline (PBS) buffer. For subsequent analysis and comparison, the ΔR units are converted to index of refraction units according to the Fresnel model as described previously [13]. All image analysis was performed using ImageJ software with additional custom script programming. Angle-dependent SPR data were analyzed using stock and custom code written in MATLAB (Mathworks, Natick, MA).

2.5 Fluorescence Staining and Image Collection

After SPR imaging, the previously fixed substrates were permeabilized with 0.05 % (v/v) Triton X-100 (Sigma) in phosphate buffered saline (PBS) for 10 minutes at room temperature and blocked in 10 % (v/v) goat serum/1 % (w/v) bovine serum albumin in PBS (blocking solution) for 30 minutes at room temperature. The samples were then stained with monoclonal anti-vinculin antibody (Sigma) diluted 1:200 in blocking solution for 1 hour at room temperature. After rinsing 3× with PBS and blocking again with blocking solution for 30 minutes at room temperature, the samples were stained with Alexa-488 goat anti-mouse secondary antibody (Invitrogen) diluted 1:100 in blocking solution for 45 minutes at room temperature. Finally, the samples were rinsed 3× in PBS and stored in PBS for imaging. Fluorescence images were acquired with a 1.3 NA, 63× objective on an upright Zeiss microscope (Zeiss, Jena, Germany) using a standard FITC filter cube set. Each fluorescence image was registered to the corresponding SPR image using 2 fiduciary marks according to the TurboReg plugin in the ImageJ software.

2.6 Focal Adhesion Image Analysis

Fluorescently stained α-vinculin images were processed and analyzed for FA area measurements using ImageJ software according to a published method [14]. The overlays generated by the α-vinculin FA analysis were used to guide the manual threshold selection for the corresponding SPRM image. The average change in reflectivity for each FA area in a SPRM image is converted into a change in refractive index as previously described [13] and then converted into mass density according to the specific refractive index increment (0.18 mL/g) for biomolecules widely used in optical live-cell mass profiling techniques [15]. The terminology typically used for optical mass measurements refers to 'dry mass' as the mass of all biomolecular components other than water. Here, our interpretation is analogous as we measure refractive index shifts compared to buffered media, however, we have adopted the nomenclature 'protein mass' and 'protein density' to describe the multi-protein component structures of FAs. The protein density value for each FA (in units of $fg/\mu m^3$) is then converted into protein mass by multiplying the measured lateral FA area (μm^2) and the measured axial penetration depth of the surface plasmon (0.15 μm) [12]. ≈450 FAs were measured for A549 cells, and ≈600 FAs were measured for A10 cells. The basal cell mass density is calculated by averaging the SPRM contrast attributed to the footprint of the cell minus the area of regions attributed to FAs. This is done as a per cell measurement with 9 cells each.

3. RESULTS AND DISCUSSION

3.1 SPRM Technique and Focal Adhesion Size

The details of our surface plasmon resonance microscopy (SPRM) apparatus has been described previously [12, 13]. Briefly, an incident arc of light at 590 nm is shaped by a digital light projector to illuminate the SPR imaging angle through a microscope objective off a gold coated coverslip and capture the reflected image onto a CCD camera (Figure 1A). This provides diffraction-limited spatial resolution for SPR imaging of cellular samples. The image contrast in reflectivity units can be converted into index of refraction units using the Fresnel model [13]. Here, we apply SPRM to quantitatively analyze cellular focal adhesions (FA) by converting index of refraction units into protein density by using the refractive index increment used for biomolecular components [15]. Subsequently, we can measure the protein mass of FAs by multiplying the measured protein density value by the measured lateral area of the FA and the previously measured axial distance of the SPR penetration depth.

Elliott, John; Halter, Michael; Peterson, Alexander; Plant, Anne; Tona, Alessandro.
"Mass Measurements of Focal Adhesions in Single Cells Using High Resolution Surface Plasmon Resonance Microscopy."
Paper presented at SPIE Photonics West BIOS: Plasmonics in Biology and Medicine XV, San Francisco, CA, United States. January 27, 2018 - February 1, 2018.

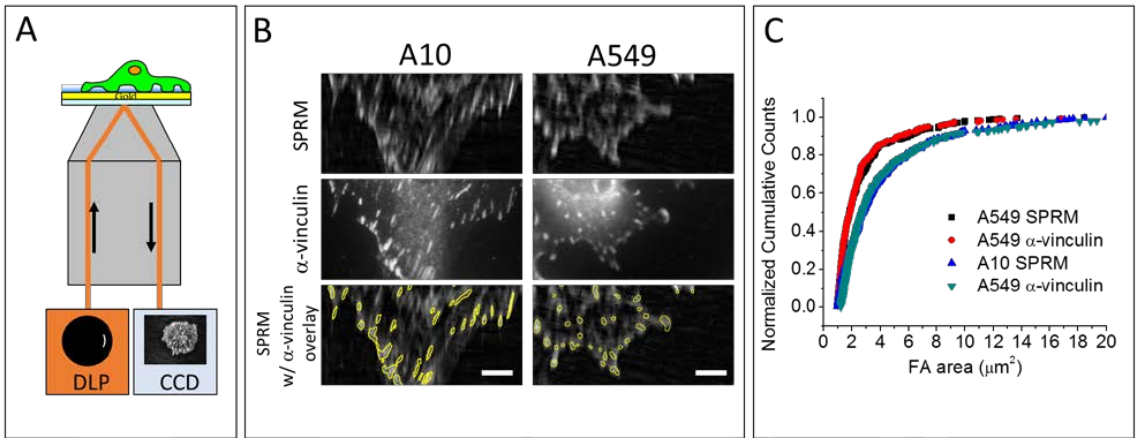


Figure 1: Surface plasmon resonance microscopy (SPRM) of A10 and A549 cellular focal adhesions (FA) show similar area measurements as that for fluorescently stained α-vinculin. A) Simplified SPRM schematic depicting an incident arc of light at the SPRM imaging angle through a microscope objective and resulting SPRM image captured on a CCD camera. B) SPRM images of an A10 and A549 cell, same cells subsequently labelled with α-vinculin and fluorescently imaged, intensity threshold of α-vinculin overlaid onto corresponding SPRM image for comparison. Scale bar = 10 μm. C) Normalized cumulative counts plots for FA area showing similar distributions for SPRM and fluorescently stained α-vinculin for each cell type, with distinct median values of 2.7 μm^2 for A10 and 1.8 μm^2 for A549 cells.

Two cell types, vascular smooth muscle cells (A10) and adenocarcinomic alveolar basal epithelial cells (A549) were seeded and fixed after 72 h on a fibronectin coated substrate as described in Methodology 2.3. The SPRM images of representative A10 and A549 cells show regions of high optical contrast that are directly related to the mass density of the cellular components within the evanescent wave (Figure 1B). For comparison, immunofluorescently labeled α-vinculin, a known focal adhesion associated protein, is displayed next to the SPRM image to visualize the cellular FAs. The bottom combination images depict the outline of a threshold performed on the fluorescence images of α-vinculin overlaid on the SPRM images. From these images, we can observe that the bright punctate regions in the SPR image match well with the α-vinculin stained regions and therefore SPRM is likely detecting FAs as regions of higher protein density. Image analysis was used to measure FA area for α-vinculin fluorescent images of A10 and A549 cells to help guide the manual thresholding of the FAs in the SPRM images. Comparing the cumulative counts of FA areas between SPRM and α-vinculin images and between A10 and A549 cells, it is revealed that SPRM can closely match the FA area distribution profile for α-vinculin, and that there is an observable difference in FA area between A10 and A549 cells. The median FA area for A10 cells is 2.7 μm^2 and 1.8 μm^2 for A549 cells. A10 cells are known to have large FAs on stiff substrates [16] and A549 cells are described as having substantially smaller FAs than normal cells [17].

3.2 Focal Adhesion Protein Density

Using the outlined FA areas for A10 and A549 cells with SPRM, the protein density for each FA is calculated as described above. In addition, the basal cell surface, the area of SPRM cellular contrast that corresponds to cellular area, minus the areas attributed to FA area, is averaged to calculate a basal cell mass density. Consequently, the average protein densities and standard deviations for the A10 and A549 FAs, and basal cell areas are compared (Fig. 2A) revealing that FAs in A10 cells have nearly twice the protein density as FAs in A549 cells. In addition, the protein densities of FAs in both cell types are significantly larger than the basal cell background densities which represents the mass density of the general cytoplasm ($p < 0.0001$).

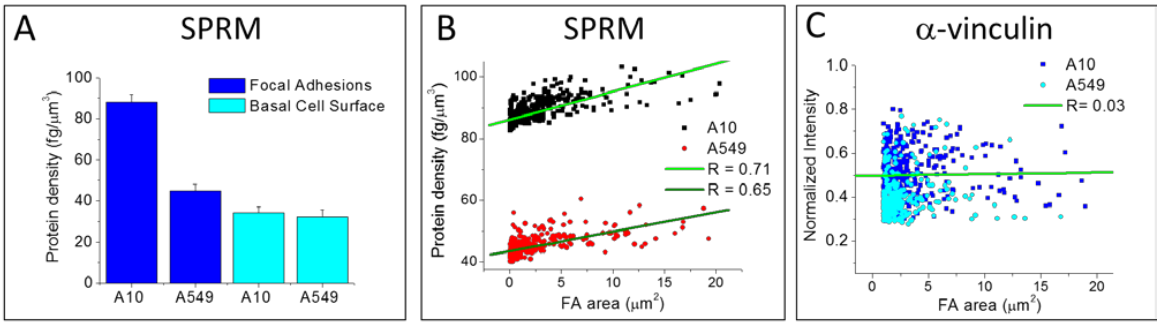


Figure 2: SPRM measured protein densities for focal adhesions of A10 and A549 cells show an increase in protein density with an increase in FA area. A) FA protein density values reported show A10 FAs are significantly denser than A549 FAs. Both cell type FAs have larger densities than that of the basal cells surfaces which have similar lower densities. B) FA protein density for A10 and A549 cells show a strong correlation with FA area (R=0.71 for A10, R=0.65 for A549). C) Normalized fluorescence intensity for α-vinculin shows no observable correlation with FA area (R=0.03 for A10, R=0.10 for A549 (fit not shown)).

In contrast, basal cell mass densities of A10 and A549 cells are statistically similar ($p > 0.65$). Protein densities of FAs for both A10 and A549 cells show a strong correlation (R=0.71 for A10, R=0.65 for A549) when plotted versus FA area, Figure 2B. This correspond to ≈ 1 fg/μm^3 gain in protein density per 1 μm^2 of FA area for A10 cells while half of that value for A549 cells. In contrast, fluorescence intensity of α-vinculin FAs plotted versus FA area reveals no correlation (R=0.03 for A10, R=0.10 for A549), Figure 2C. A possible explanation is that, in the case for α-vinculin, it is a FA associated protein that has been measured to reside in a specific plane of the FA [7], therefore it would not be predicted to change in density along with FA size, rather it will remain at constant density. However, other FA associated proteins, such as actin stress fibers, may occupy more volume in the FA [7] and may be more dynamic in abundance and density in the FA. Actin fibers in A10 cells are described as expanding when cells are on stiff substrates [16] and actin fibers are described as diminished in A549 cells compared to normal cells [17]. Regardless, SPRM measures the integrated overall protein density and large differences are observed in protein density between FAs of different cells types as well as smaller differences in FA protein density as a function of FA area.

3.3 Focal Adhesion Protein Mass

Protein mass measurements of FAs are measured by multiplying the average protein density of the FA by the volume of the FA. In our case, the volume of FA is the measured FA area multiplied by the length of the SPR penetration depth into the cell. Observing the cumulative counts of FA mass for A10 and A549 cells reveals a distinct distribution for each cell type, where the median mass for A10 FAs is 46 fg while the median mass for A549 FAs is 19 fg, Figure 3A.

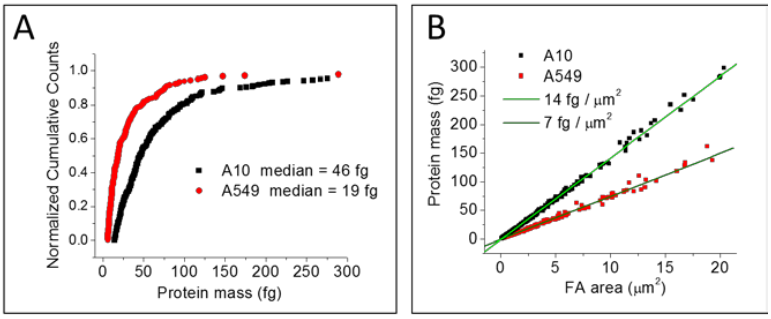


Figure 3: Focal adhesion protein mass measurements show that A10 FAs have significantly more mass than FAs for A549 cells. A) Normalized cumulative distribution plot for FA protein mass shows a median value of 46 fg for A10 cells and 19 fg for A549 cells. B) FA protein mass plotted versus FA area shows a strong positive correlation (R=0.99) for A10 and A549 cells. The slope of a linear fit reports a larger mass gain of 14 fg /μm^2 of FA area for A10 cells compared to 7 fg /μm^2 of FA area for A549 cells.

The separation in these mass distributions appears to be more distinct than that of FA area alone, Figure 1C. Evaluating FA protein mass versus FA area shows a strong correlation between FA protein mass and FA size for both A10 and A549 cells, Figure 3B. However, the measured slopes of protein mass versus FA area show distinct values of 14 fg/μm^2 for A10 FAs and 7 fg/μm^2 for A549 cells. The large differences in FA protein mass between these cell types may be due to the presence or absence of certain FA components, such as actin stress fiber attachments. The strong linear correlation of FA protein mass with FA area may indicate a level of homogeneity in FA growth response. Additionally, both A10 and A549 cells grown on flat, stiff, fibronectin coated surfaces were non-motile cells which may contribute to more homogenous FA area and mass response.

4. CONCLUSION

We have created a SPRM system that can obtain near-diffraction lateral resolution and has an evanescent wavelength of 150 nm that is able to visualize, label-free, subcellular components, such as FAs, and is ideally suited to make quantitative measurements of FA protein density and mass. The protein density and mass measurements on FAs show distinct differences between FAs for two different cells types. These protein mass measurements are tied to biophysical processes that have significant meaning for understanding of FA formation and development. The results here are very promising and continued work will be on evaluating FA dynamics in live-cells.

REFERENCES

[1] S. K. Sastry, and K. Burridge, "Focal adhesions: A nexus for intracellular signaling and cytoskeletal dynamics," Experimental Cell Research, 261(1), 25-36 (2000).

[2] K. Burridge, and M. ChrzanowskaWodnicka, "Focal adhesions, contractility, and signaling," Annual Review of Cell and Developmental Biology, 12, 463-518 (1996).

[3] B. Geiger, J. P. Spatz, and A. D. Bershadsky, "Environmental sensing through focal adhesions," Nature Reviews Molecular Cell Biology, 10(1), 21-33 (2009).

[4] M. A. Schwartz, M. D. Schaller, and M. H. Ginsberg, "Integrins: Emerging paradigms of signal transduction," Annual Review of Cell and Developmental Biology, 11, 549-599 (1995).

[5] M. E. Berginski, E. A. Vitriol, K. M. Hahn *et al.*, "High-Resolution Quantification of Focal Adhesion Spatiotemporal Dynamics in Living Cells," Plos One, 6(7), (2011).

[6] A. L. Mattheyses, S. M. Simon, and J. Z. Rappoport, "Imaging with total internal reflection fluorescence microscopy for the cell biologist," Journal of Cell Science, 123(21), 3621-3628 (2010).

[7] P. Kanchanawong, G. Shtengel, A. M. Pasapera *et al.*, "Nanoscale architecture of integrin-based cell adhesions," Nature, 468(7323), 580-U262 (2010).

[8] J. Homola, [Surface Plasmon Resonance Based Sensors] Springer, Berlin(2006).

[9] D. Altschuh, M. C. Dubs, E. Weiss *et al.*, "Determination of Kinetic Constants for the Interaction between a Monoclonal-Antibody and Peptides Using Surface-Plasmon Resonance," Biochemistry, 31(27), 6298-6304 (1992).

[10] L. K. Wolf, D. E. Fullenkamp, and R. M. Georgiadis, "Quantitative angle-resolved SPR imaging of DNA-DNA and DNA-drug kinetics," J Am Chem Soc, 127(49), 17453-17459 (2005).

[11] B. Huang, F. Yu, and R. N. Zare, "Surface plasmon resonance imaging using a high numerical aperture microscope objective," Analytical chemistry, 79(7), 2979-2983 (2007).

[12] A. W. Peterson, M. Halter, A. Tona *et al.*, "High resolution surface plasmon resonance imaging for single cells," BMC Cell Biol, 15, 35 (2014).

[13] A. W. Peterson, M. Halter, A. L. Plant *et al.*, "Surface plasmon resonance microscopy: Achieving a quantitative optical response," Review of Scientific Instruments, 87(9), (2016).

Elliott, John; Halter, Michael; Peterson, Alexander; Plant, Anne; Tona, Alessandro.
"Mass Measurements of Focal Adhesions in Single Cells Using High Resolution Surface Plasmon Resonance Microscopy."
Paper presented at SPIE Photonics West BIOS: Plasmonics in Biology and Medicine XV, San Francisco, CA, United States. January 27, 2018 - February 1, 2018.

[14] U. Horzum, B. Ozdil, and D. Pesen-Okvur, “Step-by-step quantitative analysis of focal adhesions,” MethodsX, 1, 56-9 (2014).
[15] T. A. Zangle, and M. A. Teitell, “Live-cell mass profiling: an emerging approach in quantitative biophysics,” Nature Methods, 11(12), 1221-1228 (2014).
[16] A. L. Plant, K. Bhadriraju, T. A. Spurlin *et al.*, “Cell response to matrix mechanics: Focus on collagen,” Biochimica Et Biophysica Acta-Molecular Cell Research, 1793(5), 893-902 (2009).
[17] M. M. Alam, J. Hooda, D. Cadinu *et al.*, “Comparative proteomic analysis of an isogenic pair of lung normal and lung cancer cell line,” Faseb Journal, 27, (2013).

Elliott, John; Halter, Michael; Peterson, Alexander; Plant, Anne; Tona, Alessandro.
"Mass Measurements of Focal Adhesions in Single Cells Using High Resolution Surface Plasmon Resonance Microscopy."
Paper presented at SPIE Photonics West BIOS: Plasmonics in Biology and Medicine XV, San Francisco, CA, United States. January 27, 2018 - February 1, 2018.

Development and Validation of a Mechanism for Flame Propagation in R-32/Air Mixtures

Donald R. Burgess, Jr., Jeffrey A. Manion, Robert R. Burrell, Valeri I. Babushok, Michael J. Hegetschweiler, Gregory T. Linteris

Abstract

A mechanism for the combustion of the refrigerant R-32 (CH2F2) in air mixtures was developed and validated through comparisons with measured flame speeds for a range of equivalence ratios (0.9 to 1.4) and pressures (1 to 3 bar) using a constant-volume spherical flame method. Premixed flame calculations were performed and analyzed to identify primary species and reactions contributing to flame speeds and combustion. We found that there were only three HFC reactions that contributed significantly to flame speeds. Their rate constants were optimized within uncertainty limits and the model showed excellent agreement (<3 %) with measured flames speeds.

Keywords: Chemical kinetics, Refrigerant flammability, Burning velocity, Hydrofluorocarbons

1. Introduction

The overall purpose of this work is to characterize the flammability of a set of fluoromethanes, fluoroethanes, and fluoropropenes, and their mixtures for use as refrigerant working fluids. Although it is possible to make measurements of the flammability of a single refrigerant under a limited set of conditions, it is not realistic to measure flammabilities of all possible formulations under a wide range of refrigerant-to-air ratios for different diluents, ambient temperatures, and humidity levels. The driving force for this work is the development of new refrigerant blends that simultaneously minimizes their global warming potentials (GWP) and flammabilities, while maximizing their refrigerant performance which is a function of thermodynamic and physical properties (e.g., critical temperature, vaporization enthalpy, thermal conductivity, saturation vapor pressure). The ability to have robust and accurate predictive tools that are benchmarked to high quality measurements of the flammability of these refrigerants for a set of specific mixtures under a range of conditions will allow industry to screen, optimize, and rank different blends to enable rapid development of new refrigerant formulations.

Babushok, Valeri; Burgess Jr., Donald; Burrell, Robert; Hegetschweiler, Michael; Linteris, Gregory; Manion, Jeffrey.
"Development and Validation of a Mechanism for Flame Propagation in R-32/Air Mixtures."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

In the work reported here, the flammability of the refrigerant R-32 (CH_2F_2, difluoromethane) is studied using burning velocities (flame speeds). R-32 is a widely used refrigerant that is non-ozone depleting and has a moderate GWP. It, however, is mildly flammable, and is used with less flammable refrigerants in blends. This current work on R-32, a standard refrigerant, will provide a benchmark for extending flammability models based on elementary reaction kinetics to other refrigerants (both pure and blends) such as the HFC's R-125 (pentafluoroethane), R-134a (1,1,1,2-tetrafluoroethane), R-152a (1,1,-difluoroethane), and the hydrofluoroolefins HFO-1234yf (2,3,3,3-tetrafluoropropene) and HFO-1234ze (1,3,3,3-tetrafluoropropene).

2. Methods

In this work, we developed and validated a chemical kinetic mechanism/model to predict the flammability of the refrigerant R-32 (CH_2F_2). This requires understanding the chemistry on a microscopic level from the analysis of reaction pathways and comparison of the model predictions to measured flame speeds.

The full details of the measurements in this work are reported elsewhere [1] and are only summarized here. The flame speeds of R-32/air mixtures were measured using a constant-volume spherical-flame method. A spherical chamber about 15 cm diameter (~1.8 liters) was filled with R-32/air mixtures with equivalence ratios ranging from 0.9 to 1.4, and using different initial pressures on the order of (0.87 to 1.13) bar. The mixtures were then ignited by a spark at the center of the chamber, and the pressure rise (final pressures of about 7 to 11 bar) as a function of time was monitored. The final pressures were about 7 to 11 bar, but instabilities, interaction with the chamber wall, and other effects limited the reliable data to final pressures of about 1 to 3 bar. The pressure traces were then used to calculate flame radius, and thus flame speeds S_u(T,P), using a thermodynamic spherical flame propagation model. In the data reduction procedures, corrections were made to account for thermal radiation of the burned gas [3] and flame stretch [4]. Buoyancy is unimportant for the present conditions [5] and was not considered.

In this work, the development and validation of a chemical kinetic mechanism (model) to predict the flammability of R-32 (CH_2F_2) was iterative. An initial chemical kinetic mechanism was

Babushok, Valeri; Burgess Jr., Donald; Burrell, Robert; Hegetschweiler, Michael; Linteris, Gregory; Manion, Jeffrey.
"Development and Validation of a Mechanism for Flame Propagation in R-32/Air Mixtures."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

developed based on compilations from the literature, updates to reflect more recent work, and an evaluation of reactions and their rate constants. Simulations were performed using the Sandia Premix code [6] and Cantera [7] and the results examined using reaction path analysis employing the graphics post-processor XSenkplot [8] to identify important species and reactions. The refined mechanism was used to simulate the flame speeds. Rough agreement between the computed and measured flames speeds was initially found, and the rate constants were then adjusted/optimized within their uncertainty limits (factors of about 1.3 to 2.0) to achieve best agreement.

3. Rate Constant Evaluation

We provide here a short discussion of the rate constants evaluated and utilized in this work. Key rate expressions are provided in Table 1. The rate expressions for hydrogen-oxygen chemistry and hydrocarbon/oxidized hydrocarbon chemistry were taken from GRI-Mech 3.0 [9], while those for H/O/F chemistry were based on our fits to rate constants reported in the literature [10-12].

The rate expression for $CH_2F_2 \rightarrow CHF + HF$ was derived from the shock tube measurements for the analogous reaction $CHF_3 \rightarrow CF_2 + HF$ by Schug and Wagner [13] and employing the very rough relative decomposition rates for CH_2F_2 and CHF_3 from the pyrolysis study of Politanskii and Shevchuk [14]. This scaled rate expression was then refined during our optimizations. It was found that the flame speeds could not be modeled with a pressure independent rate constant for this unimolecular reaction. It was necessary to use a pressure-dependent rate constant of the form $Rate = k_0[M] + k_1$, where the pressure dependent term $k_0[M]$ dominated – suggesting that the reaction was approaching the low pressure limit under (P, T) conditions in the flame.

The rate expressions for H abstractions from CH_2F_2 by the flame radicals H, O, OH, and F were evaluated in this work. In our evaluation, we utilized rate constants from the literature [15-21] and then fit them to extended Arrhenius rate expressions to provide rate constants over a wide range of temperatures. Depending upon the reaction, available data, and fitting procedures, we estimate uncertainty factors on the order of about (1.3 to 1.6) for these reactions.

Babushok, Valeri; Burgess Jr., Donald; Burrell, Robert; Hegetschweiler, Michael; Linteris, Gregory; Manion, Jeffrey.
"Development and Validation of a Mechanism for Flame Propagation in R-32/Air Mixtures."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

The rate constants for $CHF_2 + O_2 \rightarrow CF_2O + OH$ and $CHF + O_2 \rightarrow CHFO + O$ were estimated by analogy to those for $CH_3 + O_2 \rightarrow CH_2O + OH$ and $CF_2 + O_2 \rightarrow CF_2O + O$, respectively [22-23]. Both of the former reactions are significantly more exothermic (~100 kJ mol^{-1}) than the reference reactions, and consequently should have much smaller barriers. For the $CHF_2 + O_2$ and $CHF + O_2$ reactions, we initially estimated activation energies on the order of (15 to 20) kJ mol^{-1} and (20 to 25) kJ mol^{-1}, but our optimizations compared to our measured flame speeds suggest somewhat smaller activation energies – on the order of (5 to 15) kJ mol^{-1} and (15 to 25) kJ mol^{-1}, respectively.

The rate expression for $CHF_2 + CHF_2 \rightarrow CHF{=}CF_2 + H$ used in this work is from the earlier ab initio transition state and RRKM calculations by Burgess et al [24]. Although this is a primary pathway, this reaction has a small impact on flames speeds (< 3 %) and combustion – it simply provides a pathway to products. We utilized the rate expression for $CHF + H_2O \rightarrow CH_2O + HF$ from the ab initio transition state and RRKM calculations of Zachariah et al [25]. This reaction is a major destruction pathway for CHF and the primary pathway leading to the formation of HO_2 ($CH_2O + OH \rightarrow CHO + H_2O$, followed by $CHO + O_2 \rightarrow CO + HO_2$). This reaction, however, contributes a negligible amount (<0.1 %) to changes in flame speeds – it is largely a conduit for establishing steady state concentrations during the pre-ignition process.

For the reactions R• (CHF_2, CF_2, CHF) + X (H, O, OH, F), we utilized rate constants from the literature. These reactions are generally very fast and contribute negligible amounts to flame speeds and little to the combustion chemistry other than to provide pathways to complete combustion. Similarly, reactions involving CHFO, CF_2O, and CFO, contribute negligible amounts to changes in flame speeds, and provide pathways to combustion products.

Babushok, Valeri; Burgess Jr., Donald; Burrell, Robert; Hegetschweiler, Michael; Linteris, Gregory; Manion, Jeffrey.
"Development and Validation of a Mechanism for Flame Propagation in R-32/Air Mixtures."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

Table 1: Rate parameters for elementary reactions employed in the model. The rate constants have the Arrhenius form $k = Ae^{-E/RT}$ where A, E, and R are the pre-exponential, activation energy, and the gas constant, respectively. Units are mol, cm^3, s, kJ. Note: these are only rate expressions developed as part of this work. Rate expressions for other reactions used in the model were taken from the literature (see text).

Reaction	**A**	**E**
$CH_2F_2 + M \rightarrow CHF + HF + M$	5.9E+17	295.0
$CH_2F_2 \rightarrow CHF + HF$	4.9E+11	272.0
$CH_2F_2 + H \rightarrow CHF_2 + H_2$	2.5E+14	62.5
$CH_2F_2 + O \rightarrow CHF_2 + OH$	1.9E+14	56.3
$CH_2F_2 + OH \rightarrow CHF_2 + H_2O$	1.7E+13	30.0
$CH_2F_2 + F \rightarrow CHF_2 + HF$	3.1E+14	19.4
$CHF_2 + O_2 \rightarrow CF_2O + OH$	4.9E+10	14.6
$CHF + O_2 \rightarrow CHFO + O$	2.2E+13	20.9

Babushok, Valeri; Burgess Jr., Donald; Burrell, Robert; Hegetschweiler, Michael; Linteris, Gregory; Manion, Jeffrey.
"Development and Validation of a Mechanism for Flame Propagation in R-32/Air Mixtures."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

4. Results and Discussion

Figure 1 shows the experimentally-derived flames speeds (points) and the modeled flame speeds (curves). The solid symbols show the flame speeds corrected using an optically thin radiation model, while the open symbols show the uncorrected flames for 1 bar (2 and 3 bar data not shown for clarity). The deviations between the experimental flame speeds and the model predictions are on the order of (1 to 3) %. This is excellent agreement given that the uncertainties in the flames speeds are on the order of (10 to 20) % due to uncertainties in the rate constants which are multiplicative factors of about 1.5 to 2.0. As seen in Figure 1, the flame speeds are roughly linear with pressure suggesting the rate constant $Rate = k_0[M] + k_1$ is approaching the low pressure limit.

The optically thin (limit) radiation model correction increases flames speeds by about (19±2) % relative to those computed using adiabatic flame temperatures (shown for 1 bar) and is independent (<1 %) of equivalence ratio. When estimated radiation absorption corrections are used for each major species, flame speeds decrease by about (8±3) % relative to those using the optically thin radiation model (not shown). Given the insensitivity of the corrections to equivalence ratio, the rate constants for the primary reactions contributing to flame speeds can be easily adjusted to correct for absorption of radiation. The correction using the optically thin radiation model translates into changes in the rate constants of about 40% (this is shown for 1 bar). This is lower than the uncertainties in the rate constants which have uncertainty factors of about 1.4 to 2.0.

We note that under lean conditions (where flame stretch is most important) the measured flame speeds $S_u(T,P)$ appear to be slightly higher (3 to 6) % than the modeled flame speeds. Recently, we have reprocessed the data better fitting the burning velocity / temperature-pressure curves $S_u(T,P)$ and find lower burning velocities under lean conditions at 1 atm in good agreement (<3 %) with the model (these data are not shown, because we have not re-optimized model yet). In addition, we are currently in the process of optimizing the temperature and pressure dependencies of the rate constants by considering measurements that we have made using argon/O_2 mixtures (instead of air mixtures). The lower heat capacity of argon increases flame temperatures and thus increases flame speeds (by about a factor of 2). The refined experimental data and re-optimized rate constants are likely to provide better agreement between measured and model flame speeds.

Babushok, Valeri; Burgess Jr., Donald; Burrell, Robert; Hegetschweiler, Michael; Linteris, Gregory; Manion, Jeffrey.
"Development and Validation of a Mechanism for Flame Propagation in R-32/Air Mixtures."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

Figure 2 shows the dependence of flame speeds for the three most important reactions as a function of equivalence ratio. The dependencies are on the order of (5 to 20) %, while all other reactions are less than 3 % and relatively independent of equivalence ratios. The quantity shown here is the percent change in flame speed relative to a change in rate constant. That is, a value of 20 % would mean increasing a rate constant by 10 % would result in an increase of 2 % in flame speed. The two most important reactions are the unimolecular decomposition of the refrigerant $CH_2F_2 \rightarrow CHF + HF$ and the subsequent reaction of its decomposition product CHF with oxygen in the mixture $CHF + O_2 \rightarrow CHFO + O$. Both of these reactions contribute to increases in flame speeds. In contrast, H abstraction from the refrigerant $CH_2F_2 + H \rightarrow CHF_2 + H_2$ causes a decrease in flame speeds – it inhibits the flame by consuming the radical H atom during pre-ignition that otherwise would be involved in radical chain branching leading to ignition.

Reaction pathway analysis and flame speed dependencies of rate constants show that burning of R-32 can be described by two quasi-separate reaction pathways. The first stage is pre-ignition driven by unimolecular decomposition of CH_2F_2 forming CHF. The subsequent second stage is a combustion pathway driven by H abstraction by flame radicals H, O, OH, and F forming CHF_2. Both CHF and CHF_2 then react quickly with the reactant O_2 to form the fluorocarbonyls CHFO and CF_2O, respectively. The pre-ignition pathway is primarily responsible for generation of initial flame radicals and hence drives flame propagation, while the combustion pathway occurs after ignition, mediated by flame radicals, and drives the process to products. Interestingly, reactions in the combustion pathway actually inhibit ignition by consuming radicals needed for chain branching – tying up the chemistry in relatively stable intermediates that are slow to decompose.

Babushok, Valeri; Burgess Jr., Donald; Burrell, Robert; Hegetschweiler, Michael; Linteris, Gregory; Manion, Jeffrey.
"Development and Validation of a Mechanism for Flame Propagation in R-32/Air Mixtures."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

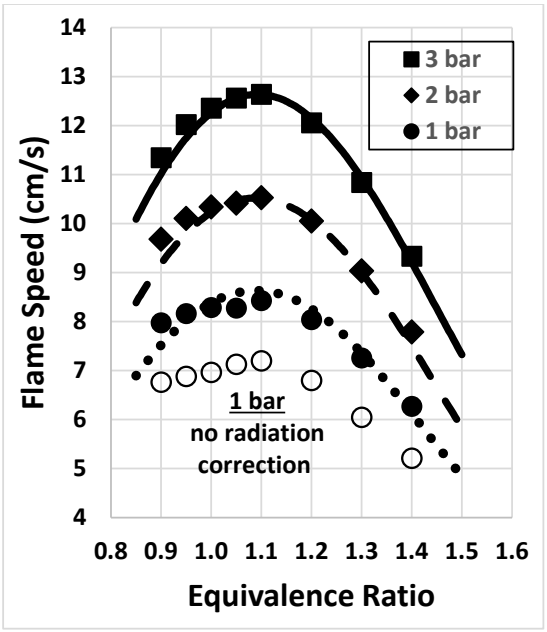


Figure 1: Comparison between measured and model flame speeds for CH_2F_2/air mixtures.

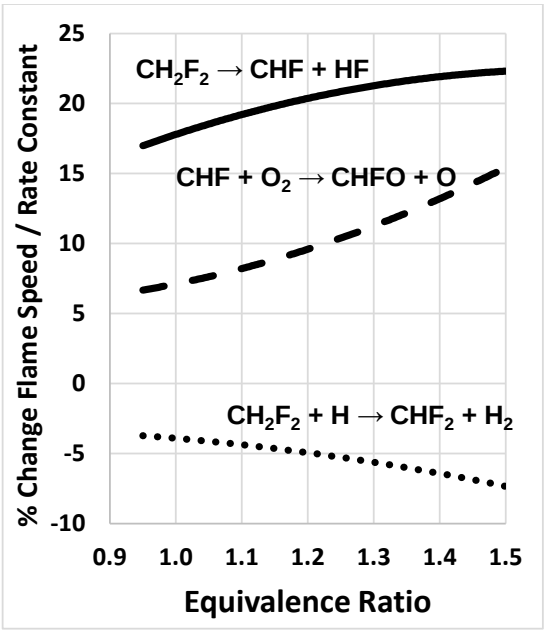


Figure 2: Dependence of flame speeds on the three most important rate constants.

Babushok, Valeri; Burgess Jr., Donald; Burrell, Robert; Hegetschweiler, Michael; Linteris, Gregory; Manion, Jeffrey.
"Development and Validation of a Mechanism for Flame Propagation in R-32/Air Mixtures."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

5. Conclusions

In this work, a mechanism for the combustion of the refrigerant R-32 (CH_2F_2) in air mixtures was developed and validated through comparisons with measured flame speeds for a range of equivalence ratios (0.9 to 1.5) and pressures (1 to 3 bar) using a constant-volume spherical flame method. Premixed flame calculations were performed and analyzed to identify primary species and reactions contributing to flame speeds and combustion. We found that there were just three HFC reactions that contributed significantly to flame speeds. Their rate constants were optimized (within their uncertainty limits) and the model showed excellent agreement (<3 %) with measured flame speeds. Ongoing experiments and modeling will refine this model using other conditions such using Ar/O_2 mixtures (instead of air) to change flame temperatures and the addition of H_2O and H_2 to change the concentration of flame radicals. The radiation parameters will also be improved. This work will be also extended to other refrigerants: the fluoroethanes R-125, R-134a, and R-152a and the hydrofluoroolefins HFO-1234yf (2,3,3,3-tetrafluoropropene) and HFO-1234ze (1,3,3,3-tetrafluoropropene).

6. Acknowledgements

This work was supported by the Buildings Technologies Office of the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy under contract no. DE-EE0007615 with Antonio Bouza serving as Project Manager.

Babushok, Valeri; Burgess Jr., Donald; Burrell, Robert; Hegetschweiler, Michael; Linteris, Gregory; Manion, Jeffrey.
"Development and Validation of a Mechanism for Flame Propagation in R-32/Air Mixtures."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

7. References

[1] R. Burrell, J. L. Pagliaro, G. T. Linteris, Effects of stretch and thermal radiation on difluoromethane-air burning velocity measurements in constant volume spherically expanding flames, Proc. Combust. Inst. 37 (2018) submitted.

[2] C. Xiouris, T.L. Ye, J. Jayachandran, F. N. Egolfopoulos, Laminar flame speeds under engine-relevant conditions: Uncertainty quantification and minimization in spherically expanding flame experiments, Combust. Flame 163 (2016) 270-283.

[3] Z. Chen, Effects of radiation absorption on spherical flame propagation and radiation-induced uncertainty in laminar flame speed measurement, Proc. Combust. Inst. 36 (2017) 1129-1136.

[4] F.J. Wu, W.K. Liang, Z. Chen, Y.G. Ju, C.K. Law, Uncertainty in Stretch Extrapolation of Laminar Flame Speed from Expanding Spherical Flames, Proc. Combustion Inst. 35 (2015) 663-670.

[5] K. Takizawa, S. Takagi, K. Tokuhashi, S. Kondo, M. Mamiya, H. Nagai, Assessment of. Burning Velocity Test Methods for Mildly Flammable Refrigerants, Part 1: Closed-Vessel Method, ASHRAE Trans. 119 (2013) 243-254.

[6] R. J. Kee, J. F. Grcar, M. D. Smooke, J. A. Miller, PREMIX: A Fortran Program for Modeling Steady Laminar One-Dimensional Premixed Flames (Version 2.5b), Report No. SAND85-8240, Sandia National Laboratories, Livermore, California, 1992.

[7] D. G. Goodwin, H. K. Moffat, R. L. Speth, Cantera, Ver. 2.3.0, 2017. http://www.cantera.org

[8] D. Burgess, J. Racek, XSenkplot: An Interactive, Graphics Postprocesser for Numerical Simulations of Chemical Kinetics, National Institute of Standards and Technology, Gaithersburg, MD, 1996.

[9] G. P. Smith, D. M. Golden, M. Frenklach, N. W. Moriarty, B. Eiteneer, M. Goldenberg, C. T. Bowman, R. K. Hanson, S. Song, W. C. Gardiner, Jr., V. V. Lissianski, Z. Qin, GRI-Mech 3.0, 1999.

[10] A. Persky, H. Kornweitz, The kinetics of the reaction $F+H_2 \rightarrow HF+H$, Inter. J. Chem. Kinet. 29 (1997) 67-71.

[11] PS Stevens, WH Brune, JG Anderson, Kinetic and mechanistic investigations of F atom + water and F atom + hydrogen over the temperature range 240-373 K, J. Phys. Chem. 93 (1989) 4068.

[12] C. D. Walther, H. G. Wagner, Uber die Reaktionen von F-atomen mit H_2O, H_2O_2 und NH_3, Ber. Bunsenges. Phys. Chem. 87 (1983) 403.

[13] Schug, K. P.; Wagner, H. Gg., Zum thermischen zerfall von CH_3F, Z. Phys. Chem. 86 (1973) 59.

[14] S. F. Politanski, U. V. Shevchuk, Thermal conversions of fluoromethanes. II. Pyrolysis of difluoromethane and trifluoromethane, Kinet. Catal. 9 (1968) 411.

[15] G. O. Pritchard, M. J. Perona, Some hydrogen atom abstraction reactions of CF_2H and CFH_2 radicals, and the C-H bond dissociation energy in CF_2H_2, Inter. J. Chem. Kinet. 1 (1969) 509-525.

Babushok, Valeri; Burgess Jr., Donald; Burrell, Robert; Hegetschweiler, Michael; Linteris, Gregory; Manion, Jeffrey.
"Development and Validation of a Mechanism for Flame Propagation in R-32/Air Mixtures."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

[16] A. A. Westenberg, N. DeHaas, Rates of H + CH_3X reactions, J. Chem. Phys. 62 (1975) 3321.

[17] A. Miyoshi, K. Ohmori, K. Tsuchiya, H. Matsui, Reaction rates of atomic oxygen with straight chain alkanes and fluoromethanes at high temperatures, Chem. Phys. Lett. 204 (1993) 241-247.

[18] A. Matsugi, H. Shiina, Kinetics of hydrogen abstraction reactions from fluoromethanes and fluoroethanes, Bull. Chem. Soc. Jpn. 87 (2014) 890-901.

[19] A. Persky, The temperature dependence of the rate constants for the reactions F+CH_3F and F+CH_2F_2, Chem. Phys. Lett. 376 (2003) 181-187.

[20] L. Wang, J. Y. Liu, Z. S. Li, Ab initio direct dynamics studies on the hydrogen abstraction reactions of CF_2H_2 and CF_3H with F atom, Chem. Phys. 351 (2008) 154-158.

[21] J. B. Burkholder, J. Sander, J. R. Abbat, R. E. Huie, C. E. Kolb, M. J. Kurylo, V. L. Orkin, D. M. Wilmouth, P. H. Wine, Chemical Kinetics and Photochemical Data for Use in Atmospheric Studies, JPL Publication 15-10, Jet Propulsion Laboratory, Pasadena, 2015. http://jpldataeval.jpl.nasa.gov

[22] N. K. Srinivasan, M. C. Su, J. W. Sutherland, J. V. Michael, Reflected shock tube studies of high-temperature rate constants for CH_3+O_2, H_2CO+O_2, OH+O_2. J. Phys. Chem. A 109 (2005) 7902-7914.

[23] EL Keating, RA Matula, The high temperature oxidation of tetrafluoroethylene, J. Chem. Phys. 66 (1977) 1237.

[24] D. Burgess Jr., M.R. Zachariah, W. Tsang and P.R. Westmoreland, Thermochemical and Chemical Kinetic Data for Fluorinated Hydrocarbons, Prog. Energ. Comb. Sci., 21 (1996) 453-529.

[25] M. R. Zachariah, P. R. Westmoreland, D. Burgess, Jr., Theoretical Prediction of Thermochemistry and Kinetics of Halocarbons, in ACS Volume 611 on Halon Replacements, 1995, pp. 358-373.

Babushok, Valeri; Burgess Jr., Donald; Burrell, Robert; Hegetschweiler, Michael; Linteris, Gregory; Manion, Jeffrey.
"Development and Validation of a Mechanism for Flame Propagation in R-32/Air Mixtures."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

Chemical Compound Classification by Elemental Signatures in Castle Dust Using SEM Automated X-ray Particle Analysis

Diana L. Ortiz-Montalvo[1], Edward P. Vicenzi[1,2], Nicholas W. Ritchie[1], Carol A. Grissom[2], Richard A. Livingston[3], Zoe Weldon-Yochim[4], Joseph M. Conny[1] and Scott A. Wight[1].

[1] National Institute of Standards and Technology, Gaithersburg, MD, USA.
[2] Museum Conservation Institute, Smithsonian Institution, Suitland, MD, USA.
[3] Department of Materials Science & Engineering, University of Maryland, College Park, MD, USA.
[4] Department of Art History, University of Delaware, Newark, DE, USA.

Discoloration on the Smithsonian Institution Building (1847-1855) and Enid A. Haupt Garden gateposts (1987) was recently revealed to be related to a Mn enriched rock varnish [1]. Mn does not appear to be derived locally from the building stone; therefore, its source is likely related to atmospheric dust transport. Minor oxygen isotopic ratios ($\delta^{17}O$) of sulfate in desert varnish demonstrate that atmospheric deposition of dust is an important component of the varnish formation process [2]. A 2017 study of architectural rock varnish determined that vehicle emissions are a likely source of Mn [3]. In this study, we evaluate airborne dust as a potential Mn source at a location where rock varnish is actively forming.

Urban dust samples were collected on polycarbonate filters using a portable sampler (Hi-Q Environmental Products, Model PSU-2-GN) [4] with a size selective inlet (10 µm particle diameter size cut-off, URG). The collection times were 24 hours, and the volumetric flow rate was 30 L/min. Samples were collected near the Haupt Garden gateposts and heavily trafficked Independence Ave. (Fig. 1). A TESCAN MIRA3 and a 20 keV/1 nA electron beam and 4 × 30 mm^2 PulseTor silicon drift detectors (SDD) were used to analyze particles. Automated analysis was performed using the SEMantics extension to NIST DTSA-II [5]. The sum of the 4 SDD spectra (dwell time of 400 ms per spectrum) were used for quantification using NIST Graf [6]. A novel algorithm was then used to cluster the data obtained for 39,491 particles [7]. Data were then reprocessed using a manually developed rule set.

Table 1 lists the major particle classes representing ≈ 92% of the particle population. Figure 2 shows representative elemental signatures of the four major particle classes: silicate, Fe oxide, vehicle-related and CaMg carbonate. Only 52 out of ≈ 40K particles had elevated levels of Mn (≥ 10 wt %). Overall, our results show low, but detectable levels of Mn in the atmosphere in the Castle area. Efforts are underway to estimate the mass of Mn transported by atmospheric dust deposition for particles under 10 µm in size. To our knowledge, this is the first study to examine the linkage between individual particle analysis of dust and active rock varnish formation.

References:
[1] Vicenzi, E.P. *et al*, Heritage Science **4** (2016), 26.
[2] Bao, H., Michalski, G.M. and Thiemens, M.H, Geochimica et Cosmochimica Acta **65** (2000), 2029.
[3] Macholdt, D.S. *et al*, Atmospheric Environment **171** (2017) 205.
[4] Any mention of commercial products is for information only; it does not imply recommendation or endorsement by NIST.
[5] Ritchie, N.W.M. and Filip, V., Microscopy and Microanalysis Proceedings **17**(S2) (2011), 896.
[6] Lindstrom, A.P. and Ritchie, N.W., Microscopy and Microanalysis Proceedings **20**(S3) (2014), 748.

Conny, Joseph; Grissom, Carol; Livingston, Richard; Ortiz-Montalvo, Diana; Ritchie, Nicholas; Vicenzi, Edward; Weldon-Yochim, Zoe; Wight, Scott.
"Chemical Compound Classification by Elemental Signatures in Castle Dust Using SEM Automated X-ray Particle Analysis."
Paper presented at Microscopy & Microanalysis 2018, Baltimore, MD, United States. August 5, 2018 - August 9, 2018.

[7] Ritchie, N.W., Microscopy and Microanalysis **21**(5) (2015), 1173.

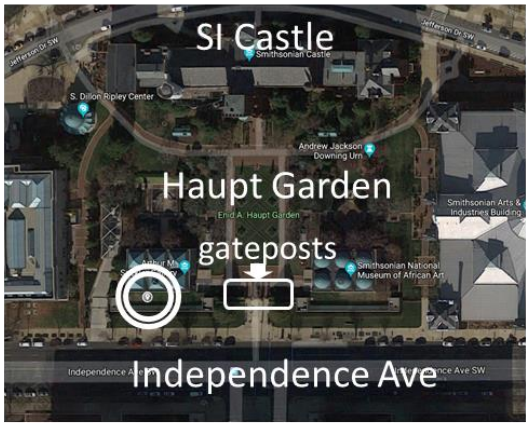


Figure 1. Plan view of the Smithsonian Institution Building (SI Castle), Enid A. Haupt Garden gateposts (white rectangle), Independence SW Ave., and dust collection site (nested white circles) in Washington, DC, USA.

Table 1. Major particle classes ranked by count fraction of 39,491 particles. (≈ 92% of particle population)

Rank	Class name	Count	Fraction (%)
1	silicate	16,868	40.7
2	Fe oxide	10,141	25.7
3	vehicle-related	2,455	6.2
4	CaMg carbonate	1,912	4.8
5	Fe(Mn) oxide	1,341	3.4
6	Ca carbonate	1,287	3.3
7	Fe oxide + silicate	1,192	3.0
8	Ca sulfate	970	2.5
9	Fe sulfide	760	1.9

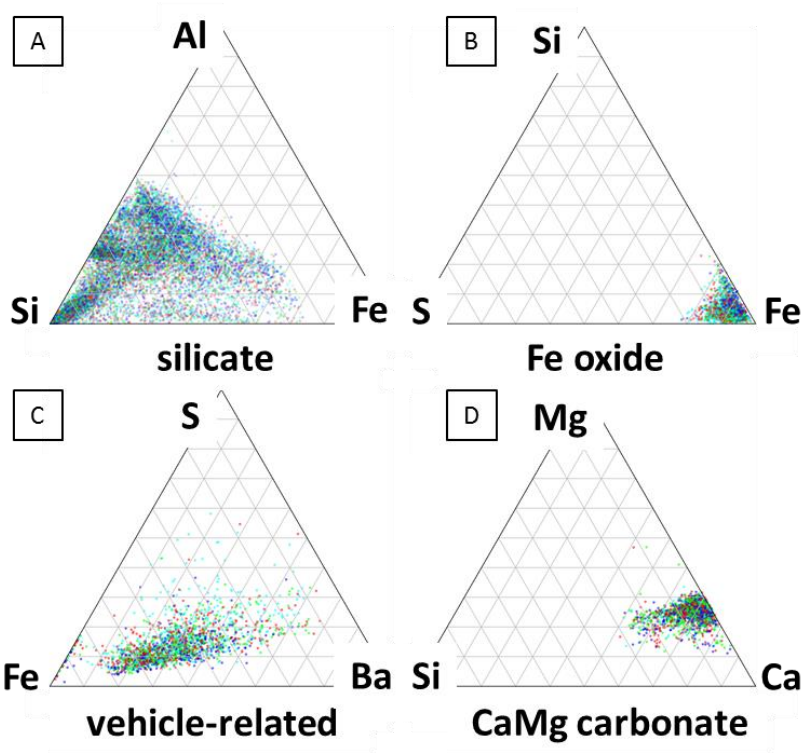


Figure 2. Elemental signatures for major particle classes by weight fraction of ternary end-members. A) silicate. B) Fe oxide. C) vehicle-related (Fe-Ba-S-Cu-Si-Sb-Zr). and D) CaMg carbonate.

Conny, Joseph; Grissom, Carol; Livingston, Richard; Ortiz-Montalvo, Diana; Ritchie, Nicholas; Vicenzi, Edward; Weldon-Yochim, Zoe; Wight, Scott.
"Chemical Compound Classification by Elemental Signatures in Castle Dust Using SEM Automated X-ray Particle Analysis."
Paper presented at Microscopy & Microanalysis 2018, Baltimore, MD, United States. August 5, 2018 - August 9, 2018.

Combinatorial Security Testing Course

Dimitris E. Simos
SBA Research
Vienna, Austria
dsimos@sba-research.org

Rick Kuhn
NIST
Gaithersburg, MD, USA
d.kuhn@nist.gov

Yu Lei
University of Texas at Arlington
Texas, USA
ylei@cse.uta.edu

Raghu Kacker
NIST
Gaithersburg, MD, USA
raghu.kacker@nist.gov

ABSTRACT

Combinatorial methods have attracted attention as a means of providing strong assurance at reduced cost, but when are these methods practical and cost-effective? This tutorial comprises two parts. The first introductory part will briefly explain the background, process, and tools available for combinatorial testing, including illustrations based on industry's experience with the method.

The main part, explains combinatorial testing-based techniques for effective security testing of software components and large-scale software systems. It describes quality assurance and effective re-verification for security testing of web applications and security testing of operating systems. It will further address how combinatorial testing can be applied to ensure proper error-handling of network security protocols and provide the theoretical guarantees for expelling Trojans injected in cryptographic hardware. Procedures and techniques, as well as workarounds will be presented and captured as guidelines for a broader audience. The tutorial is concluded with our vision for combinatorial security testing together with some current open research problems.

The tutorial is designed for participants with a solid IT security background but will not assume any prior knowledge on combinatorial security testing. Thus, we will quickly advance our discussion into core aspects of this field. This tutorial is a modified version of the tutorial held at HVC2017 [19] and QRS2016 [23]. It incorporates feedback and customized content.

KEYWORDS

combinatorial testing, security testing, software quality assurance, security vulnerabilities

ACM Reference Format:
Dimitris E. Simos, Rick Kuhn, Yu Lei, and Raghu Kacker. 2018. Combinatorial Security Testing Course: Tutorial Proposal. In *Proceedings of Hot Topics in the Science of Security (HOTSOS) conference (HOTSOS'18).* ACM, New York, NY, USA, Article 4, 3 pages. https://doi.org/10.475/123_4

1 INTRODUCTION

Identifying vulnerabilities and ensuring security functionality by security testing is a widely applied measure to evaluate and improve the security of software, which is also an inevitable part of quality assurance. Many software security exploitations result from ordinary coding flaws, rather than design or configuration errors. One study found that 64 percent of vulnerabilities are the result of such common bugs as missing or incorrect parameter checking, which leaves applications open to common vulnerabilities including buffer overflows or SQL injection [9]. Although this statistic might be discouraging, it also means that better **functionality testing** can also significantly improve security.

In the last 50 years, combinatorial methods have had profound applications in coding theory, cryptology, networking and computer science with software testing being one of the most recent ones [4]. Covering arrays (CAs) [3] are discrete mathematical structures which, with the aid of proper software engineering techniques, have been utilized in very effective test sets in order to provide strong assurance. Yet, the application of combinatorial methods to applied computer science continues to arise and it comes as no surprise that the field of software security, in particular, provides a rich source of problems that seek solutions from mathematical methods. There has been ample evidence over the last few years to support this observation. List below are several reasons that serve as the **motivation to apply combinatorial methods in order to ensure the quality of secure software**:

- The exemplary case of the Heartbleed bug[1], which allowed anyone on the Internet to read the memory of systems protected by the OpenSSL software (e.g. banking applications), highlighted even more the great need to ensure an attack-free environment of implementations of software systems [5].
- Due to the still increasing interconnectedness of such complex software systems, it is very important to strengthen activities towards assuring their security requirements by performing security testing [27].
- The latter task is not be considered an easy process, bearing in mind that software testing may consume up to half of the overall software development cost [26].

[1] heartbleed.com

Kacker, Raghu; Kuhn, David; Lei, Yu; Simos, Dimitris.
"Combinatorial Security Testing Course."
Paper presented at Hot Topics in the Science of Security, Raleigh, NC, United States. April 10, 2018 - April 11, 2018.

- Combinatorial explosion is a frequently occurring problem in testing [16], [1] where a test object is described by a number of parameters, each with many possible values. The effect of the combinatorial explosion is that it is infeasible to test every combination of parameter values.
- There exists an added level of complexity for security testing where the modelling of vulnerabilities is specific to the application domain and the identification of factors triggering such exploits is not easily done [18], [17].
- Finally, there are relatively few good methods for evaluating test set quality, after ensuring basic requirements-traceability [14]. Of particular importance is the task to develop methods that help estimate the residual risk that remains after testing.

In [24] the authors developed an ambitious research program aimed at bridging the gap between combinatorial testing (CT) and security testing and, in the process, established a new research field: *combinatorial security testing*. Several methods and case stud- ies presented in this tutorial, illustrate our experiences thus far and the success of the previously mentioned research program, which came as a result of the application of our combinatorial techniques in security testing. In summary, we showed in [24] that the devel- oped concept of CT applicable to security testing supersedes other testing approaches due to its advantages of generating minimal size test sets and revealing hard-to-spot errors in software systems in an automated way.

2 OUTLINE OF THE TUTORIAL

In this tutorial we present our work on combinatorial methods for security testing, which guarantees certain aspects of test quality
e.g. test coverage or locating faults. In particular, we formulate problems of software security testing as combinatorial problems and then use efficient algorithmic or theoretical methods to tackle them. *The central thesis of this tutorial is that combinatorial methods can make software security testing much more efficient and effective than conventional approaches - in specific application domains.*

Brief Introduction of Combinatorial Testing: This provides a quick overview of the history of combinatorial testing re- search, and their roots based on key publications in the field [15],
[12] and [13].

Web Security Interaction Testing: Here the concern is with the problem of security vulnerability detection and with the inherent, but also equally important, problem of retrieving the root cause of security vulnerabilities. We will demonstrate the process of creating attack models used for exploiting web security vulnerabilities using combinatorial methods [6], [2] and indicating methods for analyzing them [21]. The main goal of this part is to make everyone familiar with advanced combinatorial techniques for web security testing.

Security Protocol Interaction Testing: In this part of the tutorial we deal with the problem of certificate testing, which plays a central role in network security. We will present complex combinatorial models for creating test certificates to check for faults in the validation logic, which can result in impersonation attacks [11]. In addition, we will present recent efforts on the modelling of the TLS Handshake protocol using CT [20]. If time permits, the authors plan to analyze the TLS cipher suites of the aforementioned protocol using combinatorial coverage measurement techniques [22] and detail the implications of the findings for software security testing.

Combinatorial Methods for Kernel Software: The kernel of an operating system is the central authority to enforce security. The goal in this part of the tutorial is to ensure the reliability and quality assurance of kernel software. We will present two testing frameworks, ERIS [7], a combinatorial kernel testing tool, and its recent enhancement, called KERIS [8], with dynamic memory error detectors for the Linux Kernel aimed at exploiting security vulnerabili- ties. We will reproduce for the participants a security vulnerability in the Linux networking stack first discovered by Google's Project Zero team.

Detecting Hardware Trojan Horses: This part outlines the problem of malicious hardware logic detection. In particular, the concern is with cryptographic Trojans appearing as instances of malicious hardware. The exemplary scenario for this tutorial evolves around Trojans residing inside cryptographic circuits that perform encryption and decryption in FPGA technologies using the AES cryptographic algorithm. We will demonstrate that combinatorial testing constructs are capable of reducing the number of test cases needed for the Trojan excitation by several orders of magnitude, while at the same time activate the Trojan hundreds of times [10]. The authors will also briefly present similar patterns for AES software implementations based on a recent combinatorial analy- sis performed on AES validation tests [25]. Whether these latter patterns are also malicious is currently an open question.

Open challenges and outlook: The authors present and quickly discuss currently unsolved challenges.

3 INTENDED AUDIENCE

This **75 minutes** tutorial does not assume any prior knowledge of combinatorial testing methods for information security. We assume good general knowledge of information security and software engineering on a graduate CS student level with a focus on security. The goal of this tutorial is to present the knowledge from various sources in a structured way and provide researchers with the **practical fundamentals** of combinatorial methods for security testing and practitioners with the **scientific background**.

The **key takeaways** are: (I) the practical fundamentals of combinatorial methods for security testing, (II) a good understanding of the underlying mathematical, software engineering and security mechanics and, (III) an overview of the related literature and open problems in this field.

4 SHORT BIOS

Biography of Dimitris Simos. Dimitris E. Simos is a Key Researcher with SBA Research, Austria, for the "applied discrete mathematics for information security" research area where he is leading the combinatorial security testing team. He is also an Adjunct Lecturer with Vienna University of Technology and a Distinguished Guest Lecturer with Graz University of Technology. His research interests include combinatorial designs and their applications to software testing, combinatorial testing in particular, applied cryptography and optimization algorithms, and information security. He holds a Ph.D. in Discrete Mathematics and Combinatorics (2011) from

Kacker, Raghu; Kuhn, David; Lei, Yu; Simos, Dimitris.
"Combinatorial Security Testing Course."
Paper presented at Hot Topics in the Science of Security, Raleigh, NC, United States. April 10, 2018 - April 11, 2018.

the National Technical University of Athens. Prior to joining SBA Research, he was within the Project Team SECRET of INRIA Paris-Rocquencourt Research Center working on the design and analysis of cryptographic algorithms. His research was supported by a 3-year Marie Curie Fellow grant (2012-2015) awarded by the ERCIM through the EU-funded "Alain Bensoussan" Fellowship Programme. He is the author of over 70 papers in discrete mathematics and their applications to computer science and a Fellow of the Institute of Combinatorics and its Applications (FTICA). He was the general chair of MACIS 2017 and also PC chair for IWCT (2017 and 2018).

Biography of Richard Kuhn. Rick Kuhn is a computer scientist in the Computer Security Division of the National Institute of Standards and Technology. He has authored two books and more than 150 conference or journal publications on information security, empirical studies of software failure, and software assurance, and is a Fellow of the Institute of Electrical and Electronics Engineers (IEEE). His research interests are in combinatorial methods for software testing, security and access control, and empirical studies of software failure.

Biography of Yu Lei. Yu Lei is a Professor of Computer Science at the University of Texas at Arlington. He was a Member of Technical Staff in Fujitsu Network Communications, Inc. from 1998 to 2001. His current interests include combinatorial testing, concurrency testing, and security testing. He served on the Program Committee of ICST 2008 – 2009 and 2012 –2015.

Biography of Raghu Kacker. Raghu Kacker is a mathematical statistician in the Applied and Computational Mathematics Division (ACMD) of the Information Technology Laboratory (ITL) of the US National Institute of Standards and Technology (NIST). His current interests include combinatorial testing of software and system s and evaluation of uncertainty in outputs of computational models and physical measurements. Advancing the methods and tools for combinational testing is a mission of his. He has authored or co-authored over 100 papers. He has a Ph.D. in statistics. He is a Fellow of the American Statistical Association and a Fellow of the American Society for Quality.

5 ACKNOWLEDGEMENTS

This research was funded by COMET K1, FFG - Austrian Research Promotion Agency, FFG BRIDGE Early Stage SPLIT and FFG BRIDGE SecWIT projects.

Disclaimer: *Products may be identified in this document, but identification does not imply recommendation or endorsement by NIST, nor that the products identified are necessarily the best available for the purpose.*

REFERENCES

[1] Paul Ammann and Jeff Offutt. 2008. *Introduction to Software Testing* (1 ed.). Cambridge University Press, New York, NY, USA.

[2] J. Bozic, B. Garn, I. Kapsalis, D. E. Simos, S. Winkler, and F. Wotawa. 2015. Attack Pattern-Based Combinatorial Testing with Constraints for Web Security Testing. In *QRS '15: Proceedings of the 2015 IEEE International Conference on Software Quality, Reliability and Security*. 207–212.

[19] D. E. Simos. 2017. Combinatorial Security Testing: Quo Vandis?. https://www.research.ibm.com/haifa/conferences/hvc2017/tutorials.shtml.

[20] D. E. Simos, J. Bozic, F. Duan, B. Garn, K. Kleine, Y. Lei, and F. Wotawa. 2017. Testing TLS Using Combinatorial Methods and Execution Framework. In *ICTSS '17: Proceedings of the 29th International Conference on Testing Software and Systems, Lecture Notes in Computer Science*, Vol. 10533. 162–177.

[21] D. E. Simos, K. Kleine, L. Ghandehari, B. Garn, and Y. Lei. 2016. A combinatorial approach to analyzing cross-site scripting (XSS) vulnerabilities in web application security testing. In *ICTSS '16: Proceedings of the 28th International Conference on Testing Software and Systems, Lecture Notes in Computer Science*, Vol. 9976. 70–85.

[22] D. E. Simos, K. Kleine, A. G. Voyiatzis, R. Kuhn, and R. Kacker. 2016. TLS Cipher Suites Recommendations: A Combinatorial Coverage Measurement. In *QRS '16: Proceedings of the 2016 IEEE International Conference on Software Quality,*

[3] Charles J. Colbourn. 2006. Covering Arrays. In *Handbook of Combinatorial Designs* (2nd ed.), Charles J. Colbourn and Jeffrey H. Dinitz (Eds.). CRC Press, Boca Raton, Fla., 361–365.

[4] Charles J. Colbourn and Paul C. van Oorschot. 1989. Applications of Combinatorial Designs in Computer Science. *ACM Comput. Surv.* 21, 2 (1989).

[5] Zakir Durumeric, James Kasten, David Adrian, J. Alex Halderman, Michael Bailey, Frank Li, Nicolas Weaver, Johanna Amann, Jethro Beekman, Mathias Payer, and Vern Paxson. 2014. The Matter of Heartbleed. In *Proceedings of the 2014 Conference on Internet Measurement Conference (IMC '14)*. ACM, New York, NY, USA, 475–488.

[6] B. Garn, I. Kapsalis, D. E. Simos, and S. Winkler. 2014. On the Applicability of Combinatorial Testing to Web Application Security Testing: A Case Study. In *JAMAICA 14: Proceedings of the 2nd International Workshop on Joining AcadeMiA and Industry Contributions to Test Automation and Model-based Testing, collocated with ISSTA '14: International Symposium on Software Testing and Analysis, ACM*. 16–21.

[7] B. Garn and D. E. Simos. 2014. Eris: A Tool for Combinatorial Testing of the Linux System Call Interface. In *IWCT '14: Proceedings of the 3rd International Workshop on Combinatorial Testing, collocated with ICST '14: 7th IEEE International Conference on Software Testing, Verification and Validation*. 58–67.

[8] B. Garn, F. Würfl, and D. E. Simos. 2017. KERIS: A CT Tool of the Linux Kernel with Dynamic Memory Analysis Capabilities. In *HVC '17: Proceedings of the 13th Haifa Verification Conference, Lecture Notes in Computer Science*, Vol. 10629. 225–228.

[9] Jon Heffley and Pascal Meunier. 2004. Can Source Code Auditing Software Identify Common Vulnerabilities and Be Used to Evaluate Software Security?. In *HICSS*.

[10] P. Kitsos, D. E. Simos, Jose Torres-Jimenez, and A. G. Voyiatzis. 2015. Exciting FPGA Cryptographic Trojans using Combinatorial Testing. In *ISSRE '15: Proceedings of the 26th IEEE International Symposium on Software Reliability Engineering*. 69–76.

[11] K. Kleine and D. E. Simos. 2017. Coveringcerts: Combinatorial Methods for X.509 Certificate Testing. In *ICST '17: Proceedings of the 10th IEEE International Conference on Software Testing, Verification and Validation*. 69–79.

[12] D.R. Kuhn, R.N. Kacker, and Y. Lei. 2013. *Introduction to Combinatorial Testing*. Taylor & Francis.

[13] D Richard Kuhn, Renee Bryce, Feng Duan, Laleh Sh Ghandehari, Yu Lei, and Raghu N Kacker. 2015. Chapter one-combinatorial testing: Theory and practice. *Advances in Computers* 99 (2015), 1–66.

[14] D Richard Kuhn, Raghu N. Kacker, and Yu Lei. 2015. Combinatorial Coverage as an Aspect of Test Quality. *CrossTalk* 28, 2 (2015), 19–23.

[15] Rick Kuhn, Yu Lei, and Raghu Kacker. 2008. Practical combinatorial testing: Beyond pairwise. *It Professional* 10, 3 (2008).

[16] Aditya P. Mathur. 2008. *Foundations of Software Testing* (1st ed.). Addison-Wesley Professional.

[17] Gary McGraw. 2006. *Software Security: Building Security In*. Addison-Wesley Professional.

[18] B. Potter and G. McGraw. 2004. Software security testing. 2, 5 (2004), 81–85. *IEEE Security Privacy Reliability and Security*. 69–73.

[23] D. E. Simos, Rick Kuhn, Yu Lei, and Raghu Kacker. 2016. Combinatorial Security Testing. http://paris.utdallas.edu/qrs16/program/QRS-2016-Program.pdf.

[24] D. E. Simos, R. Kuhn, A. G. Voyiatzis, and R. Kacker. 2016. Combinatorial methods in security testing. *IEEE Computer* 49 (2016), 40–43.

[25] D. E. Simos, S. Mekesis, D. R. Kuhn, and R. N. Kacker. [n. d.]. Combinatorial Coverage Measurement of Test Vectors used in Cryptographic Algorithm Validation. In *STC '17: Proceedings of the 2017 IEEE Software Technology Conference*.

[26] G. Tassey. 2002. *The economic impacts of inadequate infrastructure for software testing*. National Institute of Standards and Technology.

[27] F. Wotawa. 2016. On the Automation of Security Testing. In *2016 International Conference on Software Security and Assurance (ICSSA)*. 11–16.

Kacker, Raghu; Kuhn, David; Lei, Yu; Simos, Dimitris.
"Combinatorial Security Testing Course."
Paper presented at Hot Topics in the Science of Security, Raleigh, NC, United States. April 10, 2018 - April 11, 2018.

Poster: What Proportion of Vulnerabilities can be Attributed to Ordinary Coding Errors?

Rick Kuhn[1], Mohammad Raunak[2], Raghu Kacker[1]
kuhn@nist.gov, raghu.kacker@nist.gov raunak@loyola.edu
[1]National Institute of Standards and Technology [2]Loyola University Maryland

I. Introduction

The analysis reported in this poster developed from questions that arose in discussions of the Reducing Software Vulnerabilities working group, sponsored by the White House Office of Science and Technology Policy in 2016 [1]. The key question we sought to address is the degree to which vulnerabilities arise from ordinary program errors, which may be detected in code reviews and functional testing, rather than post-release.

The analysis used 2008 - 2016 data from the US National Vulnerability Database (NVD) [2]. NVD is the US government's repository of information system security vulnerabilities, which compiles nearly all publicly reported vulnerabilities using the Common Vulnerabilities and Exposures (CVE) dictionary [3]. Each reported CVE is assigned to one or more categories called the Common Weakness Enumeration (CWE) [4], which specifies categories that may include a number of subsidiary weaknesses. For example, CWE-119, Buffer errors, includes 14 subsidiary CWEs, such as out of bounds read (CWE-125), and untrusted pointer dereference (CWE-822).

We further grouped the NVD CWE categories into primary classes of Configuration, Design, and Implementation errors. In determining the class of each CWE category, we considered the common errors in each type. *Configuration* vulnerabilities result when a system is not set up correctly with respect to security goals. A simple example would be failure to enable password checking. *Design* related vulnerabilities are those that originate in the planning and design of the system, such as selecting an outdated or weak cryptographic algorithm. *Implementation* errors occur in program construction. One of the most common implementation vulnerabilities is simple buffer overflow. Failure to check that input size is within maximum buffer size is a simple error that should almost never occur, but continues to be a widespread problem. A wide variety of implementation related vulnerabilities also result from failure to properly validate input.

II. Analysis and Results

The poster includes analysis of the following data [5]:

- Severity trends - proportion of vulnerabilities designated *low*, *medium*, and *high* by year.
- Primary CWE type trends - direction of trend for 19 primary CWE types, further classed as Configuration, Design, or Implementation vulnerabilities.

Significant findings include:

- The proportion of *high* severity vulnerabilities trends downward, declining about 15 percentage points since 2008. About two-thirds of this fraction has shifted to *medium* severity vulnerabilities.
- Implementation or coding errors account for roughly two thirds of the total. We consider the proportion of implementation vulnerabilities, rather than absolute numbers, because the number of vulnerabilities is partially a function of the number of applications released, which has increased over time. The proportion of implementation vulnerabilities for 2008-2016 is close to the 64% reported for 1998 - 2003 in an analysis of an early version of NVD [6].

The high proportion of implementation errors suggests that little progress has been made in reducing these vulnerabilities that result from simple mistakes, but also that more extensive use of static analysis tools, code reviews, and testing could lead to signficant improvement. The poster also briefly summarizes data on effectiveness of approaches to preventing and detecting errors before release.

Products may be identified in this document, but such identification does not imply recommendation by the US National Institute of Standards and Technology or the US Government, nor that the products identified are necessarily the best available for the purpose.

[1] Black, P. E., Badger, M. L., Guttman, B., & Fong, E. N. (2016). *Dramatically Reducing Software Vulnerabilities: Report to the White House Office of Science and Technology Policy.* NIST Interagency Report, NISTIR-8151.
[2] National Vulnerability Database, http://nvd.nist.gov 2017
[3] Common Vulnerabilities and Exposures, https://cve.mitre.org.
[4] Common Weakness Enumeration, https://cwe.mitre.org.
[5] Kuhn, D. R., Raunak, M. S., & Kacker, R. (2017, July). An Analysis of Vulnerability Trends, 2008-2016. *Software Quality, Reliability and Security* (QRS-C), 2017 IEEE International Conference on (pp. 587-588).
[6] Heffley, Jon, and Pascal Meunier. "Can source code auditing software identify common vulnerabilities and be used to evaluate software security?" *System Sciences, 37th Annual Hawaii Intl Conf*, IEEE, 2004.

Kacker, Raghu; Kuhn, David; Raunak, Mohammad.
"Poster: What Proportion of Vulnerabilities can be Attributed to Ordinary Coding Errors?."
Paper presented at Hot Topics in the Science of Security, Raleigh, NC, United States. April 10, 2018 - April 11, 2018.

Physical and Chemical Transformations of Silver Nanomaterial-containing Textiles After Use

D.E. Gorka*, J.M. Gorham*

*National Institute of Standards andTechnology,
Mail Stop 8520, 100 Bureau Drive, Gaithersburg, MD, USA
danielle.gorka@nist.gov, justin.gorham@nist.gov

ABSTRACT

Nanomaterials have been increasingly used in consumer products and silver nanomaterials (AgNMs) especially have been used for their antimicrobial properties. As use of AgNMs in consumer products continues to increase, a corresponding increase in silver's presence in the environment will be observed due to disposal. To better understand what materials are entering the environment, work needs to be performed to determine the chemical and physical properties of AgNM-containing consumer textiles throughout their lifecycle prior to introduction into the environment through disposal. Therefore, the aim of this work is to evaluate chemical and physical transformations that AgNM-containing textiles undergo during modeled human exposure. A commercially available AgNM-containing wound dressing was studied as our model system. To model this textile during use, the material was exposed to synthetic sweat or simulated wound fluid for varied durations up to 7 days. The textile was extracted and stored under vacuum to minimize extraneous transformations after removal from test media. Both pristine and exposed wound dressings were characterized using a variety of analytical techniques including scanning electron microscopy (SEM) with energy dispersive X-ray spectroscopy (EDS), X-ray diffraction (XRD), X-ray photoelectron spectroscopy (XPS), inductively coupled plasma mass spectrometry (ICP-MS), UV-Visible spectroscopy (UV-Vis), and dynamic light scattering (DLS). Electron microscopy revealed the formation of micron-sized structures on the surface of the commercial wound dressing after synthetic sweat exposure which spectroscopic and diffraction based techniques suggested were consistent with silver chloride. In contrast, wound dressing exposed to simulated wound fluid did not show any large structures on the surface of the material. In fact, the surface was similar, though less defined, than the pristine wound dressing. The release of silver from the wound dressing into the exposure media was also examined. Though ICP-MS found silver release, our DLS and UV-vis based results suggest that released silver was not detected in metallic form, aggregate, or on the nanoscale in the exposure media. A better understanding of the chemical and physical transformations of AgNMs in consumer products is necessary for manufacturers and regulators to make more informed decisions on product design and use.

Keywords: silver, nanomaterials, characterization, textile, acticoat

1 INTRODUCTION

Due to their unique properties, nanomaterials have undergone increasing use in consumer products. For that reason, silver nanomaterials can be found in consumer goods including clothing, personal care items, and food storage products where they act as antimicrobial agents.[1] Moreover, their antibacterial property makes AgNMs an attractive additive for biomedical products and devices such as bandages and wound dressing.[2]

During use these AgNMs will undergo physical and chemical changes. These transformations are not well understood, though several studies have examined the release of silver from these textiles. Researchers suggest that soluble silver species will be released after the textile has been exposed to synthetic sweat or skin surface film liquid.[3, 4] Fewer studies have examined physical and chemical transformations that occur on the textile. One study that examined a commerical AgNM product called Acticoat Flex3 found the formation of silver and chloride containing crystals on the fibers after saline exposure.[5] These studies suggest transformations will occur on AgNM-containing textiles that may affect their chemical and physical properties after use. A greater understanding of the physical and chemical transformations that occur during the AgNM-containing textile use phase are necessary to better understand what forms are left on the textile and are entering the environment. This is important because a recent study suggests almost all of the silver present in a commercially available AgNM-containing wound dressing remains in the product upon disposal into the environment.[6]

The focus of this work was to examine the physical and chemical transformations that AgNM-containing textiles undergo during modeled human exposure. To do this, a commerical AgNM-containing wound dressing was examined quantitatively and qualitatively using various orthogonal analytical techniques including SEM-EDS, XRD, XPS, ICP-MS, UV-Vis, and DLS.

TechConnect Briefs 2018, TechConnect.org, ISBN 978-0-9988782-4-9

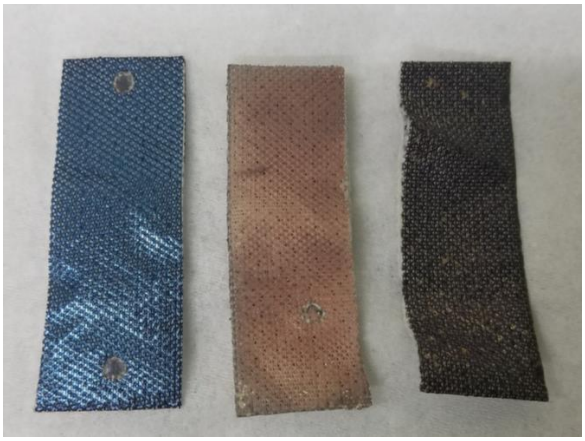


Figure 1: Photographs of Acticoat before (left) and after 168 h expsoure to synthetic sweat (middle) and simulated wound fluid (right). Exposure results in a color change from blue to light brown (synthetic sweat) or dark gray (simulated wound fluid).

2 METHODS[1]

Solutions of synthetic sweat and simulated wound fluid were prepared. Synthetic sweat was prepared following the International Standard Organization (ISO)105-E04-2008E acidic type synthetic sweat method.[7] Briefly, 0.5 g l-histidine monochloride monohydrate (VWR), 5 g sodium chloride (Alfa Aesar, 99%), and 2.2 g sodium dihydrogen orthophosphate dihydrate (Alfa Aesar, 99%) were mixed and diluted to 1 L with MilliQ water. For simulated wound fluid, an isotonic solution with an added 1% protein component was prepared.[8] Briefly, 8.27 g sodium chloride (Alfa Aesar, 99%), 0.37 g calcium chloride dihydrate (Amresco), and 10 g bovine serum albumin (BSA, SeraCare) were mixed and diluted to 1 L with MilliQ water.

A 2/3 inch x 2/3 inch of Acticoat (Acticoat 7, Smith & Nephews) was fully submerged in a 30 mL LDPE plastic bottle containing 10 mL of test solution. The bottle was wrapped in foil to prevent light exposure and rotated on a room temperature incubator at 50 rpm for varying amounts of time. The textile was removed at the following times after addition: 5 s, 1 h, 2 h, 6 h, 24 h, and 168 h. The textile was stored in a vacuum desiccator until analysis. The remaining test expsoure solution was stored in the dark at 4 °C until analysis.

2.2 Characterization

Pristine and exposed samples were analyzed using a FEI Quanta 200 (Hillsboro, OR) environmental scanning electron microscope (SEM) with energy-dispersive X-ray spectroscopy (EDS). Samples were imaged at an operating voltage of 10 kV with a unitless spot size of 3. A Bruker XFlash Detector 5030 (Billerica, MA) was used to collect EDS spectra, with an acquisition time of 300 s. EDS data was analyzed using Bruker Esprit v. 1.9.3. A small piece approximately 2 mm x 2 mm was analyzed by adhering to an Al SEM stub with carbon tape. The middle silver-containing layer was imaged to minimize possible contamination of the outer layers.

Exposure media was analyzed by UV-Visible (UV-Vis) spectrophotometry using a PerkinElmer Lambda 750 spectrophotometer (Waltham, MA) to determine if particles were released from the Acticoat or AgNM textile into the synthetic sweat and simulated wound fluid after exposure. Absorbance data was collected from 200 nm to 800 nm using a plastic microcuvette with a pathlength of 1 cm. Dynamic light scattering (DLS) using a Malvern Zetasizer Nano ZS (Westborough, MA) was performed to determine the sizes of any particles released into the exposure media. Bulk silver optical density and refractive index were used and scattering was measured at 173°. Test exposure media samples were held at 23 °C for 180 s before the first run to equilibrate the temperature and were then held at that temperature during the run. For each sample, 8 measurements were made, each measurement consisted of 11 scans, and each scan was 10 s long.

3 RESULTS AND DISCUSSION

3.1 Physical Transformations

Exposure of Acticoat to either synthetic sweat or simulated wound fluid caused a visible change to the color of the product. As shown in Figure 1, the pristine Acticoat was blue in color, however exposure to either solution resulted in a noticable color change. After exposure to synthetic sweat Acticoat was light brown (Figure 1, middle) and after exposure to simulated wound fluid Acticoat was dark gray in color (Figure 1, left). To better understand the physical transformations that were occuring, SEM images were acquired. Prior to exposure, discrete spherical crystals in the nanoscale range were found on the surface of pristine Acticoat (Figure 2, top) which agreed with previous work on the surface of Acticoat.[9] After 24 h exposure to synthetic sweat, submicron and larger micron-sized nonspherical crystals formed on the surface. This results agreed with Rigo, et al., where silver- and chloride-containing crystals were formed on Acticoat Flex3 after saline exposure.[5] In contrast to synthetic sweat, the surface of the simulated wound fluid exposed Acticoat looked similar to the pristine material. However, the

[1] Certain trade names and company products are mentioned in the text or identified in illustrations in order to adequately specify the experimental procedure and equipment used. In no case does such identification imply recommendation or endorsement by National Institute of Standards and Technology, nor does it imply that the products are necessarily the best available for the purpose.

Gorham, Justin; Gorka, Danielle.
"Physical and Chemical Transformations of Silver Nanomaterial-containing Textiles After Use."
Paper presented at TechConnect World Innovation Conference, Anaheim, CA, United States. May 13, 2018 - May 16, 2018.

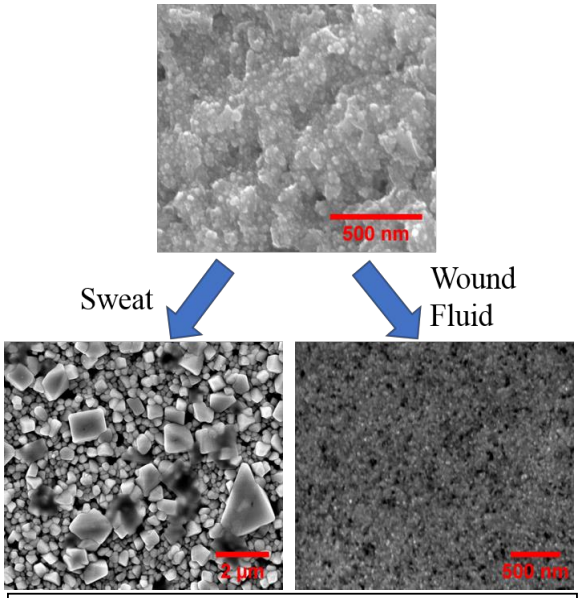


Figure 3: (SEM) images of Acticoat before (top) and after 24 h exposure to synthetic sweat (bottom left) and simulated wound fluid (bottom right).

features were less defined and the surface appeared to have less roughness than the pristine material. Importantly, there was no formation of submicron or micron-sized crystals on the surface of the wound dressing.

3.2 Chemical Transformations

To determine chemical transformations that may have occurred during modeled human exposure, EDS was first used. EDS mapping for pristine Acticoat indicated the presence of silver with peaks for carbon and oxygen present and minimal amounts of chloride. In contrast, the EDS spectra for synthetic sweat exposed Acticoat indicated both silver and chloride being present, as well as sodium (Figure 3A). The EDS map indicated overlap of both silver and chloride on the submicron and micron-sized crystals. Spots in the EDS map that indicated chloride but not silver also showed sodium, suggesting crystals containing sodium and chloride were present on the surface as well. The EDS spectra for simulated wound fluid exposed Acticoat showed the presence of silver, chloride, and sodium (Figure 3B). However, EDS mapping for this material showed almost complete overlap of silver and chloride. Other results by XRD and XPS (data not shown) suggest the presence of silver in metallic and silver chloride form after exposure to either test media suggesting a partial conversion of the

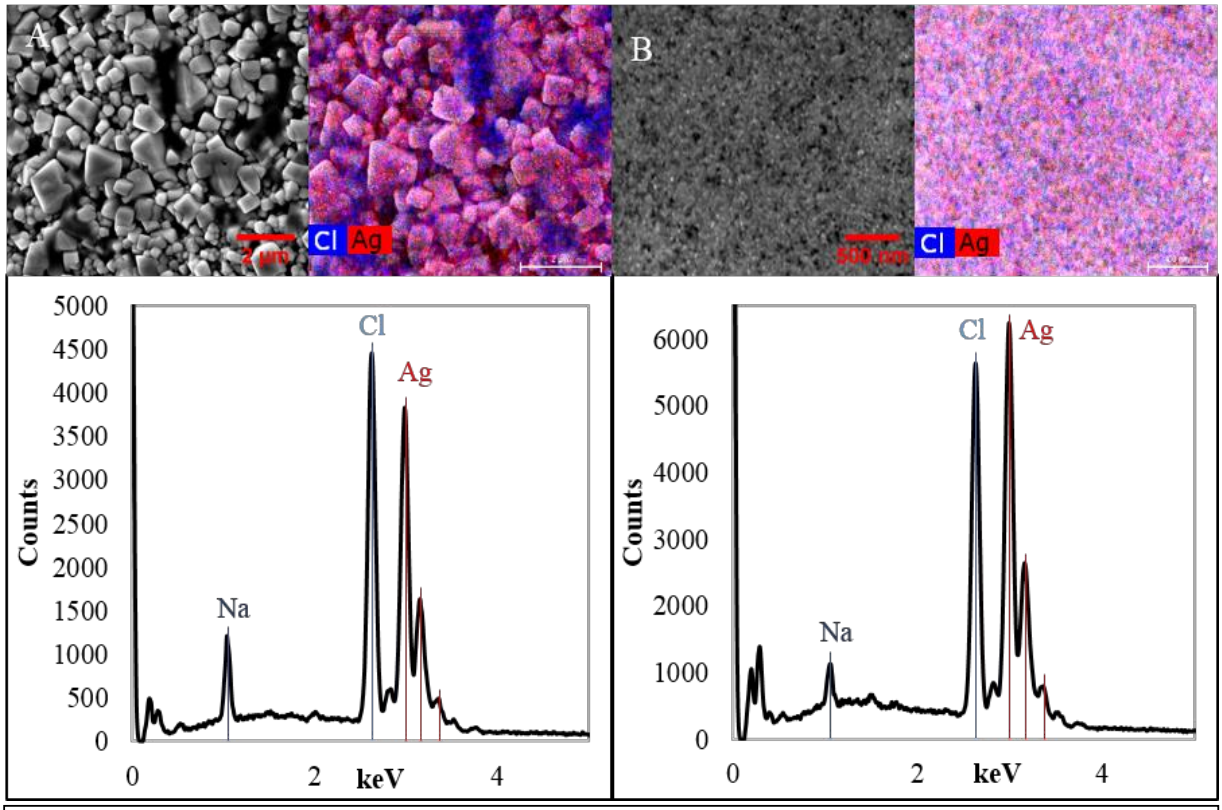


Figure 2: SEM images, energy-dispersive X-ray spectroscopy (EDS) maps, and EDS spectra of Acticoat after 24 h exposure to A) synthetic sweat and B) simulated wound fluid.

TechConnect Briefs 2018, TechConnect.org, ISBN 978-0-9988782-4-9

Gorham, Justin; Gorka, Danielle.
"Physical and Chemical Transformations of Silver Nanomaterial-containing Textiles After Use."
Paper presented at TechConnect World Innovation Conference, Anaheim, CA, United States. May 13, 2018 - May 16, 2018.

surface layer to the +1 oxidation state.

3.3 Evaluation of released silver

After the Acticoat was examined, the test media was studied to determine if any nanoscale particles were released into the fluids during exposure. Data to be published in future studies using ICP-MS revealed the presence of ppm levels of soluble silver in the wound fluid supernatant while the simulated sweat equivalent revealed only sub ppm levels.

DLS was used to determine if any nanoscale particles or aggregates were released into the test expsoure solutions. Interestingly, Acticoat exposed synthetic sweat did not result in the appearance of any peaks in the DLS spectra, suggesting that any silver released was neither in a nanoscale particulate nor an aggregated form. Similarly, Acticoat-exposed simulated wound fluid supernatant did not display any new peaks in the DLS spectra compared to the unexposed simulated wound fluid. This again suggests that any silver released from the Acticoat during exposure was not in a nanoscale particulate form nor in an aggregated form.

UV-Vis was also performed to determine if any metallic silver was released from the textile during modeled human exposure. Consistent with the DLS-based findings, synethic sweat and simulated wound fluid did not show any increases in absorbance or the appearance of any new peaks after interacting with Acticoat. This suggests that the silver released was not metallic or of a size sufficiently large enough to be detected by DLS. Therefore, the most likely form of released silver in the simulated wound fluid is as a complex with the added protein. This would agree with work by Rigo, et al., and Ostermeyer, et al. which found sequestration of silver in protein.[5, 10]

4 CONCLUSIONS

This work shows that modeled human exposure will result in both physical and chemical transformations of AgNMs on consumer textiles such as wound dressings. Exposure to synthetic sweat causes the formation of submicron and micron-sized silver and chloride containing crystals on the surface of the commerical wound dressing. Simulated wound fluid, in contrast, does not result in the formation of micron-sized crystals on the surface of the wound dressing. The wound dressing was found to release silver into the exposure media, with simulated wound fluid showing greater release than synthetic sweat.. However, our results demonstrate that silver is not present in a metallic form, in aggregate, or on the nanoscale in the exposure media. The physical and chemical transformations found in this work necessitate the study of AgNM-containing consumer textiles under more realistic conditions. This will allow for better understanding of the chemical and physical transformations that AgNMs in consumer will undergo during use and allow for manufacturers and regulators to make more informed decisions on product design and use.

5 ACKNOWLEDGEMENTS

DE Gorka acknowledges funding and support from the National Academy of Science – National Research Council Postdoctoral Research Associateship Program.

REFERENCES

[1] T. Benn, B. Cavanagh, K. Hristovski, J.D. Posner, P. Westerhoff, The Release of Nanosilver from Consumer Products Used in the Home, Journal of Environmental Quality 39(6) (2010) 1875-1882.
[2] L. Pourzahedi, M.J. Eckelman, Environmental life cycle assessment of nanosilver-enabled bandages, Environ Sci Technol 49(1) (2015) 361-8.
[3] C. Bianco, S. Kezic, M. Crosera, V. Svetlicic, S. Segota, G. Maina, C. Romano, F. Larese, G. Adami, In vitro percutaneous penetration and characterization of silver from silver-containing textiles, Int J Nanomedicine 10 (2015) 1899-908.
[4] A.B. Stefaniak, M.G. Duling, R.B. Lawrence, T.A. Thomas, R.F. LeBouf, E.E. Wade, M.A. Virji, Dermal exposure potential from textiles that contain silver nanoparticles, Int J Occup Environ Health 20(3) (2014) 220-34.
[5] C. Rigo, M. Roman, I. Munivrana, V. Vindigni, B. Azzena, C. Barbante, W.R. Cairns, Characterization and evaluation of silver release from four different dressings used in burns care, Burns 38(8) (2012) 1131-1142.
[6] R.J. Courtemanche, N.S. Taylor, D.J. Courtemanche, Initiating silver recycling efforts: Quantifying Ag from used burn dressings, Environmental Technology & Innovation 4 (2015) 29-35.
[7] ISO, Textiles — Tests for colour fastness — Part E04: Colour fastness to perspiration, ISO 105-E04:2008, 2008.
[8] T.W. Canada, KM; Cowan, ME; Lindsey, BJ, Challenging silver-influence of extraction medium on the release of silver from commercial silver dressings, 2007.
[9] L.S. Dorobantu, G.G. Goss, R.E. Burrell, Effect of light on physicochemical and biological properties of nanocrystalline silver dressings, RSC Advances 5(19) (2015) 14294-14304.
[10] A.K. Ostermeyer, C. Kostigen Mumuper, L. Semprini, T. Radniecki, Influence of bovine serum albumin and alginate on silver nanoparticle dissolution and toxicity to Nitrosomonas europaea, Environ Sci Technol 47(24) (2013) 14403-10.

Gorham, Justin; Gorka, Danielle.
"Physical and Chemical Transformations of Silver Nanomaterial-containing Textiles After Use."
Paper presented at TechConnect World Innovation Conference, Anaheim, CA, United States. May 13, 2018 - May 16, 2018.

Physical and Chemical Transformations of Silver Nanomaterials in Textiles After Use and Disposal

D.E. Gorka*, J.M. Gorham*

*National Institute of Standards andTechnology,
Mail Stop 8520, 100 Bureau Drive, Gaithersburg, MD, USA
danielle.gorka@nist.gov, justin.gorham@nist.gov

ABSTRACT

Silver nanomaterials (AgNMs) have been increasingly used in consumer products for their antibacterial properties. Textiles, including wound dressings, are just one of the many products which take advantage of AgNM's antimicrobial properties. To better understand realistic transformations that may occur to these materials upon entrance into the environment, more work needs to be performed to determine the chemical and physical properties of AgNM-containing consumer products throughout their lifecycle. Previous work demonstrated transformations to AgNM containing wound dressings during simulated use (e.g. wound fluid and sweat exposure). The aim of the current work is to evaluate transformations these same textiles undergo during modeled environmental exposure. To model textile disposal conditions, the pristine wound dressings were exposed to synthetic freshwater or artificial landfill leachate. All specimens were analyzed before and after exposure with the techniques dynamic light scattering (DLS), UV-Visible spectroscopy (UV-Vis), X-ray diffraction (XRD), X-ray photoelectron microscopy (XPS), scanning electron microscopy (SEM), energy dispersive X-ray spectroscopy (EDS), and inductively coupled plasma mass spectrometry (ICP-MS). SEM-EDS showed the formation of chloride and phosphorous containing crystals on the surface of the commercial wound dressing after synthetic fresh water exposure. The surface was found to be heterogeneous with some areas showing increased granularity while other areas were similar to the pristine material. Wound dressing exposed to artificial landfill leachate showed increased granularity compared to pristine wound dressing. Unlike synthetic fresh water, no large crystals were found on the surface of the artificial landfill leachate exposed textile. Future studies will evaluate the transformations that occur to wound dressings exposed to simulated disposal conditions *after* being treated with test media for simulating use, thus representing more realistic end-of-use scenarios.

Keywords: silver, nanomaterials, characterization, textile, acticoat

1 INTRODUCTION

Nanomaterials are increasingly being used in consumer goods due to their unique properties. As a result of their antimicrobial nature, silver nanomaterials can be found in consumer products such as athletic clothing, stuffed animals, and food storage products.[1] Additionally, AgNMs are considered an attractive additive for biomedical products and devices such as bandages and wound dressing due to the antimicrobial properties they impart upon their products.[2]

However, increased use of AgNMs in consumer products will result in an increase in their entrance into the environment. This was demonstrated recently in a study by Courtemanche, et al., which found almost all silver remains in commerical wound dressings upon product disposal.[3]

Silver nanomaterials have been shown to have negative consequences on environmental organisms. Work by Arnaout and Gunsch found decreased nitrification by *Nitrosomonas europaea*, a bacteria necessary for nitrogen cycling in wastewater treatment plants.[4] Additionally, AgNMs inhibited anaerobic digestion of waste in lab-scale bioreactors used as a modelled landfill scenario.[5] Due to the potential negative consequences of AgNM-containing consumer products after disposal, it is necessary to understand what physical and chemical transformations these products will undergo throughout their lifecycle.

Here we evaluate the transformations to AgNM-containing textiles, specifically wound dressing, before and after exposure to simulated fluids consistent with disposal in landfills. In this work, wound dressings were exposed to synthetic fresh water and artificial landfill leachate. The resulting transformations have been examined by SEM-EDS with preliminary measurements reported in the following text.

2 METHODS[1]

2.1 Exposure

[1] Certain trade names and company products are mentioned in the text or identified in illustrations in order to adequately specify the experimental procedure and equipment used. In no case does such identification imply recommendation or endorsement by National Institute of Standards and Technology, nor does it imply that the products are necessarily the best available for the purpose.

Gorham, Justin; Gorka, Danielle.
"Physical and Chemical Transformations of Silver Nanomaterials in Textiles After Use and Disposal."
Paper presented at TechConnect World Innovation Conference, Anaheim, CA, United States. May 13, 2018 - May 16, 2018.

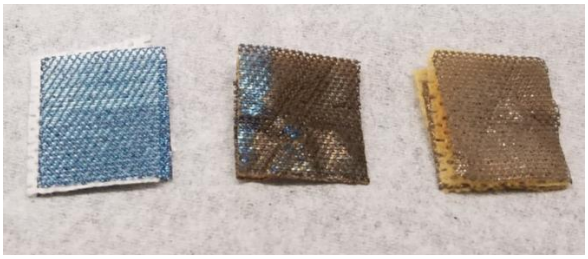


Figure 1: Photographs of Acticoat before (left) and after 24 h exposure to synthetic fresh water (middle) and artificial landfill leachate (right). The wound dressing becomes mottled blue and dark brown after exposure to synthetic fresh water and gray brown after exposure to artificial landfill leachate.

Solutions of synthetic sweat and simulated wound fluid were prepared. Artificial landfill leachate was prepared following United States Environmental Protection Agency (US EPA) guidelines Toxicity Characteristic Leaching Procedure (TCLP).[6] Briefly, 5.7 mL glacial aceitic acid (Mallinckrodt, ACS grade) was added to 500 mL MilliQ water. Then 64.3 mL 1 N sodium hydroxide (Titristar, MilliporeSigma) was added and the solution diluted to 1 L. The pH was adjusted to 4.98. EPA moderately hard water (MHW) was also prepared following EPA guidelines.[7] Briefly, 1.2 g magnesium sulfate (Sigma-Aldrich, 99.5%), 1.92 g sodium bicarbonate (Sigma, 99.5%), and 0.08 g potassium chloride (Mallinkcrodt, ACS grade) were added to 19 L MilliQ water. The solution was aerated overnight. Next 1.2 g calcium sulfate dihydrate (Sigma-Aldrich, 98%) dissolved in 1 L MilliQ water was added and the solution was aerated overnight.

Ten mL exposure media was placed into a 30 mL low-density polyethylene plastic bottle and wrapped in foil (to prevent light exposure). A 2/3 inch x 2/3 inch of Acticoat (Acticoat 7, Smith & Nephews) was then added to the bottle, capped, and placed in a room temperature incubator, rotating horizontally at 50 rpm. The textile was removed at the following times after addition: 5 s, 1 h, 2 h, 6 h, 24 h, and 168 h. All solid samples were stored under vacuum until analysis. Liquid media was stored in the dark at 4 °C until analysis.

2.2 Characterization

A FEI Quanta 200 (Hillsboro, OR) environmental scanning electron microscope (SEM) with a Bruker XFlash 5030 (Billerica, MA) energy-dispersive X-ray spectroscopy (EDS) detector was used to image samples and to collect EDS spectra and mapping data. An operating voltage of 10 kV and a unitless spot size of 3 were used to image the samples and collect EDS data. EDS acquire times were 300 s and data was analyzed using Bruker Esprit v. 1.9.3 software. The middle layer of Acticoat (approximately 2 mm x 2 mm) was adhered to an Al stub with carbon tape.

3 RESULTS AND DISCUSSION

3.1 Physical Transformations

The transformation of Acticoat after exposure to synthetic freshwater or artificial landfill leachate were visibly apparent. The pristine Acticoat was blue on the surface of the silver containing layers (Figure 1), however the color changed after exposure. At shorter exposure times (less than 2 h) there was little to no change in color from blue when exposed to the synthetic freshwater. Longer term synthetic freshwater exposure (> 2h), however, resulted in a gradual color change from blue to a dark brown color. This gradual process is perhaps best exemplified by the 24 h sample shown to the left (Figure 1, middle) which demonstrates how the majority of the Acticoat was dark brown, with the left side remaining blue. By 168 h the textile was completely dark brown. In contrast, exposure to artificial landfill leachate resulted in a color change to gray brown (Figure 1, right). At short exposure times, (less than 1 h) the wound dressing was a dark gray brown. As time increased the material became lighter in color, though it still remained gray brown. Interestingly, the white gauze layer (seen in the pristine wound dressing) between the active layers became yellowish/brown after exposure to either synthetic fresh water and artificial landfill leachate. One potential explanation is that silver redeposited and/or nucleated on this gauze layer after release from the blue wound dressing layers.

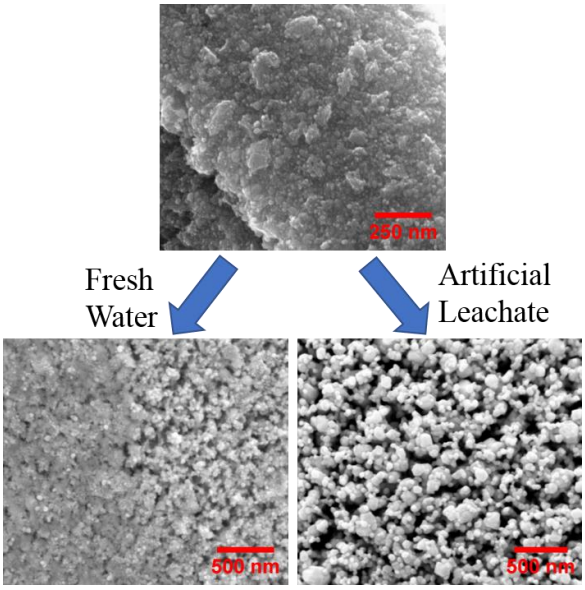


Figure 2: Scanning electron microscopy (SEM) images of pristine Acticoat (top) and Acticoat after 24 h exposure to synthetic fresh water (bottom left) and artificial landfill leachate (bottom right).

Gorham, Justin; Gorka, Danielle.
"Physical and Chemical Transformations of Silver Nanomaterials in Textiles After Use and Disposal."
Paper presented at TechConnect World Innovation Conference, Anaheim, CA, United States. May 13, 2018 - May 16, 2018.

To further examine the physical transformations of Acticoat exposure to modeled environmental media, scanning electron microscopy images were taken (see Figure 2). Pristine Acticoat consisted of discrete spherical nanoscale particles deposited onto the substrate which agrees with previous studies.[8] After exposure to synthetic fresh water, the surface of Acticoat transformed in a non-consistent fashion (Figure 2, bottom left). Some areas showed increased granularity compared to the pristine wound dressing while other areas, sometimes in the same field of view, showed no change at all. Observations from other areas included the presence of submicron and micron-sized crystals on the surface of the silver layer (images in Figure 3A). These crystals were generally triangular or square in shape. In contrast, the surface of Acticoat after exposure to artificial landfill leachate was much more uniform with respect to the type of transformations that occurred. Compared to the pristine material, exposed Acticoat had increased granularity and space between the nanocrystals (Figure 2, bottom right). Particles still retained their spherical shapes. Chemical analysis was next performed to determine the chemical transformations of Acticoat after modelled environmental exposure.

3.2 Chemical Transformations

To better understand the elemental transformations that occurred as result of modeled environmental exposure, EDS was used. Unsurprisingly, EDS mapping of pristine Acticoat revealed silver on the surface, and EDS spectra for the pristine material showed signatures predominantly of silver and also showed signatures for carbon and oxygen. After exposure to synthetic fresh water, EDS mapping showed that many of the micron-sized square crystals contained chloride while the triangular crystals contained phosphorous (Figure 3A). The EDS spectra for the entire mapped region displayed signatures for silver, as well as a signature for chloride. Signatures for phosphorous could be found in samples that contained several triangular crystals. EDS mapping for Acticoat after artificial landfill leachate exposure found silver on the surface. In some instances small crystals on the surface were found to contain chloride. The EDS spectra for Acticoat exposed to artificial landfill leachate contained signatures primarily for silver, with peaks for chloride and oxygen.

4 CONCLUSIONS

To better understand the chemical transformations that are occuring as a result of environmental exposure, further measuerments will be performed using XRD and XPS. These techniques will determine the bonding environment and the oxidation states of the exposed textiles while corroborating the compositional information from EDS. This will then be used to help elicidate the effect of environmental exposure. Similarly, ICP-MS, DLS, and UV-Vis will be used to study the release of silver from the

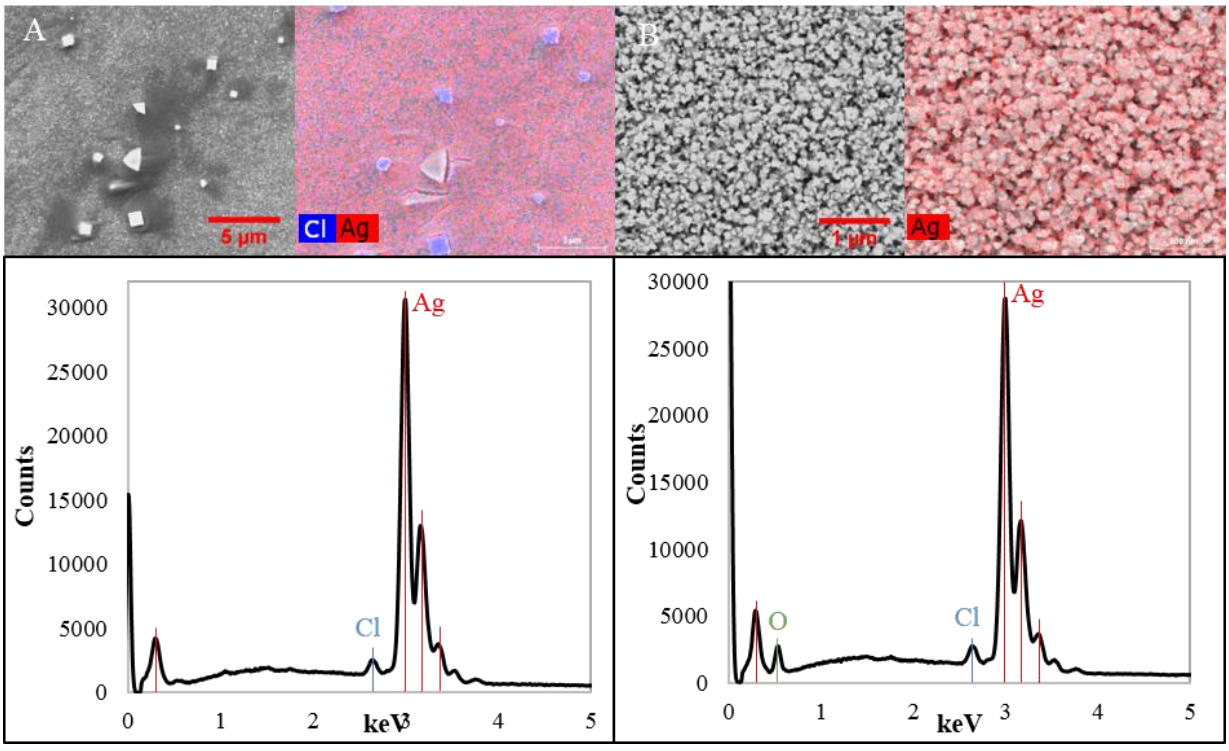


Figure 3: SEM images, energy dispersive X-ray spectroscopy (EDS) maps, and EDS spectra for Acticoat after 24 h exposure to A) synthetic fresh water and B) artificial lanfill leachate.

TechConnect Briefs 2018, TechConnect.org, ISBN 978-0-9988782-2-5

Gorham, Justin; Gorka, Danielle.
"Physical and Chemical Transformations of Silver Nanomaterials in Textiles After Use and Disposal."
Paper presented at TechConnect World Innovation Conference, Anaheim, CA, United States. May 13, 2018 - May 16, 2018.

commerical wound dressing. Knowing the amount of silver released under different environmental scenarios is necessary to determine the environmental impacts of commercial AgNM-containing textiles.

To examine more realistic use and disposal scenarios, the commerical wound dressing will be exposed to model human exposure media (e.g. synthetic sweat or simulated wound fluid) followed by model environmental exposure media. This will allow for more relevant data to be collected and for a more informed understanding of the physical and chemical transformations that occur to AgNM-containing textiles during their lifecycle.

Work here examined the physical and chemical transformations of AgNMs in consumer textiles that result after modeled environmental exposure. Exposure to synthetic freshwater resulted in the formation of chloride and phosphorous containing crystals on the surface of the textile. Additionally, exposure caused the surface to become heterogeneous, with some areas of the textile showing greater granularity than others. Exposure to artificial landfill leachate resulted in increased granularity of the commerical wound dressing compared to the pristine material. Future work will examine more realistic scenarios where the commerical AgNM-containing wound dressing is first exposed to model human exposure media (e.g. synthetic sweat or simulated wound fluid) before model environmental exposure. Data from this work will allow for a greater understanding of the transformations that will occur after disposal for AgNM-containing consumer textiles.

5 ACKNOWLEDGEMENTS

DE Gorka acknowledges funding and support from the National Academy of Science – National Research Council Postdoctoral Research Associateship Program.

REFERENCES

[1] C. Marambio-Jones, E.M.V. Hoek, A review of the antibacterial effects of silver nanomaterials and potential implications for human health and the environment, Journal of Nanoparticle Research 12(5) (2010) 1531-1551.
[2] P.L. Taylor, O. Omotoso, J.B. Wiskel, D. Mitlin, R.E. Burrell, Impact of heat on nanocrystalline silver dressings. Part II: Physical properties, Biomaterials 26(35) (2005) 7230-40.
[3] R.J. Courtemanche, N.S. Taylor, D.J. Courtemanche, Initiating silver recycling efforts: Quantifying Ag from used burn dressings, Environmental Technology & Innovation 4 (2015) 29-35.
[4] C.L. Arnaout, C.K. Gunsch, Impacts of silver nanoparticle coating on the nitrification potential of Nitrosomonas europaea, Environmental Science & Technology 46(10) (2012) 5387-95.
[5] Y. Yang, M. Xu, J.D. Wall, Z. Hu, Nanosilver impact on methanogenesis and biogas production from municipal solid waste, Waste Manag 32(5) (2012) 816-25.
[6] U.S.E.P. Agency, METHOD 1311 TOXICITY CHARACTERISTIC LEACHING PROCEDURE, in: E.P. Agency (Ed.) 1992, p. 35.
[7] U.S.E.P. Agency, Methods for measuring the acute toxicity of effluents and receiving waters to freshwater and marine organisms, Environmental Monitoring Systems Laboratory, Office of Research and Development, US Environmental Protection Agency, 1991.
[8] L.S. Dorobantu, G.G. Goss, R.E. Burrell, Effect of light on physicochemical and biological properties of nanocrystalline silver dressings, RSC Advances 5(19) (2015) 14294-14304.

Gorham, Justin; Gorka, Danielle.
"Physical and Chemical Transformations of Silver Nanomaterials in Textiles After Use and Disposal."
Paper presented at TechConnect World Innovation Conference, Anaheim, CA, United States. May 13, 2018 - May 16, 2018.

An Efficient Timestamp-Based Monitoring Approach to Test Timing Constraints of Cyber-Physical Systems

Mohammadreza Mehrabian
Arizona State University
Mohammadreza.Mehrabian@asu.edu

Mohammad Khayatian
Arizona State University
mkhayati@asu.edu

Ahmed Mousa
Arizona State University
aomoussa@asu.edu

Aviral Shrivastava
Arizona State University
Aviral.Shrivastava@asu.edu

Ya-Shian Li-Baboud
National Institute of Standards and Technology
ya-shian.li-baboud@nist.gov

Patricia Derler
National Instruments
patricia.derler@ni.com

Edward Griffor
National Institute of Standards and Technology
edward.griffor@nist.gov

Hugo A. Andrade
Xilinx
hugoa@xilinx.com

Marc Wiess
Marc Weiss Consulting
marcweissconsulting@gmail.com

John C. Eidson
University of California Berkeley
eidson@eecs.berkeley.edu

Anand Dhananjay
National Institute of Standards and Technology
dhananjay.anand@nist.gov

ABSTRACT

Formal specifications on temporal behavior of Cyber-Physical Systems (CPS) is essential for verification of performance and safety. Existing solutions for verifying the satisfaction of temporal constraints on a CPS are compute and resource intensive since they require buffering signals from the CPS prior to constraint checking. We present an online approach, based on Timestamp Temporal Logic (TTL), for monitoring the timing constraints in CPS. The approach reduces the computation and memory requirements by processing the timestamps of pertinent events reducing the need to capture the full data set from the signal sampling. The signal buffer size bears a geometric relationship to the dimension of the signal vector, the time interval being considered, and the sampling resolution. Since monitoring logic is typically implemented on Field Programmable Gate Arrays (FPGAs) for efficient monitoring of multiple signals simultaneously, the space required to store the buffered data becomes the limiting resource. The monitoring logic, for the timing constraints on the Flying Paster (a printing application requiring synchronization between two motors), is illustrated in this paper to demonstrate a geometric reduction in memory and computational resources in the realization of an online monitor.

CCS CONCEPTS

• **Computer systems organization → Embedded and cyber-physical systems**; **Embedded software**;

1 INTRODUCTION

Cyber-Physical Systems (CPS) are becoming an integral part of human life. While it is desirable to build systems with guarantees of correct behavior, it is becoming increasingly difficult, due to the increasing scale, complexity, and non-deterministic nature of applications, networks, processing platforms, and unpredictable interactions with the physical world [1]. One promising approach to ensure that the system is executing in a safe manner is to monitor the system at runtime [2]. In online monitoring, application constraints are continuously monitored during runtime. Online monitoring can be used to analyze the system behavior in the field and check for bugs in the design. In contrast, offline monitoring in real systems utilizes forensic analysis and therefore does not offer the ability for timely correction of system deviations. Although offline monitoring can be useful, online monitoring is desirable since it may be possible to detect early violations and prevent a system from reaching an unsafe state [3].

Since the correct operation of many CPS applications relies upon the correct timing of the system, both functional and temporal requirements of a CPS must be monitored [4]. This paper focuses on the monitoring of timing constraints in CPS. Many existing monitoring systems define system timing constraints using Signal Temporal Logic (STL) [5]. STL allows users to define timing constraints on real-valued signals relative to current time. For example, in the *Globally* constraint in STL, a user could specify a timing constraint $\phi := \Box_{[2,6]}(|x[t]| < 2)$, which means that a property ϕ will be true at time $t = \tau$, *iff* the real-valued signal $x[t] \in [-2, 2]\ \forall \tau \in [2, 6]$. To compute whether the timing constraint ϕ was met at time $t = \tau$, the conventional approaches [5–10] record the value of the signal $x[t]$ at all times in the interval $t \in [\tau + 2, \tau + 6]$. The signal values are compared against the constraint, $(|x[t]| < 2)$, to determine if the requirements are satisfied within the time interval $[\tau + 2, \tau + 6]$. The constraint evaluation is repeated for each sampling period.

Often, existing monitoring systems are implemented in a simulation. To test real-time systems, FPGA (Field Programmable Gate Array) implementation has the potential to minimize computational latencies and allows for simultaneous monitoring of multiple signals, while supporting the flexibility for modifications and upgrades. The scheme by Jakšić *et al.* [11] was implemented on FPGAs. However, for FPGAs, the available memory to store the signal histories and perform the computation becomes the main bottleneck.

To evaluate how practical the existing state-of-the-art monitoring schemes are, we built a model of a Flying Paster application. We specified seven timing constraints to minimize the amount of

Anand, Dhananjay; Andrade, Hugo; Derler, Patricia; Eidson, John; Griffor, Edward; Khayatian, Mohammad; Li-Baboud, YaShian; Mehrabian, Mohammadreza; Mousa, Ahmed; Shrivastava, Aviral.
"An Efficient Timestamp-Based Monitoring Approach to Test Timing Constraints of Cyber-Physical Systems."
Paper presented at Design Automation Conference, San Francisco, CA, United States. June 24, 2018 - June 28, 2018.

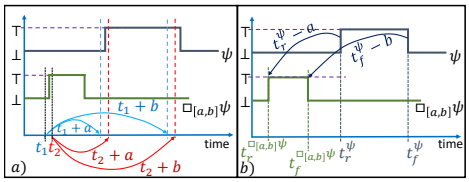


Figure 1: *Globally*: a) Conventional monitoring calculation at each time-step, b) TMA uses two subtractions per pulse.

unused paper while ensuring sufficient time to paste and splice the paper of the new roll before the first roll expires. The implemented test code to evaluate against the timing constraints using the (latest) *Counters* approach by Jakšić *et al.* [11], could not be compiled due to insufficient memory on the commercial off-the-shelf (COTS) FPGA board with 82, 000 flip-flops (FFs) and 41, 000 look-up tables (LUTs), at a sampling rate of 20 kHz. In CPS, examples of systems using high sampling rates include power systems where IEC 61869-9 specifies sampling rates at 4.8 kHz for Alternating Current (AC) and up to 96 kHz for Direct Current (DC) measurements [12].

In this paper, we propose a more efficient online approach for monitoring the timing constraints of CPS, **T**imestamp-based **M**onitoring **A**pproach. The key improvement is rather than evaluating a constraint at each sampling period, TMA only computes the constraint satisfaction at the occurrence of relevant events extracted from monitored signals. For constraint $\phi := \square_{[2,6]}(|x[t]| < 2)$, TMA identifies $x[t]$ as the signal-of-interest when $x[t]$ goes above or below 2 V. Accordingly, ϕ is re-computed only at the occurrence of the next event-of-interest. TMA can monitor all seven timing constraints of Flying Paster model application at a sampling rate of 20 kHz, using only 11% of the FFs and 11.5% of LUTs on the same FPGA.

In general, for a constraint that is defined over a time interval of T , and must be monitored for the duration of experiment d, with a sampling frequency of f, the conventional approach requires $O(Tfd)$ computation time. The requirements of both, computation time and memory, depending on the interval size and the sampling rate. In contrast, our approach has a complexity of $O(k)$, where k is the maximum number of events during time d. Case in point, both the computation and memory requirement of the monitoring logic for implementing a *Globally* constraint using the Jakšić *et al.* approach [11] increases with the interval size of the *Globally* operator, while the monitoring logic of TMA is independent of the time interval, and scales well. Another important point to note is that although event-based constraints (e.g, whenever signal s_1 *rises* above 2.5 V, the signal s_2 should *fall* below 1 V in less than 2 s.) can be specified in STL the logic that is generated can be complex and resource intensive since an event-based constraint is composed of several STL temporal operators. On the other hand, event-based temporal constraints are specified using TTL, *Simultaneity*, *Chronological*, *Phase*, *Frequency*, *Latency*, among others [13]. TTL provides for a more intuitive and simple specification (e.g., $\mathcal{L}(\langle s_1, 2.5, \nearrow\rangle, \langle s_2, 1, \nearrow\rangle) < 2$). In this paper, we apply two of the primitives, namely *Latency* and *Simultaneity*, to illustrate the online monitoring approach.

2 RELATED WORK

Conventional monitoring methods have high memory usage and processing time requirements since they evaluate timing constraints at every time-step. Offline tools for analyzing the timing requirements in CPS have been implemented in Breach[14], and S-Taliro [15]. Both tools record simulation data and evaluate timing constraints after the simulation has finished. Figure 1.a depicts the conventional monitoring approach. It plots the value of Boolean signal ψ along the time axis at the top. To evaluate the constraint $\square_{[a,b]}\psi$ at time t_1, the existing techniques look at the entire interval $[t_1 + a, t_1 + b]$. If the signal is *true* for the entire duration, the constraint is met at t_1. Since the signal ψ was *false* for some time (just after $t_1 + a$), the constraint $\square_{[a,b]}\psi$ is not met at time t_1. However, the constraint is met at t_2. The computation required to evaluate this constraint is $O(Tf^2)$, where T is the time interval over which the temporal operator is defined, which in this case, is $T = b - a$, and f is the sampling frequency. The memory buffer required for this computation will be $O(Tf)$. If there is a constraint with P temporal operators, and w signals, then the amount of computations is $O(TPwf^2)$, while the amount of buffer needed will be $O(TPwf)$. To monitor one timing constraint with four temporal operators defined over an interval of 100 s, with a system sampling rate and analog to digital converter (ADC) resolution of 20 kHz ($t_s = 50\ \mu s$) and 12-bit, we need 12 MB of memory ($M = 4 \times \frac{100}{50\ \mu s} \times \frac{12}{8}$). Primarily, because of the computational complexity, evaluating temporal constraints in real-time is not scalable. Practical CPS applications, such as power generation and distribution have numerous constraints, each containing multiple, high-frequency data signals to be monitored simultaneously, and may have evaluation time intervals over extended durations.

Recognizing the high overhead, AMT [16] proposed an incremental approach to compute the constraints at a segment granularity. However, they can reduce the complexity only by the factor of the granularity. An incremental method was proposed by Deshmukh *et al.* [10] where timing constraints are evaluated by traversing the parse tree generated for STL formulas. They optimize their calculation by eliminating repetitive computations.

While all the previous approaches were implemented in simulation, Jakšić *et al.* [11] implemented a monitoring method called *Counters* algorithm on FPGA. The *Counters* algorithm reduces the computation complexity from $O(n^2)$ to $O(n \log(n))$, where n is the size of time interval of the temporal constraints. Although Jakšić *et al.* showed a way to reduce memory usage, the storage remains a concern (even for bounded constraints). This technique converts future STL operators into past ones and translates all constraints such that their interval starts from zero. Then, a counter is dedicated for measuring the duration of a positive pulse in each interval. The number of needed counters depends on the variability of the monitored signal and the length of the interval bound (a). In each time-step, the active counter is incremented to measure the duration of positive pulses. The maximum number of counters is $\lceil \frac{2a}{b-a+1} \rceil$ where each counter has $\lceil \log_2 a \rceil$ bits. For example, $\square_{[5000,5001]}\psi$ needs 5000 counters, each with $\log_2 5000 = 13$ bits. Therefore, we need around 8 kB to monitor just one signal. In contrast, to monitor the same constraint using TMA, only the last two timestamps of the events-of-interest and the last two timestamps of the result are needed. Therefore, four $32 - bit$ variables for each operator is needed, which is independent of interval length and sampling rate, and a small memory footprint for the state machine. We need one

Anand, Dhananjay; Andrade, Hugo; Derler, Patricia; Eidson, John; Griffor, Edward; Khayatian, Mohammad; Li-Baboud, YaShian; Mehrabian, Mohammadreza; Mousa, Ahmed; Shrivastava, Aviral.
"An Efficient Timestamp-Based Monitoring Approach to Test Timing Constraints of Cyber-Physical Systems."
Paper presented at Design Automation Conference, San Francisco, CA, United States. June 24, 2018 - June 28, 2018.

state machine per operator and the size of each state machine is very small since it needs only 2 bits to store the state.

STL expressions are often combined and/or nested and must be evaluated recursively. Additionally, although STL has the capability to express event-based timing constraints, they are constructed out of a variety of level-based timing constraints. In order to represent only one event (rising or falling) in STL, we should use past and future operators together in one expression[1]. In contrast, TTL can easily express the event timing constraint so that the implementation test code can be succinct as well.

3 TIMESTAMP MONITORING APPROACH

We use TTL to specify the application timing constraints, since TTL succinctly expresses both event-based and level-based timing constraints commonly used in CPS. TTL considers temporal behavior of analog signals upon a given threshold function in level-based timing constraints. Also, this logic can express event-based timing constraints where the time at which a signal value changes (e.g. $\mathcal{L}(\langle s_1, 2.5, \nearrow\rangle, \langle s_2, 1, \nearrow\rangle) > 2$, whenever a rising s_1 signal *crosses* 2.5 V, a rising s_2 shall not *cross* 1 V earlier than 2 s). Hence, we first convert analog signals to discrete event boolean signals by the method in [13, 17]. Therefore, we have $\mathbb{R} \rightarrow \mathbb{B}$ to transform the analog to boolean signals. Then, timestamps corresponding to rising and falling edges are extracted. We define finite sets of rising edges Γ_r and falling edges Γ_f for a boolean signal, ψ, as: $\Gamma_r = \{t^{\psi}_{r_1}, ..., t^{\psi}_{r_n}\}$ and $\Gamma_f = \{t^{\psi}_{f_1}, ..., t^{\psi}_{f_n}\}$ where $t^{\psi}_{r_i}$ and $t^{\psi}_{f_i}$ are the timestamps for the i^{th} rising and falling edge on ψ. Figure 2.a depicts a boolean signal ψ, which is created when the analog signal $s(t)$ crosses a threshold, $f(t)$ that shows after threshold crossing, the boolean signal is described by $t^{\psi}_{r_i}$ and $t^{\psi}_{f_i}$, $(i = 1, ..., n)$. Now, we present a boolean signal as a tuple consisting of an initial state (ψ_{init}), a set of rising edges (Γ_r) and a set of falling edges (Γ_f): $\psi = (\psi_{init}, \Gamma_r, \Gamma_f)$

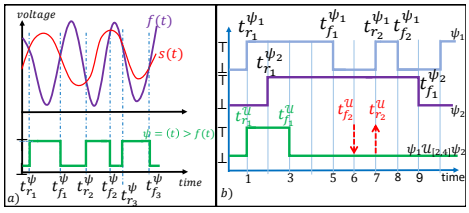


Figure 2: a) Crossing signal $s(t)$ with function $f(t)$, b) Illustration for *Until* computation by TMA.

The *differentiate* operator ($\bowtie$), $\varphi = \bowtie(\psi)$ extracts the rising edge of a boolean signal $\psi \in \mathbb{B}$ where the value of φ is 1 if $(\psi(t^+) \oplus \psi(t)) \wedge \neg\psi(t) = \top$, and $\bot$ otherwise. $\oplus$ is the XOR operator and, t^+ refers to the right neighborhood (the next time-step) of signal at time t. By applying differentiate operator on a boolean signal ψ ($\varphi = \bowtie \psi$), we provide another boolean signal, φ, which is *true* for a short period (sampling time) and *false* otherwise. Since this operator provides the event set, Θ^{φ}, it contains just the timestamps which show the time of ***events*** (not rising and falling) as: $\Theta^{\varphi} = \{\theta^{\varphi}_1, ..., \theta^{\varphi}_n\}$.

[1] $\uparrow \psi = (\psi \wedge (\neg\psi \mathcal{S} T)) \vee (\neg\psi \wedge (\psi \mathcal{U} T))$ for rising edges and
$\downarrow \psi = (\neg\psi \wedge (\psi \mathcal{S} T)) \vee (\psi \wedge (\neg\psi \mathcal{U} T))$ for falling edge.

3.1 Level-based Approach

In this section, we introduce three algorithms executed at each rising and falling edge to define the set of timestamps for online constraint evaluation of level-based TTL operators.

3.1.1 Globally Rules. Given a boolean signal (ψ) expressed with a set of rising and falling edges Γ^{ψ}_r and Γ^{ψ}_f respectively, for every new pair of timestamps, generated from the signal threshold crossings, we update the set of rising and falling edges for $\Box_{[a,b]}\psi$ ($\Gamma^{\Box}_r$ and $\Gamma^{\Box}_f$) by applying Algorithm 1 on the received timestamps. The new $\Box_{[a,b]}\psi$ rising and falling edges are computed based on the most recent t^{ψ}_r (expressed as the current rising edge timestamp on ψ) , t^{ψ}_f (expressed as the current falling edge timestamp on ψ) as well as the values of a and b. The computed rising and falling edges are only added to $\Gamma^{\Box}_r$ and $\Gamma^{\Box}_f$ if its timestamp for the rising edge is less than that of the falling edge.

Algorithm 1 Globally (t^{ψ}_r, t^{ψ}_f, a, b)

1: $t^{\Box}_{r_i} = t^{\psi}_r - a$
2: $t^{\Box}_{f_i} = t^{\psi}_f - b$
3: **if** $t^{\Box}_{r_i} < 0$ **then**
4: $\quad t^{\Box}_{r_i} = 0$
5: **end if**
6: **if** $t^{\Box}_{r_i} <= t^{\Box}_{f_i}$ **then**
7: $\quad \Gamma^{\Box}_r = \Gamma^{\Box}_r + \{t^{\Box}_{r_i}\}$
8: $\quad \Gamma^{\Box}_f = \Gamma^{\Box}_f + \{t^{\Box}_{f_i}\}$
9: **end if**

3.1.2 Eventually Rules. A boolean signal (ψ) expressed with, Γ^{ψ}_r and Γ^{ψ}_f, for every new pair of timestamps, we update the set of rising and falling edges by applying Algorithm 2. The calculated timestamps are only added to the set under the constraint; a rising edge must occur after the last falling edge. Also, if the last computed falling is in the range of new pulse, the last falling should be replaced with the new falling edge to append the last pulse on the result.

3.1.3 Until Rules. Given two boolean signals, ψ_1 and ψ_2, with new rising and falling edges $t^{\psi_1}_r$, $t^{\psi_2}_r$, $t^{\psi_1}_f$ and $t^{\psi_2}_f$, we update the set of rising and falling edges for $\psi_1 \mathcal{U}_{[a,b]} \psi_2$ ($\Gamma^{\mathcal{U}}_r$ and $\Gamma^{\mathcal{U}}_f$) using Algorithm 3 with the incoming pairs of timestamps. The new rising and falling edges for *Until* are computed in the first 2 lines. Starting at line 3, new edges are either appended or discarded, depending on whether or not they comply with the signals. For example, any negative time value and any set of edges with a falling happening before a corresponding rising edge indicate the constraint is not satisfied. Similarly, any edge with rising that comes before the falling edge of the previous set is discarded and the previous falling is replaced with the new falling since the last positive pulse should be extended to the new falling edge. A pair of timestamps appended to $\Gamma^{\mathcal{U}}_r$ and $\Gamma^{\mathcal{U}}_f$ signifies that there is a new valid interval where the constraint, $\psi_1 \mathcal{U}_{[a,b]} \psi_2$, was met. As depicted in the *Until* example in Figure 2.b, we have $t^{\mathcal{U}}_{r_1} = max(1, 2-4) = 1$ and $t^{\mathcal{U}}_{f_1} = min(5, 9)-2 = 3$. Since $t^{\mathcal{U}}_{r_1} < t^{\mathcal{U}}_{f_1}$ they can be used to update $\psi_1 \mathcal{U}_{[2,4]} \psi_2$ by being

Anand, Dhananjay; Andrade, Hugo; Derler, Patricia; Eidson, John; Griffor, Edward; Khayatian, Mohammad; Li-Baboud, YaShian; Mehrabian, Mohammadreza; Mousa, Ahmed; Shrivastava, Aviral.
"An Efficient Timestamp-Based Monitoring Approach to Test Timing Constraints of Cyber-Physical Systems."
Paper presented at Design Automation Conference, San Francisco, CA, United States. June 24, 2018 - June 28, 2018.

Algorithm 2 Eventually (t_r^{ψ}, t_f^{ψ}, a, b)

$t_{r_i}^{\diamond} = t_r^{\psi} - b$
$t_{f_i}^{\diamond} = t_f^{\psi} - a$
if $t_{r_i}^{\diamond} < 0$ **then**
  $t_{r_i}^{\diamond} = 0$
end if
if $t_{f_{i-1}}^{\diamond} < t_{r_i}^{\diamond}$ **then**
  $\Gamma_r^{\diamond} = \Gamma_r^{\diamond} + \{t_{r_i}^{\diamond}\}$
  $\Gamma_f^{\diamond} = \Gamma_f^{\diamond} + \{t_{f_i}^{\diamond}\}$
end if
if $t_{r_i}^{\diamond} <= t_{f_{i-1}}^{\diamond}$ and $t_{f_{i-1}}^{\diamond} < t_{f_i}^{\diamond}$ **then**
  $\Gamma_f^{\diamond} = \Gamma_f^{\diamond} - \{t_{f_{i-1}}^{\diamond}\}$
  $\Gamma_f^{\diamond} = \Gamma_f^{\diamond} + \{t_{f_i}^{\diamond}\}$
end if

appended to $\Gamma_r^{\mathcal{U}}$ and $\Gamma_f^{\mathcal{U}}$. The potential $\psi_1 \mathcal{U}_{[2,4]} \psi_2$ rising and falling edges obtained from the second pulse of ψ_1 are then computed as follows: $t_{r_2}^{\mathcal{U}} = max(7, 2-4) = 7$ and $t_{f_2}^{\mathcal{U}} = min(8, 9) - 2 = 6$. Since $t_{f_2}^{\mathcal{U}} \leq t_{r_2}^{\mathcal{U}}$ they must be disregarded rather than appended to $\Gamma_r^{\mathcal{U}}$ and $\Gamma_f^{\mathcal{U}}$. This concludes that $\mathcal{U}_{[2,4]}$, were met in the interval from time $t = 1$ to $t = 3$, when the first pulse of ψ_1 must hold until the rising event on ψ_2 is true at some time step between a and b[2]. The Finite State Machine (FSM) in Figure 3, calculates the result of *Until* operator with just four states (two bits)[3] .

Algorithm 3 Until ($t_r^{\psi_1}$, $t_r^{\psi_2}$, $t_f^{\psi_1}$, $t_f^{\psi_2}$, a, b)

$t_{r_i}^{\mathcal{U}} = max(t_r^{\psi_1}, t_r^{\psi_2} - b)$
$t_{f_i}^{\mathcal{U}} = min(t_f^{\psi_1}, t_f^{\psi_2}) - a$
if $t_{r_i}^{\mathcal{U}} < 0$ **then**
  $t_{r_i}^{\mathcal{U}} = 0$
end if
if $t_{f_{i-1}}^{\mathcal{U}} < t_{r_i}^{\mathcal{U}}$ and $t_{r_i}^{\mathcal{U}} < t_{f_i}^{\mathcal{U}}$ **then**
  $\Gamma_r^{\mathcal{U}} = \Gamma_r^{\mathcal{U}} + \{t_{r_i}^{\mathcal{U}}\}$
  $\Gamma_f^{\mathcal{U}} = \Gamma_f^{\mathcal{U}} + \{t_{f_i}^{\mathcal{U}}\}$
end if
if $t_{r_i}^{\mathcal{U}} <= t_{f_{i-1}}^{\mathcal{U}}$ and $t_{f_{i-1}}^{\mathcal{U}} < t_{f_i}^{\mathcal{U}}$ **then**
  $\Gamma_f^{\mathcal{U}} = \Gamma_f^{\mathcal{U}} - \{t_{f_{i-1}}^{\mathcal{U}}\}$
  $\Gamma_f^{\mathcal{U}} = \Gamma_f^{\mathcal{U}} + \{t_{f_i}^{\mathcal{U}}\}$
end if

3.2 Event-based Approach

The second category of operators in TTL is event-based. They deal with timestamps of events (Θ set) and produce boolean signals represented by rising and falling sets (Γ_r and Γ_f).

[2]In the calculations for $\psi_1 \mathcal{U}_{[a,b]} \psi_2$ operator, we just consider the overlapped pulses on ψ_1 and ψ_2.

[3]The reader can find all proofs in https://github.com/cmlasu/tma. This link also contain a simulation software, *TMA_Testing.zip*, to evaluate TTL timing constraints.

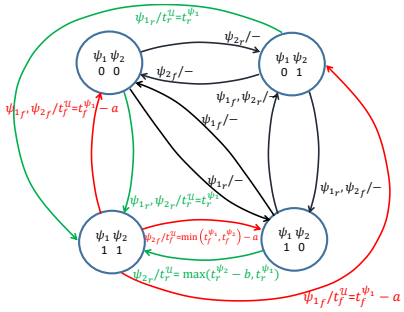


Figure 3: FSM to implement an *Until* operator. FSM captures rise and fall time of boolean signals ψ_1, ψ_2 and computes $\mathcal{U}$.

3.2.1 Simultaneity Constraint. To determine the satisfiability of the *Simultaneity* constraint, the point in time where a set of events have occurred within a time tolerance of ϵ is evaluated. Figure 4.a shows the example of three events occurring within ϵ so that the constraint is met between $\theta_{min} - b$ and $\theta_{max} - a$. We use timed-automata to evaluate this timing constraint. As Figure 4.b, if the timed-automata detects n events and ϵ duration passed after observing the first event the constraint can be calculated.

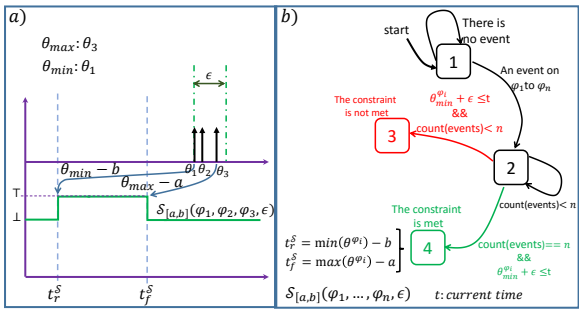


Figure 4: a) Calculation of *Simultaneity* constraint b) The timed-automata to calculate Simultaneity constraint.

3.2.2 Latency Constraint. A latency constraint specifies the time difference between the occurrence of two events. A simple example of a latency constraint is the minimum, maximum or exact time interval between two events, denoted as follows: $\mathcal{L}(\varphi_1, \varphi_2) \triangledown c$ where $\triangledown \in \{>, <, ==\}$. The test code generation takes as input two events (φ_1 and φ_2) and compares the difference between the event timestamps with a real number c. Since the signals are singletons, the sets of Θ^{φ_1} and Θ^{φ_2} each contain one element. Hence, whenever event Θ^{φ_2} is received, the latency can be calculated. The latency constraint evaluation is comprised of two steps: (1) calculating the delay Δt between two timestamps, $\theta_1^{\varphi_1}$, and $\theta_1^{\varphi_2}$, and (2) comparing Δt with c. *Latency* block, $\Delta t = \theta_1^{\varphi_2} - \theta_1^{\varphi_1}$ in comparison block if $((\Delta t \triangledown c) == \top)$ then the rising and falling edges of result are: $t_r^{\mathcal{L}} = \theta_1^{\varphi_2} - b$, $t_f^{\mathcal{L}} = \theta_1^{\varphi_1} - a$.

4 EMPIRICAL EVALUATION

In this section, we applied TMA method to monitor timing constraints in Flying Paster application and compared the required number of FFs and LUTs with the *Counters* algorithm in [11]. Also, we implemented *Globally* operator with different time intervals to show the required memory for an FPGA implementation (Table 1).

Anand, Dhananjay; Andrade, Hugo; Derler, Patricia; Eidson, John; Griffor, Edward; Khayatian, Mohammad; Li-Baboud, YaShian; Mehrabian, Mohammadreza; Mousa, Ahmed; Shrivastava, Aviral.
"An Efficient Timestamp-Based Monitoring Approach to Test Timing Constraints of Cyber-Physical Systems."
Paper presented at Design Automation Conference, San Francisco, CA, United States. June 24, 2018 - June 28, 2018.

Table 1: Memory requirement on FPGA for *Globally*.

		#FFs		#LUTs	
		Jakšić [11]	TMA	Jakšić [11]	TMA
1	$\square_{[0,100]}$	1902	1820	2981	2696
2	$\square_{[0,200]}$	3935		5895	
3	$\square_{[0,300]}$	7821		9314	
4	$\square_{[100,200]}$	1891		2875	
5	$\square_{[200,400]}$	3702		5431	
6	$\square_{[300,600]}$	6312		9612	

4.1 Case Study: Flying Paster Application

The schematic diagram of Flying Paster application [18, 19] is shown in Figure 5. The active roll B feeds the paper and should make contact with the reserve roll A before B runs out of paper, for continuous operations. C and E are idler wheels. Sensor H measures the radius of the paper on B and whenever the radius is less than a threshold, it generates an Approaching Out of Paper (AOP) event. Then, roll A starts to rotate. On the outer side of the reserve roll paper, there is a double-sided adhesive *tape*, which can be detected by sensors F and G. To calculate the angular velocity of roll A, sensors F and J are utilized. When the velocity of the paper at the edge of A becomes the same as roll B, *match* signal is generated. Once *match* is observed, after detecting two rotations, G can detect the tape. When G detects it, idler wheel E pushes the belt to the spare roll A, such that after *tapeToContactAngle*, papers on A and B attach together by the adhesive tape and then the paper on roll A follows the path. Cutter D should cut the paper on roll B after *tapeToCutAngle*. In order to ensure that the new paper is attached properly, we should have *tapeToContactAngle* < *tapeToCutAngle*.

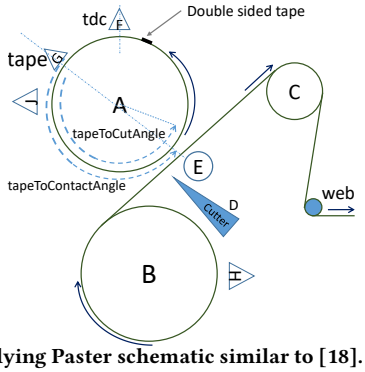


Figure 5: Flying Paster schematic similar to [18]. Active roll, B, is replace by reserve one, A, to feed the web.

To implement this application, we used two Hansen DC motors as rolls A and B. Motors are driven by an Arduino Mega2560 board to control the speed and also to generate *AOP*, *match*, *contact* and *cut* signals. On each motor, we installed a dialed disk with a drilled hole at zero degree (Figure 6). An Omron EESX970C1 sensor was installed close to each disk to detect the drilled hole and hence, measure the rotation speed of each motor. We utilized an NI-cRIO 9035 with an on-board FPGA, Xilinx Kintex-7 7K70T, containing 82,000 FFs and 41,000 LUTs with a 40 MHz clock frequency. For signal monitoring, we used a 20 kHz NI-9381 I/O module and it uses a 12-bit ADC. The pins of NI-9381 were directly connected to *photomicrosensors*, *AOP*, *match*, *cut* and *contact* signals.

Next, we express timing constraints of flying paster application based on STL. The notations for the case study variables are as follows: linear velocity (V), radius (r) and angular velocity (ω). 1) The velocity of the paper on active roll should be constant:

$$V_{active} = (r_{active} \times \omega_{active}) \pm 1\% \text{ m/s}$$

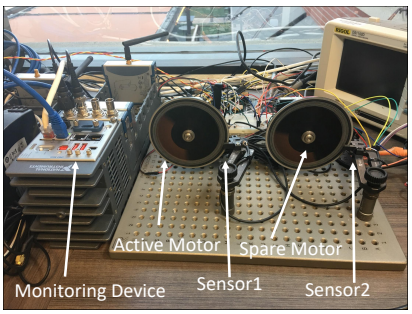


Figure 6: Implemented Flying paster comprises 2 motors and is monitored by reconfigurable data acquisition system.

$$\square_{[t_i,t_s]}(V_{active} = r_{active} \times \omega_{active} \pm 1\%)$$

2) The time interval between *AOP* rising to *match* rising edge must be no more than t_{action}: $\square(\uparrow AOP \Rightarrow \Diamond_{[0,t_{action}]}(\uparrow match))$

3) After *match*, the paper speed of the spare should remain the same as active: $V_{active} = r_{active} \times \omega_{active}$ and $V_{spare} = V_{active} \pm 1\%$

$$\square_{[t_{match},t_{cut}]}(V_{active} = r_{active} \times \omega_{active})$$
$$\square_{[t_{match},t_{cut}]}(V_{spare} = r_{spare} \times \omega_{spare})$$
$$\square_{[t_{match},t_{cut}]}(V_{spare} = V_{active} \pm 1\%)$$

4) Catch the *TDC* (2 rotations of A after *match*).

$$t_{spareTDC} - t_{matchOnSpare} < \frac{4\pi}{\omega_{spare}}$$
$$\Diamond_{[t_{match},t_{match}+\frac{4\pi}{\omega_{spare}}]}(\uparrow spareTDC)$$

5) When tape is 225 degrees after G, *contact* signal must fire.

$$t_{contact} - (t_{spareTDC} + \frac{225\ degrees}{\omega_{spare}}) < \pm 1\ ms.$$
$$\square_{[t_{spareTDC}+\frac{225\ degrees}{\omega_{spare}+1\ ms}, t_{spareTDC}+\frac{225\ degrees}{\omega_{spare}-1ms}]}(\uparrow contact)$$

6) When tape is 270 degrees after G, *cut* signal must fire.

$$t_{cut} - (t_{spareTDC} + \frac{270\ degrees}{\omega_{spare}}) < \pm 1\ ms$$
$$\square_{[t_{spareTDC}+\frac{270\ degrees}{\omega_{spare}\pm 1\ ms}, t_{spareTDC}+\frac{290\ degrees}{w_{spare}\pm 1\ ms}]}(\uparrow cut)$$

7) AOP to cut should not be more than $t_{termination}$.

$$\Diamond_{[t_{AOP},t_{AOP}+t_{termination}]}(\uparrow cut)$$

We implemented the timing constraints of Flying Paster with three approaches (conventional, Jakšić [11] and TMA) on FPGA.

4.1.1 Temporal Logic Expression. We began with the conventional method describing the constraints in STL. We changed the future STL formulas to the past ones [11], and we represented the same timing constraint in TTL for TMA. For example, in $\square(\uparrow AOP \Rightarrow \Diamond_{[0,t_{action}]}(\uparrow match))$, we have:

1) Conventional Method (which is pointed out as *Register Buffer* in [11]): Since rising and falling edges ($\uparrow$ and $\downarrow$) cannot be represented in STL, we express them as the way in [17]:

$$\uparrow \psi = (\psi \wedge (\neg\psi \ \mathcal{S}\ T)) \vee (\neg\psi \wedge (\psi \ \mathcal{U}\ T))$$
$$\downarrow \psi = (\neg\psi \wedge (\psi \ \mathcal{S}\ T)) \vee (\psi \wedge (\neg\psi \ \mathcal{U}\ T))$$

Therefore, the example is converted to:

$$\square((AOP \wedge (\neg AOP \ \mathcal{S}\ T)) \vee (\neg AOP \wedge (AOP \ \mathcal{U}\ T)) \Rightarrow$$
$$\Diamond_{[0,t_{action}]}(match \wedge (\neg match \ \mathcal{S}\ T) \vee (\neg match \wedge (match \ \mathcal{U}\ T)))$$

2) Jakšić in [11] method: Future STL should be converted into past:
$\square(\Diamond_{\{t_{action}\}} \uparrow AOP \Rightarrow \Diamond_{[0,t_{action}]}(\uparrow match))$
The edge ($\uparrow$) operator should be replaced by the equivalent constraint like the conventional method.

Anand, Dhananjay; Andrade, Hugo; Derler, Patricia; Eidson, John; Griffor, Edward; Khayatian, Mohammad; Li-Baboud, YaShian; Mehrabian, Mohammadreza; Mousa, Ahmed; Shrivastava, Aviral.
"An Efficient Timestamp-Based Monitoring Approach to Test Timing Constraints of Cyber-Physical Systems."
Paper presented at Design Automation Conference, San Francisco, CA, United States. June 24, 2018 - June 28, 2018.

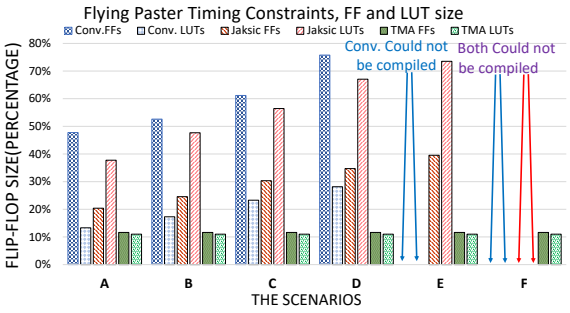


Figure 7: Comparison of FF and LUT numbers in 3 methods.

3) TMA method: Since the constraint is a latency between *AOP* and *match*, it can be easily written in TTL as:

$\mathcal{L}(\langle AOP, 2.5\ V, \nearrow\rangle, \langle match, 2.5\ V, \nearrow\rangle) \leq t_{action}$

The level threshold, 2.5 V, is the threshold to detect *true* or *false* on the boolean signal (0 V and 5 V correspond to $false$ and $true$, respectively). Next, we implemented the constraint specifications on the FPGA using the three monitoring methods.

Table 2: Six different scenarios in which the linear speed of active roll, *AOP* to *match* and time to *contact* time are varied.

	A	B	C	D	E	F
v_{active}	22 m/s	20 m/s	18 m/s	16 m/s	14 m/s	12 m/s
t_{action}	2 s	3 s	4 s	5 s	6 s	7 s
$t_{termination}$	3 s	4 s	5 s	6 s	7 s	8 s

As Figure 7 depicts, conventional and Jakšić methods required more FFs and LUTs in the case study. With increasing intervals, the FF and LUT utilization increases linearly for the Jakšić method. In $\Diamond_{[a,b]}\psi = \Diamond_{\{a\}}\Diamond_{[a,b-a]}$, the variability is $b - a + 1$. In contrast, TMA takes a constant amount of memory in all scenarios because it does not require retention of signal history. When a signal event is observed, the result can be deduced. Moreover, the computation part – that affects the LUT size – is minimal by reducing operators (either event-based or level-based) to simple computations.

4.2 Low variability signals

We evaluate the last timing constraint of flying paster application $(\Diamond_{[t_{AOP}, t_{AOP}+t_{termination}]}(\uparrow cut))$, using all three methods to see the efficiency of TMA in monitoring low variability signals for different values of $t_{termination}$ as shown in the third row of Table 2. Figure 8 compares the FF utilization based upon the conventional, Jakšić, and TMA approaches for constraint evaluation in the case study application, where TMA used the least amount of memory.

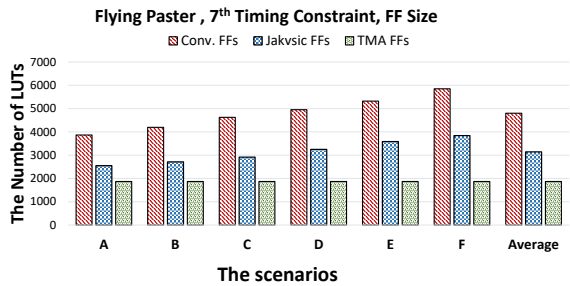


Figure 8: #FFs utilization in three methods.

5 CONCLUSION

We propose a lightweight monitoring methodology, TMA, for CPS timing constraints and demonstrated the efficiencies gained based on an initial case study. The approach utilizes signal timestamps to compute the range for a constraint, rather than processing the levels of signals, requiring data at each sample. The proposed method minimizes computation overhead compared to existing monitoring approaches. The implementation is independent of the constraint interval, allowing the memory usage to be constant for any interval. TMA is particularly geared towards monitoring constraints in TTL, which allows for the succinct description of common timing constraints in CPS, thus simplifying the description and the constraint evaluation algorithms. Future research in this area includes expansion of constraint primitives, such as duration, to fully capture and express temporal constraints in CPS.

***Disclaimer**: Certain commercial entities, equipment, or materials are identified in this document in order to describe the experimental design or to illustrate concepts. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology or the institutions of the other authors, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.*

REFERENCES

[1] Aviral Shrivastava *et al.* A Testbed to Verify the Timing Behavior of Cyber-Physical Systems. In *DAC*. ACM, 2017.
[2] Oded Maler et al. Checking Temporal Properties of Discrete, Timed and Continuous Behaviors. *LNCS*, 2008.
[3] Dejan Nickovic. *Checking Timed and Hybrid Properties: Theory and Applications*. PhD thesis, Université Joseph-Fourier-Grenoble I, 2008.
[4] Thomas Reinbacher, Matthias Függer, and Jörg Brauer. Runtime Verification of Embedded Real-time Systems. *Formal methods in system design*, 2014.
[5] Alexandre Donzé, Thomas Ferrere, and Oded Maler. Efficient Robust Monitoring for STL. In *CAV*. Springer, 2013.
[6] Georgios E Fainekos and George J Pappas. Robustness of Temporal Logic Specifications. In *FATES/RV*. Springer, 2006.
[7] Georgios E Fainekos and George J Pappas. Robustness of Temporal Logic Specifications for Continuous-time Signals. *Theoretical Computer Science*, 2009.
[8] Georgios E Fainekos *et al.* Verification of Automotive Control Applications using S-Taliro. In *American Control Conference (ACC)*. IEEE, 2012.
[9] Howard Barringer, Allen Goldberg, Klaus Havelund, and Koushik Sen. Program Monitoring with LTL in EAGLE. In *IPDPS. 18th*. IEEE, 2004.
[10] Jyotirmoy V Deshmukh et al. Robust online monitoring of signal temporal logic. In *Runtime Verification*, pages 55–70. Springer, 2015.
[11] Stefan Jakšić*et al.* From Signal Temporal Logic to FPGA Monitors. In *MEMOCODE*, 2015.
[12] Wang Mian*et al.* A Review on AC and DC Protection Equipment and Technologies: Towards Multivendor Solution. In *CIGRE INTERNATIONAL COLLOQUIUM*, 2017.
[13] Mohammadreza Mehrabian *et al.* Timestamp Temporal Logic (TTL) for Testing the Timing of Cyber-Physical Systems. In *ESWEEK*. ACM, 2017.
[14] Alexandre Donzé. Breach, A Toolbox for Verification and Parameter Synthesis of Hybrid Systems. In *CAV*, volume 10, pages 167–170. Springer, 2010.
[15] Yashwanth Annpureddy *et al.* S-TaLiRo: A Tool for Temporal Logic Falsification for Hybrid Systems. In *TACAS*. Springer, 2011.
[16] Dejan Nickovic and Oded Maler. AMT: A Property-based Monitoring Tool for Analog Systems. *Formal Modeling and Analysis of Timed Systems*, 2007.
[17] Oded Maler and Dejan Ničković. Monitoring Properties of Analog and Mixed-signal Circuits. *STTT*, 2013.
[18] Patricia Derler *et al.* Using PTIDES and Synchronized Clocks to Design Distributed Systems with Deterministic System Wide Timing. In *ISPCS*. IEEE, 2013.
[19] Drupaloge. PrintIP - Lithoman IV flying splice, last accessed nov. 2017. URL https://www.youtube.com/watch?NR=1%5C&v=wYRGiXMUzA4.

Evaluation of binary and ternary refrigerant blends as replacements for R134a in an air-conditioning system [a]

Ian BELL[1*], Piotr DOMANSKI [2], Greg LINTERIS [2], Mark McLINDEN [1]

[1] National Institute of Standards and Technology, Boulder, CO, USA
[2] National Institute of Standards and Technology, Gaithersburg, MD, USA
ian.bell@nist.gov

* Corresponding Author

ABSTRACT

We investigated refrigerant blends as possible low-GWP (global warming potential) alternatives for R134a in an air-conditioning application. We carried out an extensive screening of the binary and ternary blends possible among a list of 13 pure refrigerants comprising four hydrofluoroolefins (HFOs), eight hydrofluorocarbons (HFCs), and carbon dioxide. The screening was based on a simplified cycle model, but with the inclusion of pressure drops in the evaporator and condenser. The metrics for the evaluation were nonflammability, low-GWP, high COP (coefficient of performance), and a volumetric capacity similar to the R134a baseline system. While no mixture was ideal in all regards, we identified 14 "best" blends that were nonflammable (based on a new estimation method by Linteris, et al., presented in a companion paper at this conference) and with COP and capacity similar to the R134a baseline; the tradeoff, however, was a reduction in GWP of, at most, 51% compared to R134a. An additional eight blends that were estimated to be "marginally flammable" (ASHRAE Standard 34 classification of A2L) were identified with GWP reductions of as much as 99%. These 22 "best" blends were then simulated in a more detailed cycle model.

1. INTRODUCTION

Like all segments of society, the U.S. military is examining the options to reduce the global-warming-potential (GWP) footprint of its air-conditioning and refrigeration systems. But while much of the refrigeration industry is considering a move to refrigerants that are flammable, or at least marginally flammable, the unique operating environments of many military systems demand nonflammable replacement refrigerants. The goal of this work was to identify nonflammable, but lower-GWP, replacements for R134a in a baseline air-conditioning application while maintaining capacity and energy efficiency.

The selection of a refrigerant blend to replace refrigerant R134a is a multi-objective optimization process. There are several desired objectives:

- Minimize/eliminate flammability: As discussed in Linteris et al. (2018), the combination of the adiabatic flame temperature and the F-substitution ratio yields a prediction of the flammability class (1, 2L, 2, 3) according to the ASHRAE 34 standard (ASHRAE, 2016). It is preferred to have a flammability class designation of 1 ("no flame propagation"), but as demonstrated below, enforcing that the blend be nonflammable comes at the cost of a lower system efficiency and/or a higher GWP.
- Minimize GWP: The GWP of a blend is defined as the mass-fraction-weighted GWP of the blend's components. Several time horizons are possible for the calculation of GWP, but it is most common to consider a 100-year time horizon. The 100-year GWP values for pure fluids are tabulated in a number of sources, and here we used the values from the UN IPCC report (Myhre et al., 2013).
- Maximize COP: the coefficient of performance, or COP, characterizes the efficiency of the heat pump. The larger the COP, the better the system efficiency.

Bell, Ian; Domanski, Piotr; Linteris, Gregory; McLinden, Mark.
"Evaluation of binary and ternary refrigerant blends as replacements for R134a in an air-conditioning system."
Paper presented at 17th International Refrigeration and Air Conditioning Conference at Purdue, July 9-12, 2018, West Lafayette, IN, United States. July 9, 2018 - July 12, 2018.

- Match the volumetric capacity Q_{vol} of the baseline system: the Q_{vol} of a heat pump is a figure of merit that is related to the size of the compressor. The larger the volumetric capacity, the smaller the compressor needs to be for a given cooling capacity.

Our search for optimal R134a replacement blends involved the above four figures of merit and consisted of the following steps:

1. Selection of pure refrigerants of low toxicity that could possibly form a replacement blend.
2. Determination of flammability classification, GWP, COP, and Q_{vol} for an exhaustive matrix of possible binary and ternary mixture compositions. In this step, we evaluated COP and Q_{vol} using a simplified cycle model.
3. Selection of "best" blends based on the blend's figures of merit.
4. Determination of COP and Q_{vol} of the "best" blends using an advanced cycle model.

2. FLUID SELECTION

Based on a comprehensive search of chemical compounds that could serve as working fluids in air-conditioning systems, McLinden et al. (2017) demonstrated that there are very limited options for low-GWP refrigerants. They identified the best working fluids based on assessments of their environmental, safety, and performance characteristics. But no single-component refrigerant was ideal in all respects; that is, no fluid was simultaneously nonflammable, low-GWP, and with good performance in an air-conditioning system. Thus, in this study, we turn to blends.

For blending, we selected 13 fluids within a range of pressure, flammability, and GWP values that might produce a blend with the desired characteristics of a R-134a replacement. These included hydrofluoroolefins (HFOs), which have very low GWP values (≈1 relative to CO_2), but that are mildly flammable; hydrofluorocarbons (HFCs) with moderate-to-high GWP values that were nonflammable and thus, might serve to suppress the flammability of a blend; additional mildly flammable HFCs; and carbon dioxide (CO_2), which is nonflammable with GWP = 1, but which would raise the working pressure of a blend. All the selected fluids were of low toxicity (*i.e.*, an "A" classification under ASHRAE Standard 34 (ASHRAE, 2016)). Additional considerations were the commercial availability of the fluid and the availability of property data (in the form of an accurate equation of state), so that cycle simulations could be carried out with some measure of confidence.

The list of candidate working fluids considered in this study is summarized in Table 1. The global warming potential values (based on a 100-year horizon) were taken from the IPCC report on climate change (Myhre et al., 2013).

Table 1: Pure fluids selected in this study and some of their characteristics (T_c: critical temperature)

ASHRAE designation	long name	formula	T_c/K	GWP_{100}	ASHRAE classification
R134a	1,1,1,2-tetrafluoroethane	CF_3CH_2F	374.2	1300	A1
R227ea	1,1,1,2,3,3,3-heptafluoropropane	CF_3CHFCF_3	374.9	3350	A1
R125	pentafluoroethane	CHF_2CF_3	339.2	3170	A1
R143a	1,1,1-trifluoroethane	CF_3CH_3	345.9	4800	A2L
R32	difluoromethane	CH_2F_2	351.3	677	A2L
R152a	1,1-difluoroethane	CHF_2CH_3	386.4	138	A2
R134	1,1,2,2-tetrafluoroethane	CHF_2CHF_2	391.8	1120	Not assigned
R41	fluoromethane	CH_3F	317.3	116	Not assigned
R1234yf	2,3,3,3-tetrafluoropropene	$CF_3CF{=}CH_2$	367.9	1	A2L
R1234ze(E)	trans-1,3,3,3-tetrafluoropropene	$CHF{=}CHCF_3$ (trans)	382.5	1	A2L
R1234ze(Z)	cis-1,3,3,3-tetrafluoropropene	$CHF{=}CHCF_3$ (cis)	423.3	1	Not assigned
R1243zf	3,3,3-trifluoropropene	$CF_3CH{=}CH_2$	376.9	1	Not assigned
R744	carbon dioxide	CO_2	304.1	1	A1

Bell, Ian; Domanski, Piotr; Linteris, Gregory; McLinden, Mark.
"Evaluation of binary and ternary refrigerant blends as replacements for R134a in an air-conditioning system."
Paper presented at 17th International Refrigeration and Air Conditioning Conference at Purdue, July 9-12, 2018, West Lafayette, IN, United States. July 9, 2018 - July 12, 2018.

3. ESTIMATED FIGURES OF MERIT OF THE BLENDS

3.1 Simplified Cycle Model

The cycle model is based upon a simplified analysis of a four-component heat pump system with lumped pressure drops. A schematic of the system is shown in Fig. 1, and $\log(p)$-h and T-s property figures are shown in Fig. 2. Due to the subtle complexities of modeling blends in thermodynamic cycles, we describe the cycle model in detail below. The specification of the model parameters is as follows:

- Evaporator dew-point temperature $T_{\text{evap,dew}}$: 10 °C
- Condenser bubble-point temperature $T_{\text{cond,bub}}$: 40 °C
- Evaporator outlet superheat ΔT_{sh}: 5 K
- Condenser exit subcooling ΔT_{sc}: 7 K
- Compressor adiabatic efficiency η_{a}: 0.7
- Evaporator pressure drop: for the baseline system, a reduction in dew-point temperature of 2 K
- Condenser pressure drop: for the baseline system, a reduction in bubble-point temperature of 2 K

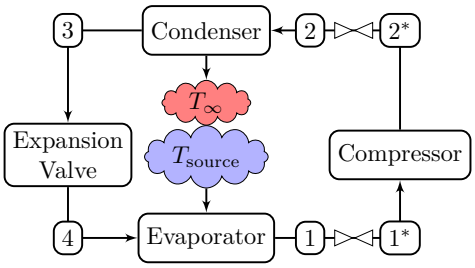


Figure 1: System schematic. The state point indices 1, 2, etc. correspond to the labeled state points in Fig. 2.

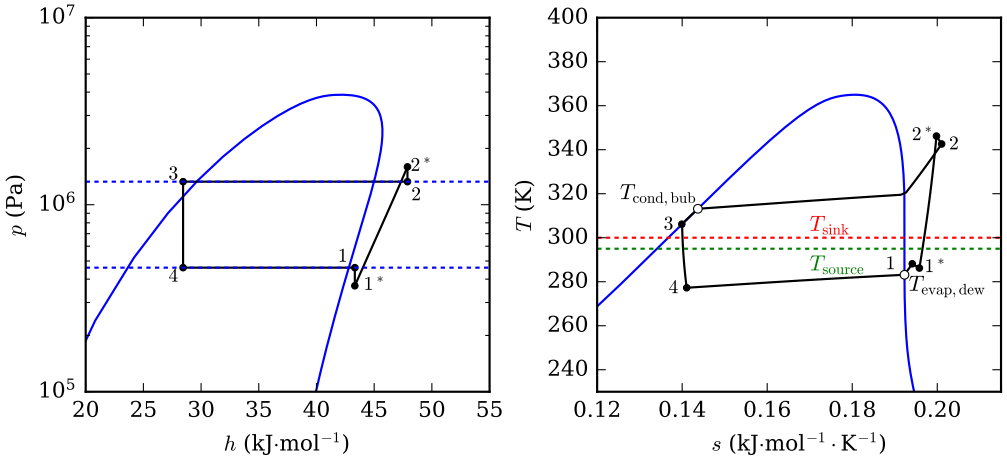


Figure 2: p-h and T-s cycle diagrams for an equimolar mixture of R125 + R1234ze(E). Calculations are carried out with NIST REFPROP (Lemmon et al., 2018).

Bell, Ian; Domanski, Piotr; Linteris, Gregory; McLinden, Mark.
"Evaluation of binary and ternary refrigerant blends as replacements for R134a in an air-conditioning system."
Paper presented at 17th International Refrigeration and Air Conditioning Conference at Purdue, July 9-12, 2018, West Lafayette, IN, United States. July 9, 2018 - July 12, 2018.

The key difference between this cycle model and other simplified cycle models is the inclusion of a simplified pressure drop model. It is assumed that the pressure drop from the high-side components and the low-side components can be lumped into pressure drops at the outlet and inlet of the compressor, respectively. Therefore the compressor sees a larger pressure lift than the pressure ratio corresponding to the pressures in the evaporator and condenser. The drop in saturation temperature for high- and low-sides of the system are specified for the baseline R134a system, and the pressure drop scaling (described below) is used to calculate the pressure drop for the refrigerant blends.

In the simplified cycle analysis, the pressures in the evaporator and the condenser are assumed to be constant, given by vapor-liquid equilibrium calculations at the respective saturation pressure

$$p_{\text{evap}} = p_{\text{dew}}(T_{\text{evap,dew}}) \tag{1}$$

$$p_{\text{cond}} = p_{\text{bub}}(T_{\text{cond,bub}}). \tag{2}$$

The selection of the saturation states used to define the low- and high-side pressures is based on a rudimentary pinch analysis. This pinch analysis assumes that the source and sink temperatures are fixed, that the condenser outlet pinch is fixed, and that the evaporator outlet pinch is fixed. Therefore, stacking up the temperature differences (plus the respective superheating or subcooling), we can arrive at the relevant saturation temperature. This method is the worst-case simplified cycle analysis option for mixtures with temperature glide (McLinden and Radermacher, 1987) because the heat-transfer irreversibilities are maximized. This represents a conservative approach in the sense that it favors drop-in replacements that would require little or no modifications of existing systems. For blends having significant temperature glide, and systems with counterflow or cross-counterflow heat exchange, the temperature profiles of the source and sink fluids and that of the working mixture may be better aligned, resulting in lower heat transfer irreversibilities and higher efficiencies.

Condenser The outlet enthalpy of the condenser is given by

$$h_3 = h(T_3, p_{\text{cond}}), \tag{3}$$

where the outlet temperature of the condenser T_3 is given by

$$T_3 = T_{\text{cond,bub}} - \Delta T_{\text{sc}}, \tag{4}$$

and where the bubble-point temperature of the condenser is given by

$$T_{\text{cond,bub}} = T_{\text{bub}}(p_{\text{cond}}). \tag{5}$$

The pressure drop in the condenser (Δp_{high}) is given by Eq. (15), in which ρ'' and μ'' are evaluated at the dew point at the condensing pressure p_{cond}.

Evaporator The dew-point temperature is imposed for the evaporator, as is its inlet enthalpy (because the outlet state of the condenser is fully specified and the throttling process is assumed to be adiabatic). Therefore the states 3, 4, and 1 can be fully specified and the enthalpies calculated from

$$h_4 = h_3 \tag{6}$$

$$h_1 = h(T_{\text{evap}} + \Delta T_{\text{sh}}, p_{\text{evap}}) \tag{7}$$

The pressure drop in the evaporator Δp_{low} is given by the relationship in Eq. (15), in which ρ'' and μ'' are evaluated at the dew point at the evaporation pressure p_{evap}.

Compressor The pressure drops in the cycle are lumped at the compressor. Therefore, the inlet state of the compressor 1^* is given by the pressure drop relative to the state point 1:

$$h_{1^*} = h_1 \tag{8}$$

$$T_{1^*} = T(h_{1^*}, p_{\text{evap}} - \Delta p_{\text{low}}) \tag{9}$$

Bell, Ian; Domanski, Piotr; Linteris, Gregory; McLinden, Mark.
"Evaluation of binary and ternary refrigerant blends as replacements for R134a in an air-conditioning system."
Paper presented at 17th International Refrigeration and Air Conditioning Conference at Purdue, July 9-12, 2018, West Lafayette, IN, United States. July 9, 2018 - July 12, 2018.

Similarly, the outlet pressure of the compressor p_{2^*} is given by $p_{2^*} = p_{\text{cond}} + \Delta p_{\text{high}}$. The classical adiabatic efficiency formulation is used for the compressor, assuming that there is no heat transfer from the compressor to the environment. Therefore, the adiabatic efficiency is defined by

$$\eta_a = \frac{h_{2\text{s}} - h_{1^*}}{h_{2^*} - h_{1^*}}, \tag{10}$$

where the isentropic enthalpy $h_{2\text{s}}$ is obtained from

$$h_{2\text{s}} = h(s_{1^*}, p_{2^*}). \tag{11}$$

Cycle metrics The COP of the air conditioner is given by

$$\text{COP} = \frac{h_1 - h_4}{h_{2^*} - h_{1^*}} \tag{12}$$

and the volumetric capacity of the heat pump is given by

$$Q_{\text{vol}} = (h_4 - h_1) \cdot \rho(T_{1^*}, p_{1^*}) \tag{13}$$

Pressure drop modeling As demonstrated by McLinden et al. (2017), the inclusion of pressure drop in the model (even if highly approximate), is crucial to yield a fair screening of refrigerants. The simplified pressure drop in our analysis is based upon scaling the system for the refrigerant blends to have the same capacity as the baseline R134a system.

The pressure drop in each of the heat exchangers is assumed to be based upon a frictional pipe flow analysis of a homogeneous fluid (making use of the Fanning friction factor f_F, and neglecting accelerational pressure drop) given by

$$\Delta p = \frac{2 f_F G^2 L}{\rho D} = \frac{2L}{D} \frac{\dot{m}^2}{A^2} \frac{1}{\rho} f_F. \tag{14}$$

For a specified pressure drop Δp and equality of system cooling capacity $Q = \dot{m}/(h_1 - h_4)$, after canceling all non-thermophysical properties and lumping them into a constant, the system specific term $C_{\Delta p}$ is given by

$$C_{\Delta p} = \frac{\Delta p \rho''(h_1 - h_4)^{1.8}}{\mu''^{0.2}}, \tag{15}$$

which is obtained for the baseline system (all units are base SI), for an imposed pressure drop given as a change in saturation temperature for R134a. The pressure drop coefficient obtained is then used for all of the blends, where the thermophysical properties (density ρ'' and viscosity μ'') are evaluated at the dew point state at the specified heat exchange pressure.

3.2 Estimation of Flammability

The refrigerant flammability prediction of Linteris et al. (2018) uses two parameters that can be readily evaluated: the adiabatic flame temperature T_{ad} and the ratio of the number of fluorine atoms to the total of fluorine plus hydrogen atoms in the refrigerant blend, F/(F + H). T_{ad} is calculated from the enthalpy of the reactants and products, via `cantera` (Goodwin et al., 2017), an open-source software package for problems involving chemical kinetics, thermodynamics, and transport properties. The calculation of the F/(F + H) ratio is a simple mathematical evaluation calculated from the chemical formulas of the blend components and their mole fractions. A plot of T_{ad} vs F/(F+H) is constructed, with each point representing a refrigerant, as shown in Fig. 3. Less flammable compounds are in the lower right of the plot, and more flammable, the upper left. The flammability of the refrigerant is represented by the slope of the line between an origin (at (F/(F + H)) = 0 and T_{ad} = 1600 K and the point in question. The origin point is based on the observation that hydrocarbons (for which (F/(F + H)) = 0) do not burn when diluted with an inert gas such that their adiabatic flame temperature falls below 1600 K. For more details, please see Linteris et al. (2018).

Bell, Ian; Domanski, Piotr; Linteris, Gregory; McLinden, Mark.
"Evaluation of binary and ternary refrigerant blends as replacements for R134a in an air-conditioning system."
Paper presented at 17th International Refrigeration and Air Conditioning Conference at Purdue, July 9-12, 2018, West Lafayette, IN, United States. July 9, 2018 - July 12, 2018.

Figure 3 shows the three ASHRAE Standard 34 flammability classes 1, 2L, and 2/3. Assessment of pure fluids and blends having an ASHRAE 34 classification are presented, based on their T_{ad} and F/(F + H), with mixtures evaluated at their nominal blend compositions. As indicated, the present flammability estimation appears to represent the ASHRAE 34 data reasonably well.

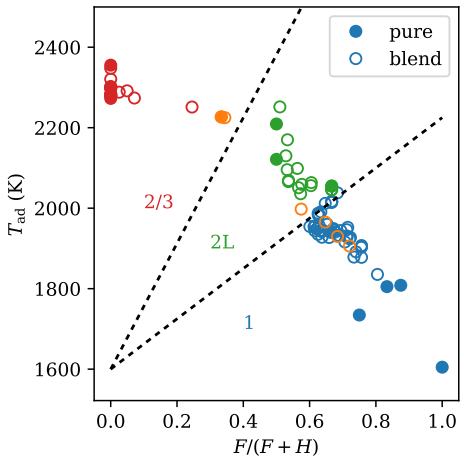


Figure 3: Estimation of blend flammability based on adiabatic flame temperature T_{ad} versus F/(F+H) for pure fluids and nominal blend compositions specified in the ASHRAE 34 standard (ASHRAE, 2016). The colors correspond to the flammability class – 1: blue, 2L: green, 2: orange, 3: red.

3.3 Screening Results

The screening involved an extensive evaluation, using the simplified cycle model described in Section 3.1, of all possible combinations of the 13 fluids listed in Table 1 taken two or three at a time (*i.e.*, all possible binary and ternary mixtures). A composition interval of 0.04 mole fraction was applied to yield a total of 100,387 mixtures to be evaluated. The simplified cycle calculations were carried out in parallel in Python with the `multiprocessing` Python package[b]. The flammability analysis of Linteris et al. (2018) was then applied to estimate the flammability class of the blend.

The screening resulted in a large dataset of binary and ternary mixtures, and for each mixture, an assessment of their figures of merit (flammability class, GWP, COP, and Q_{vol}). The production of this set of data was, in some sense, the easy part of this study; much more difficult was the determination of the "best" refrigerant blend(s). In truth, the selection of the "best" blend depends largely on how the user weights the available figures of merit.

Figure 4 provides an overview of the results for the mixtures formed of the 13 components in Table 1. This figure presents a scatter plot of the COP versus Q_{vol} results for the studied blends sorted into nine "bins" of GWP and flammability. Additional blends had GWP > 1300 and are not shown in Fig. 4. In the upper left hand corner of the figure are mixtures that are probably nonflammable according to the flammability assessment of Linteris et al. (2018) and have a GWP < 150, *i.e.*, less than 12% that of refrigerant R134a.

Figure 5 shows a graphical representation of the prevalence of each component in the different bins. The larger the radius of a wedge, the more prevalent the component is in the mixtures in that bin. In many of the bins there are certain components that dominate the bin. For instance, the low-GWP, nonflammable bin is dominated by carbon dioxide (R744), and the low-GWP, moderately flammable bin is dominated by the HFOs. Each time a component occurs in a bin, its mole fraction in the mixture is added to the running sum for that bin. The mole-fraction-weighted prevalences are then normalized within the bin in order to yield the relative prevalence of each component.

[b] \https://docs.python.org/3/library/multiprocessing.html

Bell, Ian; Domanski, Piotr; Linteris, Gregory; McLinden, Mark.
"Evaluation of binary and ternary refrigerant blends as replacements for R134a in an air-conditioning system."
Paper presented at 17th International Refrigeration and Air Conditioning Conference at Purdue, July 9-12, 2018, West Lafayette, IN, United States. July 9, 2018 - July 12, 2018.

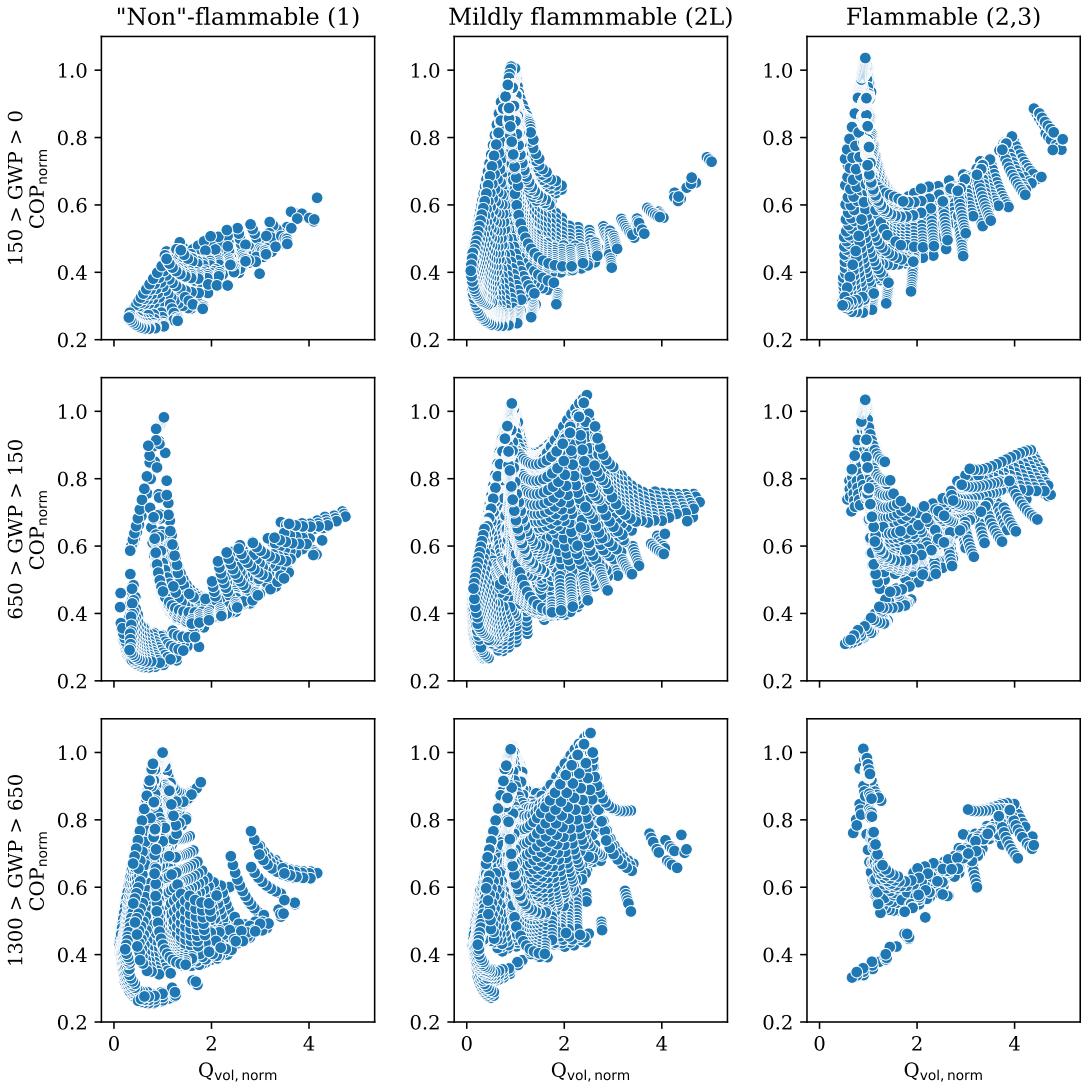


Figure 4: An overview of the cycle figures of merit for the binary and ternary blends studied, divided into bins of GWP and estimated flammability. The "best" bin is at the upper left, and the bins moving towards the lower right are worse according to our objective functions. Values of the volumetric capacity and COP are normalized by the value for the baseline R134a system.

Bell, Ian; Domanski, Piotr; Linteris, Gregory; McLinden, Mark.
"Evaluation of binary and ternary refrigerant blends as replacements for R134a in an air-conditioning system."
Paper presented at 17th International Refrigeration and Air Conditioning Conference at Purdue, July 9-12, 2018, West Lafayette, IN, United States. July 9, 2018 - July 12, 2018.

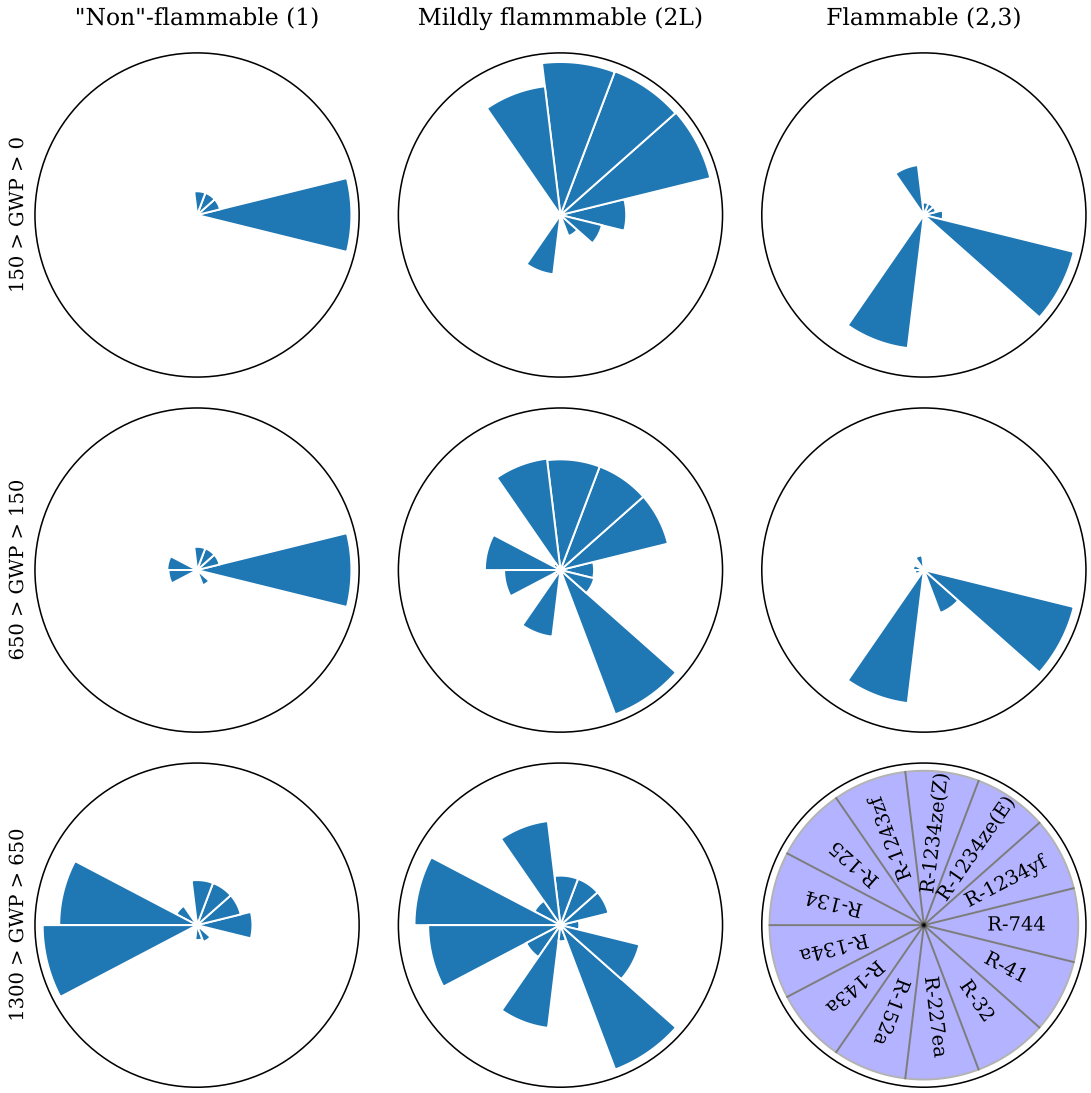


Figure 5: Radial histograms showing the prevalence of each component in each of the bins. The key in the lower right corner is aligned with the radial histograms in each bin.

Bell, Ian; Domanski, Piotr; Linteris, Gregory; McLinden, Mark.
"Evaluation of binary and ternary refrigerant blends as replacements for R134a in an air-conditioning system."
Paper presented at 17th International Refrigeration and Air Conditioning Conference at Purdue, July 9-12, 2018, West Lafayette, IN, United States. July 9, 2018 - July 12, 2018.

3.4 Selection of "Best" Blends

We were not able to identify any blends that met all of our desired constraints. The mixtures in the non-flammable/low GWP bin (upper-left corner of Fig. 4) meet two of the desired objectives, but they suffer from a much lower efficiency than the baseline R134a system and were dropped from further consideration. Thus, to define the "best" blends we selected the nonflammable blends having the highest COP within a range of GWP values from 643 to 870 (*i.e.*, from the remaining two bins in the left column of Fig. 4). These 14 "best blends" are listed in Table 2. Note that we did not separately select blends having very similar compositions to the "best" blends unless they offered a distinct advantage in one or more of our metrics.

If one is willing to tolerate a probable 2L flammability classification according to the ASHRAE 34 standard, there are low GWP options that yield efficiency near that of the baseline R134a system (*i.e.*, the top two rows of the middle column in Fig. 4). We also selected eight additional blends that were marginally flammable with GWP values ranging from 8 to 573. All this is to say that the search for the "perfect" refrigerant blend continues.

4. DETAILED CYCLE SIMULATIONS

4.1 Model Description

We performed detailed cycle simulations using the CYCLE_D-HX model (Brown et al., 2017) on the "best" blends described in Section 3.4. In contrast to the simplified vapor compression cycle model, which requires refrigerant saturation temperatures in the evaporator and condenser as inputs, CYCLE_D-HX establishes saturation temperatures in the heat exchangers using the specified temperatures profiles of the heat source and heat sink (*i.e.*, the conditioned and outdoor air) and the mean effective temperature differences (ΔT_{hx}) in the evaporator and condenser. This representation of heat exchangers facilitates the inclusion of both thermodynamic and transport properties in cycle simulations (Brown et al., 2002a,b; Brignoli et al., 2017). The evaporator and condenser can be counterflow, crossflow, or parallel flow, although only cross-flow is simulated here. During the iteration procedure, CYCLE_D-HX calculates ΔT_{hx} for each heat exchanger from (Domanski and McLinden, 1992):

$$\frac{1}{\Delta T_{hx}} = \frac{Q_1}{Q_{hx}\Delta T_1} + \frac{Q_2}{Q_{hx}\Delta T_2} + \cdots = \frac{1}{Q_{hx}} \sum_i \frac{Q_i}{\Delta T_i} \tag{16}$$

In this equation, ΔT_{hx} is a harmonic mean weighted with the fraction of heat transferred in individual sections of the heat exchanger, based on the assumption of a constant overall heat-transfer coefficient throughout the heat exchanger. Each term represents the contribution of a heat exchanger section. At the outset, the model calculates ΔT_{hx} based on sections corresponding to the subcooled liquid, two-phase, and superheated regions. Then, the model bisects each section and uses Eq. (16) to calculate a new value of ΔT_{hx}. The model repeatedly bisects each subsection until the ΔT_{hx} obtained from two consecutive evaluations agree within a convergence parameter. As an alternative to specifying ΔT_{hx}, the heat exchangers can be characterized by the overall heat conductance $UA_{hx} = 1/R_{hx}$ (R_{hx} being the total resistance to heat transfer in the heat exchanger). In this case, the model calculates the corresponding ΔT_{hx} from the basic heat-transfer relation, $\Delta T_{hx} = Q_{hx}/UA_{hx}$, where Q_{hx} is the product of refrigerant mass flow rate and enthalpy change in the evaporator or condenser, as appropriate. The representation of heat exchangers by their UA_{hx} allows for inclusion of refrigerant heat transfer and pressure drop characteristics in comparable evaluations of different refrigerants. For this purpose, CYCLE_D-HX considers R_{hx} as the summation of the resistance on the refrigerant side (R_r), and combined resistances of the heat exchanger material and heat-transfer-fluid (HTF) side, ($R_{tube} + R_{HTF}$)

$$R_{hx} = R_r + (R_{tube} + R_{HTF}) \tag{17}$$

$$R_r = 1/(\alpha_r \cdot A_r) \tag{18}$$

where α_r is the refrigerant heat-transfer coefficient in $W \cdot m^2 \cdot K^{-1}$ and A_r is the surface area on the refrigerant side in m^2.

The refrigerant heat-transfer resistance R_r varies with operating conditions and the refrigerant, but the other resistances ($R_{tube} + R_{HTF}$) are assumed to be constant. Their combined value can be calculated from UA_{hx}, α_r, and A_r values during a simulation run for the "reference" refrigerant, for which the heat exchanger's ΔT_{hx} are known from laboratory measurements and were provided as input. CYCLE_D-HX calculates ($R_{tube} + R_{HTF}$) for the evaporator and condenser within

Bell, Ian; Domanski, Piotr; Linteris, Gregory; McLinden, Mark.
"Evaluation of binary and ternary refrigerant blends as replacements for R134a in an air-conditioning system."
Paper presented at 17th International Refrigeration and Air Conditioning Conference at Purdue, July 9-12, 2018, West Lafayette, IN, United States. July 9, 2018 - July 12, 2018.

Table 2: Detailed results from CYCLE_D-HX

Blend components	Composition (molar)	GWP_{100}	COP/COP_{R134a}	$Q_{vol}/Q_{vol,R134a}$
..... *Class 1 nonflammable (predicted)*				
R134a/1234yf/134	0.48/0.48/0.04	634	0.987	0.975
R134a/1234yf/1234ze(E)	0.52/0.32/0.16	640	0.987	0.989
R134a/1234yf	0.52/0.48	640	0.989	1.029
R134a/1234yf/134	0.40/0.44/0.16	665	0.986	0.958
R134a/125/1234yf	0.44/0.04/0.52	676	0.985	1.049
R134a/227ea/1234yf	0.40/0.04/0.56	681	0.984	1.007
R134a/1234ze(E)	0.60/0.40	745	0.988	0.908
R134a/1234yf	0.60/0.40	745	0.990	1.031
R134a/1234ze(E)/1243zf	0.60/0.36/0.04	750	0.990	0.966
R134a/R1234yf/1234ze(E)	0.64/0.2/0.16	799	0.990	0.986
R134a/152a/1234yf	0.64/0.04/0.32	817	0.993	1.023
R134a/1234yf/134	0.52/0.32/0.16	825	0.990	0.966
R134a/1234ze(E)	0.68/0.32	852	0.991	0.929
R134a/1234yf/1243zf	0.68/0.2/0.12	870	0.994	1.020
..... *Class 2L flammable (predicted)*				
R152a/1234yf	0.08/0.92	8	0.980	0.957
R134a/1234yf	0.20/0.80	238	0.980	0.996
R134a/152a/1234yf	0.20/0.16/0.64	270	0.987	0.984
R152a/1234yf/134	0.16/0.48/0.36	418	0.984	0.900
R134a/1234yf	0.36/0.64	436	0.985	1.018
R134a/1234yf/1243zf	0.36/0.44/0.20	451	0.988	1.004
R134a/152a/1234yf	0.36/0.20/0.44	496	0.994	0.994
R134a/1234yf	0.468/0.532	573	0.988	1.027

this "reference run" and stores their values for use in subsequent simulation runs for calculation of UA_{hx} characterizing the heat exchangers with a new refrigerant or operating conditions.

CYCLE_D-HX requires the following operational input data for the "reference run": HTF inlet and outlet temperatures for the evaporator and condenser; ΔT_{hx} for the evaporator and condenser (to achieve the measured evaporator and condenser saturation temperatures); evaporator superheat and pressure drop; and condenser subcooling and pressure drop. Additional "reference run" inputs include compressor isentropic and volumetric efficiencies, and electric motor efficiency. Heat exchanger geometry inputs include the tube inner diameter and length, the number of refrigerant circuits, and the total length of heat exchanger tubing.

CYCLE_D-HX also optimizes the coil circuiting in the evaporator and condenser to maximize the system's COP. This option represents a design environment where the HTF and number of refrigerant tubes remains constant, but the tube connections and refrigerant mass flux can be changed. Using this option, the model provides information on the relative performance potentials of refrigerants operating in systems with serpentine air-to-refrigerant heat exchangers.

4.2 Simulation Results

The series of CYCLE_D-HX simulations of the 22 "best" blends started with R134a simulations, which served as the "reference" refrigerant. For this purpose, we established an R134a system, with operating parameters approximating those used in the simplified cycle simulations: the same evaporator outlet superheat (5 K), condenser exit subcooling (7 K) and compressor efficiency (0.7) were used; however, refrigerant pressure drop (corresponding to 2 K drop in saturation temperature) was imposed in the heat exchangers (as opposed to the compressor suction and discharge sides), and average two-phase temperatures in the heat exchangers were considered as opposed to the dew-point temperature (evaporator) and bubble-point temperature (condenser). The circuitry in the R134a system was optimized to attain the maximum COP, and the performance of this R134a optimized system became the reference for normalization of COP and Q_{vol} of the nineteen blends.

Bell, Ian; Domanski, Piotr; Linteris, Gregory; McLinden, Mark.
"Evaluation of binary and ternary refrigerant blends as replacements for R134a in an air-conditioning system."
Paper presented at 17th International Refrigeration and Air Conditioning Conference at Purdue, July 9-12, 2018, West Lafayette, IN, United States. July 9, 2018 - July 12, 2018.

Table 2 presents GWP and simulation results for the nineteen blends. For the nonflammable group, the normalized values for COP and Q_{vol} were in the 0.984 – 0.994 and 0.908 – 1.049 range, respectively, with the GWP values ranging from 634 to 870. The main component of all of these blends is R134a. The other components are the HFOs R1234yf, R1234ze(E), and R1243zf and HFCs R152a and R134; R125 and R227ea appear at a low concentration in one blend each. For the mildly flammable group, the GWP values range from 8 to 573, and the normalized COP ranges from 0.980 to 0.994. The normalized $Q_{vol}$the blends in this group are in the range 0.900 – 1.027. These blends comprise R134a as the main component along with R1234yf, and R152a; R134 and R1243zf appear in one blend each.

Keeping in mind that the main goal of this study was to find a nonflammable, low-GWP replacement with a comparable COP and Q_{vol} of that for R134a in an air-conditioning application, the lowest GWP among the suitable nonflammable blends is 634, a 51% reduction in GWP compared with R134a. The blend R134a/1234yf, with molar composition (0.468/0.532) and GWP = 573, was predicted to be marginally flammable by our estimation method; this blend is designated R513A by ASHRAE Standard 34 with a classifcation of A1 (i.e., nonflammable).

If one is willing to tolerate a probable 2L flammability classification according to the ASHRAE 34 standard, there are options that yield efficiency near that of the baseline R134a system with GWP of 8 and 4.3 % lower Q_{vol}. Similarly, if a more moderate reduction in GWP is acceptable, there are higher-pressure low-GWP options with R32 that attain a similar COP as R134a with a more than doubled Q_{vol} (i.e., the fluids making up the second COP maxima shown in the middle panel of Figure 4.)

5. CONCLUSIONS

Our search for nonflammable low-GWP replacements for R134a in an air-conditioning system yielded several blends with COP and Q_{vol} similar to those of R134a. The GWP of the identified nonflammable blends were in the 634 - 870 range. Among the mildly flammable (2L) blends, GWP reductions of up to a factor of 100 relative to R134a were identified.

The study was limited to binary and ternary blends formed from a set of 13 pure fluids currently available in NIST REFPROP (Lemmon et al., 2018). Additional pure fluids, such as those identified by McLinden et al. (2017), should be considered once sufficient experimental data become available to build the thermodynamic equations of state and mixture models required to implement them into REFPROP.

The COP and Q_{vol} values calculated from CYCLE_D-HX model present the relative performance potential of the considered fluids in a system with air-to-refrigerant heat exchangers. Experimental validation of these findings and predicted flammability classifications is merited.

Finally, flammability limits are generally device-dependent, so while the current estimation method can predict the behavior of a mixture in the ASTM E681 test protocol (for constituents which are chemically similar to those used to develop the model; i.e., hydrocarbons, HFCs, HFOs, etc.), the behavior of the mixtures in other flammability tests or actual full-scale configurations having more powerful ignition sources, clutter, turbulence, etc., may not be predicted as well. Moreover, since there is uncertainty in the flammability behavior and prediction for compounds near the flammability boundary, the actual ASHRAE Standard 34 flammability behavior predicted in the present work should ultimately be verified experimentally.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Defense, Strategic Environmental Research and Development Program (SERDP), project WP-2740.

REFERENCES

ASHRAE (2016). ANSI/ASHRAE Standard 34-2016 Designation and Safety Classification of Refrigerants.

Brignoli, R., Brown, J. S., Skye, H. M., and Domanski, P. A. (2017). Refrigerant performance evaluation including effects of transport properties and optimized heat exchangers. Int. J. Refrig, 80:52–65.

Bell, Ian; Domanski, Piotr; Linteris, Gregory; McLinden, Mark.
"Evaluation of binary and ternary refrigerant blends as replacements for R134a in an air-conditioning system."
Paper presented at 17th International Refrigeration and Air Conditioning Conference at Purdue, July 9-12, 2018, West Lafayette, IN, United States. July 9, 2018 - July 12, 2018.

Brown, J., Kim, Y., and Domanski, P. (2002a). Evaluation of Carbon Dioxide as R22 Substitute for Residential Conditioning. ASHRAE Transactions, 108(2):954–963.

Brown, J. S., Brignoli, R., and Domanski, P. A. (2017). CYCLE_D-HX: NIST vapor compression cycle model accounting for refrigerant thermodynamic and transport properties, version 1.0, user's guide. Technical Report 1974, NIST.

Brown, J. S., Yana-Motta, S. F., and Domanski, P. A. (2002b). Comparitive analysis of an automotive air conditioning systems operating with CO_2 and R134a. Int. J. Refrig., 25(1):19–32.

Domanski, P. A. and McLinden, M. O. (1992). A simplified cycle simulation model for the performance rating of refrigerants and refrigerant mixtures. Int. J. Refrig, 15(2):81–88.

Goodwin, D. G., Moffat, H. K., and Speth, R. L. (2017). Cantera: An Object-oriented Software Toolkit for Chemical Kinetics, Thermodynamics, and Transport Processes. http://www.cantera.org. Version 2.3.0.

Lemmon, E. W., Bell, I. H., Huber, M. L., and McLinden, M. O. (2018). NIST Standard Reference Database 23: Reference Fluid Thermodynamic and Transport Properties-REFPROP, Version 10.0, National Institute of Standards and Technology.

Linteris, G., Bell, I., McLinden, M., and Domanski, P. (2018). An Empirical Model For Refrigerant Flammability Based On Molecular Structure and Thermodynamics. In 17th International Refrigeration and Air Conditioning Conference at Purdue, July 9-12, 2018.

McLinden, M. and Radermacher, R. (1987). Methods for comparing the performance of pure and mixed refrigerants in the vapour compression cycle. International Journal of Refrigeration, 10(6):318 – 325.

McLinden, M. O., Brown, J. S., Brignoli, R., Kazakov, A. F., and Domanski, P. A. (2017). Limited options for low-global-warming-potential refrigerants. Nat. Commun., 8:14476.

Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J.-F., Lee, D., Mendoza, B., Nakajima, T., Robock, A., Stephens, G., Takemura, T., and Zhang, H. (2013). Anthropogenic and Natural Radiative Forcing. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

NOMENCLATURE

Variables

A	cross-sectional area (m^2)
A_r	surface area on the refrigerant side (m^2)
$C_{\Delta p}$	pressure drop constant
COP	coefficient of performance (-)
D	diameter (m)
G	mass flux ($kg^2 \cdot s^{-1} \cdot m^{-2}$)
f_F	Fanning friction factor (-)
GWP	global warming potential (-)
h	mass specific enthalpy ($J \cdot kg^{-1}$)
L	length (m)
$\dot{m}$	mass flow rate ($kg \cdot s^{-1}$)
p	pressure (Pa)
Δp	pressure difference (Pa)
Q	capacity (W)
Q_{vol}	volumetric capacity ($W \cdot m^{-3}$)
R	heat-transfer resistance ($K \cdot W^{-1}$)
s	mass specific entropy ($J \cdot kg^{-1} \cdot K^{-1}$)
ΔT	temperature difference (K)
T	temperature (K)
UA	overall heat conductance ($W \cdot K^{-1}$)
α_r	refrigerant heat-transfer coefficient ($W \cdot m^2 \cdot K^{-2}$)
μ''	viscosity of saturated vapor (Pa·s)
ρ	mass density ($kg \cdot m^{-3}$)
ρ''	mass density of saturated vapor ($kg \cdot m^{-3}$)
η_a	adiabatic efficiency (-)

Subscripts

ad	adiabatic flame temperature
bub	bubble
c	critical
cond	condenser
evap	evaporator
hx	heat exchanger
r	refrigerant
vol	volumetric
sc	subcooling
sh	superheat
1-4	state points
2s	isentropic compression
1*, 2*	state points with pressure drop

Bell, Ian; Domanski, Piotr; Linteris, Gregory; McLinden, Mark.
"Evaluation of binary and ternary refrigerant blends as replacements for R134a in an air-conditioning system."
Paper presented at 17th International Refrigeration and Air Conditioning Conference at Purdue, July 9-12, 2018, West Lafayette, IN, United States. July 9, 2018 - July 12, 2018.

Verification of Resilience Policies that Assist Attribute Based Access Control

Antonios Gouglidis
School of Computing and Communications
Lancaster, UK
a.gouglidis@lancaster.ac.uk

Vincent C. Hu
Computer Security Division
NIST, USA
vincent.hu@nist.gov

Jeremy S. Busby
Department of Management Science
Lancaster, UK
j.s.busby@lancaster.ac.uk

David Hutchison
School of Computing and Communications
Lancaster, UK
d.hutchison@lancaster.ac.uk

ABSTRACT

Access control offers mechanisms to control and limit the actions or operations that are performed by a user on a set of resources in a system. Many access control models exist that are able to support this basic requirement. One of the properties examined in the context of these models is their ability to successfully restrict access to resources. Nevertheless, considering only restriction of access may not be enough in some environments, as in critical infrastructures. The protection of systems in this type of environment requires a new line of enquiry. It is essential to ensure that appropriate access is always possible, even when users and resources are subjected to challenges of various sorts. Resilience in access control is conceived as the ability of a system not to restrict but rather to ensure access to resources. In order to demonstrate the application of resilience in access control, we formally define an attribute based access control model (ABAC) based on guidelines provided by the National Institute of Standards and Technology (NIST). We examine how ABAC-based resilience policies can be specified in temporal logic and how these can be formally verified. The verification of resilience is done using an automated model checking technique, which eventually may lead to reducing the overall complexity required for the verification of resilience policies and serve as a valuable tool for administrators.

CCS Concepts

•General and reference → Verification; •Security and privacy → Formal security models; Access control; •Software and its engineering → Model checking;

Keywords

Attribute based access control; resilience

1. INTRODUCTION

Access control is an essential technique in all computing systems. Its main role is to control and limit the actions or operations in a system that are performed by a user on a set of resources. Access control policies, models and mechanisms are considered to be three abstractions of control introduced by access control systems [10]. These levels of abstraction are responsible for the enforcement of access control policies, as well as for preventing the access policy from subversion. Specifically, a policy can be defined as a high-level requirement that specifies how a user may access a specific resource and how. Access control policies can be enforced in a system through an access control mechanism. The latter is responsible for permitting or denying a user access to a resource, and specifying the nature of access that is permitted. An access control model can be defined as an abstract container of a collection of access control mechanism implementations. These are capable of preserving support for the reasoning of the system policies through a conceptual framework. Hence, the abstraction gap between the mechanism and policy in a system is bridged by means of access control model [24].

The importance of access control in systems led to research in several directions, one of them being the investigation of properties related to the security offered by policies, e.g. secure inter-operation [11, 12, 26]. However, little attention has been paid to the verification of resilience specifications in access control policies. Resilience in access control is conceived as the ability of a system not to restrict, but to enable access to resources [21]. Most of the research work in this context is initiated around the *'resiliency checking problem'*, which examines whether a given resilience policy is satisfied by an access control state. This problem has been investigated from a generic point of view [21], and thus the proposed approaches are agnostic to the actual type of policies implemented by an underlying model. Additional research on the resiliency checking problem was performed to investigate the time complexity introduced by the various parameters used in it [8]. Moreover, the *'resiliency checking*

Busby, Jeremy; Gouglidis, Antonios; Hu, Chung Tong; Hutchison, David.
"Verification of Resilience Policies that Assist Attribute Based Access Control."
Paper presented at 2nd Workshop on Attribute Based Access Control (ABAC 2017), Scottsdale, AZ, United States. March 24, 2017 - March 24, 2017.

problem' is shown to have a connection with the *'work-flow satisfiability problem'* in [9], with the latter being investigated extensively in the literature, e.g. in [6, 7, 27] amongst others. Furthermore, information on work-flow management systems and on how to model and enforce resilience policies is available in [2, 4].

Despite of several research directions being followed to address the problem of ensuring resilience in access control policies, they assume the construction of a resilience aware policy during the design phase of a system, and not during its operational phase. This does not negate the correctness of these approaches, but questions the level of usability provided by them, as well as their applicability in real operational environments. The need for verifying the correcteness of resilience properties during the operational phase of a system should be a requirement in critical infrastructures, including utility networks and industrial control systems, where services must be provided in an uninterrupted manner. Therefore, motivated by the absence of a practical approach, we examine in this paper how resilience policies can be specified and verified in the context of an actual access control model. For this purpose, the Attribute-Based Access Control (ABAC) was selected amongst others due to its flexibility and high level of expressiveness [15]. We anticipate a practical approach that would be able to efficiently ensure the resilience of policies. This will eventually reduce the overall complexity required for verifying resilience policies. The latter requirement derives mostly from the need to apply such a process in the operational phase of a system. Such functionality is currently absent from existing approaches since, firstly, the majority of solutions appear to propose the problem be solved during the design phase of a system, and secondly, tools that would help to facilitate the process of verifying resilience are absent from existing solutions, to the best of our knowledge. Hence, for us to achieve these objectives, we specify resilience policies in the context of ABAC, and embrace an existing set of tools, viz. the Access Control Policy Tool (ACPT) by NIST[1] and the NuSMV symbolic model checker[2] in order to specify ABAC-based policies and formally verify them, respectively [16].

The structure of the remainder of this paper is as follows: in Section 2 we elaborate on the relation between access control and resilience. A formal definition of our proposed ABAC model is provided in Section 3. Section 4 provides prerequisite information on the formal verification of policies in ABAC; and, resilience policies are specified and verified in Section 5. Concluding remarks and future work are discussed in Section 6.

2. ACCESS CONTROL AND RESILIENCE

In the context of industrial control systems cyber-security, resilience is a particularly significant issue. Such systems control physical processes, often safety-critical processes, in real time, in chemical production plants, water treatment facilities, nuclear power generation installations, oil and gas facilities and so on. Losses of availability and integrity in particular can have immediate and possibly unrecoverable effects. Such systems must simultaneously exclude adversarial intervention and ensure legitimation intervention. Thus access control always has to satisfy the dual requirement of denying access to certain types of actor but guaranteeing access for others. Following other authors [21], and our own prior work [13], the second requirement is labelled as *'resilience'*. As this second requirement has to be met at the same time as the first, and the two requirements must not contradict each other, it makes sense to express access control and resilience requirements using the same basic formalism, and to find a way of verifying them under some integrated mechanism.

With regards to access control in critical infrastructures – the authors in [22] provide information on how role based access control policies may appear in SCADA systems, and argue on the fact that roles and type of access on critical resources have to be clearly defined. A subset of roles in the hierarchy described in [22] is: *'Junior operator'*, *'Senior operator'*, *'Supervisor'*, and *'Manager'*. A set of operations is assigned with each role. Briefly, a user assigned with the *'Junior operator'* role, has a very restrictive set of operations, such as monitoring screens only; the *'Senior operator'* role offer operations such as these of starting or stopping a system and the potential to acknowledge an alarm (on top of the roles inherited by the *'Junior operator'* role); and, the *'Supervisor'* role offers the operation of disabling alarms (on top of the roles inherited by the *'Senior operator'* role) – a *'Manager'* is considered to have no restrictions, and thus can perform all the above operations [22], and also can be connected with other role hierarchies. In a scenario where the operations of starting, stopping a SCADA system and disabling alarms are considered to be critical, we must ensure the presence of personnel that would own the appropriate set of permissions to accomplish these critical operations successfully. Thus, in case of assigning *'user1'* with the role of *'Supervisor'* and *'user2'* with the role of *'Manager'* this policy can be characterised as being *'resilient'* since upon the absence (or removal) of one user, there still exist one disjoint set of users that contains one user authorised for starting, stopping the SCADA system and to acknowledge alarms.

With regards to resilience – this concept is identified to be of vital importance for organisations since it ensures an organisation's survivability and prosperity [5]. One of the important processes of resilience at an organisational level is operational resilience management, which in general refers to the set of strategies that when applied are able to protect and sustain the services and assets of an organisation [23]. Access management consists one of the operations introduced in an operational resilience management strategy. Its purpose is to ensure that the access granted to the subjects of a system (i.e. assignment with organisational assets) has to be proportionate with the business and resilience requirements [23]. Hence, to fulfil the resilience requirements, the subjects of a system (e.g. users) have to have a sufficient, yet not excessive, level of access to the organisation's assets. In order to successfully achieve this goal, a series of practices needs to be applied. Such practices, as identified by major research and development centres [23], are related with (i) enabling access, (ii) managing changes to access privileges, (iii) performing periodic review and maintenance of access privileges, and (iv) correction of inconsistencies.

In this paper, we investigate the first practice, i.e. enable access, that should be used in industrial control systems to ensure their protection and orderly functionality. The

[1]http://csrc.nist.gov/groups/SNS/acpt/
[2]http://nusmv.fbk.eu/

Busby, Jeremy; Gouglidis, Antonios; Hu, Chung Tong; Hutchison, David.
"Verification of Resilience Policies that Assist Attribute Based Access Control."
Paper presented at 2nd Workshop on Attribute Based Access Control (ABAC 2017), Scottsdale, AZ, United States. March 24, 2017 - March 24, 2017.

practice of enabling access is concerned with ensuring that the appropriate level of access to organisational assets is informed by resilience requirements [23]. This is of vital importance in critical infrastructures since their operational environment is required to keep access control current and reflective of the security and resilience requirements towards maximizing their availability [17].

3. ATTRIBUTE BASED ACCESS CONTROL MODEL

Attribute-based access control (ABAC) has gained the attention of researchers because it offers a high level of flexibility – it can implement various policies, and also it is an ideal candidate for use in highly-distributed and rapidly changing environments [15]. In general, access decisions in ABAC are based on the requester's owned attributes. The advantage of this approach is that it is possible to provide access to users in a collaborative environment without the need for them to be known by the resource a priori. This results in an inherent support for distributed access control and collaboration amongst domains. An implementation of ABAC can be seen in the eXtensible Access Control Markup Language (XACML) that is an OASIS standard[3]. And, another implementation of ABAC can be found in the Next Generation Access Control standard in [1].

In the following, we provide a definition of the ABAC model. This includes a reference to the main elements that could take part in the authorisation process, and a high-level description of its main administrative operations and administrative review functions.

3.1 Proposed Model

The definition of our ABAC model is based on the recommendations proposed by NIST in [14], where a set of guidelines forms the basis of a formal definition of ABAC. Thus, we provide all the required information with regard to the specifications of ABAC. Specifically, we elaborate on its main elements and the relation between them; we provide a formal definition of the model, and provide a list of ABAC's system and administrative functional specifications.

3.2 Elements

The ABAC model consist of the following six categories of elements: attributes, subjects, objects, operations, policies, and environmental conditions. A major difference between ABAC and other access control models is that in ABAC access is not granted or not based on the subject's identity, but rather it is evaluated on the basis of a set of attributes assigned to subjects and objects, as well as on environmental conditions. Figure 1 illustrates the main elements in ABAC and the interactions amongst them.

Attributes are characteristics of the subject, object, or environment conditions. Attributes may contain information given by a name-value pair, i.e. a tuple of the form: $(NAME, VALUE)$. As depicted in Figure 1, both subject and object attributes are able to support the use of meta-attributes. The latter provides an additional index for referring to groups of subjects and objects per se. Hierarchies in ABAC are intrinsically supported via the meta-attribute

[3]http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html

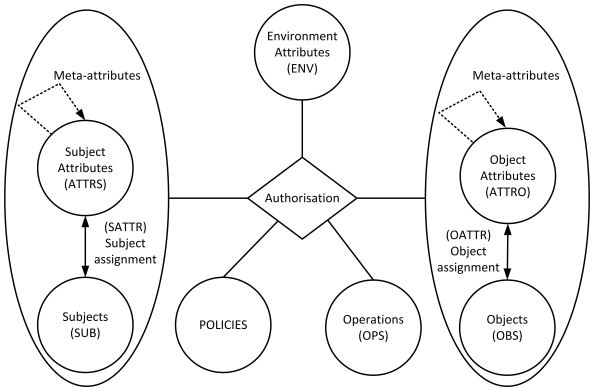


Figure 1: ABAC's main elements and relationships

functionality. This provides ABAC with the potential to express powerful hierarchies between elements of the same type.

A **subject** is usually interpreted as being a user or process that issues access requests to perform operations on objects. Subjects can be assigned with one or more attributes.

An **object** can be a system resource for which access is managed by the attribute-base access control system. These could be devices, files, records, tables, processes, programs, networks, or domains containing or receiving information. It can be the resource or requested entity, as well as any entity on which an operation may be performed by a subject including data, applications, services, devices, and networks.

An **operation** is the execution of a function at the request of a subject upon an object. Example of operations include the read, write, edit, delete, copy, execute, and modify commands.

A **policy** is the representation of rules or relationships that makes it possible to determine if a requested access should be allowed, given the values of the attributes of the subject, object, and possible environment conditions.

An **environment condition** is an operational or situational context in which access requests occur. Environment conditions are detectable environmental characteristics. Environment characteristics are independent of subject or object, and may include the current time, day of the week, location of a user, the current threat level, etc.

The provision of the above definitions subsequently helps in the provision of a reference model for ABAC and a formal specification of it. In the following, we provide information about the main elements of the model and the relations between them.

3.3 Definitions

A formal model of the ABAC is defined as follows:

- $SUB, OBJ, ATTR_S, ATTR_O, ENV, OPS, POLICIES$ consist of subjects, objects, subjects' attributes, objects' attributes, environmental conditions, operations, and policies, respectively.
- $ATTR_S, ATTR_O, ENV$ are sets of subject, object and environmental conditions attributes in tuples of the form: $(NAME, VALUE)$.
- SUB is a set of tuples of the form: $(NAME, VALUE)$.

Busby, Jeremy; Gouglidis, Antonios; Hu, Chung Tong; Hutchison, David.
"Verification of Resilience Policies that Assist Attribute Based Access Control."
Paper presented at 2nd Workshop on Attribute Based Access Control (ABAC 2017), Scottsdale, AZ, United States. March 24, 2017 - March 24, 2017.

- OBJ is a set of tuples of the form: ($NAME$, $VALUE$).
- $SATTR \subseteq (SUB \times ATTR_S)$ is a set of SUB and $ATTR_S$ mapping relation pairs.
- $OATTR \subseteq (OBJ \times ATTR_O)$ is a set of OBJ and $ATTR_O$ mapping relation pairs.
- $AssignedSubjects(a : ATTR_S) = 2^{SUB}$, or formally defined:
  $AssignedSubjects(a) = \{s \in SUB \mid (s, a) \in SATTR\}$.
- $AssignedObjects(a : ATTR_O) = 2^{OBJ}$, or formally defined:
  $AssignedObjects(a) = \{o \in OBJ \mid (o, a) \in OATTR\}$.
- OPS is a set of tuples of the form: ($NAME$, $VALUE$).
- $POLICIES \subseteq SATTR \times OATTR \times ENV \times OPS$, where $POLICIES$ is a set of rules or relationships that makes it possible to determine if a requested access should be allowed, given the values of the attributes of the subject, object, and possibly environment conditions.

3.4 ABAC system and administrative functional specifications

The ABAC system and administrative functional specifications describe the main features required by an ABAC system. This includes the specification of a set of administrative operations and administrative review functions. The former consists of a set of functions that are required to administer the main elements of the access control model. These include operations such as the creation and deletion of elements, and assignments. The administrative review functions are capable of performing query operations on ABAC elements and relations. Tables 1 and 2 in Appendix B provide the function prototypes of the proposed ABAC model, including a short description of their functionality – these have been specified using a subset of Z notation, which is standardised in ISO/IEC 13568:2002 [18], but a full description of them is omitted here since it is out of the scope of this paper.

4. PRELIMINARIES ON VERIFICATION

An authorisation mechanism may include various complex operations, viz. assembles the policy, attributes and renders a decision based on the logic provided in the policy [14]. In this section, we provide information regarding some of the basic principles in temporal logic, which we use to specify policies in ABAC, and thus express authorisations. For more information, we refer the reader to [3]. The use of temporal logic, apart from providing a language for the property specification of policies, will eventually underpin the mathematical foundation used to formally verify authorisation policies. This requires the definition of a language for expressing polices and a transition system able to describe the behaviour of the access control model, and thus for properties to be verifiable for the model.

We consider AP to be a set of atomic propositions, and α, β and γ elements of AP. The set of propositional logic formulae over AP is inductively defined as:

- *true* is a formula;
- Any atomic proposition, which is element of AP is a formula;
- If Φ, Φ_1 and Φ_2 are formulae, then are $(\neg\Phi)$ and $(\Phi_1 \wedge \Phi_2)$;
- Nothing else is a formula.

We have that the conjunction operator $\wedge$ binds stronger then the derived binary operators, such as that of disjunction, implication, etc. Specifically, we define the former two as in the following: $\Phi_1 \vee \Phi_2 := \neg(\neg\Phi_1 \wedge \neg\Phi_2)$ and $\Phi_1 \rightarrow \Phi_2 := \neg\Phi_1 \vee \Phi_2$, respectively. The $\rightarrow$ means *'imply'*.

We also assume the following notation regarding the associativity and commutativity law for disjunction and conjunction: $\bigwedge_{1 \leq i \leq n} \Phi_i$ for $\Phi_1 \wedge \ldots \wedge \Phi_n$ and $\bigvee_{1 \leq i \leq n} \Phi_i$ for $\Phi_1 \vee \ldots \vee \Phi_n$. If $I = \emptyset$, then $\bigwedge_{i \in \emptyset} \Phi_i := true$ and $\bigvee_{i \in \emptyset} \Phi_i := false$.

Furthermore, we consider the *evaluation* of atomic propositions. This is done by assigning a truth value to each of them, i.e. a function $\mu : AP \rightarrow \{0, 1\}$, where 0 is *false* and 1 is *true*. The $\rightarrow$ means *'maps to'*. Therefore, a satisfaction relation $\models$ indicates the evaluations μ for which a formula Φ is *true*. Formally, it is written as:

- $\mu \models true$
- $\mu \models \alpha \Leftrightarrow \mu(\alpha) = 1$
- $\mu \models \neg\Phi \Leftrightarrow \mu \not\models \Phi$
- $\mu \models \Phi \wedge \Psi \Leftrightarrow \mu \models \Phi$ and $\mu \models \Psi$

Further on, we define the authorisation rule, property, and transition system of the ABAC model. The definitions are based on [14], but modified to fit the requirements of ABAC. Here we use the Computation Tree Logic (CTL) in order to specify policy properties. Linear-time Temporal Logic (LTL) could alternatively be used since we do not take advantage of the different expression level of neither CTL or LTL in our defined properties [19].

With regard to CTL, the prefixed path quantifiers assert arbitrary combinations of linear-time operators. Hence, we use the universal path quantifier $\forall$ that means *'for all paths'*, and the linear temporal operators $\Box$ and $\Diamond$ that mean *'always'* and *'eventually'*, respectively. Furthermore, we use the temporal modalities $\forall\Box\Phi$ representing *invariantly* Φ, and $\forall\Diamond\Phi$ representing inevitably Φ, where Φ is a state formula.

Definition 1. An ABAC rule is an implication of type '$c \rightarrow d$', where constraint c is a predicate expression as ($sub \wedge sattr \wedge obj \wedge oattr \wedge env \wedge ops$), which when *true* implies the permission decision d. The $\rightarrow$ means *'imply'*.

Definition 2. An ABAC access control property p is an implication formula of type '$b \rightarrow d$', where the result of the access permission d depends on *quantified* predicate b on ABAC attributes and system states.

Definition 3. A transition system TS_{ABAC} is a tuple (S, Act, δ, i_0) where

- S is a set of states, $S = \{Permit, Deny\}$;
- Act is a set of actions,
  where $Act = \{(sub \wedge sattr \wedge obj \wedge oattr \wedge env \wedge ops), \ldots\}$ and $sub \in SUB$, $sattr \in ATTR$, $obj \in OBJ$, $oattr \in OATTR$, $env \in ENV$ and $ops \in OPS$;

Busby, Jeremy; Gouglidis, Antonios; Hu, Chung Tong; Hutchison, David.
"Verification of Resilience Policies that Assist Attribute Based Access Control."
Paper presented at 2nd Workshop on Attribute Based Access Control (ABAC 2017), Scottsdale, AZ, United States. March 24, 2017 - March 24, 2017.

- δ is a transition relation where $\delta : S \times Act \rightarrow S$;
- $i_0 \in S$ is the initial state.

The p in Definition 2 is expressed by the proposition $p : S \times Act^2 \rightarrow S$ of TS_{ABAC}, which can be collectively translated in terms of logical formula such that $p = (s_i \wedge \bigvee_{1 \leq i \leq n} (sub_n \wedge sattr_n \wedge obj_n \wedge oattr_n \wedge env_n \wedge ops_n)) \rightarrow d$ where $p \in P$ is a set of properties.

The behaviour of the system is defined by the ABAC rules, and they function as the transition relation δ in TS_{ABAC}. Thus, by representing an access control property using the temporal logic formula p, we can assert that model TS_{ABAC} satisfies p by $TS_{ABAC} \vDash \forall \Box (b \rightarrow \forall \Diamond d)$. Property $\forall \Box (b \rightarrow \forall \Diamond d)$ is a response pattern such that d responds to b globally (b is the cause and d is the effect) [25].

With regard to computational complexity – it is interesting to reference the computational complexity of the *'resiliency checking problem'*, which is NP-hard in the general case [21], and the computational complexity of model checking, which is P-complete for CTL and PSPACE-complete for LTL [3, 20]. However, we have to clarify that the computational complexity of the *'resiliency checking problem'* and that of verification of resilience specifications using model checking are not directly comparable. This is because in the former case a resilience policy is prepared from scratch taking into consideration resilience requirements, whereas in the latter case, resilience specifications are verified against an existing policy. We argue that in real-world cases, an initial set of resilience access control policies may be present and that they can change over time. Nevertheless, the development of a new – from scratch – resilience policy in operational environments is not always feasible due to operational requirements. Thus, the verification of resilience specifications may appear to be a more realistic and efficient solution.

5. VERIFICATION OF RESILIENCE POLICIES

5.1 Specification of resilience

In this section, we elaborate on the notion of resilience policies, and discuss how this could be interpreted in the context of the defined ABAC model. In order to do this, we embrace the definition of resilience policies defined in [21]. Specifically, a resilience policy is defined as the tuple of $ResiliencePolicy\langle P, s, d, t\rangle$, where P is the set of permissions, $s \geq 0$, $d \geq 1$ and $t \in N^+$ or $t = \infty$. Thus, a resilience policy is satisfied in an access control state *'if and only if upon removal of any set of s users, there still exist d mutually disjoint sets of users such that each set contains no more than t users and the users in each set together are authorised for all permissions in P'* [21]. The construction of a resilience policy is also known in the literature as the *'resiliency checking problem'* [8], [21]. Specifically, given a resilience policy tuple $ResiliencePolicy\langle P, s, d, t\rangle$ the solution provides an answer to the existence of binary relation between users U and permissions P, i.e. $UP \subseteq U \times P$ [21], or between users U and their authorised resources R, i.e. $UR \subseteq U \times R$ [8]. In general, permissions are considered to be operations on objects. Assuming the example in Section 2, we have the following critical operations on a SCADA system: *'monitor screen'*, *'start system'*, *'stop system'*, *'disable alarm'*, and *'change set points'*, and thus we set $P = \{Supervisor, Manager\}$. This is because $OATTR \times OPS$ is an ordered set that represents permissions P, and $ATTR_S \times P$ is also an ordered set that can be used for creating pairs of roles with permissions. Since both roles in the example are paired with all permissions, we can continue with the assumption that it is safe to use $roles \in ATTR_S$ and $permissions \in P$ interchangeably. Given P, we may have the following values for the rest of the resilience policy parameters: $s = 1$, $d = 1$, and $t = 1$. Specifically, $s = 1$ indicates that we want the policy to be resilient to the absence of any (one) user, $d = 1$ indicates that we require one set of users such that users in that set together possess all permissions; and, $t = 1$ since there is a single user that has all the permissions [21].

The definition of a resilience policy requires initially a careful definition of the different critical tasks in an organisation and subsequently identification of the main users and assigned permissions required to successfully complete these tasks. As mentioned already, this process can be performed during the early stages of the design of a system. Nevertheless, users and policies may change in a system, i.e. certain policies may be altered, deleted or new policies may be introduced. Therefore, these operations may introduce disruptions in an already existing resilience policy. Designing these policies from scratch may not be a viable solution, especially in the context of critical infrastructures, where systems must operate in an uninterrupted manner. Hence, administrators or operators in such environments may require to verify at any time the resilience offered by the active set of policies in their operational environment. Such an approach may also lead to reducing the overall complexity imposed by solving the resiliency checking problem from scratch.

Towards providing a viable solution to this requirement, we outline a process that is able to verify the resilience of a subset of ABAC policies. The resiliency checking problem is known to have various levels of complexity, introduced by the variance of each of the elements in the resilience policies tuple [21]. In this paper, we are concerned with the verification of resilience provided by a set of access control policies that are required to achieve a critical operation or task, and thus verify their resilience in the presence of several threats, e.g. absence of users that are responsible for completing - collaboratively or not - a specific critical operation or task.

In order to verify the resilience of ABAC policies, we use the response pattern as defined in Section 4. In general, we check resilience in ABAC policies between users and attributes – the latter representing permissions required to perform a task. For this, we use the satisfiability relation expressed by Formula 1.

$$TS_{RP\_ABAC} \vDash \forall \Box \Big(\bigwedge_{1 \leq i \leq n} !sub_n \bigwedge_{0 \leq i \leq m} attr_m \bigwedge_{1 \leq i \leq k} !sub_k \rightarrow \forall \Diamond Deny \Big) \quad (1)$$

where TS_{RP_ABAC} is the resilience ABAC policy transition system, sub_n, $sub_k \in SUB$, $sub_n \neq sub_k$, $attr_m \in ATTR_S : \{sub_n\} \times \{attr_m\} \in SATTR$, and $Deny \in S$ is the permission decision. In relation to the resilience policy tuple, i.e. $\langle P, s, d, t\rangle$, sub_n is mapped onto the set of users s that are considered to be absent; $attr_m$ refers to the attributes assigned with a user s and represent permissions required

Busby, Jeremy; Gouglidis, Antonios; Hu, Chung Tong; Hutchison, David.
"Verification of Resilience Policies that Assist Attribute Based Access Control."
Paper presented at 2nd Workshop on Attribute Based Access Control (ABAC 2017), Scottsdale, AZ, United States. March 24, 2017 - March 24, 2017.

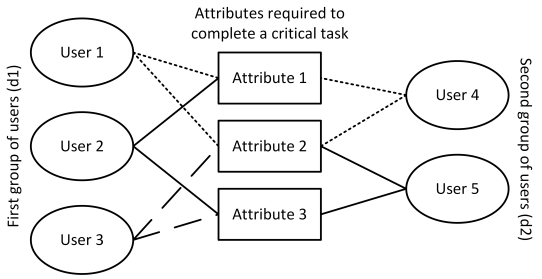


Figure 2: Example of a resilience policy

to perform a task; and, sub_k refer to the mutual disjoint set of users expressed by d. With regard to the t parameter – this can be introduced implicitly by introducing additional specifications, following Formula 1.

5.2 Example

To demonstrate the applicability of the verification process in identifying resilience in ABAC policies, we choose to elaborate a generic, yet, representative example of a resilience policy, as described in [21]. Nevertheless, we also describe in Appendix A the implementation of an RBAC policy for a SCADA system using function calls of the proposed ABAC model (described in Appendix B).

The resilience policy in Fig. 2 assumes the existence of a critical task T. In order to successfully accomplish the critical task, the users (e.g. operators in a utility organisation) have to be collaboratively authorised for all three attributes. In this example, we consider two groups of users, where the first group includes the following users and assigned attributes: $User1 \times \{Attribute1, Attribute2\}$, $User2 \times \{Attribute1, Attribute3\}$, $User3 \times \{Attribute2, Attribute3\}$; and the second group includes the following users and attributes: $User4 \times \{Attribute1, Attribute2\}$, $User5 \times \{Attribute2, Attribute3\}$. In the context of an industrial control system, the above attributes could be equivalent with: $Attribute1 \equiv$ (monitor a device), $Attribute2 \equiv$ (start or stop a device), and $Attribute3 \equiv$ (maintain a device). Thus, in case of a device malfunction, operators (i.e. users of the system) shall be in position to monitor and acknowledge the problem, stop the faulty device, maintain the device, and finally, start the device. In order to define the above policy in ABAC and to formally verify its resilience, we use ACPT for the definition of access control policies and NuSMV for the verification of the resilience policy specifications.

In the following, we provide in Listing 1 the NuSMV code that defines the ABAC policy described in Figure 2.

Listing 1: *Specification of an ABAC policy in NuSMV*

```
MODULE main
VAR
    USER : {User1, User2, User3, User4, User5};
    ATTR : {Attribute1, Attribute2, Attribute3};
    ABAC_Policy01 : ABAC_Policy01(USER, ATTR);
ASSIGN
    next (USER) := USER;
    next (ATTR) := ATTR;
MODULE ABAC_Policy01(USER, ATTR)
VAR
    decision : {Permit, Deny};
ASSIGN
    init (decision) := Deny ;
    next (decision) := case
        USER = User1 & ATTR = Attribute1 : Permit;
        USER = User1 & ATTR = Attribute2 : Permit;
        USER = User2 & ATTR = Attribute1 : Permit;
        USER = User2 & ATTR = Attribute3 : Permit;
        USER = User3 & ATTR = Attribute2 : Permit;
        USER = User3 & ATTR = Attribute3 : Permit;
        USER = User4 & ATTR = Attribute1 : Permit;
        USER = User4 & ATTR = Attribute2 : Permit;
        USER = User5 & ATTR = Attribute2 : Permit;
        USER = User5 & ATTR = Attribute3 : Permit;
        1 : Deny;
    esac;
```

Subsequently, we define the set of specifications that are required to be verified on the transition system defined in Listing 1. We examine three different scenarios, where (i) we omit the policy introduced by group two; (ii) we omit the policy introduced by group one, and (iii) we consider the existence of both policies.

In the first scenario, we define two specifications in accordance with Formula 1. Therefore, making the assumption that $User1$ is absent, the CTL specifications that will verify the resilience of the examined policy are given in Listing 2. These specifications, when verified by the model checker, will provide a counterexample indicating which of the remaining users in group one (i.e. $User2, User3$) are in position to provide attributes $Attribute1$ and $Attribute2$, respectively. The verification of the given specifications and provision of counterexamples ensures the existence of resilience in the absence of $User1$. Specifically, it holds that $P = \{Attribute1, Attribute2, Attribute3\}$, $s = 1$, $d = 1$, and $t = 2$, with the latter stating that the set of users that together possess all permissions is equal to two.

In Listings 2 to 4, AG and AF are CTL expressions that are recognised by NuSMV as *'forall globally'* (i.e. $\forall \Box$) and *'forall finally'* (i.e. $\forall \Diamond$), respectively.

Listing 2: *Specification of resilience properties in CTL considering the absence of User1 and exclusion of the second group of users*

```
SPEC AG (( !(USER = User1) & (ATTR = Attribute1) &
           !(USER = User4) & !(USER = User5)
) -> AF decision =  Deny)
-- Evaluation: false, counterexample is provided

SPEC AG (( !(USER = User1) & (ATTR = Attribute2) &
           !(USER = User4) & !(USER = User5)
) -> AF decision =  Deny)
-- Evaluation: false, counterexample is provided
```

In the second scenario, we also define two specifications to verify the resilience of the examined policy in the absence of $User4$. In this case, the verification of the specifications provided in Listing 3 is evaluated as *true*. This is interpreted as: if $User4$ is absent then the remaining users (i.e. $User5$) do not collectively possess the required set of attributes to complete the critical task. No resilience is provided in this instance.

Busby, Jeremy; Gouglidis, Antonios; Hu, Chung Tong; Hutchison, David.
"Verification of Resilience Policies that Assist Attribute Based Access Control."
Paper presented at 2nd Workshop on Attribute Based Access Control (ABAC 2017), Scottsdale, AZ, United States. March 24, 2017 - March 24, 2017.

Listing 3: *Specification of resilience properties in CTL considering the absence of User4 and exclusion of the first group of users*

```
SPEC AG (( !(USER = User4) & (ATTR = Attribute1) &
           !(USER = User1) & !(USER = User2) &
           !(USER = User3)) -> AF decision =  Deny)
-- Evaluation: true

SPEC AG (( !(USER = User4) & (ATTR = Attribute2) &
           !(USER = User1) & !(USER = User2) &
           !(USER = User3)) -> AF decision =  Deny)
-- Evaluation: false, counterexample is provided
```

The third scenario under examination is presented in Listing 4. In this scenario, we assume the existence of both groups of users, and examine the resilience provided by the policy in the absence of *User*5. This is possible by omitting $\bigwedge_{1 \leq i \leq k} !sub_k$ in Formula 1. The evaluation of the defined specifications result in providing a countermeasure in both cases, and thus indicates the resilience of the policy. Specifically, in this instance, it holds that $P = \{Attribute1, Attribute2, Attribute3\}$, $s = 1$, $d = 2$, and $t = \infty$, with the latter stating that the set of users that together possess all permissions can be of any size.

Listing 4: *Specification of a resilience property in CTL considering both group of users*

```
SPEC AG (( !(USER = User5) & (ATTR = Attribute2)
) -> AF decision =  Deny)
-- Evaluation: false, counterexample is provided

SPEC AG (( !(USER = User5) & (ATTR = Attribute3)
) -> AF decision =  Deny)
-- Evaluation: false, counterexample is provided
```

6. CONCLUSION

In this paper, we examined an automated method for the formal verification of resilience specifications in the context of a specific access control model. For this purpose, we provided a formal definition of an ABAC model based on the guidelines provided by NIST; specified resilience using propositional logic; and, formally verified resilience specifications in a set of ABAC policies. We anticipate the research presented here will provide an interesting insight towards the consideration of resilience properties in addition to that of security in access control. The level of usability provided by existing approaches could be perceived as being low since they do not offer an appropriate set of tools that can automate the verification process. This holds mostly because existing approaches are considering the development of resilience policies during the design phase of a system. On the contrary, we have proposed the use of model checking as a means for the verification of resilience specifications in a set of existing policies during the operational phase of a system. Finally, by means of an example we demonstrated the applicability and level of automation offered by a computer-aided method such as model checking in verifying formally the correct functioning of ABAC policies against resilience specifications.

In future, we aim to investigate jointly the concepts of security and resilience in access control, including the possibility of conflicts that may arise.

7. ACKNOWLEDGEMENTS

This work is sponsored by the European Union under Grant SEC-2013.2.5-4: Protection systems for utility networks - Capability Project, Project Number: 608090, Hybrid Risk Management for Utility Providers (HyRiM).

8. REFERENCES

[1] ANSI. Information technology - Next Generation Access Control - Functional Architecture, 2013.

[2] V. Atluri and J. Warner. Supporting conditional delegation in secure workflow management systems. In *Proceedings of the tenth ACM symposium on Access control models and technologies*, pages 49–58. ACM, 2005.

[3] C. Baier, J.-P. Katoen, et al. *Principles of model checking*. MIT press Cambridge, 2008.

[4] E. Bertino, E. Ferrari, and V. Atluri. The specification and enforcement of authorization constraints in workflow management systems. *ACM Transactions on Information and System Security (TISSEC)*, 2(1):65–104, 1999.

[5] BSI. BS 65000 - Guidance for organizational resilience, 2014.

[6] D. Cohen, J. Crampton, A. Gagarin, G. Gutin, and M. Jones. Iterative plan construction for the workflow satisfiability problem. *Journal of Artificial Intelligence Research*, 51:555–577, 2014.

[7] J. Crampton, G. Gutin, and D. Karapetyan. Valued workflow satisfiability problem. In *Proceedings of the 20th ACM Symposium on Access Control Models and Technologies*, pages 3–13. ACM, 2015.

[8] J. Crampton, G. Gutin, S. Pérennes, and R. Watrigant. A multivariate approach for checking resiliency in access control. *arXiv preprint arXiv:1604.01550*, 2016.

[9] J. Crampton, G. Gutin, and R. Watrigant. Resiliency policies in access control revisited. In *Proceedings of the 21st ACM on Symposium on Access Control Models and Technologies*, pages 101–111. ACM, 2016.

[10] D. Ferraiolo, D. R. Kuhn, and R. Chandramouli. *Role-based access control*. Artech House, 2003.

[11] L. Gong and X. Qian. Computational issues in secure interoperation. *Software Engineering, IEEE Transactions on*, 22(1):43–52, 1996.

[12] A. Gouglidis, I. Mavridis, and V. C. Hu. Security policy verification for multi-domains in cloud systems. *International Journal of Information Security*, 13(2):97–111, 2014.

[13] A. Gouglidis, S. N. Shirazi, S. Simpson, P. Smith, and D. Hutchison. A multi-level approach to resilience of critical infrastructures and services. In *Proc. 23rd International Conference on Telecommunications (ICT 2016)*. IEEE, 2016.

[14] V. C. Hu, D. Ferraiolo, R. Kuhn, A. Schnitzer, K. Sandlin, R. Miller, and K. Scarfone. Guide to attribute based access control (ABAC) definition and considerations. *NIST SP*, 800:162, 2014.

[15] V. C. Hu, D. R. Kuhn, and D. F. Ferraiolo. Attribute-based access control. *IEEE Computer*, 48(2):85–88, 2015.

Busby, Jeremy; Gouglidis, Antonios; Hu, Chung Tong; Hutchison, David.
"Verification of Resilience Policies that Assist Attribute Based Access Control."
Paper presented at 2nd Workshop on Attribute Based Access Control (ABAC 2017), Scottsdale, AZ, United States. March 24, 2017 - March 24, 2017.

[16] V. C. Hu and R. Kuhn. Access control policy verification. *IEEE Computer*, 49(12):80–83, Dec 2016.

[17] Intel Security. Protect critical infrastructure. Technical report, McAfee. Part of Intel Security, 2016.

[18] ISO. IEC 13568: 2002: Information technology–Z formal specification notation–syntax, type system and semantics, 2002.

[19] R. B. Krug. CTL vs. LTL. Presentation, May 2010.

[20] F. Laroussinie, N. Markey, and P. Schnoebelen. Model checking CTL+ and FCTL is hard. In *International Conference on Foundations of Software Science and Computation Structures*, pages 318–331. Springer, 2001.

[21] N. Li, Q. Wang, and M. Tripunitara. Resiliency policies in access control. *ACM Transactions on Information and System Security (TISSEC)*, 12(4):20, 2009.

[22] M. Majdalawieh, F. Parisi-Presicce, and R. Sandhu. RBAC model for SCADA. In *Innovative Algorithms and Techniques in Automation, Industrial Electronics and Telecommunications*, pages 329–335. Springer, 2007.

[23] A. C. Richard, H. A. Julia, D. C. Pamela, W. W. David, and R. Y. Lisa. CERT resilience management model, version 1.0 improving operational resilience processes. Technical report, Software Engineering Institute, 2010.

[24] R. S. Sandhu and P. Samarati. Access control: principle and practice. *Communications Magazine, IEEE*, 32(9):40–48, 1994.

[25] SAnToS Laboraroty. Specification patterns, Responce property pattern, 2012.

[26] B. Shafiq, J. B. Joshi, E. Bertino, and A. Ghafoor. Secure interoperation in a multidomain environment employing RBAC policies. *Knowledge and Data Engineering, IEEE Transactions on*, 17(11):1557–1577, 2005.

[27] Q. Wang and N. Li. Satisfiability and resiliency in workflow authorization systems. *ACM Transactions on Information and System Security (TISSEC)*, 13(4):40, 2010.

APPENDIX

A. IMPLEMENTATION OF A RESILIENCE POLICY

In this section, we implement an RBAC resilience policy using the proposed ABAC model. Roles are introduced as subject attributes, and role hierarchy is supported via the meta-attribute functionality. The seniority of a role is expressed by the name-value tuples as *(senior role, junior role)*, with the senior role inheriting all the permissions of the junior role. In Listing A.1, we implement the RBAC policy defined in [22]. The offered resilience of this policy is easy to understand when examining the values of the $ResiliencePolicy\langle P, s, d, t\rangle$ tuple, i.e. $s = 1$, $d = 1$, $t = 1$, and $P = \{$*((action, Start system), (objectid, SCADA_DEV_001)), ((action, Stop system), (objectid, SCADA_DEV_001)), ((action, Acknowledge alarm), (objectid, SCADA_DEV_001)), ((action, Disable alarm), (objectid, SCADA_DEV_001))*$\}$, as explained in Section 5.1. Function calls used in Listing A.1 are briefly explained in Appendix B.

Listing A.1: *Implement an RBAC policy*

```
// Add new operations
 AddOperation(action, Monitor any screen);
 AddOperation(action, Start system);
 AddOperation(action, Stop system);
 AddOperation(action, Acknowledge alarm);
 AddOperation(action, Disable alarm);
 AddOperation(action, Change set point);
 AddOperation(action, Can change graphics);
 AddOperation(action, See alarm logs);
 AddOperation(action, Change security codes);
 AddOperation(action, Configure graphics);
 AddOperation(action, Controller setting);
 AddOperation(action, Security codes);

// OPS includes the following tuples
 OPS = {(action, Monitor any screen),
        (action, Start system),
        (action, Stop system),
        (action, Acknowledge alarm),
        (action, Disable alarm),
        (action, Change set point),
        (action, Can change graphics),
        (action, See alarm logs),
        (action, Change security codes),
        (action, Configure graphics),
        (action, Controller setting),
        (action, Security codes)}

// Add new subjects
 AddSubject(userid, user1);
 AddSubject(userid, user2);

// SUB includes the following tuples
 SUB = {(userid, user1), (userid, user2)}

// Add roles as new attributes
 AddAttribute(subject, (role, Junior operator));
 AddAttribute(subject, (role, Senior operator));
 AddAttribute(subject, (role, Supervisor));
 AddAttribute(subject, (role, Technician));
 AddAttribute(subject, (role, Engineer));
 AddAttribute(subject, (role, Manager));

// ATTRS includes the following tuples
 ATTRS = {(role, Junior operator),
          (role, Senior operator),
          (role, Supervisor), (role, Technician),
          (role, Engineer), (role, Manager)}

// Assign users with roles
 AssignSubject((userid, user1), (role, Supervisor));
 AssignSubject((userid, user2), (role, Manager));

// Introduce hierarchy relations (senior, junior)
// using the meta-attribute functionality
 AssignSubject((role, Supervisor),
               (role, Senior operator));
 AssignSubject((role, Senior operator),
               (role, Junior operator));
 AssignSubject((role, Engineer), (role, Technician));
 AssignSubject((role, Manager), (role, Supervisor));
 AssignSubject((role, Manager), (role, Engineer));
```


Busby, Jeremy; Gouglidis, Antonios; Hu, Chung Tong; Hutchison, David.
"Verification of Resilience Policies that Assist Attribute Based Access Control."
Paper presented at 2nd Workshop on Attribute Based Access Control (ABAC 2017), Scottsdale, AZ, United States. March 24, 2017 - March 24, 2017.

```
// SATTR includes the following tuples
 SATTR = {((userid, user1), (role, Supervisor)),
          ((userid, user2), (role, Manager)),
          ((role, Manager),
           (role, Supervisor)),
          ((role, Supervisor),
           (role, Senior operator)),
          ((role, Senior operator),
           (role, Junior operator))}

// Add new objects
 AddObject(objectid, SCADA_DEV_001);

// OBS includes the following tuples
 OBS = { (objectid, SCADA_DEV_001) };

// Policy definition
 AddPolicy(({}, (role, Junior operator)),
           ((objectid, SCADA_DEV_001), {}), {},
           (action, Monitor any screen));
 AddPolicy(({}, (role, Senior operator)),
           ((objectid, SCADA_DEV_001), {}), {},
           (action, Start system));
 AddPolicy(({}, (role, Senior operator)),
           ((objectid, SCADA_DEV_001), {}), {},
           (action, Stop system));
 AddPolicy(({}, (role, Senior operator)),
           ((objectid, SCADA_DEV_001), {}), {},
           (action, Acknowledge alarm));
 AddPolicy(({}, (role, Supervisor)),
           ((objectid, SCADA_DEV_001), {}), {},
           (action, Disable alarm));
 AddPolicy(({}, (role, Supervisor)),
           ((objectid, SCADA_DEV_001), {}), {},
           (action, Change set point));
 AddPolicy(({}, (role, Technician)),
           ((objectid, SCADA_DEV_001), {}), {},
           (action, Can change graphics));
 AddPolicy(({}, (role, Technician)),
           ((objectid, SCADA_DEV_001), {}), {},
           (action, See alarm logs));
 AddPolicy(({}, (role, Technician)),
           ((objectid, SCADA_DEV_001), {}), {},
           (action, Change security codes));
 AddPolicy(({}, (role, Engineer)),
           ((objectid, SCADA_DEV_001), {}), {},
           (action, Configure graphics));
 AddPolicy(({}, (role, Engineer)),
           ((objectid, SCADA_DEV_001), {}), {},
           (action, Controller setting));
 AddPolicy(({}, (role, Engineer)),
           ((objectid, SCADA_DEV_001), {}), {},
           (action, Security codes));
```

B. ADMINISTRATIVE OPERATIONS AND REVIEW FUNCTIONS IN ABAC

Tables 1 and 2 include a list of administrative operations and review functions respectively of the proposed ABAC model. Although not all of them are demonstrated in this paper, we include them for completeness.

Busby, Jeremy; Gouglidis, Antonios; Hu, Chung Tong; Hutchison, David.
"Verification of Resilience Policies that Assist Attribute Based Access Control."
Paper presented at 2nd Workshop on Attribute Based Access Control (ABAC 2017), Scottsdale, AZ, United States. March 24, 2017 - March 24, 2017.

Table 1: Administrative operations in ABAC

Operation	Description
AddSubject (subject?: SUB)	Creates a new subject
DeleteSubject (subject?: SUB)	Deletes an existing subject from the ABAC database
AddObject (object?: OBJ)	Creates a new object
DeleteObject (object?: OBJ)	Deletes an existing object from the ABAC database
AddOperation (operation?: OPS)	Creates a new operation
DeleteOperation (operation?: OPS)	Deletes an existing operation from the ABAC database
AddAttribute (type?: subject | object, attr?: $ATTR_S$ | $ATTR_O$)	Creates a new attribute
ModifyAttribute (type?: subject | object, attr?: $ATTR_S$ | $ATTR_O$)	Modifies the value of an existing attribute
DeleteAttribute (type?: subject | object, attr?: $ATTR_S$ | $ATTR_O$)	Deletes an existing attribute from the $ATTR$ set
AddEnvironment (attr?: ENV)	Creates a new environment condition attribute
ModifyEnvironment (attr!: ENV, value?: VALUE)	Modifies the value of an existing environment attribute
DeleteEnvironment (attr?: ENV)	Deletes an existing attribute from the ENV set
AssignSubject (subject?: SUB, attr?: $ATTR_S$)	Assigns a subject to an attribute
DeassignSubject (subject?: SUB, attr?: $ATTR_S$)	Deassigns a subject from an attribute
AssignObject (object?: OBJ, attr?: $ATTR_O$)	Assigns an object to an attribute
DeassignObject (object?: OBJ, attr?: $ATTR_O$)	Deassigns an object from an attribute
AddPolicy (sattr?: SATTR, oattr?: OATTR, env?: ENV, ops?: OPS)	Adds an action to a subject to perform an operation on an object given the subject's and object's attribute values, and potential environment attribute values
DeletePolicy (sattr?: SATTR, oattr?: OATTR, env?: ENV, ops?: OPS)	Deletes the action from a subject to perform an operation on an object given the subject's and object's attribute values, and potential environment attribute values

Table 2: Administrative review functions in ABAC

Function	Description
AssignedSubjects ($attr?: ATTR_S$, result!: 2^{SUB})	Return the set of subjects assigned to an attribute
AssignedObjects ($attr?: ATTR_O$, result!: 2^{OBJ})	Return the set of objects assigned to an attribute
SubjectAttributes (sub?: SUB, result!: 2^{ATTR_S})	Return the set of attributes assigned to a subject
ObjectAttributes (obj?: OBJ, result!: 2^{ATTR_O})	Return the set of attributes assigned to an object
SubjectAllAttributes (sub?: SUB, result!: 2^{ATTR_S})	Return the set of all attributes a subject may be eligible for, including attributes inherited from potential hierarchies implemented using the meta-attribute functionality
ObjectAllAttributes (obj?: OBJ, result!: 2^{ATTR_O})	Return the set of all attributes an object may be eligible for, including attributes inherited from potential hierarchies implemented using the meta-attribute functionality

Busby, Jeremy; Gouglidis, Antonios; Hu, Chung Tong; Hutchison, David.
"Verification of Resilience Policies that Assist Attribute Based Access Control."
Paper presented at 2nd Workshop on Attribute Based Access Control (ABAC 2017), Scottsdale, AZ, United States. March 24, 2017 - March 24, 2017.

Impact of Sampling and Augmentation on Generalization Accuracy of Microscopy Image Segmentation Methods

Michael Majurski, Petre Manescu, Joe Chalfoun, Peter Bajcsy, and Mary Brady
National Institute of Standards and Technology
100 Bureau Drive, Gaithersburg, MD 20899
POC: peter.bajcsy@nist.gov

1. Motivation

Terascale microscopy imaging requires automated software-based measurements of objects found via segmentation. Convolutional Neural Network (CNN) models have become a popular and successful supervised segmentation method incorporating the domain expert knowledge via annotations. To alleviate the annotation effort, sampling and augmentation methods have been leveraged to generate the large numbers of representative examples required for CNN training. As CNNs frequently report very high accuracies, there is a need to understand the impact of sampling/augmentation methods and their parameters on the generalization accuracy of image segmentation to improve our confidence in the reported accuracy. The confidence problem is illustrated in Figure 1.

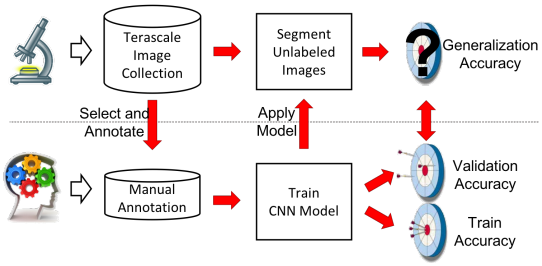


Figure 1 The confidence problem in segmentation generalization accuracy for terascale image collections

We approach the problem of estimating our confidence in CNN-based segmentation accuracy by performing several quantitative evaluations varying sampling and augmentation methods and their parameters over large image collections. The collections were acquired by time-lapse microscopy imaging of cell colonies. The ground truth segmentation was obtained by (a) using a special stain and a fluorescent imaging channel, and (b) segmenting generated high contrast images via thresholding.

2. Methodology

To deliver a sufficiently large number of representative samples for training complex CNN models with millions of parameters, one must analyze (1) the sampling method, (2) the sampling size (count), (3) the augmentation method, and (4) the augmentation parameters. Random sampling was chosen since it is the most frequently used method in the literature. Sample size range was selected based on its statistical relationship to estimation confidence spanning a 95 % confidence interval.

We classified the widely-used augmentation methods based on their types of transformations. Each image augmentation consisted of applying parametrized label-preserving transformations (e.g., affine, reflection, noise) to the annotated examples. To simplify the augmentation parametrization, we used at most one parameter per augmentation transformation, deriving the parameter range from the image data (transformation severity).

We used the Dice metric[1] for evaluating segmentation accuracy and focused primarily on the improvement in our accuracy confidence due to augmentation over the accuracy confidence provided by a selected sampling size.

3. Conclusions

We quantified the impact of sampling and augmentation models and their parametrization on the validation and generalization accuracies of CNN-based segmentation as the generalization error gaps (see the deltas in Figure 2) over 60 configurations of sampling size, augmentation model + parameter, and image object. We observed that training the CNN-based segmentation using rotation, reflection, and jitter lowered the generalization error gap the most (improved our accuracy confidence). We hypothesize that these quantitative results indicate that the augmentation configurations are closely mimicking the imaging variations.

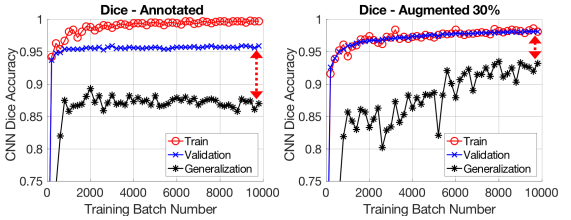


Figure 2: Constant vs. decreasing gap between train, validation, and generalization CNN segmentation accuracy.

[1] Measures spatial overlap between two segmentations. Dice, L. (1945) Measures of the amount of ecologic association between species. Ecology 26, 297–302

DISCLAIMER: Any mention of commercial products or reference to commercial organizations is for information only; it does not imply recommendation or endorsement by NIST nor does it imply that the products mentioned are necessarily the best available for the purpose.

Bajcsy, Peter; Brady, Mary; Chalfoun, Joe; Majurski, Michael; Manescu, Petru.
"Impact of Sampling and Augmentation on Generalization Accuracy of Microscopy Image Segmentation Methods."
Paper presented at Computer Vision for Microscopy Image Analysis (CVMI), Salt Lake City, UT, United States. June 18, 2018 - June 22, 2018.

Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds

Nawaf Alhebaishi[1,2], Lingyu Wang[1], Sushil Jajodia[3], and Anoop Singhal[4]

[1] Concordia Institute for Information Systems Engineering, Concordia University
[2] Faculty of Computing and Information Technology, King Abdulaziz University
{n_alheb,wang}@ciise.concordia.ca
[3] Center for Secure Information Systems, George Mason University
jajodia@gmu.edu
[4] Computer Security Division, National Institute of Standards and Technology
anoop.singhal@nist.gov

Abstract. As today's cloud providers strive to attract customers with better services and less downtime in a highly competitive market, they increasingly rely on remote administrators including those from third party providers for fulfilling regular maintenance tasks. In such a scenario, the privileges granted for remote administrators to complete their assigned tasks may allow an attacker with stolen credentials of an administrator, or a dishonest remote administrator, to pose severe insider threats to both the cloud tenants and provider. In this paper, we take the first step towards understanding and mitigating such a threat. Specifically, we model the maintenance task assignments and their corresponding security impact due to privilege escalation. We then mitigate such impact through optimizing the task assignments with respect to given constraints. The simulation results demonstrate the effectiveness of our solution in various situations.

1 Introduction

The widespread adoption of cloud leads to many unique challenges in terms of security and privacy [13]. As the cloud service market becomes more and more competitive, cloud providers are striving to attract customers with better services and less downtime at a lower cost. The search for an advantage in cost and efficiency will inevitably lead cloud providers to follow a similar path as what has been taken by their tenants, i.e., outsourcing cloud maintenance tasks to remote administrators including those from specialized third party maintenance providers [9]. Such an approach may also lead to many benefits due to resource sharing, e.g., the access to specialized and experienced domain experts, the flexibility (e.g., less need for full-time onsite staff), and the lower cost (due to the fact such remote administrators are shared among many clients).

However, such benefits come at an apparent cost in terms of increased security threats. Specifically, the remote administrators must be provided with necessary privileges, which may involve direct accesses to the underlying cloud infrastructure, in order to complete their assigned maintenance tasks. Armed with such privileges, a dishonest remote administrator, or an attacker with the stolen credentials of an administrator, can pose severe insider threats to both the cloud tenants (e.g., causing a large scale leak of

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

confidential user data) and the provider (e.g., disrupting the cloud services or abusing the cloud infrastructure for illegal activities) [12]. On the other hand, cloud providers are under the obligation to prevent such security or privacy breaches caused by insiders [14], either as part of the service level agreements, or to ensure compliance with security standards (e.g., ISO 27017 [19]). Therefore, there is a pressing need to better understand and mitigate such insider threats.

Dealing with the insider threat of remote administrators in clouds faces unique challenges. First, there is a lack of public access to the detailed information regarding cloud infrastructure configurations and typical maintenance tasks performed in clouds. Evidently, most existing works on insider attacks in clouds either stay at a high level or focus on individual nodes instead of the infrastructure [9, 20, 31] (a more detailed review of related work will be given in Section 6). Second, cloud infrastructures can be quite different from typical enterprise networks in terms of many aspects of security. For instance, multi-tenancy means there may co-exist different types of insiders with different privileges, such as administrators of a cloud tenant, those of the cloud provider, and third party remote administrators. Also, virtualization means a more complex attack surface consisting of not only physical nodes but also virtual or hypervisor layers. To the best of our knowledge, there is a lack of any concrete study in the literature on the insider attack of remote administrators in cloud data centers.

In this paper, we take the first step towards understanding and mitigating such insider threats. Specifically, we first model the maintenance tasks and their corresponding privileges. We then model the insider threats posed by remote administrators assigned to maintenance tasks by applying the existing k-zero day safety metric as follows; remote administrators possess elevated privileges due to the assigned maintenance tasks, and those privileges correspond to initially satisfied security conditions, which are normally only accessible by external attackers after exploiting certain vulnerabilities. Such model allows us to formulate the mitigation of the insider threats of remote administrators as an optimization problem and solve it using standard optimization techniques. We evaluate our approach through simulations and the results demonstrate the effectiveness of our solution under various situations. In summary, the main contribution of this paper is twofold:

- To the best of our knowledge, this is the first study on the insider threat of remote administrators in cloud infrastructures. As cloud providers leverage third parties for better efficiency and cost saving, our study demonstrates the need to also consider the security impact, and our model provides a way for quantitatively reasoning about the tradeoff between such security impact with other related factors.
- By formulating the optimization problem of mitigating the insider threat of remote administrators through optimal task assignments, we provide a relatively effective solution, as evidenced by our simulation results, for achieving the optimal tradeoff between security and other constraints using standard optimization techniques.

The remainder of this paper is organized as follows. Section 2 presents a motivating example and discusses maintenance tasks and privileges. In Section 3, we present our models of task assignment and insider threat. Section 4 formulates the optimization problem and discusses several use cases. Section 5 gives simulation results. Section 6 discusses related work. Section 7 concludes the paper.

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

2 Preliminaries

This section gives a motivating example and discusses maintenance tasks and privileges.

2.1 Motivating Example

A key challenge to studying security threats in cloud data centers is the lack of public accesses to detailed information regarding hardware and software configurations deployed in real cloud data centers. Existing work mainly focus on either high level frameworks and guidelines for risk and impact assessment [1, 27, 21], or specific vulnerabilities or threats in clouds [15, 29], with a clear gap between the two. To overcome such a limitation, we choose to devise our own fictitious, but realistic cloud data center designs, by piecing together publicly available information gathered from various cloud vendors and providers [5], as shown in Figure 1.

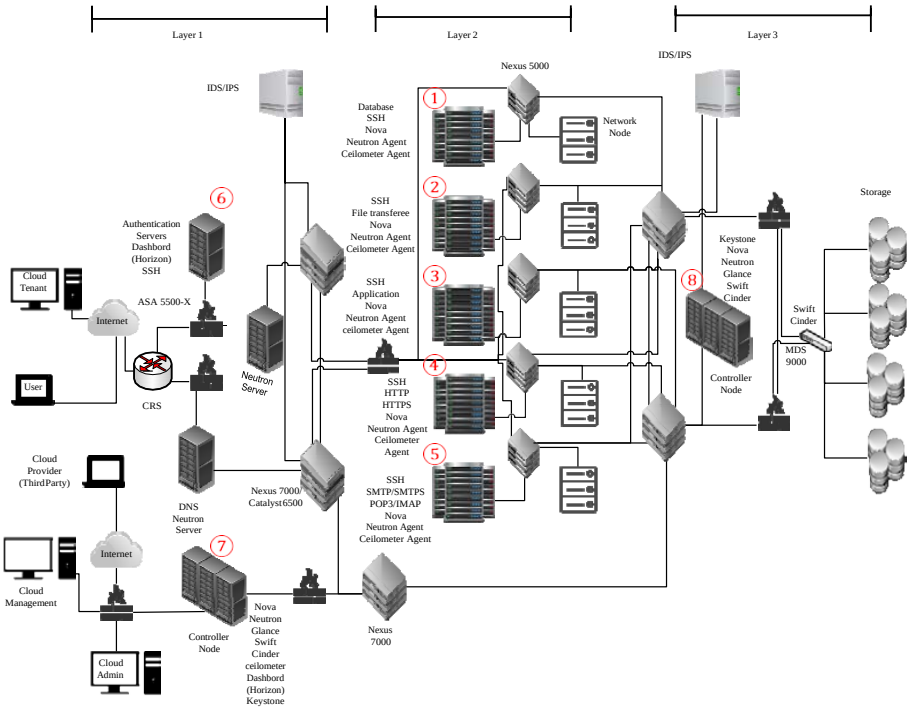


Fig. 1: An example of cloud data center

To make our design more representative, we devise this configuration based on concepts and practices borrowed from major cloud vendors and providers. For example,

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

we borrow the multi-layer concept and some hardware components, e.g., Carrier Routing System (CRS), Nexus (7000,5000,2000), Catalyst 6500, and MDS 9000, from the cloud data center design of Cisco [7]. We synthesize various concepts of the VMware vSphere [18] for main functionality of hardware components in our cloud infrastructure (e.g., authentication servers, DNS, and SAN). We also assume the cloud employs OpenStack as its operating system [24]. The infrastructure provides accesses to both cloud users and remote administrators through the three layer design. Layer 1 connects the cloud to the internet and includes the authentication servers, DNS, and Neutron Server. Layer 2 includes the rack servers and compute nodes. Layer 3 includes the storage servers. OpenStack components run on the authentication servers, DNS server (a Neutron component provides address translation to machines running the requested services), and compute nodes (Nova to host and manage VMs, Neutron to connect VMs to the network, and Ceilometer to calculate the usage) to provide cloud services.

Such a cloud data center may require many maintenance tasks to be routinely performed to ensure the normal operation of the hardware and software components. Such maintenance tasks may be performed by both internal staff working onsite and remote administrators, including those from specialized third party providers. In our example, assume the cloud provider decides to rely on third party remote administrators for the regular maintenance of the five compute nodes (nodes #1-5 in Figure 1), the authentication servers (node #6), and the two controllers (nodes #7 and 8). Table 1, shows the maintenance tasks need to be performed on those nodes. For simplicity, we only consider three types of tasks here (more discussions about maintenance tasks will be given in next section).

Node Number (in Figure 1)	Maintenance Tasks		
	Read log files	Modify configuration files	Install a new system
1	×	×	
2	×		×
3	×	×	×
4		×	×
5	×		×
6	×	×	
7	×		
8	×		

Table 1: An example of required maintenance tasks

In such a scenario, the cloud provider would naturally raise security concerns due to the fact that necessary privileges must be granted in order to allow the third party remote administrators to perform their assigned maintenance tasks. For instance, the task *read log files* needs certain read privilege to be granted, whereas modifying configuration files and installing a new system would demand much higher levels of privileges. Such privileges may allow a dishonest remote administrator, or attackers with stolen credentials of a remote administrator, to launch an insider attack and cause significant damage to the cloud provider and its tenants. Even though the cloud provider may (to some extent) trust the third party maintenance provider as an organization, it is in its best interest to understand and mitigate such threats from individual administrators.

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

However, as demonstrated by this example, there are many challenges in modeling and mitigating such insider threats.

- First, as demonstrated in Table 1, there may exist complex relationships between maintenance tasks and corresponding privileges needed to fulfill such tasks, and also relationships between different privileges (e.g., a root privilege implies many other privileges). Those relationships will determine the extent of an insider threat.
- Second, the insider threat will also depend on which nodes in the cloud infrastructure are involved in the assigned tasks, e.g., an insider with privileges on the authentication servers (node #6 in Figure 1) or on the compute nodes (nodes #1-5) may have very different security implications.
- Third, the extent of the threat also depends on the configuration (e.g., the connectivity and firewalls), e.g., an insider having access to the controller node #8 would have a much better chance to compromise the storage servers than one with access to the other controller node #7).
- Finally, while an obvious way to mitigate the insider threat is through assigning less tasks to each remote administrator such as to limit his/her privileges, our study will show that the effectiveness of such an approach depends on other factors and constraints, e.g., the amount of tasks to be assigned, the number of available remote administrators, constraints like each administrator may only be assigned to a limited number of tasks due to availability, or a subset of tasks due to his/her skill set, etc.

Clearly, how to model and mitigate the insider threat may not be straightforward even for such a simplified example (we will give the solution for this example scenario in Section 4.2), and the scenario might become far more complex in practice than the one demonstrated here. The remainder of the paper will tackle those challenges.

2.2 Remote Administrators, Maintenance Tasks, and Privileges

A cloud provider may hire different types of administrators to perform maintenance tasks onsite or through remote accesses [9]. First, *hardware administrators* have physical access to the cloud data center to perform maintenance on the physical components. Second, *security team administrators* are responsible for maintaining the cloud security policies. Third, *remote administrators* (RAs) perform maintenance tasks on certain nodes inside the infrastructure. The first two types can be considered relatively more trustworthy due to their limited quantity and the fact they work onsite, and directly for the cloud provider. The last type is usually considered riskier due to two facts, i.e., they work through remote access which is susceptible to attacks (e.g., via stolen credentials), and they may be subcontracted through third party companies which means less control by the cloud provider. In this paper, we focus on such remote administrators (RAs), even though our models and mitigation solution may equally work for dealing with other types of users if necessary.

There exists only limited public information about the exact maintenance tasks performed at major cloud providers. We have collected such information from various sources, and our findings are summarized on the left-hand side of Table 2, which shows sample maintenance tasks mentioned by Amazon Web Service [2], Google Cloud [3],

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

and Microsoft Azure [4]. As to privileges required for typical maintenance tasks, Bleikertz et al. provided five sample privileges required for maintaining the compute nodes in clouds [9], which we will borrow for our further discussions, as shown on the right-hand side of Table 2.

Maintenance Task	AWS [2]	GCP [3]	Azure [4]	Privilege	Restriction
Review Logs	×	×	×	No privilege	No access
Hard Disk Scan		×	×	Read	Cannot read VM-related data
Update Firmware	×	×	×	Write_L1	The restriction of read privilege applies, software modification restricted to trusted repository
Patch Operating System	×	×	×		
Update Operating System	×	×	×		
System Backup	×	×	×	Write_L2	Bootloader, kernel, policy enforcement, maintenance agent, file system snapshots, package manager transaction logs, and certain dangerous system parameters
Upgrades System	×	×	×		
Maintain Automated Snapshots	×				
Bug Fix	×	×	×		
Update Kernel	×	×		Write_L3	No restriction

Table 2: Maintenance tasks in popular cloud platforms (left) and the privileges (right)

To simplify our discussions, our running example will be limited to ten maintenance tasks on three compute nodes with corresponding privileges on such nodes, as shown in Table 3. Later in Section 4.2, we will expand the scope to discuss the solution for our motivating example which involves all the eight nodes.

Task Number	Node Number (in Figure 1)	Task Description	Privilege
1	4 (*http*)	Read log files for monitoring	Read
2	4 (*http*)	Modifying configuration files	Write L1
3	4 (*http*)	Patching system files	Write L3
4	3 (*app*)	Read log files for monitoring	Read
5	3 (*app*)	Modifying configuration files	Write L1
6	3 (*app*)	Update kernel	Write L3
7	1 (*DB*)	Read log files for monitoring	Read
8	1 (*DB*)	Modifying configuration files	Write L1
9	1 (*DB*)	Update kernel	Write L3
10	1 (*DB*)	Install new systems	Write L2

Table 3: Maintenance tasks and privileges for the running example

3 Models

This section presnts out threat model and models of the maintenance task assignment and insider threat.

3.1 Threat Model and Maintenance Task Assignment Model

Our work is intended to assist the cloud provider in understanding and mitigating the insider threat from dishonest remote administrators or attackers with stolen credentials of a remote administrator. To this end, we assume the majority of remote administrators is trusted, and if there are multiple dishonest administrators (or attackers with their credentials), they do not collude (a straightfoward extension of our models by considering

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

each possible combination of administrators as one insider can accommodate such colluding administrators, which is considered as future work). We assume the third party provider is trusted as an organization and will collaborate with the cloud provider to implement the intended task assignment. We assume the cloud provider is concerned about certain critical assets inside the cloud, and it is aware of the constraints about task assignments such as the number of remote administrators, their availability and skill set, etc. Finally, as a preventive solution, our mitigation approach is intended as a complementary solution to existing vulnerability scanners, intrusion detection systems, and other solutions for mitigating insider threats.

The cloud provider assigns the maintenance tasks to remote administrators (RAs) based on given constraints (e.g., which tasks may be assigned each RA), and consequently the RA will obtain privileges required by those tasks. This can be modeled as follows (which has a similar syntax as [26]).

Definition 1 (Maintenance Task Assignment Model). *Given*

- *a set of remote administrators RA,*
- *a set of maintenance task T,*
- *a set of privileges P,*
- *the remote administrator task relation* $RAT \subseteq RA \times T$ *which indicates the maintenance tasks that are allowed to be assigned to each remote administrator, and*
- *the task privilege relation* $TP \subseteq T \times P$ *which indicates the privileges required for each task,*

a maintenance task assignment is given by function $ta(.)\colon RA \rightarrow 2^T$ *that satisfies* $(\forall ra \in RA)(ta(ra) \subseteq \{t \mid (ra, t) \in RAT\})$ *(meaning a remote administrator is only assigned with the tasks to which he/she is allowed), and the corresponding set of privileges given to the remote administrator is given by function* $pa(ra) = \bigcup_{t \in ta(ra)} \{p \mid (t, p) \in TP\}$.

3.2 Insider Threat Model

We given an overview of our model for the insider threat, which will be demonstrated through an example shown in Figure 2. First, we borrow the resource graph concept [30] to represent the causal relationships between different resources inside the given cloud configuration. Second, we map the privileges given to RAs through maintenance task assignments (Definition 1) to exploits of corresponding resources in the resource graph. Third, we apply the *k*-zero day safety metric [32] to quantify the insider threat of each RA through his/her *k* value. Finally, we take the average (and minimum) of all RAs' *k* values as the average (and worst) case indication of insider threat.

Figure 2 shows an example resource graph for our running example (the dashed lines and shades can be ignored and will be discussed later in Section 4.2; also, only a small portion of the resource graph is shown here due to space limitations). Each triplet inside an oval indicates a potential zero day or known exploit in the format <service or vulnerability, source host, destination host> (e.g. <Xen, RA, 4> indicates an exploit on Xen), and the plaintext pairs indicate the pre- or post-conditions of those exploits in the format <condition, host> where condition can be either a privilege on the host (e.g., <W1,4> means the level 1 write privilege and <R,4> means the read privilege

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

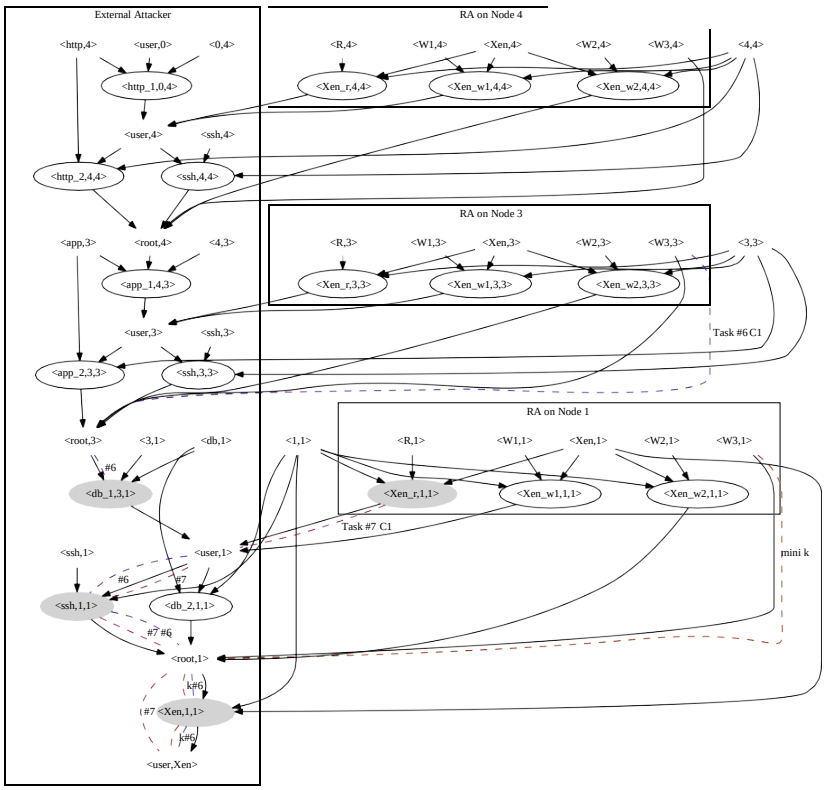


Fig. 2: Modeling insider threat using the resource graph

which are both explained in Section 2.2), the existence of a service on the host (e.g., <Xen,4>), or a connectivity (e.g., <0,4>means attacker can connect to host 4 and <4,4> means a local exploit on host 4). The edges point from pre-conditions to an exploit and then to its post-conditions, which indicate that any exploit can be executed if and only if all of its pre-conditions are satisfied, whereas executing an exploit is enough to satisfy all its post-conditions.

In Figure 2, the left-hand side box indicates the normal resource graph which depicts what an external attacker may do to compromise the critical asset <user, Xen>. The right-hand side boxes depict the insider threats coming from RAs assigned to each of the three compute nodes. The gray color exploits are what captures the consequences of granting privileges to remote administrators. For example, an RA with the level 1 write privilege <W1,4> can potentially exploit Xen (i.e., <Xen_w1,4,4>) to escalate his/her privilege to the user privilege on host 4 (i.e.., <user,4>), whereas a higher level privilege <W2,4> can potentially lead to the root privilege <root,4> through an exploit <Xen _w2,4,4>, and the highest privilege <W3,4> can even directly lead to that priv-

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

ilege. Those examples show how the model can capture the different levels of insider threats as results of different privileges obtained through maintenance task assignments.

Next, given the maintenance task assignment for each RA, we can obtain all the possible paths he/she may follow in the resource graph, starting from all the initially satisfied conditions (e.g., <Xen,4>) and those implied by the task assignment (e.g., <W1,4>) to the critical asset (i.e., <user,Xen>). To quantify the relative level of such threats, we apply the *k*-zero day safety metric (*k*0d) [32] which basically counts the number of zero day exploits (known exploits are not counted, and exploits of the same service are only counted once) along the shortest path. The metric value of each RA provides an estimation for the relative level of threat of each RA, since a larger number of distinct zero day exploits on the shortest path means reaching the critical asset is (exponentially, if those exploits are assumed to be independent) more difficult. For example, an RA with privilege <W3,1> would have a *k*0d value of 1 since only one zero day exploit <Xen,1,1> is needed to reach the critical asset, whereas an RA with <W2,1> would have a *k* value of 2 since an additional exploit <Xen_w2,1,1> is needed. Finally, once we have calculated the *k* values of all RAs based on their given maintenance task assignments, we take the average (and minimum) of those *k* values as the average (and worst) case indication of the overall insider threat of the given maintenance task assignments. The above discussions are formally defined as follows.

Definition 2 (Insider Threat Model). *Given the maintenance task assignment (i.e., RA, T , P, RAT, TP, ta, and pa, as given in Definition 1) let* $C_r = \bigcup_{ra \in RA} pa(ra)$ *be the set of privileges implied by the assignment and* E_r *be the set of new exploits enabled by* C_r*. Denote by* $G(E \cup E_r, C \cup C_r, R)$ *the resource graph (where E and C denote the original set of exploits and conditions, respectively, and R denote the edges) and let* $k0d(.)$ *be the k zero day safety metric function. We say* $k0d(ra)$, $\frac{\sum_{ra \in RA} k0d(ra)}{|RA|}$, *and* $min(\{k0d(ra) : ra \in RA\})$ *represent the insider threat of* ra*, the average case insider threat of the maintenance task assignment, and the worst case insider threat of the maintenance task assignment, respectively.*

4 The Mitigation

In this section, we formulate the optimization-based solution for mitigate the insider threat during maintenance task assignment and discuss several use cases.

4.1 Optimization-based Mitigation

Based on our definitions of the maintenance task assignment model and the insider threat model, we can define the problem of optimal task assignment as follows. Note the remote administrator task relation RAT basically gives the constraints for optimization since it states which tasks may be assigned to which RA (in some cases the constraints may also be modeled differently for convenience, e.g., as the maximum number of tasks for each RA).

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

Definition 3 (The Optimal task assignment problem). *Given a resource graph G, the remote administrators RA, maintenance tasks T , privileges P, the remote administrator task relation RAT, and the task privilege relation TP, find a maintenance task assignment function ta which maximizes the insider threat* $\frac{\sum_{ra \in RA} k0d(ra)}{|RA|}$ *(or* $min(\{k0d(ra) : ra \in RA\})$*).*

Theorem 1. *The Optimal task assignment problem (Definition 3) is NP-hard.*

Proof: First, calculating the $k0d$ function is already NP-hard w.r.t. the size of the resource graph [32]. On the other hand, we provide a sketch of proof to show the problem is also NP-hard from the perspective of the maintenance task assignment. Specifically, given any instance of the well known NP-complete problem, exact cover by 3-sets (i.e., given a finite set X containing exactly $3n$ elements, and a collection C of subsets of X each of which contains exactly 3 elements, determine whether there exists $D \subseteq C$ such that every $x \in X$ occurs in exactly one $d \in D$), we can construct an instance of our problem as follows. We use X for the set of maintenance tasks, and C for the set of RAs, such that the three elements of each $c \in C$ represent three tasks which can be assigned to c. In addition, no RA can be assigned with less than three tasks, and an RA already assigned with three tasks can choose any available task to be assigned in addition. We can then construct a resource graph in which the critical asset can be reached through any combination of four privileges. It then follows that, the insider threat is maximized if and only if there exists an exact cover D due to the following. If the exact cover exists, then every RA $d \in D$ is assigned with exactly three tasks and therefore the k value of every RA, and hence the insider threat, will be equal to infinity since the critical asset cannot be reached with less than four privileges; if the cover does not exist, then to have every task assigned, we will have to assign at least one RA with more than three tasks, and hence the k value will decrease. □

In our study, we use the genetic algorithm to optimize the maintenance task assignments by maximizing k. Specifically, the resource graph is taken as input to the optimization algorithm, with the (either average case or worst case) insider threat value k as the fitness function. We try to find the best task assignment for maximizing the value k within a reasonable number of generations. The constraints can be given either through defining the remote administrator task relation RAT in the case of specific tasks that can be assigned to each RA, or as a fixed number of tasks for each RA. Other constraints can also be easily applied to the optimization algorithm. In our simulations, we choose the probability of 0.8 for crossover and 0.2 for mutation based on our experiences.

4.2 Use Cases

We demonstrate our solution through several use cases with different constraints. The first three use cases are based on the five remote administrators and ten maintenance tasks presented in Table 3 and the last use case is based on the motivating example shown in Section 2.1.

– Use Case A: In this case, each RA should be assigned with two tasks. The three tables shown in Table 4 show three possible assignments and the corresponding k values. Also, Figure 2 shows an example path (dashed lines) for tasks assigned to

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

User	A_1	B_1	C_1	D_1	E_1
Tasks Number	4	5	6	8	9
	1	10	7	3	2
$\underline{k}$	3	1	2	2	1
$\bar{k}$	1.8				
Minimum k	1				

User	A_2	B_2	C_2	D_2	E_2
Tasks Number	6	4	7	8	5
	9	3	10	1	2
$\underline{k}$	1	3	1	2	3
$\bar{k}$	2				
Minimum k	1				

User	A_3	B_3	C_3	D_3	E_3
Tasks Number	4	5	6	8	9
	1	2	7	3	10
$\underline{k}$	3	3	2	2	1
$\bar{k}$	2.2				
Minimum k	1				

Table 4: Maintenance tasks assignments for use case A

RA C_1 based on the top table, and also the shortest path yielding the minimum k value. We use the GA to find the optimal task assignment that meets the constraint given in this case, as shown in the last table, the maximal average of k values among all RAs is $\bar{k} = 2.2$. It can also be seen that the minimum k value among all RAs is always $k = 1$ in this special case.

- Use Case B: In this case, each RA should be assigned with at least one task. The optimal task assignment under this constraint is (RA1: 8,9,10 , RA2: 4,5 , RA3: 3, RA4: 1,2 and RA5: 6,7). This relaxed constraint improves the average of k from 2.2 in the previous example to 2.8, which shows relaxing the constraint may increase k (which means less threat).
- Use Case C: In this case, each RA can handle a fixed subset of tasks. In our example, we assume RA1 can be assigned to any task requiring the read privilege, RA2 to tasks requiring write level 1 privilege, RA3 to tasks requiring write level 1 and 2, RA4 to tasks requiring write level 3, and RA5 can be assigned to any task. After applying our solution, the optimal assignment yields the maximal average of k values to be $\bar{k} = 2.2$.
- Use Case D: This case shows the optimal maintenance task assignment for tasks discussed in our motivating example in Section 2.1. We have eight RAs and each RA can handle maximum two tasks. The upper table in Table 5 shows the 15 maintenance tasks to be assigned. In Table 5, the four tables on the bottom show four different tasks scenarios assigned to RAs and each table shows different average k. The bottom table on the right side shows the optimal task assignment in term of the average $\bar{k} = 3.125$.

5 Simulations

This section shows simulation results on applying our mitigation solution under various constraints. All simulations are performed using a virtual machine equipped with a 3.4 GHz CPU and 4GB RAM in the Python 2.7.10 environment under Ubuntu 12.04 LTS and the MATLAB R2017bs GA toolbox. To generate a large number of resource graphs for simulations, we start with seed graphs with realistic configurations similar to Figure 1 and then generate random resource graphs by injecting new nodes and edges into those seed graphs. Those resource graphs were used as the input to the optimization toolbox where the fitness function is to maximize the average or worst case insider threat value k (given in Definition 2) with various constraints, e.g., the number of available RAs and maintenance tasks and how many task may be assigned to each RA. We repeat each simulation on 300 different resource graphs to obtain the average result.

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

Task#	Maintenance task
1	Read log files for node 1
3	Read log files for node 2
5	Read log files for node 3
7	Install a new system for node 3
9	Install a new system for node 4
11	Install a new system for node 5
13	Modify configuration file for node 6
15	Read log files for node 8

Task#	Maintenance task
2	Modify configuration file for node 1
4	Install a new system for node 2
6	Modify configuration file for node 3
8	Modify configuration file for node 4
10	Read log files for node 5
12	Read log files for node 6
14	Read log files for node 7

User	RA1	RA2	RA3	RA4	RA5	RA6	RA7	RA8
Tasks Number	14	1	4	8	2	3	7	6
	5	9	15	12	10	11	13	
k	1	3	2	3	2	3	2	3
$\overline{k}$	2.375							
Minimum k	1							

User	RA1	RA2	RA3	RA4	RA5	RA6	RA7	RA8
Tasks Number	1	2	3	4	5	6	7	8
	9	10	11	12	13	14	15	
k	3	2	3	3	3	1	2	5
$\overline{k}$	2.75							
Minimum k	1							

User	RA1	RA2	RA3	RA4	RA5	RA6	RA7	RA8
Tasks Number	1	2	3	5	6	15	13	8
	4	7	9	10	11	12	14	
k	3	2	4	4	3	2	1	5
$\overline{k}$	3							
Minimum k	1							

User	RA1	RA2	RA3	RA4	RA5	RA6	RA7	RA8
Tasks Number	1	2	3	5	6	14	4	8
	12	7	9	10	11	15	13	
k	3	2	4	4	3	1	3	5
$\overline{k}$	3.125							
Minimum k	1							

Table 5: Maintenance task assignments for use case D (the motivating example)

The objective of the first two simulations is to study how the average case insider threat (i.e., the average of k values among all RAs) may be improved through our mitigation solution under constraints about the number of tasks and RAs, respectively. In Figure 3, the number of available RAs is fixed at 500, while the number of maintenance tasks is varied between 500 and 2,000 along the X-axis. The Y -axis shows the average of k among all RAs. The solid lines represent the results after applying our mitigation solution under constraints about the maximum number of tasks assigned to each RA. The dashed lines represent the results before applying the mitigation solution.

Results and Implications: From the result, we can make following observations. First, the mitigation solution successfully reduces the insider threat (increasing the average of k values) in all cases. Second, the results before and after applying the solution decrease (meaning increased insider threat) following similar linear trends, as the number of maintenance tasks increases until each RA reaches its full capacity. Finally, the result of maximum four tasks per RA after applying the solution is close to the result of maximum ten tasks per RA before applying the solution, which means the mitigation solution may allow more (more than double) tasks to be assigned to the same number of RAs while yielding the same level of insider threat.

In Figure 4, the number of maintenance tasks is fixed at 2,500 while the number of RAs is varied between 400 and 1,000 along the X-axis. The Y -axis shows the average of k among all RAs. The solid lines represent the results after applying the mitigation solution and the dashed lines for the results before applying the solution. All the lines start with sufficient numbers of RAs for handling all the tasks since we only consider one round of assignment. We apply the same constraint as in previous simulation.

Results and Implications: Again we can see the mitigation solution successfully reduces the insider threat (increasing the average of k values) in all cases. More interestingly, we can observe the trend of the lines as follows. The dashed lines all follow a similar near linear trend, which is expected since a larger number of RAs means less insider threat since each RA will be assigned less tasks and hence given less privileges.

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

On the other hand, most of the solid lines follow a similar trend of starting flat then increasing almost linearly before reaching the plateau. This trend indicates that, the mitigation solution can significantly reduce the insider threat when the number of RAs is within certain ranges past which it becomes less effective (because each RA already receives minimum privileges). The trend of 4 tasks per RA is slightly different mostly due to the limited number of RAs.

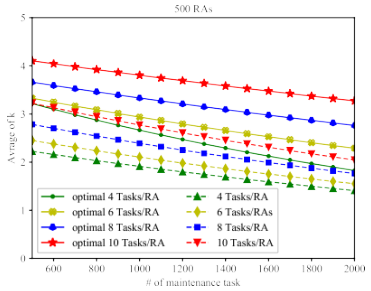


Fig. 3: Average of k among 500 RAs before and after applying the mitigation solution

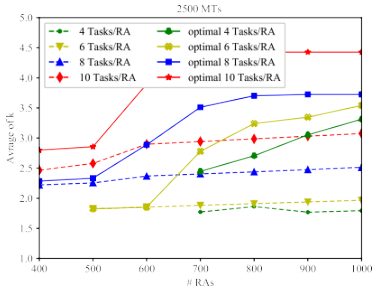


Fig. 4: Average of k among different number of RAs before and after the solution

The objective of the next two simulations is to study how the worst case insider threat (i.e., the minimum k values among all RAs) behaves under the mitigation solution. Figure 5 and Figure 6 are based on similar X-axis and constraints as previous two simulations, whereas the Y -axis shows the minimum k among all RAs (averaged over 300 simulations).

Results and Implications: In Figure 5, we can see that the minimum k values also decrease (meaning more insider threat) almost linearly as the number of tasks increases. In contrast to previous simulation, we can see the minimum k values are always lower than the average k values, which is expected. In Figure 6, we can see the minimum k values also increase almost linearly before reaching the plateau as the number of RAs increases. In contrast to previous simulation, we can see the increase here is slower, which means the worst case results (minimum k values) are more difficult to improve with a increased number of RAs. Also, we can see that the worst case results reach the plateau later (e.g., 900 RAs for 8 tasks per RA) than the average case results (700 RAs).

6 Related Work

The insider threat is a challenging issue for both traditional networks and clouds. Ray and Poolsapassit proposed an alarm system to monitor the behavior of malicious insiders using the attack tree [25]. Mathew et al. used the capability acquisition graphs (CAG) to monitor the abuse of privileges by malicious insiders [23]. Sarkar et al. proposed DASAI to analyze if a process contains a step that meet the insider attack condi-

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

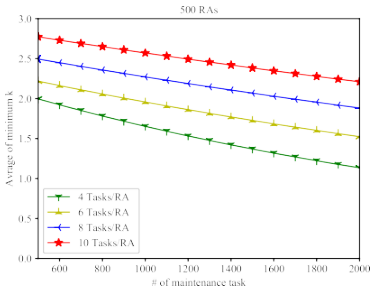


Fig. 5: Minimum k for 500 RAs

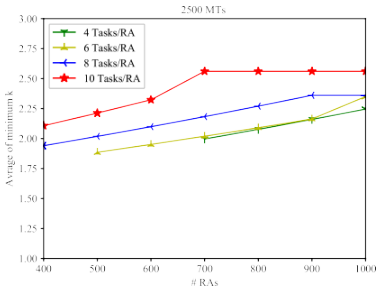


Fig. 6: Minimum k for varying # of RAs

tion [28]. Chinchani et al. proposed a graph-based model for insider attacks and measure the threat [11]. Althebyan and Panda proposed predication and detection model for insider attacks based on knowledge gathered by the internal users during work time in the organization [6]. Bishop et al. presented insider threat definition based on security policies and determine source of risk [8].

There is lack of work focusing on the cloud security metrics in general and for insider attacks especially. Our previous work focus on applying threat modeling to cloud data center infrastructures with a focus on external attackers [5]. Gruschka and Jensen devise a high level attack surface framework to show from where the attack can start [16]. The NIST emphasizes the importance of security measuring and metrics for cloud providers in [1]. A framework is propose by Luna et al. for cloud security metrics using basic building blocks [22].

Besides threat modeling, mitigating insider attackers in clouds is also a challenging task. There are many works discuss securing the cloud from insider attack by limiting the trust on the compute node [31]. Li et al. focuses on supporting users to configure privacy protection in compute node [20]. Closest to our work, Bleikertz et al. focus on securing the cloud during maintenance time by limiting the privilege grant to the remote administrator based on the tasks assigned to that administrator [9]. We borrow their categorization of the privileges. Our mitigation approach is also inspired by the network hardening approaches using genetic algorithms [17, 10].

7 Conclusion

In this paper, we have modeled the insider threat during maintenance task assignment for cloud providers to better understand such threat posed by third party remote administrators, and we have formulated the optimal assignment problem as an optimization problem and applied standard optimization algorithm to derive a solution under different constraints. We have also conducted simulations whose results show our solution can significantly reduce the insider threat of remote administrators. Our future work will focus on following directions. First, we will improve our solution to handle more realistic scenarios, e.g., incremental assignment for streams of new maintenance tasks, and handling dynamics (joining or leaving) of RAs, giving priority or weight to tasks.

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

Second, we will consider explicit cost models for assignments and incorporate the cost into the mitigation solution, e.g., based on the number of RAs, the amount or duration of tasks, and privileges needed.

Disclaimer Commercial products are identified in order to adequately specify certain procedures. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the identified products are necessarily the best available for the purpose.

References

1. National Institute of Standards and Technology: Cloud Computing Service Metrics Description. http://www.nist.gov/itl/cloud/upload/RATAX-CloudServiceMetricsDescription-DRAFT-20141111.pdf, 2015. [Online; accessed 17/06/2015].
2. Amazon Web Services. https://aws.amazon.com/, 2018. [Online; accessed 28/02/2018].
3. Google Cloud Platform. https://cloud.google.com/, 2018. [Online; accessed 28/02/2018].
4. Microsoft Azure. https://azure.microsoft.com, 2018. [Online; accessed 28/02/2018].
5. N. Alhebaishi, L. Wang, S. Jajodia, and A. Singhal. Threat modeling for cloud data center infrastructures. In *Foundations and Practice of Security - 9th International Symposium, FPS 2016, Québec City, QC, Canada, October 24-25, 2016, Revised Selected Papers*, pages 302–319, 2016.
6. Q. Althebyan and B. Panda. A knowledge-base model for insider threat prediction. In *2007 IEEE SMC Information Assurance and Security Workshop*, pages 239–246, June 2007.
7. K. Bakshi. Cisco cloud computing-data center strategy, architecture, and solutions. *DOI= http://www. cisco. com/web/strategy/docs/gov/CiscoCloudComputing_WP. pdf*, 2009.
8. M. Bishop, S. Engle, S. Peisert, S. Whalen, and C. Gates. We have met the enemy and he is us. In *Proceedings of the 2008 New Security Paradigms Workshop*, NSPW '08, pages 1–12, New York, NY, USA, 2008. ACM.
9. S. Bleikertz, A. Kurmus, Z. A. Nagy, and M. Schunter. Secure cloud maintenance: Protecting workloads against insider attacks. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, ASIACCS '12, pages 83–84, New York, NY, USA, 2012. ACM.
10. D. Borbor, L. Wang, S. Jajodia, and A. Singhal. Diversifying network services under cost constraints for better resilience against unknown attacks. In *Data and Applications Security and Privacy XXX - 30th Annual IFIP WG 11.3 Conference, DBSec 2016, Trento, Italy, July 18-20, 2016. Proceedings*, pages 295–312, 2016.
11. R. Chinchani, A. Iyer, H. Q. Ngo, and S. Upadhyaya. Towards a theory of insider threat assessment. In *2005 International Conference on Dependable Systems and Networks (DSN'05)*, pages 108–117, June 2005.
12. W. R. Claycomb and A. Nicoll. Insider threats to cloud computing: Directions for new research challenges. In *2012 IEEE 36th Annual Computer Software and Applications Conference*, pages 387–394, July 2012.
13. Cloud Security Alliance. Security guidance for critical areas of focus in cloud computing v 3.0, 2011.
14. Cloud Security Alliance. Top threats to cloud computing, 2018. Available at: https://cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf.

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

15. K. Dahbur, B. Mohammad, and A. B. Tarakji. A survey of risks, threats and vulnerabilities in cloud computing. In *Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications*, ISWSA '11, pages 12:1–12:6, New York, NY, USA, 2011. ACM.
16. N. Gruschka and M. Jensen. Attack surfaces: A taxonomy for attacks on cloud services. In *2010 IEEE 3rd international conference on cloud computing*, pages 276–279. IEEE, 2010.
17. M. Gupta, J. Rees, A. Chaturvedi, and J. Chi. Matching information security vulnerabilities to organizational security profiles: a genetic algorithm approach. *Decision Support Systems*, 41(3):592 – 603, 2006. Intelligence and security informatics.
18. M. Hany. VMware VSphere In The Enterprise. `http://www.hypervizor.com/diags/HyperViZor-Diags-VMW-vS4-Enterprise-v1-0.pdf`. [Online; accessed 05/02/2015].
19. ISO Std IEC. ISO 27017. *Information technology- Security techniques- Code of practice for information security controls based on ISO/IEC 27002 for cloud services (DRAFT)*, `http://www.iso27001security.com/html/27017.html`, 2012.
20. M. Li, W. Zang, K. Bai, M. Yu, and P. Liu. Mycloud: Supporting user-configured privacy protection in cloud computing. In *Proceedings of the 29th Annual Computer Security Applications Conference*, ACSAC '13, pages 59–68, New York, NY, USA, 2013. ACM.
21. J. Luna, H. Ghani, D. Germanus, and N. Suri. A security metrics framework for the cloud. In *Security and Cryptography (SECRYPT), 2011 Proceedings of the International Conference on*, pages 245–250, July 2011.
22. J. Luna, H. Ghani, D. Germanus, and N. Suri. A security metrics framework for the cloud. In *Security and Cryptography (SECRYPT), 2011 Proceedings of the International Conference on*, pages 245–250. IEEE, 2011.
23. S. Mathew, S. Upadhyaya, D. Ha, and H. Q. Ngo. Insider abuse comprehension through capability acquisition graphs. In *2008 11th International Conference on Information Fusion*, pages 1–8, June 2008.
24. Openstack. Openstack Operations Guide. `http://docs.openstack.org/openstack-ops/content/openstack-ops\_preface.html`. [Online; accessed 27/08/2015].
25. I. Ray and N. Poolsapassit. *Computer Security – ESORICS 2005: 10th European Symposium on Research in Computer Security, Milan, Italy, September 12-14, 2005. Proceedings*, chapter Using Attack Trees to Identify Malicious Attacks from Authorized Insiders, pages 231–246. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.
26. R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman. Role-based access control models. *Computer*, 29(2):38–47, Feb.1996.
27. P. Saripalli and B. Walters. Quirc: A quantitative impact and risk assessment framework for cloud security. In *2010 IEEE 3rd International Conference on Cloud Computing*, pages 280–288, July 2010.
28. A. Sarkar, S. Khler, S. Riddle, B. Ludaescher, and M. Bishop. Insider attack identification and prevention using a declarative approach. In *2014 IEEE Security and Privacy Workshops*, pages 265–276, May 2014.
29. F. B. Shaikh and S. Haider. Security threats in cloud computing. In *Internet Technology and Secured Transactions (ICITST), 2011 International Conference for*, pages 214–219, Dec 2011.
30. O. Sheyner, J. Haines, S. Jha, R. Lippmann, and J. M. Wing. Automated generation and analysis of attack graphs. In *Security and Privacy, 2002. Proceedings. 2002 IEEE Symposium on*, pages 273–284, 2002.
31. W. K. Sze, A. Srivastava, and R. Sekar. Hardening openstack cloud platforms against compute node compromises. In *Proceedings of the 11th ACM on Asia Conference on Computer*

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

and Communications Security, ASIA CCS '16, pages 341–352, New York, NY, USA, 2016. ACM.

32. L. Wang, S. Jajodia, A. Singhal, P. Cheng, and S. Noel. k-zero day safety: A network security metric for measuring the risk of unknown vulnerabilities. *IEEE Transactions on Dependable and Secure Computing*, 11(1):30–44, Jan 2014.

Alhebaishi, Nawaf; Jajodia, Sushil; Singhal, Anoop; Wang, Lingyu.
"Modeling and Mitigating the Insider Threat of Remote Administrators in Clouds."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs

Chen Cao[1], Lun-Pin Yuan[1], Anoop Singhal[2], Peng Liu[1], Xiaoyan Sun[3], and Sencun Zhu[1]

[1] The Pennsylvania State University
[2] National Institute of Standards and Technology
[3] California State University, Sacramento
caochen11@mails.ucas.ac.cn,lunpin@psu.edu,anoop.singhal@nist.gov, pliu@ist.psu.edu,xiaoyan.sun@csus.edu,szhu@cse.psu.edu

Abstract. Cyber-defense and cyber-resilience techniques sometimes fail in defeating cyber-attacks. One of the primary causes is the ineffectiveness of business process impact assessment in the enterprise network. In this paper, we propose a new business process impact assessment method, which measures the impact of an attack towards a business-process-support enterprise network and produces a numerical score for this impact. The key idea is that all attacks are performed by exploiting vulnerabilities in the enterprise network. So the impact scores for business processes are the function result of the severity of the vulnerabilities and the relations between vulnerabilities and business processes. This paper conducts a case study systematically and the result shows the effectiveness of our method.

1 Introduction

Although enterprises and organizations have been paying ever more attention to cyber defense, today's cyber-attacks towards enterprise networks often undermine the security of business processes. The reason is directly related to several main limitations of existing cyber-defense practice, because the security of business processes heavily relies on the deployed cyber-defense measures and procedures.

Although a fundamental limitation of existing cyber-defenses is that zero-day attacks cannot be prevented, this limitation is clearly *not the only reason* why cyber-attacks can undermine security. In many, if not most, real-world cyber-security incidents, the security of business processes is actually undermined by known attacks.

Regarding why known attacks could sigificantly undermine the security of business processes, the following main reasons have been recognized in the research community. First, enterprises and organizations do not have the resources needed to patch all the known vulnerabilities. As a result, although the security

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

administrators are working hard to patch as many vulnerabilities as possible and as soon as possible, many vulnerabilities are actually in the "not yet patched" status when cyber-attacks happen. Another contributing factor to the result is that the time a vulnerability becomes known is *not* the time the corresponding patch becomes available.

Second, when cyber-attacks are happening, even if the intrusion detection system accurately detects the intrusions, the intrusion alerts and alert correlation results are still not able to directly tell "what should I do?" in terms of intrusion response. (In real-world enterprises new intrusion alerts keep on being raised, and the security administrators are already fully loaded.) It has been widely recognized in the research community [8, 18, 20] that there is a wide *semantic gap* between the information contained in intrusion alerts and how the cost-effectiveness of intrusion response is evaluated. On one hand, the cost-effectiveness of intrusion response is usually evaluated based on business process-level metrics (e.g., the number of customers affected by a cyber-attack, the number of tasks that need to be undone) and measurements. On the other hand, business process-level metrics are not really measured by intrusion detection systems.

Therefore, to achieve cost-effective intrusion response, this semantic gap must be bridged. To bridge the semantic gap, impact assessment is necessary. Although researchers have found the necessity of using entity dependency graphs [8] to assess the impact of attacks on business processes for quite a few years, the existing impact assessment techniques still face a key challenge. The challenge is two-fold: (1) impact assessment results cannot be automatically used to make recommendations on taking active cyber-defense actions; and (2) existing active cyber-defense techniques cannot be business-process-aware. That is, these techniques will not be able to directly state their effectiveness using business process-level measurements such as how much of what tasks will be accomplished by when.

In [19] it has been perceived that attack graphs and entity dependency graphs could be interconnected to address the above key challenge; however, no realistic case study has been conducted to validate the perceived method. As a result, the intrusion response research community still lacks essential understanding about (a) how to efficiently implement the perceived method; (b) whether it really works; and (c) how well it works.

The goal of this work is to efficiently design and implement the perceived method and conduct a realistic case study to assess the impact of attacks on business processes using not only system-level metrics (e.g., how many files are corrupted, which processes are compromised) but also business process-level metrics. We believe that this case study is a solid step forward towards bridging the aforementioned semantic gap.

The main contributions of this work are as follows.

– We propose the first efficient implementation of the method perceived in [19]. We extend the perceived method to make use of CVSS scores. We invent an algorithm to prune the raw interconnected graph. Through logic

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

programming, the implemented tool can automatically generate an interconnected graph, which interconnects an attack graph and an entity dependency graph, and calculate the impact scores of an attack on tasks in a business process.
– The first realistic case study is systematically conducted to show how the perceived method and our implementation can assess the impact of attacks on business processes using not only system-level metrics but also business process-level metrics.
– Through the case study, we also evaluate our implementation in several aspects such as scalability and running time.

2 Background

2.1 CVSS score

The Common Vulnerability Scoring System (CVSS) provides a way to measure the impacts of vulnerabilities and produce a numerical score for the attack impact [9]. The current version of this score system is version three, which is released in 2015. The system contains three metric groups: base score metrics, temporal score metrics, and environmental score metrics. A base score ranging from 0 to 10 is assigned to a vulnerability according to the base score metrics. The temporal score metrics and environmental score metrics can be used to refine the base score to better reflect the risks caused by a vulnerability to the user's environment. However, the temporal score metrics and environmental score metrics are optional. Therefore, in this paper we only use base score for impact analysis and still refer it as CVSS score. The National Vulnerability Database (NVD) provides a CVSS base score for almost all known vulnerabilities. A higher CVSS base score of a vulnerability implies that: 1) the vulnerability is easier to be exploited due to more vulnerable components and available technical means for exploitation; or 2) more impact on the availability, confidentiality, and integrity upon successful exploitation. Therefore, the base score can be leveraged to assess the impact of vulnerability exploitation on business processes in terms of both exploitability and impact.

2.2 Attack Graph

To analyze the impact of attacks on business processes, it's necessary to first understand how the vulnerabilities in an enterprise network can be used to compromise the host machines. Attack graph [1, 7, 10, 12, 17] is a very effective way to generate potential attack paths. Given the vulnerabilities, the attack graph is able to show the possible attack sequences to the final attack target.

MulVAL (Multihost, multistage Vulnerability Analysis) is an attack graph generation tool that models the interaction between software vulnerabilities and the system and network configurations [11]. It leverages Datalog [14] to model network system information (such as the vulnerabilities, configurations of each

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

machine, etc.) as facts and the interaction of various network components as rules. With these facts and rules, MulVAL can generate an attack graph showing the potential attack paths from the vulnerabilities to the attack goal. In the attack graph, facts and rules are represented by nodes with different shapes. There are two types of fact nodes: primitive fact nodes and derived fact nodes. The primitive facts nodes are denoted with boxes, which represents host and network configuration information. The derived fact nodes are denoted with diamonds, which are generated according to certain rules. The interaction rules are denoted with ellipses.

Fig. 1 shows a very simple attack graph containing only 5 nodes. In Fig. 1, if the conditions in node 1, 2 and 3 are satisfied, then the rule in node 4 can be applied. The eventual consequence is that the attacker is able to execute arbitrary code on the host machine (shown in node 5).

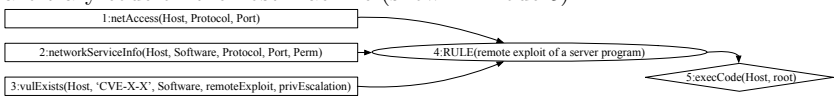


Fig. 1. An Example Attack Graph

The attack graph is essential for business process impact assessment, as it shows how the vulnerabilities can be leveraged to compromise the host machines. If the host machines are involved in the business processes, the impact of vulnerabilities on business processes can then be further analyzed.

2.3 Entity Dependency Graph

In an enterprise network, a business process is supported by a number of entities at several abstraction layers: asset layer, service layer and business process task layer. At the asset layer, an asset is (part of) a persistent disk and the file stored on the disk, a computer (hypervisors, desktops or servers), or a peripheral device. At the service layer, services represent the functionalities provided by hosts, such as web services, database services, etc. At the business process task layer, a business process is composed of one or more tasks.

An entity dependency graph [2] can be established due to the dependencies between the abstraction layers and the dependencies on each individual layer. Generally, the higher layer depends on the function of the lower layer. The business process task layer depends on the functionality provided by the services at service layer. One task may even depend on several services. The services further depends on the assets at the services layer. In addition, dependencies also exist at an individual layer. For example, at the business process task layer, a task may depend on another task.

3 Approach Overview

The primary goal of our paper is to assess the attack impact on business processes. Since attacks essentially exploit vulnerabilities in the enterprise network,

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

the attack impact heavily relies on the intrinsic characteristics of each individual vulnerability. Considering that the characteristics of vulnerabilities have been measured using the CVSS scores, the impact towards a business process can also be measured based on the scoring system. That is, an impact score can be generated for a business process to indicate the impact of attacks towards the business process. Therefore, the key problem need to be addressed is how to generate the impact score for a business process given the CVSS scores of involving vulnerabilities.

In this paper, we propose an three-step approach for business process impact assessment. The general idea is to generate an interconnected graph by analyzing the dependency relationships between vulnerabilities and attacks on hosts, between services and hosts, and between tasks and services. The approach takes three sets of knowledge units as the inputs and generates the business impact score as the output.

The three sets of knowledge units are respectively 1) Common Vulnerability and Exposure (CVE) system that provides information of publicly known vulnerabilities and their CVSS scores, 2) the vulnerability information generated by the vulnerability scanner, and 3) the business process dependency graph. The business impact assessment approach mainly involves the following steps:

Step 1: Instantiate the knowledge units with Datalog as facts and rules in MulVAL. Utilize MulVAL to generate an interconnected graph which consists of impact paths from the vulnerabilities.

Step 2: Prune the interconnected graph to get a more clear relationship between business processes and vulnerabilities.

Step 3: Calculate the impact score based on the CVSS scores of the vulnerabilities exploited in this attack.

3.1 Instantiate Knowledge Units

CVE system refers to the vulnerability database which contains all information about publicly known vulnerabilities. From this system, we can get the CVSS score of each vulnerability. The vulnerability information generated by vulnerability scanner contains the exact CVE IDs of each vulnerability and where these vulnerabilities are located in the enterprise network. By combining these two sources of knowledge, we can easily get the whole picture of these vulnerabilities, including CVSS score, CVE ID, and location in the enterprise network, etc. Such vulnerability information can be used to analyze the potential attacks that might happen, which may further impact the business processes. As the information represents facts about vulnerabilities in the network, we crafted fact nodes in MulVAL to instantiate the information.

Business process dependency graph describes how entities in the network depend on each other. Sun et al. [19] summarizes and bridges the semantic gap between the attack graph generated by MulVAL and the business process dependency graph. Hence, in this paper, we extend MulVAL to craft new fact nodes and new rule nodes to interconnect the attack graph and the business process dependency graph.

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

Listing 1.1. Example Interaction Rules Describing Three Dependency Relationships

```
interaction_rule(       /* And depends */
    (nodeImpact(Task):-
        node(Task, and, Task1, Task2), nodeImpact(Task1)
    ),
    rule_desc('An impacted child task affects an And task')
).
interaction_rule(       /* Or depends */
    (nodeImpact(Task):-
        node(Task, or, Task1, Task2),
        nodeImpact(Task1), nodeImpact(Task2)
    ),
    rule_desc('Both impacted child task affects an Or task')
).
interaction_rule(       /* Flow depends */
    (nodeImpact(Task):-
        node(Task, flow, Task1, Task2), nodeImpact(Task2)
    ),
    rule_desc('A flow node is impacted from its flow')
).
```

First of all, entities in a business process dependency graph become primitive fact nodes or derived fact nodes. Primitive fact nodes usually represent already known information, such as host configuration, network configuration, etc. Derived fact nodes are computed information by applying interaction rules towards primitive fact nodes.

Secondly, rule nodes are added to model the causality relationships among fact nodes. For example, if a service S runs on a machine H and an attacker has exploited a vulnerability to *execute arbitrary code on the machine*, then this service S can be impacted by the attacker. This relation can be interpreted as a rule "A compromised machine impacts a service running on it". In other words, when two fact nodes "S runs on machine H" and "attacker executes arbitrary code on the machine" are both present, this rule node will take effect and the derived fact node "S is impacted" will become present. In this example, machine H has a vulnerability. The attack graph generated by MulVAL can only tell "attacker executes arbitrary code on the machine," but it is not able to tell "S is impacted". Therefore, interconnecting the attack graph and the business process dependency graph can help infer the impact of attacks on business process.

Thirdly, the dependency relationships among entities in the business processes become rule nodes. There are three dependency relationships in the business process dependency graph: Or-depends, And-depends and Flow-depends. Listing 1.1 shows a set of example interaction rules crafted to depict the impact propagation among tasks when different types of dependency relationships exist among these tasks. That is, if a task *and-depends* on task 1 and task 2, then this task is impacted by the attacker when either of the two tasks are impacted. if a task *or-depends* on task 1 and task 2, then this task is impacted only when both tasks are impacted. if a task *flow-depends* on task 1 and task 2, then this task will be impacted when task 2 is impacted. In this case, task 2 can be completed only after task 1 is completed. So if task 1 is impacted,

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

then task 2 is impacted. We will explain more about the dependency relationships in section 5.1.

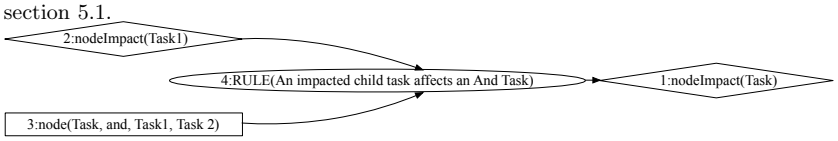


Fig. 2. And-dependency in the graph

With all the fact nodes and rule nodes set up, MulVAL can be used to generate the interconnected graph. For example, Fig. 2 shows the first and-depends example in Listing 1.1. In the interconnected graph, different nodes are represented by different shapes, i.e., box, ellipse and diamond. The ellipse shape represents rule node, which is applied only if all needed precondition fact nodes are present. Hence, the ellipse shape represents AND-relation for all precondition fact nodes. The diamond shape represents derived fact node, which is generated as long as one deriving rule node is present. Therefore, the diamond shape represents OR-relation between the deriving rule nodes. In other words, the interconnected graph reflects the relationship between vulnerabilities and the business processes. However, the interconnected graph is too complicated for generating the impact assessment score for a business process. To enable computation of the impact score, we prune the graph to reduce the complexity.

3.2 Prune Raw Interconnected Graph

Impact score is a function result of CVSS scores of the vulnerabilities involved in the interconnected graph. When we prune the graph, we must preserve the vulnerability node and the impacted business process node. We apply all the five rules below to prune the graph. The entire process of pruning may take several rounds by applying different rules in each round. In addition, based on different circumstances, we also deal with the edges connecting to the reduced nodes correspondingly.

Prune all the non-vulnerability leaf nodes. In this interconnected graph generated by MulVAL, derivation nodes (rule nodes) imply AND relations and derived fact nodes imply OR relations. The primitive fact node in this graph represents the facts in this network, such as the vulnerabilities and deployment configuration. They are represented as leaf nodes in the graph with a shape of box. These non-vulnerability leaf nodes do not participate in the function of CVSS scores. So if a node is not a vulnerability node and is not an AND or OR node, we can prune it. Then each edge derived from these nodes can also be pruned.

Prune the nodes that have only one ancestor node. If a node has only one ancestor node, no matter how many child nodes it has, it does nothing but directly deliver impact from its ancestor node to its child nodes. This node is an intermediate impact deliverer for its ancestor node and can be directly pruned without information loss. This kind of nodes is usually the derivation nodes which have only one ancestral vulnerability node, or derived fact nodes which have only one rule to be generated. By pruning one node, the edges from the ancestor node to this node and from this node to the child nodes are removed. A new edge is added directly between the ancestor node and the child node. This operation of pruning one-ancestor nodes may be done several times in the graph-pruning process, as more of them may be produced in other rounds of pruning.

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

Prune the nodes, except the vulnerability nodes, which have no ancestors. Because all left nodes are relation nodes, vulnerability nodes, and the impacted business process nodes. If a node has no ancestor node and is not a vulnerability node, it is a relation node and does not contain any valuable information. This kind of nodes are produced by pruning their ancestor nodes that are usually non-vulnerability nodes. As their ancestor nodes have been pruned, no impact information is delivered to them. Therefore, they can be pruned without impact information loss. The edges from these nodes can also be pruned.

Find the shortest path from one vulnerability node to the target impacted business process node and merge these paths. The impact assessment for an attack is to find the relationship between vulnerabilities and the target impacted business processes. If a vulnerability can be exploited in an easy way to affect a business process, there is no need to make it more complex. The assumption in our paper is that attackers always choose the easiest way to achieve the attack goal. Based on this assumption, if there are different paths between a vulnerability node and the impacted business process node in the interconnected graph, the shortest path that has least nodes should be chosen. As a result, each vulnerability node has a shortest path to the target business process node. All other nodes and edges that are not on these paths should be pruned. In some cases, one vulnerability node may have more than one shortest paths to the target business process node. In this case, these paths should also be preserved. To simplify these circumstances, if there are two or more equal shortest paths between one vulnerability node and the impacted business process node, we convert this interconnected graph to two or more interconnected graphs to ensure there is only one shortest path for a vulnerability in one interconnected graph. Finally we calculate each graph's impact score to get the average score.

Leave only one edge for linked nodes and prune the other edges between them. In some cases, there are more than one edges between two nodes. The extra edges could be produced by the previous rounds of pruning. They are not needed and thus should be removed too.

These five ways are applied sequentially to the raw interconnected graph generated by MulVAL until the graph does not change again. Two or more graphs could possible be generated as one vulnerability may have two or more equal shortest paths to a target business process node.

3.3 Calculate Impact Score

Step 2 can prune the raw interconnected graph to the simplified graph which contains only the vulnerability nodes, the target business process node and their relations. The impact score of the vulnerability node and the target business process node can be represented by V and M respectively. The impact score calculations based on AND-relations and OR-relations are called AND-calculation and OR-calculation. We take the following steps to generate the impact score.

First, we value V by a number between 0 and 1, i.e.,

$$V_i = \frac{CVSS_i}{10}. \tag{1}$$

Second, we define AND-calculation as:

$$V_i \ \ AND \ \ V_j = V_i \times V_j. \tag{2}$$

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

and OR-calculation as:

$$V_i \ \ OR \ \ V_j = V_i + V_j - V_i \times V_j. \tag{3}$$

Finally, M can be easily calculated by above mentioned calculation methods. For example,

$$M = FUNC(V_1,\ V_2,\ V_3) = (V_1 \ \ OR \ \ V_2) \ \ AND \ \ V_3 = (V_1 + V_2 - V_1 \times V_2) \ \times \ V_3 \tag{4}$$

In this paper, we use the above definitions of AND-calculation and OR-calculation to compute the impact score. However, the administrators of an enterprise network can change the definitions of AND-calculation and OR-calculation based upon different situations and scenarios.

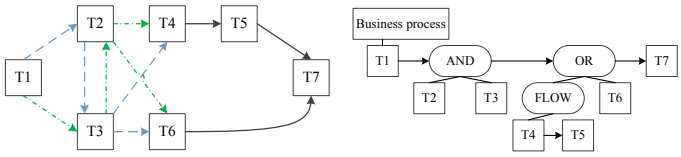
The results of AND-calculation and OR-calculation are directly influenced by the CVSS score of the vulnerabilities. Higher CVSS score usually leads to higher impact score towards the business process, which implies more impact the attack can bring to the business process.

4 Case Description

To demonstrate the method for attack impact assessment, we describe a concrete case in this section. We will illustrate the application of our method to this case in section 5.1.

Business Process Scenario. This case is a travel reservation system supporting a business process of "providing customers with a web interface for reserving tickets and hotel". This business process consists of seven tasks: T_1: Search travel information; T_2: Reserve tickets and hotel options; T_3: Prompt for signing in or signing up; T_4: If signed in, load preference and promotion code; T_5: If signed in, reserve a hotel and tickets as a member; T_6: If not signed in, reserve a hotel and tickets as a guest; T_7: Prompt for payment and confirm the reservation.

From T_1 (start of the business process) to T_7 (end of the business process), the business process may be executed through four different workflows (i.e. execution paths) as shown in Fig. 3a : P_1: $T_1T_2T_3T_4T_5T_7$; P_2: $T_1T_3T_2T_4T_5T_7$; P_3: $T_1T_2T_3T_6T_7$; and P_4: $T_1T_3T_2T_6T_7$. The difference between P_1 and P_2 and between P_3 and P_4 is the order of T_2 and T_3. The customer can either first make reservations (T_2) and then be prompted to sign in (T_3), or first sign in and then make reservations. If the customer chooses not to sign in during T_3, she is recognized as a guest. The difference between P_1 and P_3 and between P_2 and P_4 is whether the customer has signed in. If signed in, the system loads customer preference and promotion code (T_4) for reserving a hotel (T_5). Since T_5 depends on the information obtained from T_4, T_5 should come after T_4.



(a) Execution paths (b) Dependency tree

Fig. 3. Inter-task dependency

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

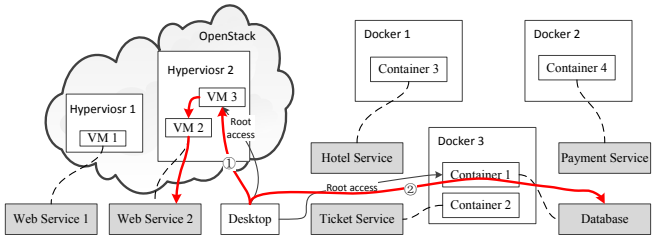


Fig. 4. Software Architecture

This travel reservation system can be viewed as a complicated business-process-support enterprise network shown in Fig. 4. The services provided by the network are hosted on different hosts. VM 1, VM 2 and VM 3 are three virtual machines. Web service 1 is hosted in VM 1 which runs in Hypervisor 1. Web service 2 is hosted in VM 2 which runs in Hypervisor 2. VM 3 also runs in Hypervisor 2.

Database service runs in Container 1 which is hosted by Docker 3. Ticket service, which processes ticket-related business, runs in Container 2 which is also hosted by Docker 3. Hotel service, which processes hotel-related business, runs in Container 3 which is hosted by Docker 1. Payment service, which is responsible for monetary transaction, runs in Container 4 hosted by Docker 2. These dockers run in different workstations. A developer's desktop can access the VM 3 and has a root account credential. It can also access Container 1 as a root user. This desktop has a dashboard which displays through HTTP protocol, i.e. it runs a web service. It can also be accessed through SSH protocol from Internet.

Table 1. Vulnerability Information

Vulnerability	CVSS Score	Exploited Result
CVE-2016-0777	6.5	Privilege Escalation
CVE-2016-7479	9.8	Privilege Escalation
CVE-2016-6325	7.8	Privilege Escalation
CVE-2014-3499	7.2	Container Escape
CVE-2016-6258	8.8	Virtual Machine Escape

Attack Scenario. We assume this network has five vulnerabilities and their related information is displayed in table 1. CVE-2016-0777, CVE-2016-7479 and CVE-2016-6325 locate in the developer's desktop and allow attackers to escalate privilege. CVE-2014-3499 locates in the docker software and can enable an attacker to escape from the container. CVE-2016-6258 locates in the Kernel-based Virtual Machine(KVM) software and can also be used to break the virtual machine.

There are two attack paths in Fig. 4. One attack path is denoted as red line 1 in Fig. 4. The attacker firstly exploits the vulnerability in the web application or the SSH application to compromise the developer desktop, which has the log-in credential for VM 3. By leveraging the vulnerability in the KVM software, the attacker can directly access the host, i.e. Hypervisor 2, by breaking the isolation between the virtual machine and the host. The attacker can then access VM2 which hosts Web service 2 and execute arbitrary code on this virtual machine. Once Web service 2 is compromised, all tasks depend on this service are impacted. The other attack path is denoted as red line 2

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

in the Fig. 4. As the developer's desktop has the log-in credential for Container 1, the attacker can also access this container. With the database running in this container, the attacker can execute arbitrary code in the database process and then affect all tasks depending on the database.

5 Case Study and Evaluation

5.1 Case Study Results

In this section, we applied the impact assessment method to the case described above and demonstrate the experiment results.

First, we obtained the CVSS scores for the five vulnerabilities in this case according to their CVE IDs.

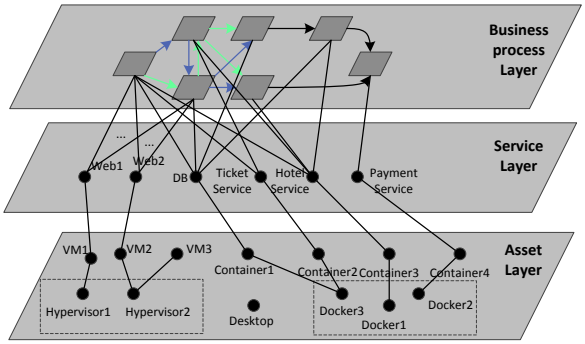


Fig. 5. The Entity Dependency Graph

Second, we constructed a entity dependency graph for this network, as shown in Fig. 5 (as web services are depended on by each task, some edges from the tasks to Web1 and Web2 are ignored in this figure). The entity dependency graph contains three layers: asset layer, service layer and business process task layer. Among these tasks, T_1 and-depends on the web services, database service, ticket service and hotel service. T_2 and-depends on web services, ticket service and hotel service. T_3 and-depends on web services, and database service. T_4 and-depends on web services, and database service. T_5 and-depends on web services, database service, and hotel service. T_6 and-depends on web services, and hotel service. T_7 and-depends on web services, and payment service.

At the business process layer, we specified the dependency relationships among tasks. To better understand the relationships, we firstly define three special tasks: T_{or}, T_{and} and T_{flow}. As the name implies, these tasks represent three relationships: Or-dependency, And-dependency, and Flow-dependency. That is, if a task T_{or} or-depends on sub-tasks T_i and T_j, then T_{or} is impacted only when T_i and T_j both are impacted. If a task T_{and} and-depends on sub-tasks T_i and T_j, then T_{and} is impacted when T_i or T_j is impacted. If a task T_{flow} flow-depends on sub-tasks T_i and then T_j, then T_{flow} is impacted when T_j is impacted. In addition, the impact on T_i will cause an impact on T_j, which leads to an impact on T_{flow}. The relationships of the seven tasks of this business process can be depicted in Fig. 3b. In other words, this business process viewed as one T_{flow} flow-depends on T_1, T_{and}, T_{or} and then T_7. T_{and} and-depends on T_2 and T_3. T_{or} or-depends on T_6 and T_{flow}, which flow-depends on T_4 and then T_5.

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

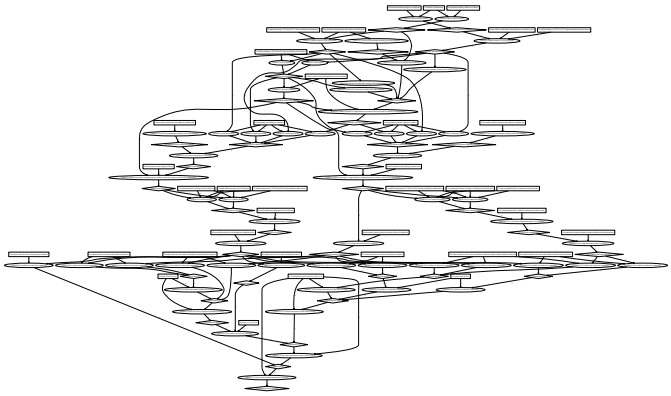


Fig. 6. Interconnected Graph

After instantiating the knowledge units, we can get the interconnected graph as shown in Fig. 6. In this graph, the ellipse represents AND-calculation and the diamond represents OR-calculation. By applying the five pruning rules described in section 3.2 against the raw graph, we generated the pruned graph, as shown in Fig. 7, to show the relationship between vulnerabilities and the target business process. The expression "nodeImpact(X)"means "X" is impacted, e.g. "nodeImpact(business process)" means the target business process is impacted. The CVSS scores of these vulnerabilities are shown in table 1. Therefore, the final impact score of this attack can be calculated as:

$$\begin{aligned} M &= (((V_{CVE-2016-0777} \;\; OR \;\; V_{CVE-2016-7479}) \;\; AND \;\; V_{CVE-2016-6325}) \\ &\quad AND \;\; V_{CVE-2016-6258}) \;\; OR \;\; V_{CVE-2016-3499} \\ &= 0.91. \end{aligned}$$

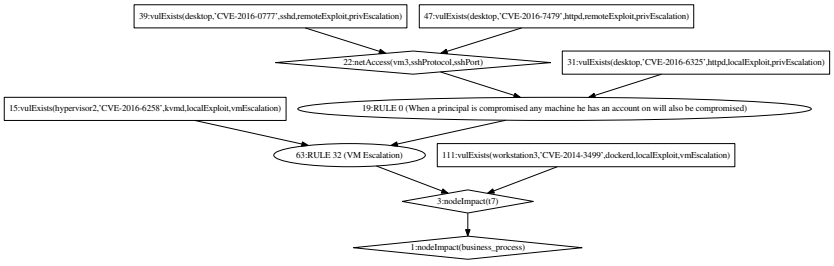


Fig. 7. Pruned Interconnected Graph

Apart from the impact score calculated from the pruned interconnected graph, there is more information about whether the services and tasks are impacted or not from the raw interconnected graph. By searching through the raw interconnected graph showed in Fig. 6, we can get that all tasks are impacted by this attack. Three services including Web service 2, Ticket service and Database service are also impacted. All tasks are impacted as they all and-depend on Web service 2. These three services are

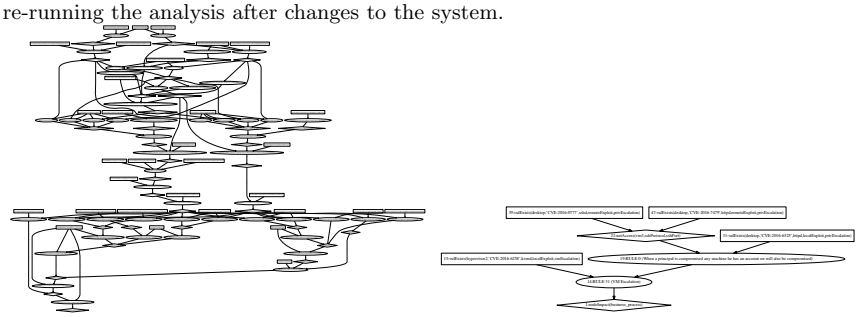
Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

impacted as they can be accessed by the developer's desktop which can be controlled by the attacker. That is, the impact on these services match the attack path described in Section 4. Moreover, we can also get the impact score for each task through the same process: pruning the graph and calculating the score based on the AND-calculation and OR-calculation. The impact score for each task is 0.992 for task 1, 0.91 for task 2, 0.973 for task 3, 0.973 for task 4, 0.973 for task 5, 0.682 for task 6, and 0.91 for task 7. We can see some scores are higher than the impact score for the whole business process. This is because some task are easily attacked by the attacker from the Internet. For example, task 3 and-depends on web service 1, web service 2 and database service. The attacker can impact task 3 without exploiting the vulnerability "CVE-2014-3499", which lowers the requirement for the attacker.

There are three services that are not impacted by the attack, including Web service 1, Hotel service and Payment service. They cannot be found as the impacted nodes in the raw interconnected graph. This is because they are not involved in the attack path. Therefore, the raw interconnected graph can precisely present the attack path in the real world.

5.2 Analysis of Different Cases

Section 5.1 has shown a successful application of our impact assessment method to the case described in section 4. However, in the real world, the enterprise network is not static. For example, a vulnerability can be patched or a host can be removed. In this section, we will show that our method can still handle the dynamic changes in the enterprise network and generate new impact scores for the business processes by re-running the analysis after changes to the system.



(a) The raw interconnected graph (b) The pruned interconnected graph

Fig. 8. Interconnected graphs with a vulnerability is patched

A vulnerability is patched. When a vulnerability is patched, it means a fact node should be deleted. As a consequence, the interconnected graph will be different and so is the pruned graph. For instance, we assume the vulnerability "CVE-2014-3499" is patched as this vulnerability is the oldest one in these five vulnerabilities. Fig. 8 shows the new raw interconnected graph and pruned graph without "CVE-2014-3499." By analyzing this pruned graph, the new impact score towards the business process is 0.682, which is much smaller than 0.91.

Whether a task or a service is impacted can also be acquired through the raw interconnected graph. By searching this graph, we can see all tasks are still impacted.

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

Two services including Web service 2 and Database service are impacted. The other four services, including Web service 1, Hotel service, Payment service and Ticket service, are not impacted. Compared with section 5.1, ticket service is not impacted in this case. This is because patching the vulnerability "CVE-2014-3499" prevents the escape from Container 1. The attacker cannot access Container 2 any more so that the ticket service running in Container 2 is free from the impact.

The developer desktop is removed. When the developer desktop is removed, several fact nodes should be deleted. For example, three vulnerabilities in this desktop no longer impact the network, so these vulnerability nodes are deleted. When generating the interconnected graph with MulVAL, we found no graph was generated. This means although there are vulnerabilities in this network, the attacker located in the Internet cannot impact this business process. The reason is that all attack paths start from this desktop as the entry point. Removing this desktop prevents the attacker from exploiting the vulnerabilities inside the network. Therefore, the interconnected graph can precisely reflect the real-world impact circumstances.

5.3 Evaluation of Scalability

Table 2. Time consumed to generate interconnected graphs according to different Number of Units (NoU) and different Connectivity Level (CL)

CL \ NoU	100	200	400	600
5	1m2.45s	7m44.71s	67m55.64s	228m42.53s
10	1m0.33s	7m49.49s	65m4.48s	253m9s
100	0m59.67s	7m48.85s	65m18.60s	224m33.49s

Section 5.1 illustrates how to leverage our impact assessment method to calculate the impact score for an attack targeting a particular business process. The key idea is to extend MulVAL to generate an interconnected graph and calculate the impact score based on the pruned graph. In this process, generating the interconnected graph is the most time-consuming part. It directly affects the scalability of our impact assessment method. Therefore, in this section, we evaluate the scalability of our method in terms of how fast interconnected graphs can be generated for different scopes of network.

In order to get different scopes of network, we view the small network of the aforementioned case in section 4 as one unit and duplicate it. These units are then combined on the basis of different connectivity levels. Because different connectivity levels differ the network complexity, which may affect the time used to generate the interconnected graph. We define connectivity level as how widely one web server is shared, i.e., how many units share one web server. These units sharing one web server constitute one group and each group is connected by the database server of one unit in the group. Therefore, the scope of a network generated through this method can be measured by number of units and connectivity level.

Table 2 describes the time consumed to generate interconnected graphs for different scopes of network according to different number of units and different connectivity level. The first column indicates connectivity level and the first row presents the total number of duplicated units. The other grids in the table indicate how much time is used to generate one graph. For example, with 100 duplicated units in the network and every 5 units sharing one web server, generating the interconnected graph for this scope of network consumes 1 minute and 2.45 seconds.

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

From table 2, we can see the time used to generate an interconnected graph is mainly determined by the number of connected units, not the connectivity level. This is because when generating the interconnected graph, the time is mainly consumed by finding new path from one node to another node. As sharing web server does not increase paths in the graph, the consumed time does not affected by the connectivity level. Furthermore, the time increases non-linearly, i.e., the time increases faster than the number of connected units increases. In summary, our method cannot scale well in a very large network. However, it does not mean our solution is not practical in the real world. Taking a university as an example, the scope of one unit is similar to a network of a department. Therefore, for a big university with 100 departments, the time consumed to generate an interconnected graph is less than 2 minutes, which means our solution is feasible in practice.

6 Related Work

Little research has been done on business process impact assessment in recent years. Jakobson [8] presents a business process impact assessment that quantifies impact by using Operational Capacity (OC), and considers intra and inter dependencies between assets, services, and business processes. Dai et al. [3] propose a cross-layer Situation Knowledge Reference Model (SKRM) which considers intra and inter-dependencies between instruction layer, OS layer, app/service layer, and workflow (task) layer. Sun et al. [18] introduce a novel probabilistic impact assessment method which leverages Bayesian networks. Sun et al. [20] also propose a multi-layer impact evaluation model which includes four layers, namely vulnerability layer, asset layer, service layer, and mission layer. They measure impacts by OC and impact factor. Poolsappasit et al. [13] leverages attack graph (called Bayesian Attack Graph) and attack tree to revise the likelihoods in the event of attack incidents and identify the vulnerable points in the network system. Frigault et al. [5] use attack graph as a special Bayesian network to model probabilistic risks in a network. They also introduce Dynamic Bayesian Networks [6] with attack graphs to model the security of dynamically changing networks. Dewri et al. [4] leverage an attack tree model with multi-objective optimization to solve the problem, i.e. balance between security hardening and limited budget for an enterprise network. Ray et al. [15] also utilize an attack tree model with an algorithm simplifying the tree to locate the malicious insiders in a network. Saripalli et al. [16] present QUIRC which utilizes Microsoft's STRIDE to assess the security risk in a cloud computing environment and define risk as a combination of the Probability of a security thread event and its severity.

Our method uses the interconnected graph, which interconnects attack graph and entity dependency graph, to demonstrate the relationships between vulnerabilities and the impacted business process. By pruning the interconnected graph, we can get simplest relationships and calculate the impact score based on vulnerabilities' CVSS score. For different cases in one network, our method can handle these changes and generate related impact scores. With these impact scores, the network operator may do further security hardening for the network.

7 Conclusion

In this paper, we propose a new business process impact assessment method, which measures the impact of an attack towards a business process in an enterprise network.

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

Our method produces a numerical score for the attack impact. We extend MulVAL, a logic-based network security analyzer, to support more fact nodes and rule nodes for business process impact assessment. With the facts and rules, our approach generates an interconnected graph for an attack and prunes the interconnected graph to show the simplified relation between vulnerabilities and business processes. In the end, the impact score can be calculated by analyzing the pruned graph and following the relation calculation rules. According to our case study, this business process impact assessment method is effective and can facilitate the cyber-defense and cyber-resilience in an enterprise network that supports business processes.

Acknowledgment

We thank the anonymous reviewers for their valuable comments. This work was supported by NIST 60NANB17D279, NSF CNS-1505664, ARO W911NF-13-1-0421 (MURI), and NSF CNS-1618684.

Disclaimer

References

1. Ammann, P., Wijesekera, D., Kaushik, S.: Scalable, graph-based network vulnerability analysis. In: Proceedings of the 9th ACM Conference on Computer and Communications Security. pp. 217–224. ACM (2002)
2. Chen, X., Zhang, M., Mao, Z.M., Bahl, P.: Automating network application dependency discovery: Experiences, limitations, and new solutions. In: OSDI. vol. 8, pp. 117–130 (2008)
3. Dai, J., Sun, X., Liu, P., Giacobe, N.: Gaining big picture awareness through an interconnected cross-layer situation knowledge reference model. In: Cyber Security (CyberSecurity), 2012 International Conference on. pp. 83–92. IEEE (2012)
4. Dewri, R., Poolsappasit, N., Ray, I., Whitley, D.: Optimal security hardening using multi-objective optimization on attack tree models of networks. In: Proceedings of the 14th ACM conference on Computer and communications security. pp. 204–213. ACM (2007)
5. Frigault, M., Wang, L.: Measuring network security using bayesian network-based attack graphs. In: Proceedings of the 2008 32nd Annual IEEE International Computer Software and Applications Conference. pp. 698–703. IEEE Computer Society (2008)
6. Frigault, M., Wang, L., Singhal, A., Jajodia, S.: Measuring network security using dynamic bayesian network. In: Proceedings of the 4th ACM workshop on Quality of protection. pp. 23–30. ACM (2008)

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

7. Jajodia, S., Noel, S., OBerry, B.: Topological analysis of network attack vulnerability. In: Managing Cyber Threats, pp. 247–266. Springer (2005)
8. Jakobson, G.: Mission cyber security situation assessment using impact dependency graphs. In: Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on. pp. 1–8. IEEE (2011)
9. NIST: Cvss score. https://nvd.nist.gov/vuln-metrics/cvss (2017)
10. Noel, S., Jajodia, S., O'Berry, B., Jacobs, M.: Efficient minimum-cost network hardening via exploit dependency graphs. In: Computer security applications conference, 2003. proceedings. 19th annual. pp. 86–95. IEEE (2003)
11. Ou, X., Boyer, W.F., McQueen, M.A.: A scalable approach to attack graph generation. In: Proceedings of the 13th ACM conference on Computer and communications security. pp. 336–345. ACM (2006)
12. Phillips, C., Swiler, L.P.: A graph-based system for network-vulnerability analysis. In: Proceedings of the 1998 workshop on New security paradigms. pp. 71–79. ACM (1998)
13. Poolsappasit, N., Dewri, R., Ray, I.: Dynamic security risk management using bayesian attack graphs. IEEE Transactions on Dependable and Secure Computing **9**(1), 61–74 (2012)
14. Racket: Datalog. https://docs.racket-lang.org/datalog/ (2017)
15. Ray, I., Poolsapassit, N.: Using attack trees to identify malicious attacks from authorized insiders. In: European Symposium on Research in Computer Security. pp. 231–246. Springer (2005)
16. Saripalli, P., Walters, B.: Quirc: A quantitative impact and risk assessment framework for cloud security. In: Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on. pp. 280–288. Ieee (2010)
17. Sheyner, O., Haines, J., Jha, S., Lippmann, R., Wing, J.M.: Automated generation and analysis of attack graphs. In: Security and privacy, 2002. Proceedings. 2002 IEEE Symposium on. pp. 273–284. IEEE (2002)
18. Sun, X., Singhal, A., Liu, P.: Who touched my mission: Towards probabilistic mission impact assessment. In: Proceedings of the 2015 Workshop on Automated Decision Making for Active Cyber Defense. pp. 21–26. ACM (2015)
19. Sun, X., Singhal, A., Liu, P.: Towards actionable mission impact assessment in the context of cloud computing. In: IFIP Annual Conference on Data and Applications Security and Privacy. pp. 259–274. Springer (2017)
20. Sun, Y., Wu, T.Y., Liu, X., Obaidat, M.S.: Multilayered impact evaluation model for attacking missions. IEEE Systems Journal **10**(4), 1304–1315 (2016)

Cao, Chen; Liu, Peng; Singhal, Anoop; Sun, Xiaoyan; Yuan, Lunpin; Zhu, S.
"Assessing Attack Impact on Business Processes by Interconnecting Attack Graphs and Entity Dependency Graphs."
Paper presented at IFIP International Conference on Database and Application Security and Privacy (DBSEC 2018), Bergamo, Italy. July 16, 2018 - July 18, 2018.

An Updated Security Analysis of PFLASH

Ryann Cartor[1] and Daniel Smith-Tone[1,2]

[1]Department of Mathematics, University of Louisville,
Louisville, Kentucky, USA
[2]National Institute of Standards and Technology,
Gaithersburg, Maryland, USA

ryann.cartor@louisville.edu, daniel.smith@nist.gov

Abstract. One application in post-quantum cryptography that appears especially difficult is security for low-power or no-power devices. One of the early champions in this arena was SFLASH, which was recommended by NESSIE for implementation in smart cards due to its extreme speed, low power requirements, and the ease of resistance to side-channel attacks. This heroship swiftly ended with the attack on SFLASH by Dubois et al. in 2007. Shortly thereafter, an old suggestion re-emerged: fixing the values of some of the input variables. The resulting scheme known as PFLASH is nearly as fast as the original SFLASH and retains many of its desirable properties but without the differential weakness, at least for some parameters.

PFLASH can naturally be considered a form of high degree HFE^- scheme, and as such, is subject to any attack exploiting the low rank of the central map in HFE^-. Recently, a new attack has been presented that affects HFE^- for many practical parameters. This development invites the investigation of the security of PFLASH against these techniques.

In this vein, we expand and update the security analysis of PFLASH by proving that the entropy of the key space is not greatly reduced by choosing parameters that are provably secure against differential adversaries. We further compute the complexity of the new HFE^- attack on instances of PFLASH and conclude that PFLASH is secure against this avenue of attack as well. Thus PFLASH remains a secure and attractive option for implementation in low power environments.

Key words: Multivariate Cryptography, HFE, PFLASH, Discrete Differential, MinRank

1 Introduction

In December of 2016, the National Institute of Standards and Technology (NIST) published an open call for proposals for new post-quantum standards for some of the most critical security applications in digital communication infrastructure, see [1]. The post-quantum technologies this project aspires to vet and standardize

Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

are designed to be secure against adversaries with access to quantum computing devices— machines capable of acheiving exponential speed-up over classical computers on certain problems, see [2].

Many avenues to post-quantum security are developing, including techniques from lattice theory, coding theory and algebraic geometry. Each of these areas enjoy hard computational problems that have been studied extensively and have histories going back many decades. They also share the common trait that the fundamental computational problems in these fields have no known significant speed-up in the quantum paradigm.

One of the hard computational problems on which the security of many post-quantum cryptosystems is based is the problem of solving systems of multivariate equations. Generically, solving systems of multivariate quadratic equations is hard, so a valid technique for constructing a cryptosystem is to find a class of quadratic vector-valued functions on a vector space that is easy to invert, and transform it into a system that appears random.

Both of these tasks present challenges. The standard technique for the second task is computing a morphism of the system in an attempt to remove the properties allowing the system to be inverted. Techniques for the prior task are more varied, and in this work our focus is on a particular big field scheme.

1.1 Prior Work

The progenitor of all "big field" schemes is commonly known as C^*, or the Matsumoto-Imai scheme, see [3]. This scheme exploits the vector space structure of extension fields to provide two versions of a function— a vector-valued version which is quadratic over the base field, and a monomial function whose input and output lie in the extension field. The cryptanalysis of this scheme by Patarin in [4] inspired many big field constructions.

In [5], Patarin introduced the Hidden Field Equations (HFE) cryptosystem, a natural generalization of the monomial based C^* in which the monomial map is replaced with a low degree polynomial. Also described in the above work is the minus modifier— the removal of public equations— which can be applied to both HFE, producing HFE$^-$, and to C^*, creating C^{*-}.

A popular iteration of C^{*-} was SFLASH, see [6], which was very efficient, but unfortunately insecure. An attack by Dubois et al. in [7] broke SFLASH by way of a symmetric differential relation present in the central monomial map.

In [8], a way to resist the attack on SFLASH is presented. The augmentation of the scheme, known as projection, fixes the value of d of the input variables producing a scheme we now call PFLASH. PFLASH is still a very fast signature scheme and is amenable to low-power environments without sacrificing side-channel resistance. This projected C^{*-} system is shown to resist differential cryptanalysis for restricted parameters, that is, when the degree is bounded by $q^{n/2-d}$, in [9] and is fully specified with paractical parameters in [10].

Since the design of PFLASH there have been a number of cryptanalytic developments in the big field venue. The development of differential invariant attacks in [11] and their further application in [12] are examples of advancement

Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

in this active area. Furthermore, the improved efficiency of the Kipnis-Shamir (KS) attack of [13] presented in [14] is directly impactful to PFLASH, as one can consider PFLASH as a possibly high degree but still low rank version of HFE$^-$.

1.2 Our Contribution

We expand and update the analysis in [9] and [10] proving resistance to differential and rank techniques for the vast majority of parameters, and verifying that the provably secure key spaces are not as severely limited as the previous works suggest. This improvement is directly impactful, providing further assurance that attacks based on equivalent keys cannot weaken PFLASH.

The degree bound restriction in [9] reduces the dimension of possible private keys by a factor of more than two. Our updated differential analysis verifies the security of the scheme when the central map has no degree bound, and thus assures us that very little entropy is lost in the key space when restricting to parameters that are provably secure against differential adversaries.

In [10], an argument for the resistance of PFLASH to the technique of [14, Section 8.2] when PFLASH is considered as a low degree projected HFE$^-$ scheme is provided. We make this assessment more robust by also considering the possibility of an adversary attempting to remove the projection modifier from PFLASH considering it to be a higher rank HFE$^-$ scheme. Whereas in the former case, the attack is impossible, in the latter case, the algebraic structure allows the possibility that the attack can succeed; however, the complexity of the attack is directly computed and shown to be infeasible.

1.3 Organization

The paper is organized as follows. The next section introduces the notion of big field schemes and provides the description of those schemes relevant to this work, namely, C^*, $PFLASH$ and HFE. In the following section, we review the cryptanalytic techniques that have proven most successful in attacking big field schemes. The subsequent two sections provide a new proof of security against differential attacks for PFLASH, first by analyzing the projected C^* primitive and then by extending these results to the full scheme. We then conclude, noting parameter choices for PFLASH and discussing applications of the scheme.

2 Big Field Schemes

Many multivariate cryptosystems utilize the structure of a degree n extension $\mathbb{K}$ of a finite field $\mathbb{F}_q$ as an $\mathbb{F}_q$-algebra. Such cryptosystems are collectively known as "big field" schemes. To emphasize a choice of basis, one chooses an $\mathbb{F}_q$-vector space isomorphism $\phi : \mathbb{F}_q^n \to \mathbb{K}$. There is then an equivalence between systems F of n quadratic polynomials in n variables over $\mathbb{F}$ and univariate polynomials of the form

$$f(x) = \sum_{0 \leq i \leq j < n} \alpha_{ij} x^{q^i + q^j}$$

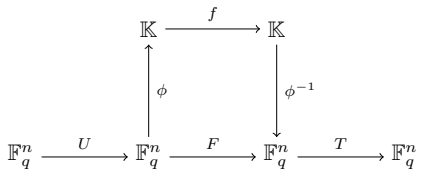
Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

over $\mathbb{K}$ given by $F = \phi^{-1} \circ f \circ \phi$.

To hide the structure of an easily invertible map, the standard technique is to apply an isomorphism of polynomials to mask the choice of basis for the input and output of f.

Definition 1 *A* polynomial morphism *between two systems of polynomials is a pair of affine maps (T, U) such that $G = T \circ F \circ U$. If both T and U are invertible, then the morphism is said to be an isomorphism and F and G are said to be isomorphic.*

Thus, for big field schemes, the construction of a public key can be summarized with the following diagram.



2.1 *C**

Matsumoto and Imai discovered massively multivariate cryptography, introducing the scheme now known as C^* at Eurocrypt '88. The $C^*(q, n)$ scheme is a big field construction in which the vector-valued representation of a quadratic monomial map $f(x) = x^{q^\theta+1}$ is hidden by an isomorphism. Thus the public key is given by $P = T \circ \phi^{-1} \circ f \circ \phi \circ U$.

The C^* scheme was originally envisioned for encryption, but could quite apparently be applied in either encryption or digital signatures. To encrypt (or to verify a signature), one simply computes the output of the public function P. To decrypt (or to sign), the preimage must be determined successively for each of the components of the private key, all of which can be computed efficiently. The interesting step, the inversion of f can be accomplished by noticing that if $b(q^\theta + 1) = 1 \ (mod \ q^n - 1)$, then $(x^{q^\theta+1})^b = x$.

2.2 PFLASH

The PFLASH scheme is a particular parametrization of a projected C^{*-} scheme. The projection and minus modifiers were both originally suggested in reference to C^* in [15]. The idea of projection is to fix the value of d input variables to change the simplicity of the central map. Thus the composition of the projection and the affine map U form a projection onto a codimension d hyperplane. The minus modification removes r equations from the public key. Thus the composition of this projection with T has corank r. The public key of PFLASH(q, n, r, d) is given by $P = \pi_r \circ T \circ \phi^{-1} \circ f \circ \phi \circ U \circ \pi_d$.

Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

We note that the public key is no longer isomorphic to the private monomial function. Instead there is merely a polynomial morphism between the central map and the public key. Since it is well-known that the morphism of polynomials problem is NP-hard, see [16], there is some hope that the information lost to the public key may secure the scheme.

Mechanically, the scheme works as a digital signature primitive as follows. Verification is accomplished by evaluating the public polynomials at the signature. Signing is done by finding preimages of each of the private maps. To find a preimage of $\pi_r \circ T\phi^{-1}$, randomly append r values to the message, then apply T^{-1} and ϕ. Once f is inverted, an element in the preimage of $\phi \circ U$ and in the image of π_d is selected as the signature.

2.3 HFE

Hidden Field Equation (HFE) scheme of [5] is a generalization of the C^* construction, in which the monomial map is replaced by a more general polynomial with a degree bound D. Given the degree n extension $\mathbb{F} \subseteq \mathbb{K}$ we choose a quadratic polynomial $f : \mathbb{K} \to \mathbb{K}$ of degree bound D. Thus f has the form:

$$f(x) = \sum\nolimits_{q^i+q^j \leq D}^{i \leq j} \alpha_{i,j} x^{q^i+q^j} + \sum_{\substack{i \\ q^i \leq D}} \beta_i x^{q^i} + \gamma,$$

where $\alpha_{i,j}, \beta_i, \gamma \in \mathbb{K}$. The public key is then constructed via the isomorphism:

$$P = T \circ \phi^{-1} \circ f \circ \phi \circ U.$$

Inversion is accomplished by first taking a ciphertext $y = P(x)$, computing $v = T^{-1}(y)$, solving $v = f(u)$ for u via the Berlekamp algorithm, see [17], and then recovering $x = U^{-1}(u)$.

3 Cryptanalyses of Big Field Schemes

The big field multivariate cryptosystems have an extensive history in cryptanalysis. Several techniques have been developed that illustrate that it is very difficult to hide efficient inversion of a system. These techniques can largely be grouped into two categories: those based on differential propertys and those based on rank properties.

3.1 Differential Techniques

By breaking "big field" schemes and also inspiring modifiers, differential attacks have been instrumental in the development and analysis of multivariate public key cryptography. Given a field map f, the discrete differential is defined by $Df(a, x) = f(a + x) - f(a) - f(x) + f(0)$. As an operator on $\mathbb{K}$, D is $\mathbb{K}$-linear and reduces the complexity while increasing the dimension of a function. For

Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

example, the differential of an affine map is zero, the differential of a quadratic map is bilinear, the differential of a cubic map is bi-quadratic, etc.

Patarin's linearization equations attack of [4] can be viewed as a differential attack as follows. The differential of the C^* monomial $f(x) = x^{q^\theta+1}$ is symmetric in characteristic two; hence, it is zero on the diagonal, $Df(x,x) = 0$. Therefore setting $v = f(u)$ we have

$$\begin{aligned} 0 = Df(v, f(u)) &= vu^{q^{2\theta}+q^\theta} + v^{q^\theta}u^{q^\theta+1} \\ &= u^{q^\theta}(vu^{q^{2\theta}} + uv^{q^\theta}), \end{aligned}$$

and whether or not $u = 0$, the right factor must be zero; thus, we obtain a bilinear relation between u and v. Setting $u = Ux$ and $v = T^{-1}y$, we obtain a bilinear relation between plaintext and ciphertext pairs: the linearization equations. Indeed even the higher order linearization equations (HOLEs) attacks pioneered in [18] can similarly be derived via differentials.

Another notable application of symmetric differential techniques in cryptanalysis is the attack on SFLASH of [7]. This attack exploits the fact that C^* polynomials are multiplicative. Specifically, $f(x) = x^{q^\theta+1}$ exhibits a differential symmetry.

Definition 2 *A function $f : \mathbb{K} \to \mathbb{K}$ has a differential symmetry if there exists a pair of $\mathbb{F}$-linear maps $L, \Lambda_L : \mathbb{K} \to \mathbb{K}$ such that*

$$Df(La, x) + Df(a, Lx) = \Lambda_L Df(a, x).$$

The attack uses the fact that left-multiplication maps of elements in $\mathbb{K}$ satisfy the above relation. This equality provides a criterion for the derivation of such maps, and via a linear algebra distillation technique, such a map can be efficiently recovered, and a full rank key derived.

It is important to note that once such a symmetry inducing linear map is discovered, there is no need to recover a full rank private key; an attack can be mounted directly with the recovered representation of the extension field multiplicative structure. Thus, even if a central map does not have a differential symmetry, it is possible that a minus-modified version of the scheme might; thus, an attack may be mounted directly on the choice of representation of the big field. This fact is the basis for the direct analysis of minus-modified schemes of [19] and [20].

It was shown in [21] that a quadratic map can only have the symmetry of Definition 2 with L a representation of left-multiplication by a field element when f is multiplicative; that is, when f has only one quadratic monomial. Later it was shown in [9] that the only linear maps L satisfying the above relation for C^* are the multiplication maps.

This famous cryptanalysis incited a more careful analysis of a technique originally proposed at ASIACRYPT 1998 in [15] and further suggested after the attack in [8]. The idea is to use projection, that is, to fix some of the input values, to make U singular. PFLASH, whose parameters are defined in [10], is

Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

a particular parametrization of this structure. This change nullifies the basis of the differential symmetric attack as proven in [9] for a certain parameter set. In the resulting scheme, a pC^{*-} scheme, the central map can be made to no longer admit any symmetry. The parameter set which is provably secure against a differential adversary appears quite small, however, and considering the fact that such a scheme can be considered a special case of HFE$^-$ with perhaps a larger degree bound but an even smaller rank, it is necessary to review the rank structure of such schemes as well.

3.2 Rank Techniques

The first significant cryptanalysis of HFE was the Kipnis-Shamir (KS) attack of [13]. The attack is based on the fact that as a quadratic form over the extension field, the public key has low rank. This attack was significantly improved in [14], where minors modeling, instead of the original modeling of the rank property by Kipnis and Shamir, and Gröbner basis techniques are employed. The result is that the security of HFE is polynomial in the degree of the extension $\mathbb{K}$ over $\mathbb{F}_q$.

PFLASH can easily be characterized as an HFE$^-$ scheme with a more efficient inversion process. This characterization is possible by absorbing the projection into the central monomial map to make a more general polynomial. As an HFE$^-$ scheme, the rank of the central map is still 2, thus the central map has a very strong property. The minus modifier, however, provably increases the rank of the public key.

One may even consider PFLASH to be an HFE instance if we append zero polynomials to the public key. In this case, one should suspect that the rank of the central map would be quite high, rendering attacks such as [13] and [14] infeasible. Still, a theoretical verification of this intuition is absent in the literature.

4 Updated Differential Analysis of Projected Primitive

As discussed in [9], we may assume that the projection mapping is tied to f and consider differential symmetries of $f \circ \pi$ where π is chosen in a basis such that $deg(\pi) = q^d$. Clearly, if $f \circ \pi$ has a differential symmetry then the equation $Df(Ma, \pi x) + Df(\pi a, Mx) = \Lambda_M Df(\pi a, \pi x)$ is satisfied for some M. We can express this relation with matrix multiplication, namely

$$a^\top(\Pi^\top \mathbf{Df} M)x + a^\top(M^\top \mathbf{Df}\Pi)x = \Lambda_M[a^\top(\Pi^\top \mathbf{Df}\Pi)x],$$

where $\mathbf{Df}$ is the matrix representing Df as a bilinear form over $\mathbb{K}$, having one in the $(0, \theta)$ and $(\theta, 0)$ coordinates and zero elsewhere, where $\Pi x = \sum_{i=0}^{d} \beta_i x^{q^i}$ and where $Mx = \sum_{i=0}^{n-1} m_i x^{q^i}$

Examining this equation, we see that $a^\top(\Pi^\top \mathbf{Df} M)x + a^\top(M^\top \mathbf{Df}\Pi)x$ will have nonzero entries restricted to certain coordinates depending only on d and

Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

θ, see Figure 1. Similarly, the right hand side of the equation, $\Pi^\top \mathbf{Df} \Pi$, has a structure dependent upon d and θ, see Figure 2. Notice, the graphs may look different depending on the choice of θ and d.

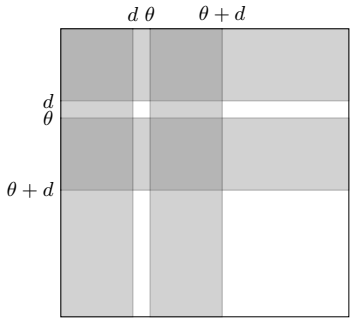


Fig. 1. The shape of the matrix representation over $\mathbb{K}$ of $Df(Ma, \pi x) + Df(\pi a, Mx)$. Shaded regions correspond to possibly nonzero values.

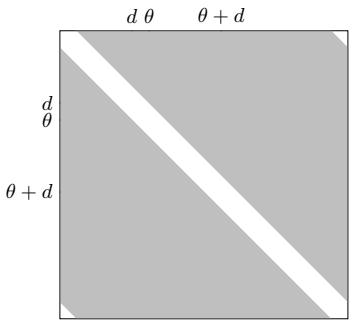


Fig. 2. The shape of the matrix representation of $\Lambda_M Df(\pi a, \pi x)$ over $\mathbb{K}$. Shaded regions correspond to possibly nonzero values.

The strategy for finding conditions on π, M and Λ_M for the existence of such a symmetry is then to find coordinates in which one side of this matrix equation is zero while the other side involves only a single unknown coefficient of M or Λ_M. While this system of equations is nonlinear in the coefficients of π, it is linear in both the unknown coefficients of M and those of Λ_M.

The system contains many more equations than variables, but certainly generates a positive dimensional ideal. The reason is that for any fixed π, $M = a\pi$ for any $a \in \mathbb{F}_q$ generates a solution. On the other hand, for a fixed π and a fixed θ, the above system becomes linear with the number of nonzero equations depending on both d and θ. Even in the best case, the number of equations is far larger than the number of variables. Since the coefficients of π are the only source of randomness for this system of linear equations, the great number of equations are not independent in a probabilistic sense. Therefore, probabilistic arguments are difficult, though extensive experiments show that the solution space is generally one dimensional.

Luckily, we can do better by bootstrapping the result of [9]. Specifically, we examine the case when $\theta > \frac{n}{2}$.

Lemma 1. *$f(x^{q^\rho}) = f(x)^{q^\rho}$ when $f(x) = x^{q^\theta+1}$*

Proof. $f(x^{q^\rho}) = (x^{q^\rho})^{q^\theta+1} = x^{(q^\theta+1)q^\rho} = \left(x^{q^\theta+1}\right)^{q^\rho} = f(x)^{q^\rho}$

Consider the special case of Lemma 1 when $\rho = -\theta$. After applying this map to the output of $\mathbf{Df}$, the nonzero terms, originally in the $(\theta, 0)$ and $(0, \theta)$

Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

coordinates, are transported to the $(0, -\theta)$ and $(-\theta, 0)$ coordinates, respectively. This observation leads to the following theorem, revealing that most parameters of PFLASH are provably secure against a differential adversary.

Theorem 1. *Let $f(x) = x^{q^\theta+1}$ be a C^* map, and let M and $\pi x := \sum_{i=0}^{d} x^{q^i}$ be linear. Suppose that f satisfies the symmetric relation:*

$$Df(Ma, \pi x) + Df(\pi a, Mx) = \Lambda_M Df(\pi a, \pi x).$$

If $d < min\{\frac{n}{2} - \theta, |n - 3\theta|, \theta - 1\}$, or if $d < \{\theta - \frac{n}{2}, |2n - 3\theta|, n - \theta - 1\}$, then $M = M_\sigma \circ \pi$ for some $\sigma \in k$.

Proof. Assume $Df(Ma, \pi x) + Df(\pi a, Mx) = \Lambda_M Df(\pi a, \pi x)$ holds true. Then, we have two cases.

1.) $\theta < \frac{n}{2}$
By [9, Theorem 3], we are done.

2.) $\theta > \frac{n}{2}$
Let $\tilde{f}(x) = f(x)^{q^{-\theta}} = f\left(x^{q^{-\theta}}\right)$
We have,

$$Df(Ma, \pi x) + Df(\pi a, Mx) = \Lambda_M Df(\pi a, \pi x)$$
$$[Df(Ma, \pi x) + Df(\pi a, Mx)]^{q^{-\theta}} = [\Lambda_M Df(\pi a, \pi x)]^{q^{-\theta}}$$
$$[Df(Ma, \pi x) + Df(\pi a, Mx)]^{q^{-\theta}} = L_\theta^{-1} \Lambda_M Df(\pi a, \pi x)$$

Let L_θ represent the map that raises terms to the θ^{th} power. We can use the definition of the discrete differential to expand the left hand side of the equation. By linearity, we can distribute the exponent $q^{-\theta}$ to each term. After applying our lemma we get the following,

$$\tilde{f}(Ma+\pi x)+\tilde{f}(Ma)+\tilde{f}(\pi x)+\tilde{f}(\pi a+Mx)+\tilde{f}(\pi a)+\tilde{f}(Mx) = L_\theta^{-1} \Lambda_M Df(\pi a, \pi x)$$

By adding $0 = 2\tilde{f}(0)$ to the left and applying $I = L_\theta L_\theta^{-1}$ to the right we get,

$$D\tilde{f}(Ma, \pi x) + D\tilde{f}(\pi a, Mx) = L_\theta^{-1} \Lambda_M (L_\theta L_\theta^{-1}) Df(\pi a, \pi x)$$

And by the lemma we have,

$$D\tilde{f}(Ma, \pi x) + D\tilde{f}(\pi a, Mx) = L_\theta^{-1} \Lambda_M L_\theta D\tilde{f}(\pi a, \pi x)$$

We now have a relation on $\widetilde{f}(x)$ where $-\theta + d < \frac{n}{2}$. Now we can apply [9, Theorem 3] to conclude that $M = M_\sigma \circ \pi$ for some $\sigma \in k$.

We note that the existence of a differential symmetry on $f \circ \pi$ implies a solution of the equation in Theorem 1 as well as the commutativity of M_σ and π. Since the commutativity of M_σ and π requires that π is L-linear, where $\mathbb{F}_q \subseteq L \subseteq k$ and $\sigma \in L$, for any nontrivial differential symmetry to exist,

Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

$(d, n) > 1$. Thus, there is a most desirable value of d from an efficiency and security standpoint: $d = 1$.

Let us specifically consider this most desired value $d = 1$. Then the only restriction on θ for provable differential security is

$$\theta \in \left(2, \frac{n-1}{3}\right) \cup \left(\frac{n+1}{3}, \frac{n}{2} - 1\right) \cup \left(\frac{n}{2} + 1, \frac{2n-1}{3}\right) \cup \left(\frac{2n+1}{3}, n - 2\right).$$

Furthermore, since $\theta = \frac{n}{2}$ always produces a many-to-one map in any characteristic, the restriction to provably secure parameters for PFLASH eliminates at most four possible values for θ for all extension degrees n.

5 Extension to PFLASH

We now generalize the analysis of the previous section in application to PFLASH. First we derive a heuristic argument for bootstrapping the provable security of the composition $f \circ \pi$ to statistical security for the projected primitive. We then clarify the resistance of PFLASH to analysis as an HFE$^-$ scheme. Finally, we derive security bounds for various PFLASH parameters.

5.1 Differential Analysis

As mentioned in Section 3, proof that differential symmetries do not exist for the central map of a scheme verifies that a differential adversary cannot recover a full rank key. Such a proof does not, however, verify that a differential adversary cannot find a symmetry revealing the extension field multiplicative structure and directly attack the scheme.

To illustrate this principal, imagine a high degree variant of HFE in which the central map has the form $f(x) = x^{q^\theta+1} + \pi_2(Q(x))$ over an extension of degree $2n$, where π_2 is a rank n projection onto the complement of the subfield of size q^n and Q is an arbitrary quadratic. Then any minus variant in which the image of π_2 is the kernel of T is a C^{*-} public key, but one with multiplicative symmetry. In particular, any map L representing muliplication by an element in the intermediate extension of degree n would satisfy

$$D(T \circ f \circ U)(U^{-1}La, x) + D(T \circ f \circ U)(a, U^{-1}Lx) = (L^{q^\theta} + L)D(T \circ f \circ U)(a, x).$$

Thus the minus scheme has a multiplicative symmetry even though the original scheme provably does not. In fact, even more strongly, we have computed functions of the form of f above over a degree 6 extension of $GF(2)$ for which no linear differential symmetry of any form exists, but under projection onto the degree 3 subfield, the *multiplicative* symmetry is exhibited.

In the case of PFLASH, we may attempt the strategy of the previous section for proving security. We may always model the removal of r equations as the application of a polynomial $\pi(x) = \sum_{i=0}^{r} a_i x^{q^i}$ to the central map. If only a few equations are removed, then the analysis proceeds just like in [19], because $f \circ \pi$

Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

is a low rank albeit high degree polynomial. Since no parameters suggested for PSFLASH are near this range, however, this analysis does not apply. When we perform this analysis with $r \approx \frac{n}{3}$ and $f \circ \pi$, however, the methods of the previous section fail to generate a provably secure class of private keys.

Fortunately, there is an easy heuristic argument revealing a simple relationship between symmetries of the central map and symmetries of a map with the minus modifier that shows that symmetry should be statistically no more likely for any minus modified scheme than for the original. Let T' be the minus projection composed with the inclusion mapping with domain $\mathbb{F}_q^{n-r}$ and codomain $\mathbb{K}$. Suppose that $T' \circ f \circ \pi$ has a differential symmetry. Then

$$D(T' \circ f)(\pi a, Mx) + D(T' \circ f)(Ma, \pi x) = \Lambda_M D(T' \circ f)(\pi a, \pi x)$$
$$T' \left[Df(\pi a, Mx) + Df(Ma, \pi x) \right] = \Lambda_M T' Df(\pi a, \pi x).$$

Since the left is clearly in $T'\mathbb{K}$, the right must be as well. Thus, with high probability, that is, when $Span_{a,x}(Df(\pi a, \pi x)) = \mathbb{K}$, we have that $\Lambda_M T'\mathbb{K} = T'\mathbb{K}$. We know from linear algebra that in this case there exists at least one invertible transformation Λ'_M such that $\Lambda_M T' = T'\Lambda'_M$. Therefore, we obtain the relation

$$Df(\pi a, Mx) + Df(Ma, \pi x) = \Lambda'_M Df(\pi a, \pi x) \pmod{ker(T')}. \tag{1}$$

Clearly, this argument is not reversible for any Λ'_M satisfying (1); therefore, we cannot in general conclude that the scheme with the minus modifier inherits any differential symmetry from the central map. On the other hand, satisfying (1) imposes $n - r$ constraints on Λ_M, while the "commuting" of Λ_M with T' imposes another r constraints. Thus, the existence of a symmetry in the minus case imposes the same number of constraints on Λ_M as for the central map and so we expect the probability of the existence of a differential symmetry to be no higher than for the central map.

5.2 Rank Analysis

One can consider PFLASH to be a high degree version of HFE$^-$ by absorbing the projection of the variables into the central map. Notice that the rank of the composition is still only two, thus PFLASH must achieve its security from the minus modifier.

Recently, in [22], a key recovery attack valid for all parameters of HFE$^-$ is presented. For an HFE$^-$ instance with parameters (q, n, D, r), the complexity is noted as $\mathcal{O}(\binom{n+\lceil log_q(D) \rceil + 1}{\lceil log_q(D) \rceil + r + 1}^{\omega})$.

In application to PFLASH, there are two things to note about this attack. First, the attack produces an equivalent HFE$^-$ key, not a pC^{*-} key. This fact may not limit the attack, because it will still recover a central map of rank two of the form $f \circ \pi$ which we may then attack as a pC^* scheme in the manner of [23]. Second, the quantity $\lceil log_q(D) \rceil$ in the complexity estimate is derived from the rank structure that the degree bound of HFE implies, not directly from the

Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

degree bound itself. Thus, the rank of the C^* monomial, which is two, plays the role of $\lceil log_q(D) \rceil$ in the application of the techniques of [22] to PFLASH.

In fact, instances of PFLASH with quite inappropriate but still large parameters can be broken with this method. In particular we note that for a PFLASH$(256, 44, 3, 1)$ that the complexity of the attack is roughly estimated $44^{(3+2+1)\omega} \sim 2^{78}$. For large values of r, however, such as in all parameter sets in [10], this attack is infeasible. For example, the smallest parameters suggested in [10] still resist this attack to dozens of orders of magnitude beyond brute force. Thus, for sensible parameters with r sufficiently large, PFLASH is secure.

5.3 Security Estimates

Now with a refined security analysis, we can eliminate differential attacks for a larger set of parameters, thus doubling the entropy of the key space for PFLASH. In addition, with the complexity estimate of $\mathcal{O}(n^{(r+3)\omega})$ and practical values of r, PFLASH is quite secure against the new attack on HFE$^-$ schemes. In conjunction with the invariant analysis of [10], we conclude that the security of PFLASH is determined by its resistance to algebraic and brute force attacks.

Viewing PFLASH as an HFE$^-$ scheme, we may use the bound in [24] to estimate the degree of regularity of PFLASH. This upper bound can be computed

$$\frac{(q-1)(R+r)}{2} + 2,$$

where R is the rank of the central map; in the case of PFLASH, this quantity is two. Though this is an upper bound, empirical evidence suggests that it is tight for random systems of rank R. Thus the degree of regularity is far too high for practical schemes to be weakened. Furthermore, direct algebraic attacks for large schemes are impractical even with smaller complexity bounds because the space complexity of the best algorithms are too large to be practical.

Therefore, we corroborate the claims of [10] that brute force collision attacks are the greatest threat to PFLASH schemes. The evidence from our increase of the entropy of the key space and the verification that PFLASH resists recent weaknesses revealed in HFE$^-$ suggest the security levels in Table 1 (all of which are in agreement with [10]).

Scheme	Public Key (Bytes)	Security (bits)
PFLASH$(16, 62, 22, 1)$	39,040	80
PFLASH$(16, 74, 22, 1)$	72,124	104
PFLASH$(16, 94, 30, 1)$	142,848	128

Table 1. Security levels for standard parameters of PFLASH

Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

6 Conclusion

The history of PFLASH intersects with most of the major advances in design and cryptanalysis in asymmetric multivariate cryptography. Interestingly, essentially all of the major cryptanalytic techniques that have proven successful in attacking multivariate schemes are relevant for PFLASH, and so any security metric for the scheme must inherently be complex. In spite of all of the tools available to an adversary, PFLASH remains secure.

Our analysis expands upon and complements previous analysis of PFLASH. We verify that the entropy of the key space is not significantly reduced by selecting parameters for which differential security is provable. We further verify security against new developments in rank analysis relevant to schemes employing the minus modifier. We conclude that any attack that fundamentally reduces the security of PFLASH below the brute force bound must include techniques as of yet undeveloped.

In venues for which speed, digest size, storage and power are severe limitations PFLASH seems to be one of the most performant options. When one considers devices in which no public key needs to be transported, such as some applications of smart cards, PFLASH is a leading candidate. In light of the security assurance this analysis provides, PFLASH appears ready for deployment.

References

1. Cryptographic Technology Group: Submission requirements and evaluation criteria for the post-quantum cryptography standardization process. NIST CSRC (2016) http://csrc.nist.gov/groups/ST/post-quantum-crypto/documents/call-for-proposals-final-dec-2016.pdf.
2. Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J. Sci. Stat. Comp. **26, 1484** (1997)
3. Matsumoto, T., Imai, H.: Public Quadratic Polynominal-Tuples for Efficient Signature-Verification and Message-Encryption. In: EUROCRYPT. (1988) 419–453
4. Patarin, J.: Cryptoanalysis of the Matsumoto and Imai Public Key Scheme of Eurocrypt'88. In Coppersmith, D., ed.: CRYPTO. Volume 963 of Lecture Notes in Computer Science., Springer (1995) 248–261
5. Patarin, J.: Hidden Fields Equations (HFE) and Isomorphisms of Polynomials (IP): Two New Families of Asymmetric Algorithms. In: EUROCRYPT. (1996) 33–48
6. Patarin, J., Courtois, N., Goubin, L.: Flash, a fast multivariate signature algorithm. CT-RSA 2001, LNCS **2020** (2001) 297–307
7. Dubois, V., Fouque, P.A., Shamir, A., Stern, J.: Practical Cryptanalysis of SFLASH. In Menezes, A., ed.: CRYPTO. Volume 4622 of Lecture Notes in Computer Science., Springer (2007) 1–12
8. Ding, J., Dubois, V., Yang, B.Y., Chen, C.H.O., Cheng, C.M.: Could SFLASH be Repaired? In Aceto, L., Damgård, I., Goldberg, L.A., Halldórsson, M.M., Ingólfsdóttir, A., Walukiewicz, I., eds.: ICALP (2). Volume 5126 of Lecture Notes in Computer Science., Springer (2008) 691–701

Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

9. Smith-Tone, D.: On the differential security of multivariate public key cryptosystems. In Yang, B.Y., ed.: PQCrypto. Volume 7071 of Lecture Notes in Computer Science., Springer (2011) 130–142
10. Chen, M.S., Yang, B.Y., Smith-Tone, D.: Pflash - secure asymmetric signatures on smart cards. Lightweight Cryptography Workshop 2015 (2015) http://csrc.nist.gov/groups/ST/lwc-workshop2015/papers/session3-smith-tone-paper.pdf.
11. Moody, D., Perlner, R.A., Smith-Tone, D.: An asymptotically optimal structural attack on the ABC multivariate encryption scheme. [25] 180–196
12. Moody, D., Perlner, R.A., Smith-Tone, D. In: Key Recovery Attack on the Cubic ABC Simple Matrix Multivariate Encryption Scheme. Springer (2017)
13. Kipnis, A., Shamir, A.: Cryptanalysis of the hfe public key cryptosystem by relinearization. Advances in Cryptology - CRYPTO 1999, Springer **1666** (1999) 788
14. Bettale, L., Faugère, J., Perret, L.: Cryptanalysis of hfe, multi-hfe and variants for odd and even characteristic. Des. Codes Cryptography **69** (2013) 1–52
15. Patarin, J., Goubin, L., Courtois, N.: C^*_{-+} and HM: variations around two schemes of t. matsumoto and h. imai. In Ohta, K., Pei, D., eds.: Advances in Cryptology - ASIACRYPT '98, International Conference on the Theory and Applications of Cryptology and Information Security, Beijing, China, October 18-22, 1998, Proceedings. Volume 1514 of Lecture Notes in Computer Science., Springer (1998) 35–49
16. Patarin, J., Goubin, L., Courtois, N.: Improved algorithms for isomorphisms of polynomials. In Nyberg, K., ed.: Advances in Cryptology - EUROCRYPT '98, International Conference on the Theory and Application of Cryptographic Techniques, Espoo, Finland, May 31 - June 4, 1998, Proceeding. Volume 1403 of Lecture Notes in Computer Science., Springer (1998) 184–200
17. Berlekamp, E.R.: Factoring polynomials over large finite fields. Mathematics of Computation **24** (1970) pp. 713–735
18. Ding, J., Hu, L., Nie, X., Li, J., Wagner, J. In: High Order Linearization Equation (HOLE) Attack on Multivariate Public Key Cryptosystems. Springer Berlin Heidelberg, Berlin, Heidelberg (2007) 233–248
19. Daniels, T., Smith-Tone, D.: Differential properties of the HFE cryptosystem. [25] 59–75
20. Cartor, R., Gipson, R., Smith-Tone, D., Vates, J.: On the differential security of the hfev- signature primitive. In Takagi, T., ed.: Post-Quantum Cryptography - 7th International Workshop, PQCrypto 2016, Fukuoka, Japan, February 24-26, 2016, Proceedings. Volume 9606 of Lecture Notes in Computer Science., Springer (2016) 162–181
21. Smith-Tone, D.: Properties of the discrete differential with cryptographic applications. In Sendrier, N., ed.: PQCrypto. Volume 6061 of Lecture Notes in Computer Science., Springer (2010) 1–12
22. Vates, J., Smith-Tone, D.: Key recovery attack for all parameters of hfe-. In Current Submission (2017)
23. Billet, O., Macario-Rat, G.: Cryptanalysis of the square cryptosystems. ASIACRYPT 2009, LNCS **5912** (2009) 451–486
24. Ding, J., Kleinjung, T.: Degree of regularity for HFE-. IACR Cryptology ePrint Archive **2011** (2011) 570
25. Mosca, M., ed.: Post-Quantum Cryptography - 6th International Workshop, PQCrypto 2014, Waterloo, ON, Canada, October 1-3, 2014. Proceedings. Volume 8772 of Lecture Notes in Computer Science., Springer (2014)

Cartor, Ryann; Smith-Tone, Daniel.
"An Updated Security Analysis of PFLASH."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

Key Recovery Attack for All Parameters of HFE-

Jeremy Vates[1] and Daniel Smith-Tone[1,2]

[1]Department of Mathematics, University of Louisville,
Louisville, Kentucky, USA
[2]National Institute of Standards and Technology,
Gaithersburg, Maryland, USA

jeremy.vates@louisville.edu, daniel.smith@nist.gov

Abstract. Recently, by an interesting confluence, multivariate schemes with the minus modifier have received attention as candidates for multivariate encryption. Among these candidates is the twenty year old HFE$^-$ scheme originally envisioned as a possible candidate for both encryption and digital signatures, depending on the number of public equations removed.

HFE has received a great deal of attention and a variety of cryptanalyses over the years; however, HFE$^-$ has escaped these assaults. The direct algebraic attack that broke HFE Challenge I is provably more complex on HFE$^-$, and even after two decades HFE Challenge II is daunting, though not achieving a security level we may find acceptable today. The minors modeling approach to the Kipnis-Shamir (KS) attack is very efficient for HFE, but fails when the number of equations removed is greater than one. Thus it seems reasonable to use HFE$^-$ for encryption with two equations removed.

This strategy may not be quite secure, however, as our new approach shows. We derive a new key recovery attack still based on the minors modeling approach that succeeds for all parameters of HFE$^-$. The attack is polynomial in the degree of the extension, though of higher degree than the original minors modeling KS-attack. As an example, the complexity of key recovery for HFE$^-(q = 31, n = 36, D = 1922, a = 2)$ is 2^{52}. Even more convincingly, the complexity of key recovery for HFE Challenge-2, an HFE$^-(16, 36, 4352, 4)$ scheme, is feasible, costing around 2^{67} operations. Thus, the parameter choices for HFE$^-$ for both digital signatures and, particularly, for encryption must be re-examined.

Key words: Multivariate Cryptography, HFE, encryption, MinRank, Q-rank

1 Introduction

In the 1990s, several important developments in the history of asymmetric cryptography occured. Among these discoveries, and of the greatest significance to

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

forward-thinking cryptographers, was the discovery by Peter Shor of polynomial time algorithms for factoring and computing discrete logarithms on a quantum computer, see [1]. In the years since that time, we have witnessed quantum computing become a reality, while *large-scale* quantum computing has transmogrified from a dream into what many of us now see as an inevitability, if not an impending phenomenon. The call for proposals by the National Institute of Standards and Technology (NIST), see [2], charges our community with the task of protecting the integrity and confidentiality of our critical data in this time of tremendous change.

The 1990s also beheld an explosive development in public key technologies relying on mathematics of a less linear character than number theory. In particular, multivariate public key cryptography (MPKC) produced numerous schemes for public key encryption and digital signatures in the late 1990s. These schemes further fuelled the development of computational algebraic geometry, and seem to have inspired the advancement of some of the symbolic algebra techniques we now apply to all areas of post-quantum cryptography, that is, cryptography designed with quantum computers in mind.

Armed with new tools and a more developed theory, many multivariate schemes were cryptanalyzed; in particular, secure multivariate encryption seemed particularly challenging. The purpose of this disquisition is to cryptanalyze an old digital signature scheme that has been repurposed to achieve multivariate encryption.

1.1 Recent History

While the ancestor of all of the "large structure" schemes is the C^* scheme of Matsumoto and Imai, see [3], the more direct parent of multivariate encryption schemes of today is HFE, see [4]. The idea behind such systems is to define a large associative algebra over a finite field and utilize its multiplication to construct maps that are quadratic when expressed over the base field.

There have been many proposals in this area in the last five years. The Simple Matrix Schemes, see [5] for the quadratic version and [6] for the cubic version, are constructed via multiplication in a large matrix algebra over the base field. ZHFE, see [7] and Extension Field Cancellation, see [8], just as HFE, utilize the structure of an extension field in the derivation of their public keys.

Many of these "large structure" schemes have effective cryptanalyses that either break or limit the efficiency of the schemes. HFE, in its various iterations, has been cryptanalyzed via direct algebraic attack, see [9], via an attack exploiting Q-rank known as the Kipnis-Shamir, or KS, attack, see [10], and via a fusion of these techniques utilizing an alternative modeling of the Q-rank property, see [11]. The Quadratic Simple Matrix Scheme is made less efficient for parameters meeting NIST's current suggested security levels in [12], while the Cubic Simple Matrix Scheme is broken for such parameters in [13]. In addition, a low Q-rank property for ZHFE is discovered in [14] which calls in to question the security of the scheme. In light of such an array of cryptanalyses for multivariate encryption

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

schemes, the question of whether the correct strategy is being employed is very relevant.

Interestingly, at PQCRYPTO 2016 and the winter school prior to the conference, three independent teams of researchers in MPKC related the same idea: the idea of using the minus modifier in encryption. In fairness, the concept of using the minus modifier in encryption is not new; it was suggested as early as in the proposal of HFE. The convergence on this strategy is surprising because it is common knowledge that either the number of equations removed is too large for effective, or even fault-tolerant, encryption, or that the scheme must have parameters that are too large for the system to be efficient. The three techniques are presented in the articles [14] and [8] and in the presentation [15].

While both of the techniques in [14] and [8] are very new schemes, HFE$^-$ has been well studied for over twenty years. Using HFE$^-$ for encryption is more complicated than using the scheme for digital signatures, so careful review of theory is critical for this application.

1.2 Previous Analysis

There are a few results in the literature that are relevant in the analysis of HFE$^-$. These articles address the security of the scheme against algebraic, differential and rank attacks.

In [16], the degree of regularity for the public key of HFE$^-$ schemes is derived. The result shows that the upper bound on the degree of regularity of the public key when a equations is removed is about $\frac{a(q-1)}{2}$ higher than the same bound for a comparable HFE scheme over $GF(q)$.

In [17], information theoretic proofs of security against differential adversaries are derived for HFE$^-$. The consequence of this work is that attacks of the flavor of the attack on SFLASH, see [18], using symmetry and attacks in the manner of the attack on the Simple Matrix Scheme, see [12], exploiting invariants are not relevant for HFE$^-$.

In the other direction, in [11, Section 8.1], an attack on weak parameters of HFE$^-$ with asymptotic complexity of $\mathcal{O}(n^{(\lceil log_q(D)\rceil+1)\omega})$ is derived, where n is the degree of the extension, D is the degree bound for HFE and ω is the linear algebra constant. The caveat here is that the attack is only successful against HFE$^-$ if only a single equation is removed. This restriction on the attack technique is fundamental and is due to theory, not computational feasibility. The existence of the attack, however, implies that at least two equations must be removed for reasonable parameters, and thus q must be quite small for encryption.

1.3 Our Contribution

We present a key recovery attack on HFE$^-$ that works for any HFE$^-$ public key. The attack is based on the Q-rank of the public key instead of the Q-rank of the private central map as in [11].

The attack works by performing key extraction on a related HFE scheme and then converting the private key of the related scheme into an equivalent private

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

key for the HFE$^-$ scheme. The complexity of the attack is dominated by the HFE key extraction phase and is on the order of $\mathcal{O}(\binom{n+\lceil log_q(D)\rceil+1}{\lceil log_q(D)\rceil+a+1}^{\omega})$, where D is the degree bound of the central HFE polynomial, a is the number of removed equations and ω is the linear algebra constant, for all practical parameters. We note that this value implies that the minus modification of HFE adds at most $a\omega log_2(n)$ bits of security for any parameters, though we find that it is much less for many practical parameters.

1.4 Organization

The paper is organized as follows. In the next section, we present isomorphisms of polynomials and describe the structure of HFE and HFE$^-$. The following section reviews the Q-rank of ideals in polynomial rings and discusses invariant properties of Q-rank and min-Q-rank. In section 4, we review more carefully the previous cryptanalyses of HFE and HFE$^-$ that are relevant to our technique. The subsequent section contains our cryptanalysis of HFE$^-$. Then, in section 6, we conduct a careful complexity analysis of our attack, followed by our experimental results in the following section. Finally, we conclude, noting the affect these results have on parameter selection for HFE$^-$.

2 HFE Variants

Numerous multivariate cryptosystems fall into a category known as "big field" schemes exploiting the vector space structure of a degree n extension $\mathbb{K}$ over $\mathbb{F}_q$. Let $\phi : \mathbb{F}_q^n \to \mathbb{K}$ be an $\mathbb{F}_q$-vector space isomoprhism. Since a generator of $Gal_{\mathbb{F}_q}(\mathbb{K})$ is the Frobenius automorphism, $x \mapsto x^q$, for every monomial map of the form $f(x) = x^{q^i+q^j}$ in $\mathbb{K}$, $\phi^{-1} \circ f \circ \phi$ is a vector-valued quadratic function over $\mathbb{F}_q$. By counting, one can see that any vector-valued quadratic map on $\mathbb{F}_q^n$ is thusly isomorphic to a sum of such monomials. Consequently, any quadratic map f over $\mathbb{K}$ can be written as a vector-valued map, F, over $\mathbb{F}_q$. Throughout this work, for any map $g : \mathbb{K} \to \mathbb{K}$, we denote by G the quantity $\phi^{-1} \circ g \circ \phi$.

This equivalence allows us to construct cryptosystems in conjunction with the following concept, the of isomorphisms of polynomials.

Definition 1 *Two vector-valued multivariate polynomials F and G are said to be* isomorphic *if there exist two affine maps T, U such that $G = T \circ F \circ U$.*

The equivalence and isomorphism marry in a method commonly referred to as the butterfly construction. Given a vector space isomorphism $\phi : \mathbb{F}_q^n \to \mathbb{K}$ and an efficiently invertible map $f : \mathbb{K} \to \mathbb{K}$, we compose two affine transformations $T, U : \mathbb{F}_q^n \to \mathbb{F}_q^n$ in order to obscure our choice of basis for the input and output. This construction generates a vector-valued map $P = T \circ \phi^{-1} \circ f \circ \phi \circ U$.

$$\begin{array}{ccccccc} & & \mathbb{K} & \xrightarrow{f} & \mathbb{K} & & \\ & & \uparrow{\scriptstyle \phi} & & \downarrow{\scriptstyle \phi^{-1}} & & \\ \mathbb{F}_q^n & \xrightarrow{U} & \mathbb{F}_q^n & \xrightarrow{F} & \mathbb{F}_q^n & \xrightarrow{T} & \mathbb{F}_q^n \end{array}$$

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

The Hidden Field Equation Scheme was first introduced by Patarin in [4]. This scheme is an improvement on the well known C^* construction of [19], where a general polynomial with degree bound D is used in place of the C^*'s central monomial map.

Explicitly, one chooses a quadratic map $f : \mathbb{K} \to \mathbb{K}$ of the form:

$$f(x) = \sum_{\substack{i \leq j \\ q^i + q^j \leq D}} \alpha_{i,j} x^{q^i + q^j} + \sum_{\substack{i \\ q^i \leq D}} \beta_i x^{q^i} + \gamma,$$

where the coefficients $\alpha_{i,j}, \beta_i, \gamma \in \mathbb{K}$ and the degree bound D is sufficiently low for efficient inversion.

The public key is computed as $P = T \circ F \circ U$. Inversion is accomplished by first taking a cipher text $y = P(x)$, computing $v = T^{-1}(y)$, solving $\phi(v) = f(u)$ for u via the Berlekamp algorithm, see [20], and then recovering $x = U^{-1}(\phi^{-1}(u))$.

HFE$^-$ uses the HFE primitive f along with a projection Π that removes a equations from the public key. The public key is $P_\Pi = \Pi \circ T \circ F \circ U$.

3 Q-Rank

A critical quantity tied to the security of big field schemes is the Q-rank (or more correctly, the min-Q-rank) of the public key.

Definition 2 *The Q-rank of any quadratic map $f(\overline{x})$ on $\mathbb{F}_q^n$ is the rank of the quadratic form $\phi^{-1} \circ f \circ \phi$ in $\mathbb{K}[X_0, \ldots, X_{n-1}]$ via the identification $X_i = \phi(\overline{x})^{q^i}$.*

Quadratic form equivalence corresponds to matrix congruence, and thus the definition of the rank of a quadratic form is typically given as the minimum number of variables required to express an equivalent quadratic form. Since congruent matrices have the same rank, this quantity is equal to the rank of the matrix representation of this quadratic form, even in characteristic 2, where the quadratics x^{2q^i} are additive, but not linear for $q > 2$.

Q-rank is invariant under one-sided isomorphisms $f \mapsto f \circ U$, but is not invariant under isomorphisms of polynomials in general. The quantity that is often meant by the term Q-rank, but more properly called min-Q-rank, is the minimum Q-rank among all nonzero linear images of f. This min-Q-rank is invariant under isomorphisms of polynomials and is the quantity relevant for cryptanalysis.

4 Previous Cryptanalysis of HFE

HFE has been cryptanalyzed via a few techniques in the over twenty years since its inception. The principal analyses are the Kipnis-Shamir (KS) attack of [10], the direct algebraic attack of [9], and the minors modeling approach of the KS-attack of [11].

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

The KS-attack is a key recovery attack exploiting the fact that the quadratic form representing the central map F over $\mathbb{K}$ is of low rank. Specfically, considering an odd characteristic case, we may write the homogeneous quadratic part of F as

$$\begin{bmatrix} x & x^q & \cdots & x^{q^{n-1}} \end{bmatrix} \begin{bmatrix} \alpha_{0,0} & \alpha'_{0,1} & \cdots & \alpha'_{0,d-1} & 0 \cdots 0 \\ \alpha'_{0,1} & \alpha_{1,1} & \cdots & \alpha'_{1,d-1} & 0 \cdots 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \ddots \vdots \\ \alpha'_{0,d-1} & \alpha'_{1,d-1} & \cdots & \alpha_{d-1,d-1} & 0 \cdots 0 \\ 0 & 0 & \cdots & 0 & 0 \cdots 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \ddots \vdots \\ 0 & 0 & \cdots & 0 & 0 \cdots 0 \end{bmatrix} \begin{bmatrix} x \\ x^q \\ \vdots \\ x^{q^{n-1}} \end{bmatrix},$$

where $\alpha'_{i,j} = \frac{1}{2}\alpha_{i,j}$ and $d = \lceil log_q(D) \rceil$. Using polynomial interpolation, the public key can be expressed as a quadratic polynomial G over a degree n extension, and it is known that there is a linear map T^{-1} such that $T^{-1} \circ G$ has rank d, thus there is a rank d matrix that is a $\mathbb{K}$-linear combination of the Frobenius powers of G. This turns recovery of the transformation T into the solution of a MinRank problem over $\mathbb{K}$.

In contrast to the KS-attack, the Gröbner basis attack of Faugère in [9], is a direct algebraic attack on HFE using the F4 Gröbner basis algorithm. The attack succeeds in breaking HFE Challenge 1, see [4]. The success is primarily due to the fact that the coefficients of the central map in HFE Challenge 1 were very poorly chosen. The scheme is defined over $GF(2)$ and uses only a degree 80 extension. Thus the scheme fails to brute force analysis with complexity at worst 2^{80}. The very small base field drastically limits the number of monomials of degree d and makes Gröbner basis techniques extremely powerful.

The key recovery attack of [11] combines these two approaches with some significant improvements. First, via a very clever construction, it is shown that a $\mathbb{K}$-linear combination of the *public* polynomials has low rank as a quadratic form over $\mathbb{K}$. Second, setting the unknown coefficients in $\mathbb{K}$ as variables, the polynomials representing $(d+1) \times (d+1)$ minors of such a linear combination, which must be zero due to the rank property, reside in $\mathbb{F}_q[t]$. Thus a Gröbner basis needs to be computed over $\mathbb{F}_q$ and the variety computed over $\mathbb{K}$. This technique is called minors modeling and dramatically improves the efficiency of the KS-attack. The complexity of the KS-attack with minors modeling is asymptotically $\mathcal{O}(n^{(\lceil log_q(D) \rceil + 1)\omega})$, where $2 \leq \omega \leq 3$ is the linear algebra constant.

The effect of the minus modifier on these schemes is worthy of notice. For the direct algebraic attack, the fact that the degree of regularity for a subsystem is lower bounded by the degree of regularity of the entire system shows that the minus modifier introduces no weakness. In particular, the degree of regularity of HFE$^-$ is investigated in [16] where it is shown that the best known upper bound on the degree of regularity for HFE increases with each equation removed. For the KS-attack with either the original modeling or the minors modeling, it suffices to note that though there is a method of reconstructing a single removed equation,

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

it is not true in general that there is a rank $\lceil log_q(D) \rceil$ $\mathbb{K}$-quadratic form in the linear span of the public key; thus, these attacks fail if the number of equations removed is at least two.

5 Key Recovery for HFE$^-$

In this section we explain our key recovery attack on HFE$^-$. The process is broken down into two main steps. The first is finding a related HFE instance of the HFE$^-$ public key. This related instance will then be the focus. Then we discuss how to systematically solve for an equivalent private key for the orignal HFE$^-$ scheme.

5.1 Reduction of HFE$^-$ to HFE

Recall that by imposing the field equations we may always assume that any affine variety associated with HFE is contained in the finite field $\mathbb{K}$. Then we may use the following definition.

Definition 3 (see Definition 1, [17]) *The* minimal polynomial*, of the algebraic set $V \subseteq \mathbb{K}$ is given by*

$$\mathcal{M}_V := \prod_{v \in V} (x - v).$$

Equivalently, $\mathcal{M}_V$ is the generator of the principal ideal $I(V)$, the intersection of the maximal ideals $\langle x - v \rangle$ for all $v \in V$.

Recall that the public key of an HFE$^-$ scheme is constructed by truncating a full rank linear combination of the central polynomials. That is, with parenthetical emphasis, $P = \Pi(T \circ F \circ U)$. We now show that this singular linear transformation can be transported "past" the invertible transformation T and "absorbed" by the central map.

Lemma 1 *Let $\Pi \circ T$ be a corank a linear transformation on $\mathbb{F}_q^n$. There exist both a nonsingular linear transformation S and a degree q^a linear polynomial π such that $\Pi \circ T = S \circ \phi^{-1} \circ \pi \circ \phi$.*

Proof. Let V be the kernel of $\Pi \circ T$ and let $\pi = \mathcal{M}_V$. Note that $|V| = q^a$, thus $\mathcal{M}_V(x)$ has degree q^a and is of the form

$$x^{q^a} + c_{a-1}x^{q^{a-1}} + \cdots + c_1 x^q + c_0 x \text{ where } c_i \in \mathbb{K} \tag{1}$$

Now let $B_V = \{b_{n-a}, b_{n-a+1}, \ldots, b_{n-1}\}$ be a basis for V and extend this to a basis $B = \{b_0, \ldots, b_{n-1}\}$ of $\mathbb{F}_q^n$. Let M be the matrix transporting from the standard basis to B. Clearly the matrix representations of both $M^{-1}(\Pi \circ T)M$ and $M^{-1}(\phi^{-1} \circ \pi \circ \phi)M$ have the last a columns of 0.

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

Observe that there exist invertible matrices A and A', corresponding to row operations, such that both $AM^{-1}(\Pi \circ T)M$ and $A'M^{-1}(\phi^{-1} \circ \pi \circ \phi)M$ are in reduced echelon form; that is:

$$AM^{-1}(\Pi \circ T)M = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & 0 \end{array}\right] = A'M^{-1}(\phi^{-1} \circ \pi \circ \phi)M \tag{2}$$

Solving for $\Pi \circ T$, we obtain

$$MA^{-1}A'M^{-1}(\phi^{-1} \circ \pi \circ \phi) = \Pi \circ T. \tag{3}$$

Let $S = MA^{-1}A'M^{-1}$ and the lemma is proven.

Lemma 1 suggests the possibility of considering an HFE$^-$ public key as a full rank basis for the low rank image of a quadratic map. In fact, Lemma 1 is powerful enough to maintain a low degree bound for this map.

Theorem 1 *Let P be the public key of an HFE$^-(q, n, D, a)$ scheme. Then*

$$P' := P \| \{p_{n-a}, p_{n-a+1} \ldots, p_{n-1}\}$$

is a public key of an HFE$(q, n, q^a D)$ scheme for any choice of $p_i \in Span(P)$ where $i \in \{n-a, n-a+1, \ldots, n-1\}$.

Proof. Let P be a public key for HFE$^-(q, n, D, a)$. Observe that P has the following form, $P = \Pi \circ T \circ F \circ U$ where $T, U : \mathbb{F}_q^n \to \mathbb{F}_q^n$ are affine transformations applied to an HFE(q, n, D) central map F. Let Π' be the natural embedding of Π as a linear map $\mathbb{F}_q^n \to \mathbb{F}_q^n$ obtained by composing the inclusion mapping $\mathbb{F}_q^{n-a} \hookrightarrow \mathbb{F}_q^n$. By Lemma 1, we can rewrite $P||\{0, 0, \ldots 0\}$ in the following way:

$$P||\{0, 0, \ldots 0\} = \Pi' \circ T \circ \phi^{-1} \circ f \circ \phi \circ U = S \circ \phi^{-1} \circ (\pi \circ f) \circ \phi \circ U, \tag{4}$$

where S is nonsingular and π is a linear polynomial of degree q^a.

Observe that $P||\{0, 0, \ldots 0\}$ now has the structure of an HFE$(q, n-a, q^a D)$, since the degree bound is increased by a factor of q^a; that is, $deg(\pi(f)) = deg(\pi)deg(f)$. Finally, construct $P' = P||\{p_{n-a}, p_{n-a+1}, \ldots, p_{n-1}\}$ where $p_i \in Span(P)$, possibly 0. Since the composition A of elementary row operations produces P' from $P||\{0, 0 \ldots, 0\}$, we obtain an HFE$(q, n, q^a D)$ key, $(AS, \pi \circ f, U)$.

Theorem 1 indicates that HFE$^-$, in some sense, *is* HFE with merely a slightly higher degree bound. Thus it is sensible to discuss recovering an equivalent key for an instance of HFE$^-$ as an HFE scheme. We can, in fact, do more and recover an equivalent HFE$^-$ key.

5.2 Key Recovery

Any HFE key recovery oracle $\mathcal{O}$, when given a public key P of an HFE instance recovers a private key of HFE "shape." By Theorem 1, such an oracle can recover a private key for the augmented public key P' which is also of HFE shape. We now show, however, that in this case, the key derived from $\mathcal{O}$ must preserve more structure.

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

Theorem 2 *Let P be a public key for an instance of HFE$^-(q,n,D,a)$ and let $P' = P \| \{p_{n-a}, p_{n-a+1} \dots, p_{n-1}\}$ be a corresponding HFE(q,n,q^aD) public key. Further, let (T', f', U') be any private key of P'. Then the representation of f' as a quadratic form over $\mathbb{K}$ is block diagonal of the form:*

$$\mathbf{F}' = \begin{bmatrix} F_1' & 0 \\ 0 & 0 \end{bmatrix}, \tag{5}$$

where $F_1' = [f_{i,j}]_{i,j}$ is $(\lceil log_q(D) \rceil + a) \times (\lceil log_q(D) \rceil + a)$ and has the property that $f_{i,j} = 0$ if $|i-j| \geq \lceil log_q(D) \rceil$. That is, F_1' has only a diagonal "band" of nonzero values of width $2\lceil log_q(D) \rceil - 1$.

Proof. Let (T, f, U) be a private key for P as an instance of HFE$^-(q,n,D,a)$. By Theorem 1, one private key of P' has the form (T', f', U') where $f' = \pi \circ f$ and

$$\pi(x) = \sum_{i=0}^{a} b_i x^{q^i}.$$

Therefore,

$$\begin{aligned} f'(x) = \pi \circ f(x) &= \sum_{\substack{i \leq j \\ q^i + q^j \leq D}} \sum_{\ell=0}^{a} b_\ell \alpha_{i,j}^{q^\ell} x^{q^{i+\ell} + q^{j+\ell}} \\ &= \sum_{\substack{i,j \leq \lceil log_q(D) + a \rceil \\ |i-j| < \lceil log_q(D) \rceil}} f_{i,j} x^{q^i + q^j} \end{aligned}$$

Thus there exists one private key of the required form.

Denote by Frob_i the map raising all entries of a vector to the power q^i and let M_b be the linear map $x \mapsto bx$ for $b \in \mathbb{K}$. By the homogeneous case of [11, Theorem 4], for any second private key (T'', f'', U'') of P', we have for some integer $0 \leq k < n$ and for some $a, b \in \mathbb{K}$ that

$$F'' = \mathrm{Frob}_k \circ M_b \circ F' \circ M_a \circ \mathrm{Frob}_{n-k}.$$

It is straightforward to check that the representation of F'' as a quadratic form has the shape of (5) with nonzero entries restricted to $|i-j| < \lceil log_q(D) \rceil$.

Armed with Theorem 2, we are prepared to perform a full key recovery for an instance $P = \Pi \circ T \circ \phi^{-1} \circ f \circ \phi \circ U$ of HFE$^-$. The strategy is simple. By way of Theorem 1, there exists an HFE instance with an equivalent public key. That is, there exists a $P' = T' \circ \phi^{-1} \circ f' \circ \phi \circ U'$ with T', U' invertible, f' of degree bounded by q^aD, and where the first $n-a$ public equations in P' form P while the remaining a equations are in the $\mathbb{F}_q$-linear span of P. We perform a key recovery on this instance of HFE via the best known attack, the KS-attack with minors modeling of [11]. Finally, we can recover a central map of degree bound D by way of the following theorem.

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

Theorem 3 *Let (T, f, U) be an HFE$^-(q, n, D, a)$ private key and let (T', f', U') be an equivalent HFE$(q, n, q^a D)$ key. Then a linear map T'' and a quadratic map f'' of degree bound D such that $\Pi \circ T'' \circ \phi^{-1} \circ f'' \circ \phi \circ U' = \Pi \circ T \circ \phi^{-1} \circ f \circ \phi \circ U$ can be recovered by solving two linear systems, the first of dimension a and the second of dimension $\binom{\lceil log_q(D) \rceil}{2}$.*

Proof. Let (T, f, U) be an HFE$^-(q, n, D, a)$ private key and let (T', f', U') be an equivalent HFE$(q, n, q^a D)$ key. Let $\mathbf{F}'$ denote the matrix representation of f' as a quadratic form over $\mathbb{K}$. Finally, let $d = \lceil log_q(D) \rceil$. By Theorem 2, $\mathbf{F}'$ has the diagonal band shape of width $2d - 1$. From the proof of Theorem 1, there exists a linear map $\pi(x) = \sum_{i=0}^{a} p_i x^{q^i}$, where we may sacrifice monicity and insist $p_0 = 1$ for convenience, and a degree bound D quadratic function f'' such that the composition $\pi(f'') = f'$. Let $\mathbf{F}'' = (f''_{i,j})_{i,j}$ and $\widehat{\pi\mathbf{F}''}$ denote the matrix representations of f'' and $\pi \circ f''$, respectively, as quadratic forms over $\mathbb{K}$. Then we have $\mathbf{F}' = \widehat{\pi\mathbf{F}''}$. The (i, j)th entry of $\widehat{\pi\mathbf{F}''}$ is of the form

$$\sum_{\ell=0}^{a} p_\ell (f''_{i-\ell, j-\ell})^{q^\ell},$$

thus, since $\mathbf{F}'$ is known, we obtain a bilinear system of equations in the unknowns p_i and $f''_{i,j}$.

The insistence that $p_0 = 1$ allows us to recover the values of $f''_{0,j}$ without cost. We then note that due to the fact that $f''_{i,j} = 0$ when $max\{i, j\} \geq d$, the $(i, i + d - 1)$th coefficients of $\widehat{\pi\mathbf{F}''}$ are $p_i (f''_{0,d-1})^{q^i}$ for $0 \leq i \leq a$. Thus, since $f''_{0,d-1}$ is known, we obtain *a linear* system of equations $f'_{i,i+d-1} = p_i (f''_{0,d-1})^{q^i}$ for $1 \leq i \leq a$ in the unknowns p_i, and can therefore solve for π. Once the values of p_i are known, the system of equations becomes linear in $f''_{i,j}$ for $i > 0$. Solving for the remaining unknown values can be done simply with the upper triangular segment from $(1, 1)$ to $(d - 1, d - 1)$, of size $\binom{d}{2}$.

To illustrate the attack in all of its steps, we have prepared a toy example in Appendix A.

6 Complexity of Attack

In this section we derive a tight complexity estimate of the key recovery attack for HFE$^-$ of Section 5. First, we expound upon the relationship between the computational complexity of of HFE$^-$ key recovery and that of HFE key recovery.

Theorem 4 *Let $\mathcal{O}$ be an HFE key recovery oracle that can recover a private key for any instance of HFE(q, n, D) in time $t(q, n, D)$. Then an equivalent HFE key for HFE$^-(q, n, D, a)$ can be recovered by $\mathcal{O}$ in time $t(q, n, q^a D)$.*

Proof. Let P be the public key for an instance of HFE$^-(q, n, D, a)$. Then make the following construction: $P' = P \| \{p_{n-a}, p_{n-a+1} \ldots, p_{n-1}\}$ where $p_i \in Span(P)$.

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

$\lceil log_q(D) \rceil$	2	3	4	5	6
d_{reg}	5	6	7	8	9

Table 1. The degree of regularity of the system arising from minors modeling on $\mathrm{HFE}^-(q, n, D, a)$ with $a = 2$, $\lceil log_q(D) \rceil$ as indicated, and n sufficiently large.

By Theorem 1, P' is an instance of $HFE(q, n, q^a D)$. Thus $\mathcal{O}$ recovers an equivalent HFE key in time $t(q, n, q^a D)$.

Thus, the complexity of deriving a key for the associated HFE scheme is bounded by the complexity of the best key recovery algorithm for HFE with a degree bound a factor of q^a larger. By Theorem 3, converting the recovered specially structured $\mathrm{HFE}(q, n, q^a D)$ key into an equivalent $\mathrm{HFE}^-(q, n, D, a)$ scheme is of complexity on the order of $\lceil log_q(D) \rceil^{2\omega}$. Since this quantity is very small, the key conversion is instantaneous for all practical parameters. Hence the complexity of the entire attack is bounded by $t(q, n, q^a D)$ from Theorem 4.

We can achieve a tight practical bound when specifying the oracle. Using the minors modeling approach to the KS-attack, which is the currently most successful algebraic attack on HFE, we can accurately determine the complexity of HFE^- key recovery. Just as in HFE, the complexity of the attack is dominated by the MinRank calculation.

Proposition 1 *Let $d = \lceil log_q(D) \rceil$. The degree of regularity of the MinRank instance with parameters $(n, a + d, n - a)$ arising from minors modeling on the public key of $HFE^-(q, n, D, a)$ is the degree of the first negative term in the series*

$$H_r(t) = (1 - t)^{(n-a-d)^2-n+a} \frac{det(\mathbf{A_{a+d}})}{t^{\binom{a+d}{2}}},$$

where $\mathbf{A_{a+d}}$ is the $(a + d) \times (a + d)$ matrix whose (i, j)-th entry is

$$a_{i,j} = \sum_{\ell=0}^{n-max\{i,j\}} \binom{n-i}{\ell} \binom{n-j}{\ell} t^\ell.$$

Proposition 1 follows immediately from [21, Corollary 3], which relies on the genericity conjecture [21, Conjecture 1] which is related to Fröberg's Conjecture, see [22]. With this proposition we can derive the degree of regularity for the MinRank instances for larger systems as well. Focusing on the case in which $a = 2$ we summarize the data in Table 1.

From these data we are prepared to make the following conjecture:

Conjecture 1 *The degree of regularity of the MinRank instance with parameters $(n, a+d, n-a)$ arising from minors modeling on the public key of $HFE^-(q, n, D, a)$ is*

$$d_{reg} = a + d + 1,$$

for all sufficiently large n.

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

Finally, under the above conjecture, we derive the complexity of our key recovery technique for HFE^-.

Theorem 5 *The complexity of key recovery for $HFE^-(q, n, D, a)$ using the minors modeling variant of the KS-attack is*

$$\mathcal{O}\left(\binom{n-a+d_{reg}}{d_{reg}}^{\omega}\right) \sim \mathcal{O}\left(\binom{n+\lceil log_q(D)\rceil+1}{\lceil log_q(D)\rceil+a+1}^{\omega}\right).$$

7 Experimental Results

We ran a series of experiments with Magma, see [23], on a 3.2 GHz Intel® Xeon™ CPU, testing the attack for a variety of values of q, n and D. In all cases, a valid private key was recovered. Table 2 summarizes some of our results for the asymptotically most costly step, the MinRank attack. The data support our complexity estimate of $\mathcal{O}\left(\binom{n+\lceil log_q(D)\rceil+1}{\lceil log_q(D)\rceil+a+1}^{\omega}\right)$.

a	$n=8$	$n=9$	$n=10$	$n=11$	$n=12$
0	37	94	235	575	1269
1	166	535	1572	3653	3374
2	764	1254	6148	26260	97838

Table 2. Average time (in ms) for 100 instances of the MinRank attack on $\text{HFE}^-(3, n, 3^2+3^2=18, a)$ for various values of n and a.

8 Conclusion

The HFE^- scheme is a central figure in the development of multivariate cryptography over the last twenty years, inspiring the development of several cryptostystems. Finally, the scheme has revealed a vulnerability significant enough to affect the necessary parameters for the signature algorithm. For example, our attack breaks the $\text{HFE}^-(31, 36, 1922, 2)$ primitive in about 2^{52} operations. For an even characteristic example, consider HFE Challenge-2, $\text{HFE}^-(16, 36, 4352, 4)$. Our attack breaks HFE Challenge-2 in roughly 2^{67} operations. This efficiency far outperforms any other cryptanalysis and implies that even larger parameters are needed for security. Considering the 2015 suggestion of NIST in [24] that we migrate to 112-bit security, secure parameters for such an HFE^- scheme will be very large, indeed.

Moreover, the use of HFE^- for encryption, in light of this attack, seems very tricky. Presumably the choice of very large and very inefficient instances of HFE^- over very large and very inefficient instances of HFE for encryption is to slightly enhance the efficiency of the scheme by lowering the degree bound. Against our attack, however, lowering $\lceil log_q(D)\rceil$ by x requires a corresponding increase in a by x to achieve *a slightly smaller* security level. This is due to the fact that this transformation preserves the degree of regularity of the MinRank system, but reduces the number of variables by a. Thus, it is reasonable to question the extent of the benefit of using HFE^- over HFE for encryption.

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

References

1. Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J. Sci. Stat. Comp. **26, 1484** (1997)
2. Group, C.T.: Submission requirements and evaluation criteria for the post-quantum cryptography standardization process. NIST CSRC (2016) http://csrc.nist.gov/groups/ST/post-quantum-crypto/documents/call-for-proposals-final-dec-2016.pdf.
3. Matsumoto, T., Imai, H.: Public Quadratic Polynominal-Tuples for Efficient Signature-Verification and Message-Encryption. In: EUROCRYPT. (1988) 419–453
4. Patarin, J.: Hidden Fields Equations (HFE) and Isomorphisms of Polynomials (IP): Two New Families of Asymmetric Algorithms. In: EUROCRYPT. (1996) 33–48
5. Tao, C., Diene, A., Tang, S., Ding, J.: Simple matrix scheme for encryption. In Gaborit, P., ed.: PQCrypto. Volume 7932 of Lecture Notes in Computer Science., Springer (2013) 231–242
6. Ding, J., Petzoldt, A., Wang, L.: The cubic simple matrix encryption scheme. [25] 76–87
7. Porras, J., Baena, J., Ding, J.: ZHFE, A new multivariate public key encryption scheme. [25] 229–245
8. Szepieniec, A., Ding, J., Preneel, B.: Extension field cancellation: A new central trapdoor for multivariate quadratic systems. [26] 182–196
9. Faugere, J.C.: Algebraic cryptanalysis of hidden field equations (HFE) using grobner bases. CRYPTO 2003, LNCS **2729** (2003) 44–60
10. Kipnis, A., Shamir, A.: Cryptanalysis of the HFE public key cryptosystem by relinearization. Advances in Cryptology - CRYPTO 1999, Springer **1666** (1999) 788
11. Bettale, L., Faugère, J., Perret, L.: Cryptanalysis of HFE, multi-HFE and variants for odd and even characteristic. Des. Codes Cryptography **69** (2013) 1–52
12. Moody, D., Perlner, R.A., Smith-Tone, D.: An asymptotically optimal structural attack on the ABC multivariate encryption scheme. [25] 180–196
13. Moody, D., Perlner, R.A., Smith-Tone, D.: Key recovery attack on the cubic abc simple matrix multivariate encryption scheme. In: Selected Areas in Cryptography – SAC 2016: 23rd International Conference, Revised Selected Papers, LNCS, Springer (2017)
14. Perlner, R.A., Smith-Tone, D.: Security analysis and key modification for ZHFE. [26] 197–212
15. Perret, L.: Grobner basis techniques in post-quantum cryptography. Presentation - Post-Quantum Cryptography - 7th International Workshop, PQCrypto 2016, Fukuoka, Japan, February 24-26, 2016 (2016) https://www.youtube.com/watch?v=0q957wj6w2I.
16. Ding, J., Kleinjung, T.: Degree of regularity for HFE-. IACR Cryptology ePrint Archive **2011** (2011) 570
17. Daniels, T., Smith-Tone, D.: Differential properties of the HFE cryptosystem. [25] 59–75
18. Dubois, V., Fouque, P.A., Shamir, A., Stern, J.: Practical Cryptanalysis of SFLASH. In Menezes, A., ed.: CRYPTO. Volume 4622 of Lecture Notes in Computer Science., Springer (2007) 1–12

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

19. Matsumoto, T., Imai, H.: Public quadratic polynomial-tuples for efficient signature verification and message-encryption. Eurocrypt '88, Springer **330** (1988) 419–545
20. Berlekamp, E.R.: Factoring polynomials over large finite fields. Mathematics of Computation **24** (1970) pp. 713–735
21. Faugère, J., Din, M.S.E., Spaenlehauer, P.: Computing loci of rank defects of linear matrices using gröbner bases and applications to cryptology. In Koepf, W., ed.: Symbolic and Algebraic Computation, International Symposium, ISSAC 2010, Munich, Germany, July 25-28, 2010, Proceedings, ACM (2010) 257–264
22. Fröberg, R.: An inequality for Hilbert series of graded algebras. Math. Scand. **56** (1985) 117–144
23. Bosma, W., Cannon, J., Playoust, C.: The Magma algebra system. I. The user language. J. Symbolic Comput. **24** (1997) 235–265 Computational algebra and number theory (London, 1993).
24. Barker, E., Roginsky, A.: Transitions: Recommendation for transitioning the use of cryptographic algorithms and key lengths. NIST Special Publication (2015) http://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-131Ar1.pdf.
25. Mosca, M., ed.: Post-Quantum Cryptography - 6th International Workshop, PQCrypto 2014, Waterloo, ON, Canada, October 1-3, 2014. Proceedings. Volume 8772 of Lecture Notes in Computer Science., Springer (2014)
26. Takagi, T., ed.: Post-Quantum Cryptography - 7th International Workshop, PQCrypto 2016, Fukuoka, Japan, February 24-26, 2016, Proceedings. Volume 9606 of Lecture Notes in Computer Science., Springer (2016)

A Toy Example

To illustrate the attack, we present a complete key recovery for a small odd prime field instance of HFE^-. We simplify the exposition by considering a homogeneous key.

Let $q = 7$, $n = 8$, $D = 14$ and $a = 2$. We construct the degree n extension $\mathbb{K} = \mathbb{F}_7[x]/\left\langle x^8 + 4x^3 + 6x^2 + 2x + 3 \right\rangle$ and let $b \in \mathbb{K}$ be a fixed root of this irreducible polynomial.

We randomly select $f : \mathbb{K} \to \mathbb{K}$ of degree D,

$$f(x) = b^{4100689}x^{14} + b^{1093971}x^8 + b^{5273323}x^2,$$

and two invertible linear transformations T and U:

$$T = \begin{bmatrix} 2&1&0&3&5&0&3&2 \\ 6&2&1&3&4&2&5&1 \\ 0&2&5&1&3&1&4&3 \\ 3&2&6&4&5&3&4&4 \\ 6&4&2&1&0&5&0&0 \\ 0&3&3&6&5&1&1&3 \\ 0&3&0&4&3&6&1&5 \\ 4&3&2&6&1&1&6&3 \end{bmatrix}, \text{ and } U = \begin{bmatrix} 5&1&4&1&4&2&5&3 \\ 0&6&1&5&3&5&3&2 \\ 3&3&5&0&3&4&2&2 \\ 4&0&5&4&0&6&4&1 \\ 2&6&4&0&0&5&3&5 \\ 0&2&4&0&2&0&6&5 \\ 4&3&0&3&3&2&2&6 \\ 6&2&5&3&5&4&0&0 \end{bmatrix}.$$

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

Since $b^{1093971}/2 = b^{4937171}$, we have

$$\mathbf{F} = \begin{bmatrix} b^{5273323} & b^{4937171} & 0\,0\,0\,0\,0\,0 \\ b^{4937171} & b^{4100689} & 0\,0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0\,0 \end{bmatrix}.$$

We fix $\Pi : \mathbb{F}_q^8 \to \mathbb{F}_q^6$, the projection onto the first 6 coordinates. Then the public key $P = \Pi \circ T \circ F \circ U$ in matrix form over $\mathbb{F}_q$ is given by:

$$\mathbf{P_0} = \begin{bmatrix} 5&6&3&6&6&0&4&2 \\ 6&0&1&3&3&5&2&1 \\ 3&1&4&0&6&0&4&4 \\ 6&3&0&3&0&2&3&1 \\ 6&3&6&0&4&2&2&4 \\ 0&5&0&2&2&2&5&1 \\ 4&2&4&3&2&5&1&5 \\ 2&1&4&1&4&1&5&2 \end{bmatrix}, \mathbf{P_1} = \begin{bmatrix} 1&6&1&5&4&2&2&2 \\ 6&5&4&4&0&1&6&2 \\ 1&4&3&5&6&2&1&1 \\ 5&4&5&2&2&3&1&5 \\ 4&0&6&2&2&1&2&4 \\ 2&1&2&3&1&6&2&6 \\ 2&6&1&1&2&2&5&6 \\ 2&2&1&5&4&6&6&2 \end{bmatrix}, \mathbf{P_2} = \begin{bmatrix} 2&5&2&2&2&3&3&2 \\ 5&1&2&1&3&2&5&4 \\ 2&2&2&1&6&2&1&0 \\ 2&1&1&4&4&5&2&3 \\ 2&3&6&4&4&5&2&2 \\ 3&2&2&5&5&3&4&6 \\ 3&5&1&2&2&4&5&5 \\ 2&4&0&3&2&6&5&2 \end{bmatrix},$$

$$\mathbf{P_3} = \begin{bmatrix} 1&6&6&4&0&0&3&4 \\ 6&2&5&5&4&5&5&6 \\ 6&5&4&6&3&6&4&2 \\ 4&5&6&4&5&2&4&5 \\ 0&4&3&5&6&3&6&0 \\ 0&5&6&2&3&2&4&1 \\ 3&5&4&4&6&4&4&4 \\ 4&6&2&5&0&1&4&0 \end{bmatrix}, \mathbf{P_4} = \begin{bmatrix} 4&4&5&2&6&6&5&2 \\ 4&4&0&0&3&4&1&6 \\ 5&0&5&3&3&0&1&0 \\ 2&0&3&4&1&3&3&2 \\ 6&3&3&1&6&5&0&1 \\ 6&4&0&3&5&4&6&0 \\ 5&1&1&3&0&6&2&6 \\ 2&6&0&2&1&0&6&4 \end{bmatrix}, \mathbf{P_5} = \begin{bmatrix} 0&2&6&1&6&2&3&4 \\ 2&4&2&0&3&1&5&0 \\ 6&2&5&1&4&3&1&1 \\ 1&0&1&5&0&0&3&0 \\ 6&3&4&0&1&4&1&4 \\ 2&1&3&0&4&5&5&5 \\ 3&5&1&3&1&5&1&2 \\ 4&0&1&0&4&5&2&6 \end{bmatrix}.$$

A.1 Recovering a Related HFE Key

This step in key recovery is a slight adaptation of the program of [11]. First, we recover the related private key of Theorem 2. To do this, we solve the MinRank instance on the above $6 = n-2$ $n \times n$ matrices with target rank $\lceil log_q(D) \rceil + a = 2 + 2 = 4$. We may fix one variable to make the ideal generated by the 5×5 minors zero-dimensional. There are $n = 8$ solutions, each of which consists of the Frobenius powers of the coordinates of

$$v = (1, b^{5656746}, b^{3011516}, b^{3024303}, b^{1178564}, b^{1443785}).$$

The combination $L = \sum_{i=0}^{5} v_i \mathbf{P_i}$ is now a rank 4 matrix with entries in $\mathbb{K}$.

We next form $\widehat{v}$ from v by appending $a = 2$ random nonzero values from $\mathbb{K}$ to v. Now we compute

$$\phi^{-1} T'^{-1} \circ \phi = \sum_{i=0}^{8} \widehat{v}_i x^{q^i}.$$

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

Next we let K_i be the left kernel matrix of the $n-i$th Frobenius power of L for $i = 0, 1, \ldots, a+1$. We then recover a vector w simultaneously in the right kernel of K_i for all i. For this example, each such element is a multiple in $\mathbb{K}$ of

$$w = (b^{4849804}, b^{3264357}, b^{4466027}, b^{638698}, b^{2449742}, b^{4337472}, b^{2752502}, b^{1186132}).$$

Then we may compute

$$\phi^{-1} \circ U \circ \phi = \sum_{i=0}^{8} w_i x^{q^i}.$$

At this point we can recover $\phi^{-1} \circ f' \circ \phi = T'^{-1} \circ P \circ U'^{-1}$, and have a full private key for the related instance HFE$(7, 8, 686)$. The transformations T' and U' and the matrix representation of f' as a quadratic form over $\mathbb{K}$ are given by

$$T' = \begin{bmatrix} 1&4&4&5&4&5&5&2 \\ 0&6&6&0&4&4&5&5 \\ 0&5&0&4&2&0&0&3 \\ 0&4&4&2&5&6&6&6 \\ 0&3&6&2&5&6&0&0 \\ 0&2&0&4&4&6&2&2 \\ 0&1&5&5&0&5&2&6 \\ 0&3&3&3&6&5&2&2 \end{bmatrix}, U' = \begin{bmatrix} 6&2&1&4&4&4&1&6 \\ 1&6&0&2&3&0&4&2 \\ 2&5&3&6&3&3&0&4 \\ 0&5&6&5&4&1&4&2 \\ 6&5&3&5&4&6&3&2 \\ 0&4&6&1&4&0&1&5 \\ 6&0&2&3&6&5&6&3 \\ 5&2&0&4&1&2&4&5 \end{bmatrix}$$

$$\mathbf{F}' = \begin{bmatrix} b^{416522} & b^{5402526} & 0 & 0 & 0\,0\,0\,0 \\ b^{5402426} & b^{3093518} & b^{5177024} & 0 & 0\,0\,0\,0 \\ 0 & b^{5177024} & b^{5689467} & b^{5706144} & 0\,0\,0\,0 \\ 0 & 0 & b^{5706144} & b^{3464750} & 0\,0\,0\,0 \\ 0 & 0 & 0 & 0 & 0\,0\,0\,0 \\ 0 & 0 & 0 & 0 & 0\,0\,0\,0 \\ 0 & 0 & 0 & 0 & 0\,0\,0\,0 \\ 0 & 0 & 0 & 0 & 0\,0\,0\,0 \end{bmatrix}.$$

A.2 Recovery of Equivalent HFE$^-$ Key

Now we describe the full key recovery given the related HFE key. We know that there exists a degree $D = 14$ map $f''(x) = f''_{0,0}x^2 + 2f''_{0,1}x^8 + f''_{1,1}x^{14}$ with associated quadratic form

$$\mathbf{F}'' = \begin{bmatrix} f''_{0,0} & f''_{0,1} & 0\,0\,0\,0\,0\,0 \\ f''_{0,1} & f''_{1,1} & 0\,0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0\,0 \\ 0 & 0 & 0\,0\,0\,0\,0\,0 \end{bmatrix},$$

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

and a polynomial $\pi(x) = x + p_1 x^7 + p_2 x^{49}$ such that $f' = \pi \circ f''$. Thus we obtain the bilinear system of equations by equating $\mathbf{F}'$ to:

$$\widehat{\pi\mathbf{F}''} = \begin{bmatrix} f''_{0,0} & f''_{0,1} & 0 & 0 & 0\,0\,0\,0 \\ f''_{0,1} & f''_{1,1} + p_1(f''_{0,0})^7 & p_1(f''_{0,1})^7 & 0 & 0\,0\,0\,0 \\ 0 & p_1(f''_{0,1})^7 & p_1(f''_{1,1})^7 + p_2(f''_{0,0})^{49} & p_2(f''_{0,1})^{49} & 0\,0\,0\,0 \\ 0 & 0 & p_2(f''_{0,1})^{49} & p_2(f''_{1,1})^{49} & 0\,0\,0\,0 \\ 0 & 0 & 0 & 0 & 0\,0\,0\,0 \\ 0 & 0 & 0 & 0 & 0\,0\,0\,0 \\ 0 & 0 & 0 & 0 & 0\,0\,0\,0 \\ 0 & 0 & 0 & 0 & 0\,0\,0\,0 \end{bmatrix}.$$

We clearly have the values of $f''_{0,0}$ and $f''_{0,1}$. Then the equations on the highest diagonal are linear in p_i. We obtain $\pi = x + b^{1948142}x^7 + b^{398370}x^{49}$ and continue to solve the now linear system to recover $f''(x) = b^{416522}x^2 + b^{1559326}x^8 + b^{1121420}x^{14}$.

We then obtain the matrix form of π over $\mathbb{F}_q$ and compose with T':

$$\widehat{\pi} = \begin{bmatrix} 2\,6\,6\,0\,2\,2\,5\,5 \\ 6\,3\,5\,3\,1\,4\,5\,0 \\ 5\,2\,6\,0\,6\,6\,6\,1 \\ 1\,1\,3\,6\,4\,1\,1\,6 \\ 5\,6\,2\,4\,6\,6\,1\,6 \\ 5\,3\,1\,5\,0\,1\,0\,4 \\ 3\,2\,1\,3\,3\,1\,3\,5 \\ 4\,2\,1\,1\,1\,4\,4\,2 \end{bmatrix}, T' \circ \widehat{\pi} = \begin{bmatrix} 0\,0\,1\,2\,0\,5\,4\,0 \\ 1\,2\,4\,4\,2\,1\,0\,4 \\ 0\,2\,2\,1\,1\,6\,1\,0 \\ 3\,3\,1\,0\,6\,3\,2\,0 \\ 0\,1\,3\,1\,0\,2\,2\,2 \\ 3\,4\,5\,0\,1\,3\,4\,2 \\ 0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \end{bmatrix}.$$

Replacing the last two rows of $T' \circ \widehat{\pi}$ to make a full rank matrix produces T''. Then the original public key P is equal to $\Pi \circ T'' \circ \phi^{-1} \circ f'' \circ \phi \circ U'$.

Smith-Tone, Daniel; Vates, Jeremy.
"Key Recovery Attack for All Parameters of HFE-."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

Practical Key Recovery Attack for ZHFE

Daniel Cabarcas[1], Daniel Smith-Tone[2,3], and Javier A. Verbel[1]

[1]Universidad Nacional de Colombia, Sede Medellín, Colombia: dcabarc@unal.edu.co
[2]University of Louisville, USA: daniel-c.smith@louisville.edu
[3]National Institute of Standards and Technology, USA: daniel.smith@nist.gov

Abstract. At PQCRYPTO 2014, Porras, Baena and Ding introduced ZHFE, an interesting new technique for multivariate post-quantum encryption. The scheme is a generalization of HFE in which a single low degree polynomial in the central map is replaced by a pair of high degree degree polynomials with a low degree cubic polynomial contained in the ideal they generate. ZHFE was constructed with the philosophy that a statistically injective multivariate expansion map may have less rigid a structure than a bijection, and may be more resistant to cryptanalysis. We show that in the case of ZHFE, this intuition is false.
We present a practical key recovery attack for ZHFE based on the independent discoveries of the low rank property of ZHFE by Verbel and by Perlner and Smith-Tone. Thus, although the two central maps of ZHFE have high degree, their low rank property makes ZHFE vulnerable to the Kipnis-Shamir(KS) rank attack. We adapt the minors modeling approach to the KS attack pioneered by Bettale, Faugère and Perret in application to HFE, and break ZHFE for practical parameters. Specifically, our attack recovers a private key for ZHFE$(7, 55, 105)$ in approximately 2^{64} operations.

Keywords: Multivariate public key cryptography, encryption schemes, ZHFE

1 Introduction

The fundamental problem of solving systems of nonlinear equations is thousands of years old and has been very influential in the development of algebra and number theory. In the realm of cryptography, the task of solving systems of nonlinear, often quadratic, equations is a principal challenge which is relevant in the analysis of many primitives, both in the symmetric and asymmetric setting. This basic problem is the basis of numerous public key schemes, which, in principle, add to the diversity of public key options. The subdiscipline of cryptography concerned with this family of cryptosystems is usually called Multivariate Public Key Cryptography (MPKC).

In addition to the benefit of creating a more robust toolkit of public key primitives, the advent of MPKC offers a potential solution to the problem of securing communication against quantum adversaries, adversaries with access to a sophisticated quantum computer. Since Peter Shor discovered in the mid

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

90s, see [27], algorithms for factoring and computing discrete logarithms on a quantum computer, a dedicated community has been emmersed in the challenge of securing data from quantum adversaries. In December 2016 the National Institute of Standards and Technology (NIST) published a call for proposals for post-quantum standards from the international community, putting a figurative spotlight on public key cryptography useful in an era with quantum computing technology. In light of this focus from NIST, the cryptometry and cryptanalysis of post-quantum schemes is not simply an academic matter.

While there are several secure, performant, and well-studied multivariate signature schemes, see [9,5,16,21], for example, there are very few unbroken multivariate encryption schemes in the current cryptonomy. Surprisingly, this general absence of secure and long-lived encryption schemes is primarily due to a small array of extremely effective cryptanalytic techniques.

Broadly, we can categorize attacks on multivariate cryptosystems as either direct algebraic, directly inverting the multivariate public key via Gröbner basis calculation, differential, exploiting some symmetric or invariant structure exhibited by the differential of the private key, or rank, recovering a low rank equivalent private key structure by solving an instance of MinRank, i.e. finding a low rank map in a space of linear maps derived from the public key. These basic tools form the core of modern multivariate cryptanalysis and the algebraic objects related to them are of great interest, not only theoretically, but also for use in cryptometry, see for example, [4,13,6,18,2,22,28,11,7,10].

In the last few years, a few novel techniques for the construction of multivariate encryption schemes have been proposed. The idea is to retain statistical injectivity while relaxing the structure of the public key by doubling the dimension of the codomain. The schemes *ABC* Simple Matrix, and Cubic Simple Matrix, proposed in [30,8], are based on a large matrix algebra over a finite field. The ZHFE scheme, proposed in [25] (with a significant key generation improvement from [1]) is based on high degree polynomials F and $\tilde{F}$ over an extension field. Decryption in the later is possible, by the existence of a low degree polynomial Ψ in the ideal generated by F and $\tilde{F}$.

The *ABC* Simple Matrix and Cubic Simple Matrix encryption schemes have been shown vulnerable to differential attacks, see [18,19]. Moreover, in [23] and independently in [31] a trivial upper bound on the Q-rank, or quadratic rank, of ZHFE is provided, further calling into question whether the design strategy of enlarging the dimension of the codomain of the public key is an effective way of achieving multivariate encryption.

On the other hand, in [32], a new security estimate is provided for the original parameters of ZHFE. The paper not only purports to prove the security of ZHFE against the attack methodology of Kipnis and Shamir on low Q-rank schemes, see [17], it also improves the estimate of the degree of regularity of the public key of ZHFE, indicating that the security level of the original parameters is at least 2^{96} instead of the original claim of 2^{80}. In particular, their bound on the complexity of the KS attack on ZHFE is 2^{138}, placing this attack well out of the realm of possibility.

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

In this paper, we make the impossible practical. We detail a full key recovery attack, that works with high probability from the public key alone, and test its effectiveness for small parameters. Our attack adapts the techniques first introduced in [17] and later improved in [2], to recover low rank central maps. Furthermore, we show how to recover a low degree polynomial equivalent to Ψ, that can be used to decrypt.

Our complexity analysis of the attack demonstrates that ZHFE is also asymptotically broken, revealing an error in the analysis of [32]. Specifically, we find that the expected complexity of the Kipnis-Shamir attack on this scheme is $\mathcal{O}\left(n^{(\lceil log_q(D)\rceil+2)\omega}\right)$, where D is the degree bound in ZHFE and ω is the linear algebra constant, instead of the complexity $\mathcal{O}\left(n^{2(\lceil log_q(D)\rceil+2)\omega}\right)$ as reported in [32]. Our empirical data from an implementation of this attack support our complexity estimate. This correction in the complexity estimate reveals that the attack is feasible for the original parameters; instead of a complexity of 2^{138} as claimed in [32], we find the complexity is 2^{64} (A.3). We thus consider ZHFE to be broken.

The article is organized as follows. In the next section, we describe the ZHFE construction and discuss the encryption scheme. In the subsequent section, we outline our attack, describing our notation, our proof of the existence of a low rank equivalent private key, reduce the task of recovering a low rank central polynomial to a MinRank problem, and state how to construct a fully functional equivalent key from it. In the following section, we derive the complexity of our attack, and present our experimental data supporting our complexity bound. A detailed comparison of our analysis to previous MinRank analysis and a toy example are provided in the appendices for space reasons. In the last section, we conclude that ZHFE is broken and discuss the current landscape of multivariate public key encryption.

2 The ZHFE encryption scheme

The ZHFE encryption scheme was introduced in [25] based on the idea that a high degree central map may resist cryptanalysis in the style of [2]. The hope of the authors was that having a high degree central map may result in high Q-rank.

Let $\mathbb{F}$ be a finite field of order q. Let $\mathbb{K}$ be a degree n extension of $\mathbb{F}$. Large Roman letters near the end of the alphabet denote indeterminants over $\mathbb{K}$. Small Roman letters near the end of the alphabet denote indeterminants over $\mathbb{F}$. An underlined letter denotes a vector over $\mathbb{F}$, e.g. $\underline{v} = (v_1, \ldots, v_n)$. A small bold letter denotes a vector over $\mathbb{K}$, e.g. $\mathbf{u} = (u_0, \ldots, u_{n-1})$. Small Roman letters near $f, g, h, \ldots$ denote polynomials over $\mathbb{F}$. Large Roman letters near $F, G, H, \ldots$ denote polynomials over $\mathbb{K}$. Large bold letters denote matrices; the field in which coefficients reside will be specified, but may always be considered to be included in $\mathbb{K}$. The function $\text{Frob}_k()$ takes as argument a polynomial or a matrix. For polynomials it returns the polynomial with its coefficient raised to k-th Frobenius power, and for matrices it raises each entry of the matrix to k-th Frobenius power.

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

Fix an element $y \in \mathbb{K}$ whose orbit under the Frobenius map $y \mapsto y^q$ is of order n. We define the canonical $\mathbb{F}$-vector space isomorphism $\phi : \mathbb{F}^n \to \mathbb{K}$ defined by $\phi(\underline{a}) = \sum_{i=0}^{n-1} a_i y^{q^i}$. We further define $\phi_2 = \phi \times \phi$.

The construction of the central map of ZHFE is quite simple. Without loss of the generality of analysis, we focus on the homogeneous case. One formally declares the following relation over $\mathbb{K}$:

$$\Psi = X\left(\alpha_1 F^{q^0} + \cdots + \alpha_n F^{q^{n-1}} + \beta_1 \tilde{F}^{q^0} + \cdots + \beta_n \tilde{F}^{q^{n-1}}\right)\Bigg(+ X^q\left(\alpha_{n+1} F^{q^0} + \cdots + \alpha_{2n} F^{q^{n-1}} + \beta_{n+1} \tilde{F}^{q^0} + \cdots + \beta_{2n} \tilde{F}^{q^{n-1}}\right)\Bigg($$

where juxtaposition represents multiplication in $\mathbb{K}$ and where Ψ is constrained to have degree less than a bound D. By its construction, Ψ has the form

$$\Psi(x) = \sum_{i=0}^{1} \sum_{\substack{i \le j \le k \\ q^i + q^j + q^k \le D}} \Bigg(a_{i,j,k} x^{q^i + q^j + q^k}.$$

One may then arbitrarily choose the coefficients α_i and β_i and solve the resulting linear system for the unknown coefficients of F and $\tilde{F}$. Even making an arbitrary selection of the coefficients $a_{i,j,k}$ of Ψ, we have an underdefined system and have a large solution space for maps F and $\tilde{F}$.

The private key is given by $\Pi = (G, S, T)$ where $G = (F, \tilde{F})$, $S \in End(\mathbb{F}^n)$ and $T \in End(\mathbb{F}^{2n})$. The public key is constructed via

$$P = T \circ \phi_2 \circ G \circ \phi^{-1} \circ S.$$

Encryption is accomplished by simply evaluating P at the plaintext $\underline{x} \in \mathbb{F}^n$. The interesting step in decryption is inverting the central map, G. Notice that if $G(X) = (Y_1, Y_2)$ then the following relation holds:

$$\Psi(X) = X(\alpha_1 Y_1 + \alpha_2 Y_1^q + \cdots \alpha_n Y_1^{q^{n-1}} + \beta_1 Y_2 + \cdots + \beta_n Y_2^{q^{n-1}} + X^q(\alpha_{n+1} Y_1 + \alpha_{n+2} Y_1^q + \cdots \alpha_{2n} Y_1^{q^{n-1}} + \beta_{n+1} Y_2 + \cdots + \beta_{2n} Y_2^{q^{n-1}}.$$

Since this equation is of degree bounded by D, solutions X can be found efficiently using Berlekamp's Algorithm. While it is possible that there may be multiple solutions to this equation, it is very unlikely; furthermore, the public key can be used to determine the actual preimage.

3 Key Recovery Attack for ZHFE

In this section describe a key recovery attack for ZHFE using the MinRank approach. We first show that with high probability there exist linear combinations of Frobenius powers of the core polynomials F and $\tilde{F}$ of low rank. Then, we show that such linear combinations can be efficiently extracted from the public key. Finally, we describe how to construct a low degree polynomial Ψ' from those low rank polynomials.

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

3.1 Existence of a low rank equivalent key

Fix the representation $a \overset{\Phi}{\mapsto} (a, a^q, \ldots, a^{q^{n-1}})$ of $\mathbb{K}$. Then the image $\Phi(\mathbb{K}) = \mathbb{A} = \{(a, a^q, \ldots, a^{q^{n-1}}) : a \in \mathbb{K}\}$ is a one-dimensional $\mathbb{K}$-algebra. We define $\mathbf{M}_n$ by $\mathbf{M}_n = \Phi \circ \phi$. Using the element y defined in Section 2, we recover an explicit representation of $\mathbf{M}_n \in \mathcal{M}_{n\times n}(\mathbb{K})$:

$$\mathbf{M}_n = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ y & y^q & & y^{q^{n-1}} \\ \vdots & & \ddots & \\ y^{n-1} & y^{(n-1)q} & & y^{(n-1)q^{n-1}} \end{pmatrix} \cdot \begin{pmatrix} & & & \\ & & & \\ & & & \\ & & & \end{pmatrix}$$

It is well known that the matrix $\mathbf{M}_n$ is invertible. The following proposition is a particular case of Proposition 4 in [2].

Proposition 1. *Let* $\boldsymbol{M}_{2n} = \begin{pmatrix} \boldsymbol{M}_n & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_n \end{pmatrix}$. *Then the function* $\varphi_2 = \mathbb{K}^2 \to \mathbb{F}^{2n}$ *can be expressed as* $(X, Y) \mapsto (X, X^q, \ldots, X^{q^{n-1}}, Y, Y^q, \ldots, Y^{q^{n-1}})\boldsymbol{M}_{2n}^{-1}$, *and its inverse* $\varphi_2^{-1} : \mathbb{F}^{2n} \to \mathbb{K}^2$ *as* $(x_1, \ldots, x_{2n}) \mapsto (X_1, X_{n+1})$, *where* $(X_1, \ldots, X_{2n}) = (x_1, \ldots, x_{2n})\boldsymbol{M}_{2n}$

Two private keys are equivalent if they build the same public key, that is:

Definition 1. *Let* $\Pi = (G, S, T)$, *and* $\Pi' = (G', S', T')$ *be private ZHFE keys. We say that* Π *and* Π' *are equivalent if*

$$T' \circ \varphi_2 \circ G' \circ \varphi^{-1} \circ S' = T \circ \varphi_2 \circ G \circ \varphi^{-1} \circ S.$$

We show that given an instance of ZHFE with public key $P = T \circ (\varphi \times \varphi) \circ (F, \tilde{F}) \circ \varphi^{-1} \circ S$ and private key $\Pi = (G, S, T)$, with high probability, there exists an equivalent key $\Pi' = (G', S', T')$, where the polynomials $G' = (F', \tilde{F}')$ have low rank associated matrices. We only consider linear transformations and homogeneous polynomials. This case can be easily adapted to affine transformations and general HFE polynomial.

It was noted by Perlner and Smith-Tone [23] and independently by Verbel [31] that there exists a linear transformation of ZHFE's core map $G = (F, \tilde{F})$ with low rank associated matrices. Recall that for each ZHFE private key (G, S, T), $G = (F, \tilde{F})$, there are scalars $\alpha_1, \ldots, \alpha_{2n}, \beta_1, \ldots, \beta_{2n}$ in the big field $\mathbb{K}$ such that the function

$$\begin{aligned} \Psi = X &\left(\alpha_1 F_0 + \cdots + \alpha_n F_{n-1} + \beta_1 \tilde{F}_0 + \cdots + \beta_n \tilde{F}_{n-1}\right) \\ + X^q &\left(\alpha_{n+1} F_0 + \cdots + \alpha_{2n} F_{n-1} + \beta_{n+1} \tilde{F}_0 + \cdots + \beta_{2n} \tilde{F}_{n-1}\right), \end{aligned}$$

has degree less than a small integer D. Notice that for $s \in \{0, 1\}$ the polynomial,

$$\alpha_{sn+1} F_0 + \cdots + \alpha_{sn+n} F_{n-1} + \beta_{sn+1} \tilde{F}_0 + \cdots + \beta_{sn+n} \tilde{F}_{n-1}$$

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

has HFE shape and its non-zero monomials with degree greater than D have the form $ZX^{q^0+q^1+q^j}$, with $Z \in \mathbb{K}$ and j an integer. Consequently, in each case the matrix associated with that polynomial has rank less than or equal to $\lceil \log_q D \rceil + 1$ and a particular form of tail shown in A.2.

Let L be the function from $\mathbb{K}^2$ to $\mathbb{K}^2$ given by $L(X,Y) = (L_1(X,Y), L_2(X,Y))$, such that

$$L_1(X,Y) = \sum_{i=1}^{n} \alpha_i X^{q^{i-1}} + \sum_{i=1}^{n} \beta_i Y^{q^{i-1}},\ L_2(X,Y) = \sum_{i=1}^{n} \alpha_{n+i} X^{q^{i-1}} + \sum_{i=1}^{n} \beta_{n+i} Y^{q^{i-1}}.$$

Notice that L is a linear transformation of the vector space $\mathbb{K}^2$ over $\mathbb{F}$. From the above observation, the matrices associated with the polynomials in $L \circ G$ are of low rank (less than or equal to $r+1 = \lceil \log_q D \rceil + 1$). Furthermore, if L is invertible, then $(L \circ G, S, T \circ R)$ is an equivalent key to (G, S, T), with $R = \varphi_2 \circ L^{-1} \circ \varphi_2^{-1}$ and the matrices associated with the core polynomials $L \circ G$ are of low rank. Indeed

$$\begin{aligned}(T \circ R) \circ \varphi_2 \circ (L \circ G) \circ \varphi^{-1} \circ S &= T \circ \varphi_2 \circ (L^{-1} \circ \varphi_2^{-1} \circ \varphi_2 \circ L) \circ G \circ \varphi^{-1} \circ S \\ &= T \circ \varphi_2 \circ G \circ \varphi^{-1} \circ S.\end{aligned}$$

For the above assertion to make sense, the function R must be an invertible linear transformation from $\mathbb{F}^{2n}$ to $\mathbb{F}^{2n}$, and this is only possible if L^{-1} is well defined. It is easy to see that if the coefficients $\alpha_1, \ldots, \alpha_{2n}, \beta_1, \ldots, \beta_{2n}$ are chosen uniformly at random in $\mathbb{K}$, the probability that L is invertible is very high (see [31] for more details).

Were L singular as suggested in [24], a different approach is also possible. Defining the linear transformation $R' = \varphi_2 \circ L \circ \varphi_2^{-1} \circ T^{-1}$ and with the public key $P = T \circ \varphi_2 \circ G \circ \varphi^{-1} \circ S$, we have $R' \circ P = \varphi_2 \circ (L \circ G) \circ \varphi^{-1} \circ S$. Thus, $R' \circ P$ has low rank core polynomials $L \circ G$, hence we can attack $R' \circ P$ and find R' in the process. We do not further discuss this approach, and instead, from now on, we assume L is invertible which happens with high probability for the scheme as originally proposed.

3.2 Finding a low rank core polynomial

In the previous section we saw that, with high probability a ZHFE public key P has at least one private key (G', S', T') such that the matrices associated with the polynomials in G' have low rank. We now discuss how from the public key P, we can obtain such an equivalent key and how to further exploit it to decrypt without knowing the secret key.

Let P be a ZHFE public key and let us assume there exists an equivalent key (G', S', T'), with low rank core map $G' = (H, \tilde{H})$, so that $P = T' \circ \varphi_2 \circ G' \circ \varphi^{-1} \circ S'$. Let $\mathbf{H}$ and $\tilde{\mathbf{H}}$ be the low rank ($r+1$, with $r = \lceil \log_q D \rceil$) matrices associated with H and $\tilde{H}$.

Note that the above relation implies that, algebraically, ZHFE is similar to a high degree (but still low rank) version of multi-HFE. Thus, we may suspect

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

that all of the consequences of low rank derived in [2] apply. In fact, our attack on ZHFE, though related, has some subtle but significant distinctions from the cryptanalysis of multi-HFE. The details of the MinRank attack follow.

Using the notation $\mathbf{H}^{*k} \in \mathcal{M}_{n\times n}(\mathbb{K})$ to represent the matrix associated with the k-th Frobenius power of a polynomial H with matrix $\mathbf{H} = [a_{i,j}]$, it is easy to see that the (i,j)-th entry of $\mathbf{H}^{*k}$ is $a_{i-k,j-k}^{q^k}$ (indices are modulo n).

Now we use the property on the matrices $\mathbf{M}_n$ and $\mathbf{M}_{2n}$ to deduce a useful relation between the matrices associated with the low rank polynomials $\mathfrak{H} = \varphi_2 \circ (H, \tilde{H}) \circ \varphi^{-1} = (h_1, \ldots, h_{2n})$ and the matrices $\mathbf{H}^{*k\prime}$s. The following Lemma follows from Lemma 2 in [2].

Lemma 1. *Let $(\boldsymbol{H}_1, \ldots, \boldsymbol{H}_{2n}) \in (\mathcal{M}_{n\times n}(\mathbb{F}))^{2n}$ be the matrices associated with the quadratic polynomials $\varphi_2 \circ (H, \tilde{H}) \circ \varphi^{-1} = (h_1, \ldots, h_{2n}) \in (\mathbb{F}[x_1, \ldots, x_n])^{2n}$, i.e. $h_i = \underline{x}\boldsymbol{H}_i\underline{x}^\top$ for all i, $1 \le i \le n$. It holds that*

$$(\boldsymbol{H}_1, \ldots, \boldsymbol{H}_{2n}) = \\ (\boldsymbol{M}_n\boldsymbol{H}^{*0}\boldsymbol{M}_n^\top, \ldots, \boldsymbol{M}_n\boldsymbol{H}^{*n-1}\boldsymbol{M}_n^\top, \boldsymbol{M}_n\tilde{\boldsymbol{H}}^{*0}\boldsymbol{M}_n^\top, \ldots, \boldsymbol{M}_n\tilde{\boldsymbol{H}}^{*n-1}\boldsymbol{M}_n^\top)\boldsymbol{M}_{2n}^{-1}$$

Let $(\mathbf{P}_1, \ldots, \mathbf{P}_{2n}) \in (\mathcal{M}_{n\times n}(\mathbb{F}))^{2n}$ be the matrices associated with the quadratic public polynomials. Then,

$$\begin{aligned} P(\underline{x}) &= T(\mathfrak{H}(S(\underline{x}))) \\ (\underline{x}\mathbf{P}_1\underline{x}^\top, \ldots, \underline{x}\mathbf{P}_{2n}\underline{x}^\top) &= (h_1(\underline{x}\mathbf{S}), \ldots, h_{2n}(\underline{x}\mathbf{S}))\mathbf{T} \\ (\underline{x}\mathbf{P}_1\underline{x}^\top, \ldots, \underline{x}\mathbf{P}_{2n}\underline{x}^\top) &= (\underline{x}\mathbf{S}\mathbf{H}_1\mathbf{S}^\top\underline{x}^\top, \ldots, \underline{x}\mathbf{S}\mathbf{H}_{2n}\mathbf{S}^\top\underline{x}^\top)\mathbf{T}, \end{aligned} \tag{1}$$

where $\mathbf{S} \in \mathcal{M}_{n\times n}(\mathbb{F})$ and $\mathbf{T} \in \mathcal{M}_{2n\times 2n}(\mathbb{F})$. Using this relation and Lemma 1, we can derive a simultaneous MinRank problem on the matrices associated with the public polynomials, which lie in $\mathcal{M}_{n\times n}(\mathbb{F})$, the solutions of which lie in the extension field $\mathbb{K}$. This result is similar to, but has consequential differences from, [2, Theorem 2].

Theorem 1. *Given the notation above, for any instance of ZHFE, calculating $\boldsymbol{U} = \boldsymbol{T}^{-1}\boldsymbol{M}_{2n} \in \mathcal{M}_{2n\times 2n}(\mathbb{K})$ for some equivalent key (G', S', T') reduces to solving a MinRank instance with rank $r+1$ and $k = 2n$ on the public matrices $(\boldsymbol{P}_1, \ldots, \boldsymbol{P}_{2n}) \in \mathcal{M}_{n\times n}(\mathbb{F})$. The solutions of this MinRank instance lie in $\mathbb{K}^n$.*

Proof. By Equation (1) and Lemma 1,

$$(\mathbf{P}_1, \ldots, \mathbf{P}_{2n})\mathbf{U} = \\ (\mathbf{W}\mathbf{H}^{*0}\mathbf{W}^\top, \ldots, \mathbf{W}\mathbf{H}^{*n-1}\mathbf{W}^\top, \mathbf{W}\tilde{\mathbf{H}}^{*0}\mathbf{W}^\top, \ldots, \mathbf{W}\tilde{\mathbf{H}}^{*n-1}\mathbf{W}^\top), \tag{2}$$

where, $\mathbf{W} = \mathbf{S}\mathbf{M}_n \in \mathcal{M}_{n\times n}(\mathbb{K})$ and $\mathbf{U} = \mathbf{T}^{-1}\mathbf{M}_{2n} \in \mathcal{M}_{2n\times 2n}(\mathbb{K})$. If $\mathbf{U} = [u_{i,j}]$, by (2) we get the following equations

$$\sum_{i=0}^{2n-1} u_{i,0}\mathbf{P}_{i+1} = \mathbf{W}\mathbf{H}\mathbf{W}^\top, \text{ and } \sum_{i=0}^{2n-1} u_{i,n}\mathbf{P}_{i+1} = \mathbf{W}\tilde{\mathbf{H}}\mathbf{W}^\top. \tag{3}$$

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

Since $\mathbf{H}$ has rank $r+1$ and $\mathbf{W}$ is an invertible matrix, the rank of $\mathbf{WHW}^\top$ is also $r+1$ (similarly for $\tilde{\mathbf{H}}$). Consequently, the last equation implies that the vectors $\mathbf{u} = (u_{0,0}, \ldots, u_{2n-1,0})$ and $\mathbf{v} = (u_{0,n}, \ldots, u_{2n-1,n})$ are solutions (called the original solutions) for the MinRank problem associated with the $k = 2n$ public symmetric matrices $(\mathbf{P}_1, \ldots, \mathbf{P}_{2n})$ and the integer $r+1$.

An immediate consequence of Theorem 1 is that if we solve that MinRank problem we get the matrix associated with a linear combination of the Frobenius powers of H and $\tilde{H}$ composed with $\varphi^{-1} \circ S$. We must next analyze the space of solutions of the MinRank problem for ZHFE and complete the key extraction.

3.3 Finding solution from a MinRank problem

From the previous section we know that there are at least two solution $\mathbf{u}$ and $\mathbf{v}$ (the original solutions) for the MinRank problem associated with ZHFE. In this part we show that every nonzero linear combination of a Frobenius power of the original solutions, i.e, $\alpha\mathbf{u}^{q^k} + \beta\mathbf{v}^{q^k}$, is also solution for the MinRank problem associated with ZHFE.

First of all, note that for each nonzero vector $(a_{00}, a_{10}) \in \mathbb{K} \times \mathbb{K}$ there is another vector $(a_{01}, a_{11}) \in \mathbb{K} \times \mathbb{K}$ such that the matrix $A^* = \begin{bmatrix} a_{00} & a_{10} \\ a_{01} & a_{11} \end{bmatrix}$ is an invertible matrix. If $\mathcal{A}$ is the linear transformation associated with A^*, the private key (G'', S'', T'') with

$$\begin{aligned} G'' &= \text{Frob }_k \circ \mathcal{A} \circ (H, \tilde{H}) \circ \text{Frob }_{n-k} \\ T'' &= T' \circ \varphi_2 \circ \mathcal{A}^{-1} \circ \text{Frob}_{n-k} \circ \varphi_2^{-1} \\ S'' &= \varphi \circ \text{Frob}_k \circ \varphi^{-1} \circ S', \end{aligned}$$

is equivalent to (G', S', T'). From Proposition 8 in [2], we know that the matrix associated with $\varphi_2 \circ \mathcal{A} \circ \varphi_2^{-1}$ is $\mathbf{M}_{2n}\widehat{A^*}\mathbf{M}_{2n}^{-1}$, where $\widehat{A^*} = \left[\begin{array}{c|c} A_{00} & A_{01} \\ \hline A_{10} & A_{11} \end{array}\right]$ and $A_{ij} = \text{Diag}(a_{ij}, a_{ij}^q, \ldots, a_{ij}^{q^{n-1}})$.

Also, from Proposition 10 in [2], the matrix associated with $\varphi_2 \circ \text{Frob }_{n-k} \circ \varphi_2^{-1}$ is $\mathbf{M}_{2n}\mathbf{P}_{2,n-k}\mathbf{M}_{2n}^{-1}$, where $\mathbf{P}_{N,k} = \text{Diag}(\mathbf{R}_{n,k}, ..., \mathbf{R}_{n,k})(N \text{ times})$, and $\mathbf{R}_{n,k}$ is the $n \times n$ matrix of a k positions left-rotation. So the matrices associated with H', $\tilde{H}'$(where $G'' := (H', \tilde{H}')$), T''^{-1} and S'' are respectively

$$\begin{aligned} \mathbf{H}' &= a_{00}\,\text{Frob}_k(\mathbf{H}) + a_{01}\,\text{Frob}_k(\tilde{\mathbf{H}}), \\ \tilde{\mathbf{H}}' &= a_{10}\,\text{Frob}_k(\mathbf{H}) + a_{11}\,\text{Frob}_k(\tilde{\mathbf{H}}), \\ \mathbf{T''}^{-1} &= \mathbf{T'}^{-1}\mathbf{M}_{2n}\mathbf{P}_{2,k}\widehat{A^*}\mathbf{M}_{2n}^{-1}, \\ \mathbf{S}'' &= \mathbf{S}'\mathbf{M}_{2n}\mathbf{P}_{1,k}\mathbf{M}_{2n}^{-1}. \end{aligned}$$

As $\text{Rank}(\mathbf{H}') \leq r+1$, (similarly for $\tilde{\mathbf{H}}'$), from equation (3) we get that all columns of $\mathbf{T''}^{-1}\mathbf{M}_{2n}$ are solutions of the MinRank problem associated with the public matrices $(\mathbf{P}_1, \ldots, \mathbf{P}_{2n})$ and $r+1$. Note that $\mathbf{T''}^{-1}\mathbf{M}_{2n} = \mathbf{U}\mathbf{P}_{2,k}\widehat{A^*}$, so

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

the first column of $\mathbf{U}\mathbf{P}_{2,k}\widehat{A^*}$, namely $a_{00}\mathbf{u}^{q^k} + a_{10}\mathbf{v}^{q^k}$, is in particular a solution for such MinRank problem. Moreover, we expect most solutions to be of this form because the system is very overdetermined. Our experiments confirm this latest claim (see Section 4).

So far we know that there are many equivalent keys like (G'', T'', S''). In the following, we explain how we can find one of them. First, we solve the MinRank problem, and use the vector solution $\mathbf{u}' = a_{00}\mathbf{u}^{q^k} + a_{10}\mathbf{v}^{q^k} = (u'_1, \ldots, u'_{2n})$ to compute $\mathbf{K}' = \ker\left(\sum_{i=0}^{2n-1} u'_i \mathbf{P}_{i+1}\right)$. Next, we find another solution $\mathbf{v}' = (v'_0, \ldots, v'_{2n-1})$ to the MinRank problem by solving the linear system,

$$\mathbf{K}' \left(\sum_{i=0}^{2n-1} x_i \mathbf{P}_{i+1} \right) = \mathbf{0}_{(n-r)\times n}.$$

Again, we expect that the new solution $\mathbf{v}'$ preserves the form as a linear combination of the Frobenius power of the original solutions, i.e, $\mathbf{v}' = a_{01}\mathbf{u}^{q^{k_1}} + a_{11}\mathbf{v}^{q^{k_1}}$. Moreover, we claim that both founded solutions come from the same Frobenius power, i.e, $k_1 = k$. Indeed, if $\mathbf{u} = (u_0, \ldots, u_{2n-1})$ (one of the original solutions) and we set $\mathbf{K} = \ker\left(\sum_{i=0}^{2n-1} u_i \mathbf{P}_{i+1}\right)$, Theorem 6 in [2] give us $\mathbf{K}' = \mathrm{Frob}_k(\mathbf{K}) = \mathrm{Frob}_{k_1}(\mathbf{K})$, hence, if $\mathbf{K}$ has at least one entry in $\mathbb{K} \setminus \mathbb{F}$, then $k_1 = k$.

It is easy to see that the probability that $\mathbf{A} = [a_{ij}]$, $i, j = 0, 1$ is invertible is high. In that case, we already know that the matrix $\mathbf{T}''$, such that, $\mathbf{T}''^{-1} = \mathbf{U}''\mathbf{M}_{2n}^{-1}$, with $\mathbf{U}'' := [\mathbf{u}'|\cdots|\mathbf{u}'^{q^{n-1}}|\mathbf{v}'|\cdots|\mathbf{v}'^{q^{n-1}}]$ is part of an equivalent key. In the rest of this section we show how to find the other two elements of the already fixed equivalent key.

Once an equivalent key has been fixed, our second target is to find $\mathbf{W}'' := \mathbf{S}''\mathbf{M}_n$. Keeping in mind that $\sum_{i=0}^{2n-1} u'_i \mathbf{P}_{i+1} = \mathbf{W}''\mathbf{H}'\mathbf{W}''^{\top}$, and $\mathbf{W}''$ is invertible, we get $\ker(\mathbf{H}') = \mathbf{K}'\mathbf{W}''$. Assuming $\mathbf{H}'$ has the shape

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{B}^T \\ \hline \mathbf{B} & 0_{(n-r)\times(n-r)} \end{array}\right],$$

where $\mathbf{A}$ is a full rank $r \times r$ matrix, and $\mathbf{B}$ is a rank one $(n-r) \times r$ matrix, it is easy to see that $\ker(\mathbf{H}')$ is of the form $\left[0_{(n-r-1)\times r} \mid \mathbf{C}\right]$, where $\mathbf{C}$ is a full rank $(n-r-1) \times (n-r)$ matrix. Thus $\mathbf{K}'\mathbf{W}''$ has its first r columns set to zero. In particular, if $\mathbf{w}$ is the first column of $\mathbf{W}''$, then $\mathbf{K}'\mathbf{w} = 0$ leads to a linear system of $n-r-1$ equations in n variables. Such a system might have spurious solutions that do not correspond to a matrix of the form $\mathbf{W}'' = \mathbf{S}''\mathbf{M}_n$. In order to get more equations we can use Frobenius powers of $\mathbf{K}'$. For $j = 0, \ldots, n-1$,

$$\mathrm{Frob}_j(\mathbf{K}') = \ker\left(\sum_{i=0}^{2n-1} u_{i,j}\mathbf{P}_{i+1}\right) = \ker\left(\mathbf{W}''\mathbf{H}'^{*j}\mathbf{W}''^{\top}\right) = \ker\left(\mathbf{W}''\mathbf{H}'^{*j}\right),$$

hence $\ker(\mathbf{H}'^{*j}) = \mathrm{Frob}_j(\mathbf{K}')\mathbf{W}''$. Moreover, $\ker(\mathbf{H}'^{*j})$ has r zero columns indexed by $j+1, \ldots, j+r+1 \mod n$. Therefore, for $j = n-r, \ldots, n-1$,

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

$\mathrm{Frob}_j(\mathbf{K}')\mathbf{w} = 0$ and each of these contributes $n - r - 1$ equations in the same n variables. Note that we only need one column of $\mathbf{W}''$ to build the rest of the matrix.

Once $\mathbf{U}''$ and $\mathbf{W}''$ are recovered, we might find the core polynomials by using the following equations

$$\mathbf{H}' = \mathbf{W}''^{-1} \left(\sum_{i=0}^{2n-1} {u'}_i \mathbf{P}_{i+1} \right) \mathbf{W}''^{-t} \text{ and } \tilde{\mathbf{H}}' = \mathbf{W}''^{-1} \left(\sum_{i=0}^{2n-1} {v'}_i \mathbf{P}_{i+1} \right) \mathbf{W}''^{-t}.$$

At this point, we are not able to decrypt a ciphertext because the recovered core polynomials $\mathbf{H}'$ and $\tilde{\mathbf{H}}'$ would have high degree. But fortunately $\mathbf{H}'$ and $\tilde{\mathbf{H}}'$ satisfy the following equations

$$\begin{aligned} a_{11}\mathbf{H}' - a_{01}\tilde{\mathbf{H}}' &= (a_{11}a_{00} - a_{01}a_{10})\,\mathrm{Frob}_k(\mathbf{H}) = \det(A^*)\,\mathrm{Frob}_k(\mathbf{H}), \text{ and} \\ -a_{10}\mathbf{H}' + a_{00}\tilde{\mathbf{H}}' &= (-a_{01}a_{10} + a_{11}a_{00})\,\mathrm{Frob}_k(\tilde{\mathbf{H}}) = \det(A^*)\,\mathrm{Frob}_k(\tilde{\mathbf{H}}), \end{aligned}$$

where the a'_{ij}s are the ones given by the equivalent key already fixed by $\mathbf{T}''$. Consequently, if we would know the a'_{ij}s, we could derive a low degree polynomial (useful to invert H' and $\tilde{H}'$) as shown in the next equation

$$\begin{aligned} X(a_{11}H' - a_{01}\tilde{H}') + X^q(-a_{10}H' + a_{00}\tilde{H}') = \\ \det(A^*)\left[X\,\mathrm{Frob}_k(H) + X^q\,\mathrm{Frob}_k(\tilde{H})\right] = \\ \det(A^*)\,\mathrm{Frob}_k(\Psi). \end{aligned}$$

Setting $\mathbf{H}' = [h_{ij}]$ and $\tilde{\mathbf{H}}' := [\tilde{h}_{ij}]$, we try to find a_{00}, a_{01}, a_{10}, and a_{11} by first solving the overdetemined systems

$$\begin{aligned} \begin{bmatrix} h_{1,r+1} & h_{1,r+2} & \cdots & h_{1,n-1} & h_{1,n} \\ \tilde{h}_{1,r+1} & \tilde{h}_{1,r+2} & \cdots & \tilde{h}_{1,n-1} & \tilde{h}_{1,n} \end{bmatrix}^\top \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \mathbf{0} \text{ , and} \\ \begin{bmatrix} h_{2,r+1} & h_{2,r+2} & \cdots & h_{2,n-1} & h_{2,n} \\ \tilde{h}_{2,r+1} & \tilde{h}_{2,r+2} & \cdots & \tilde{h}_{2,n-1} & \tilde{h}_{2,n} \end{bmatrix}^\top \begin{bmatrix} y_0 \\ y_1 \end{bmatrix} = \mathbf{0}. \end{aligned}$$

For n large enough we expect that both solution spaces are one-dimensional, i.e, our expected solution are of the form

$$\begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \alpha \begin{bmatrix} a_{11} \\ -a_{01} \end{bmatrix}, \quad \begin{bmatrix} y_0 \\ y_1 \end{bmatrix} = \beta \begin{bmatrix} -a_{10} \\ a_{00} \end{bmatrix}$$

for some $\alpha, \beta \in \mathbb{K}$. Then, we compute

$$\begin{aligned} \alpha a_{11}H' - \alpha a_{01}\tilde{H}' &= \alpha \det(A^*)\,\mathrm{Frob}_k(H), \quad \text{and} \\ -\beta a_{10}H' + \beta a_{00}\tilde{H}' &= \beta \det(A^*)\,\mathrm{Frob}_k(\tilde{H}), \end{aligned}$$

and by solving a linear system, we can get α, β, and our low degree polynomial

$$\Psi'' := \gamma \det(A^*)\,\mathrm{Frob}_k(\Psi), \quad \text{with } \gamma \in \mathbb{K}.$$

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

4 Experimental Results and Complexity

In order to experimentally verify our attack, we generated ZHFE instances for different parameters and carried out the full attack. We were able to solve the MinRank problem associated with each instance of ZHFE and then we recovered an equivalent key for every solved MinRank problem. We also recovered the low degree polynomial Ψ''. Every time we successfully solved the MinRank problem, we were able to carry out the rest of the attack. This confirms that most solutions for such MinRank problem are of the form $a_{00}\mathbf{u}^{q^k}+a_{10}\mathbf{v}^{q^k}$. For these experiments we used the fast key generation method propposed by Baena, et al. [26], so we need to keep in mind that n must be even and the relation $q+2q^{r-1}<D\leq q^r$ must be satisfied. The experiments were performed using Magma v2.21-1 [3] on a server with a processor Intel(R) Xeon(R) CPU E5-2609 0 @ 2.40GHz, running Linux CentOS release 6.6.

			Minors		KS	
q	r	n	CPU time [s]	Memory [MB]	CPU time [s]	Memory [MB]
7	2	8	255	4216	280	439
7	2	12	3111	59651	1272	752
7	2	16			5487	2537
17	2	8	277	5034	299	503
17	2	12	3584	68731	1330	817
17	2	16			6157	2800

Table 1. MinRank attack to ZHFE

Table 1 shows the time and memory required for the attacks using either the Kipnis-Shamir modeling or the minors modeling for solving the MinRank. These few data measures suggests that the Kipnis-Shamir modeling is more efficient. The Kipnis-Shamir modeling yields a bilinear system of $n(n-r-1)$ equations in $(n-r-1)(r+1)+2n$ variables. The Groebner Basis computation on every reported instance with $r=2$ had a falling degree of 4. It follows that under this modeling the resulting system is not bi-regular as defined in [14]. To the best of our knowledge, there is no tight bound in the literature for the falling degree for the system that arises from the Kipnis-Shamir modeling.

Alternatively, the minors modeling yields a system of $\binom{n}{r+2}^2$ equations in $2n$ variables, whose complexity can be studied as in [2]. Assuming the conjecture about regularity in [12], the Hilbert series of the minors model ideal is

$$HS(t)=(1-t)^{(n-R)^2-2n}\frac{\det A(t)}{t^{\binom{R}{2}}},$$

where R is the target matrix rank (in our case $R=r+1$) and $A(t)=[a_{i,j}(t)]$ is the $R\times R$ matrix defined by $a_{i,j}=\sum_{\ell=0}^{n-\max(i,j)}\binom{n-i}{\ell}\binom{n-j}{\ell}t^{\ell}$. The degree of regularity is then given by the index of the first negative coefficient of $HS(t)$.

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

In comparison to the Hilbert Series in the multi-HFE case discussed in [2], the only difference is the $2n$ term in the exponent of $1-t$ (simply n in their case). This does not affect significantly the analysis thereafter. For example, if we define $H_R(t) = (1-t)^{(n-R)^2-2n} \det A(t)$, we can compute

$$H_1(t) = 1 + nt - \frac{1}{4}n(n-4)(n+1)^2t^2 + \mathcal{O}(t^3).$$

Note that the coefficient of 1 and t are positive and that the coefficient of t^2 is negative for $n > 4$. So the degree of regularity is $2 = R + 1 = r + 2$. Similarly, with $r = 1$, $R = 2$, the degree of regularity is $3 = R + 1 = r + 2$ for $n > 5.88$, with $r = 2$, $R = 3$, the degree of regularity is 4 for $n > 7.71$, and with $r = 3$, $R = 4$, the degree of regularity is 5 for $n > 9.54$. We thus adventure to claim that the degree of regularity of the minors modeling of the min-rank problem arising from the attack on ZHFE is less or equal to $r + 2$ for all cases of interest. It follows that the complexity is $\mathcal{O}\left(\binom{2n+r+2}{r+2}\right)^{\omega} \sim \mathcal{O}\left(n^{(r+2)\omega}\right)$, where $2 < \omega < 3$ is the linear algebra constant. This is polynomial in n for r constant. Even if r is a logarithmic function of n, the complexity is barely superpolynomial in n.

It is worth spelling out the practical consequences of the above analysis. The expected degree of regularity $r + 2$ is also the degree of the minors. Thus, for n large enough, these minors span the whole degree $r + 2$ polynomial ring's subspace. Therefore, to solve this system it suffices to gather enough minors and linearly reduce them among themselves. No Groebner basis algorithm is necessary. Moreover, in practice two variables can be fixed to 0 and 1, thus we just need to row-reduce a $\binom{2n+r}{r+2}$ square matrix.

5 Conclusion

We have shown a practical and asymptotic key recovery attack on the ZHFE encryption scheme. The details provided leave no doubt about its effectiveness. The asymptotic analysis shows the scheme vulnerable even for larger parameters. The rank structure of the central polynomials has proven too difficult to mask. Though the concept of ZHFE was directly inspired by a desire to avoid rank weakness, ZHFE succombed to rank weaknesses.

Nevertheless, the idea of an injective multivariate trapdoor function may be viable, though ZHFE is not the correct technique. The landscape for multivariate public key encryption remains fairly bleak at this time. Fundamentally new ideas must emerge to realize the goal of secure multivariate encryption.

References

1. Baena, J.B., Cabarcas, D., Escudero, D.E., Porras-Barrera, J., Verbel, J.A.: Efficient ZHFE key generation. In: Takagi [29], pp. 213–232, http://dx.doi.org/10.1007/978-3-319-29360-8_14

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

2. Bettale, L., Faugère, J.C., Perret, L.: Cryptanalysis of HFE, multi-HFE and variants for odd and even characteristic. Designs, Codes and Cryptography 69(1), 1–52 (2013)
3. Bosma, W., Cannon, J., Playoust, C.: The Magma algebra system. I. The user language. J. Symbolic Comput. 24(3-4), 235–265 (1997), http://dx.doi.org/10.1006/jsco.1996.0125, computational algebra and number theory (London, 1993)
4. Cartor, R., Gipson, R., Smith-Tone, D., Vates, J.: On the differential security of the hfev- signature primitive. In: Takagi [29], pp. 162–181, http://dx.doi.org/10.1007/978-3-319-29360-8_11
5. Chen, M.S., Yang, B.Y., Smith-Tone, D.: Pflash - secure asymmetric signatures on smart cards. Lightweight Cryptography Workshop 2015 (2015), http://csrc.nist.gov/groups/ST/lwc-workshop2015/papers/session3-smith-tone-paper.pdf
6. Daniels, T., Smith-Tone, D.: Differential properties of the HFE cryptosystem. In: Mosca [20], pp. 59–75, http://dx.doi.org/10.1007/978-3-319-11659-4_4
7. Ding, J., Hodges, T.J.: Inverting HFE systems is quasi-polynomial for all fields. In: Rogaway, P. (ed.) Advances in Cryptology - CRYPTO 2011 - 31st Annual Cryptology Conference, Santa Barbara, CA, USA, August 14-18, 2011. Proceedings. Lecture Notes in Computer Science, vol. 6841, pp. 724–742. Springer (2011), http://dx.doi.org/10.1007/978-3-642-22792-9_41
8. Ding, J., Petzoldt, A., Wang, L.: The cubic simple matrix encryption scheme. In: Mosca [20], pp. 76–87, http://dx.doi.org/10.1007/978-3-319-11659-4_5
9. Ding, J., Schmidt, D.: Rainbow, a new multivariable polynomial signature scheme. In: Ioannidis, J., Keromytis, A.D., Yung, M. (eds.) Applied Cryptography and Network Security, Third International Conference, ACNS 2005, New York, NY, USA, June 7-10, 2005, Proceedings. Lecture Notes in Computer Science, vol. 3531, pp. 164–175 (2005), http://dx.doi.org/10.1007/11496137_12
10. Ding, J., Yang, B.Y.: Degree of regularity for hfev and hfev-. In: Gaborit [15], pp. 52–66, http://dx.doi.org/10.1007/978-3-642-38616-9
11. Dubois, V., Gama, N.: The degree of regularity of HFE systems. In: Abe, M. (ed.) Advances in Cryptology - ASIACRYPT 2010 - 16th International Conference on the Theory and Application of Cryptology and Information Security, Singapore, December 5-9, 2010. Proceedings. Lecture Notes in Computer Science, vol. 6477, pp. 557–576. Springer (2010), http://dx.doi.org/10.1007/978-3-642-17373-8_32
12. Faugère, J.C., El Din, M.S., Spaenlehauer, P.J.: Computing loci of rank defects of linear matrices using grÖbner bases and applications to cryptology. In: Proceedings of the 2010 International Symposium on Symbolic and Algebraic Computation. pp. 257–264. ISSAC '10, ACM, New York, NY, USA (2010)
13. Faugère, J., Gligoroski, D., Perret, L., Samardjiska, S., Thomae, E.: A polynomial-time key-recovery attack on MQQ cryptosystems. In: Katz, J. (ed.) Public-Key Cryptography - PKC 2015 - 18th IACR International Conference on Practice and Theory in Public-Key Cryptography, Gaithersburg, MD, USA, March 30 - April 1, 2015, Proceedings. Lecture Notes in Computer Science, vol. 9020, pp. 150–174. Springer (2015), http://dx.doi.org/10.1007/978-3-662-46447-2_7
14. FaugÃ¨re, J.C., Din, M.S.E., Spaenlehauer, P.J.: GrÃ¶bner bases of bihomogeneous ideals generated by polynomials of bidegree : Algorithms and complexity. Journal of Symbolic Computation 46(4), 406 – 437 (2011)
15. Gaborit, P. (ed.): Post-Quantum Cryptography - 5th International Workshop, PQCrypto 2013, Limoges, France, June 4-7, 2013. Proceedings, Lecture Notes

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

in Computer Science, vol. 7932. Springer (2013), http://dx.doi.org/10.1007/978-3-642-38616-9
16. Kipnis, A., Patarin, J., Goubin, L.: Unbalanced oil and vinegar signature schemes. In: Stern, J. (ed.) Advances in Cryptology - EUROCRYPT '99, International Conference on the Theory and Application of Cryptographic Techniques, Prague, Czech Republic, May 2-6, 1999, Proceeding. Lecture Notes in Computer Science, vol. 1592, pp. 206–222. Springer (1999), http://dx.doi.org/10.1007/3-540-48910-X_15
17. Kipnis, A., Shamir, A.: Cryptanalysis of the HFE public key cryptosystem by relinearization. In: Advances in cryptology—CRYPTO '99 (Santa Barbara, CA), Lecture Notes in Computer Science, vol. 1666, pp. 19–30. Springer, Berlin (1999)
18. Moody, D., Perlner, R.A., Smith-Tone, D.: An asymptotically optimal structural attack on the ABC multivariate encryption scheme. In: Mosca [20], pp. 180–196, http://dx.doi.org/10.1007/978-3-319-11659-4_11
19. Moody, D., Perlner, R.A., Smith-Tone, D.: Key Recovery Attack on the Cubic ABC Simple Matrix Multivariate Encryption Scheme. Springer (2017)
20. Mosca, M. (ed.): Post-Quantum Cryptography - 6th International Workshop, PQCrypto 2014, Waterloo, ON, Canada, October 1-3, 2014. Proceedings, Lecture Notes in Computer Science, vol. 8772. Springer (2014), http://dx.doi.org/10.1007/978-3-319-11659-4
21. Patarin, J., Courtois, N., Goubin, L.: Quartz, 128-bit long digital signatures. In: Naccache, D. (ed.) Topics in Cryptology - CT-RSA 2001, The Cryptographer's Track at RSA Conference 2001, San Francisco, CA, USA, April 8-12, 2001, Proceedings. Lecture Notes in Computer Science, vol. 2020, pp. 282–297. Springer (2001), http://dx.doi.org/10.1007/3-540-45353-9_21
22. Perlner, R.A., Smith-Tone, D.: A classification of differential invariants for multivariate post-quantum cryptosystems. In: Gaborit [15], pp. 165–173, http://dx.doi.org/10.1007/978-3-642-38616-9
23. Perlner, R.A., Smith-Tone, D.: Security analysis and key modification for ZHFE. In: Takagi [29], pp. 197–212, http://dx.doi.org/10.1007/978-3-319-29360-8_13
24. Perlner, R.A., Smith-Tone, D.: Security analysis and key modification for ZHFE. In: Post-Quantum Cryptography - 7th International Conference, PQCrypto 2016, Fukuoka, Japan, February 24-26, 2016. Proceedings (2016)
25. Porras, J., Baena, J., Ding, J.: Zhfe, a new multivariate public key encryption scheme. In: Mosca [20], pp. 229–245, http://dx.doi.org/10.1007/978-3-319-11659-4_14
26. Porras, J., Baena, J., Ding, J.: ZHFE, a new multivariate public key encryption scheme. In: Mosca, M. (ed.) Post-Quantum Cryptography, Lecture Notes in Computer Science, vol. 8772, pp. 229–245. Springer International Publishing (2014)
27. Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM Rev. 41(2), 303–332 (electronic) (1999)
28. Smith-Tone, D.: On the differential security of multivariate public key cryptosystems. In: Yang, B.Y. (ed.) PQCrypto. Lecture Notes in Computer Science, vol. 7071, pp. 130–142. Springer (2011)
29. Takagi, T. (ed.): Post-Quantum Cryptography - 7th International Workshop, PQCrypto 2016, Fukuoka, Japan, February 24-26, 2016, Proceedings, Lecture Notes in Computer Science, vol. 9606. Springer (2016), http://dx.doi.org/10.1007/978-3-319-29360-8
30. Tao, C., Diene, A., Tang, S., Ding, J.: Simple matrix scheme for encryption. In: Gaborit [15], pp. 231–242, http://dx.doi.org/10.1007/978-3-642-38616-9

31. Verbel, J.A.: Efficiency and Security of ZHFE. Master's thesis, Universidad Nacional de Colombia, sede Medellín (2016)
32. Zhang, W., Tan, C.H.: On the security and key generation of the zhfe encryption scheme. In: Advances in Information and Computer Security - 11th International Workshop on Security, IWSEC 2016, Tokyo, Japan, September 12-14, 2016, Proceedings (2015)

A Appendix

A.1 Toy Example

We provide a small example of the MinRank attack for ZHFE with parameters $n = 8$, $q = 3$ and $D = 9$. The small field is $\mathbb{F} = \mathbb{F}_q$, the extension field is $\mathbb{K} = \mathbb{F}/\langle g(y)\rangle$, where $g(y) = y^8 + 2y^5 + y^4 + 2y^2 + 2y + 2 \in \mathbb{F}[y]$, and b is a primitive root of the irreducible polynomial $g(y)$.

For ease of presentation, we consider a homogeneous public key and linear transformations. An easy adaptation for the general case can be done following the ideas expressed in [2].

The matrices associated with our private key $\left((F, \tilde{F}), S, T\right)$ are

$$\mathbf{F} = \begin{pmatrix} b^{827} & b^{4873} & b^{3298} & b^{211} & b^{1824} & b^{5374} & b^{6155} & b^{2404} \\ b^{4873} & b^{5172} & b^{1526} & b^{1317} & b^{2727} & b^{1863} & b^{3546} & b^{5876} \\ b^{3298} & b^{1526} & b^{1842} & b^{3540} & b^{2647} & b^{2349} & b^{4599} & b^{2987} \\ b^{211} & b^{1317} & b^{3540} & b^{5242} & b^{5758} & b^{4705} & b^{2663} & b^{4097} \\ b^{1824} & b^{2727} & b^{2647} & b^{5758} & b^{4629} & b^{5792} & b^{5196} & b^{666} \\ b^{5374} & b^{1863} & b^{2349} & b^{4705} & b^{5792} & b^{6318} & b^{4937} & b^{6150} \\ b^{6155} & b^{3546} & b^{4599} & b^{2663} & b^{5196} & b^{4937} & b^{2275} & b^{1436} \\ b^{2404} & b^{5876} & b^{2987} & b^{4097} & b^{666} & b^{6150} & b^{1436} & b^{4721} \end{pmatrix}, \mathbf{S} = \begin{pmatrix} 0 & 0 & 2 & 2 & 2 & 2 & 1 & 2 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 2 & 1 & 1 & 0 & 0 & 2 & 2 & 2 \\ 1 & 2 & 2 & 1 & 1 & 1 & 0 & 2 \\ 0 & 2 & 1 & 2 & 0 & 1 & 0 & 2 \\ 0 & 1 & 1 & 0 & 1 & 2 & 2 & 0 \\ 0 & 1 & 2 & 1 & 2 & 1 & 0 & 0 \\ 2 & 1 & 0 & 2 & 2 & 1 & 0 & 2 \end{pmatrix}$$

$$\tilde{\mathbf{F}} = \begin{pmatrix} b^{5574} & b^{2257} & b^{4540} & b^{880} & b^{2073} & b^{4932} & b^{3441} & b^{5482} \\ b^{2257} & b^{301} & b^{5824} & b^{5391} & b^{1155} & b^{1678} & b^{572} & b^{3108} \\ b^{4540} & b^{5824} & b^{5208} & b^{3763} & b^{6074} & b^{2097} & b^{3074} & b^{139} \\ b^{880} & b^{5391} & b^{3763} & b^{125} & b^{2055} & b^{1763} & b^{1168} & b^{4512} \\ b^{2073} & b^{1155} & b^{6074} & b^{2055} & b^{5080} & b^{1720} & b^{5820} & b^{5832} \\ b^{4932} & b^{1678} & b^{2097} & b^{1763} & b^{1720} & b^{5850} & b^{1822} & b^{5443} \\ b^{3441} & b^{572} & b^{3074} & b^{1168} & b^{5820} & b^{1822} & b^{2857} & b^{939} \\ b^{5482} & b^{3108} & b^{139} & b^{4512} & b^{5832} & b^{5443} & b^{939} & b^{1665} \end{pmatrix}$$

and $\mathbf{T} = [\mathbf{T}_1|\mathbf{T}_2]$, where

$$\mathbf{T1} = \begin{pmatrix} 0 & 2 & 2 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 1 \\ 2 & 2 & 0 & 2 & 2 & 2 & 1 & 2 & 2 & 2 & 2 & 0 & 1 & 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 & 2 & 0 & 2 & 2 & 2 & 1 & 1 & 0 & 1 & 2 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 2 & 2 & 1 & 2 & 0 & 2 & 2 & 1 & 0 & 0 & 0 & 1 & 2 \\ 0 & 1 & 1 & 1 & 2 & 1 & 2 & 0 & 0 & 0 & 1 & 2 & 1 & 1 & 0 & 0 \\ 0 & 1 & 2 & 2 & 1 & 0 & 2 & 2 & 1 & 2 & 2 & 1 & 0 & 2 & 2 & 1 \\ 2 & 0 & 0 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}^{\top}, \mathbf{T2} = \begin{pmatrix} 2 & 1 & 0 & 2 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 0 & 1 & 0 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 2 & 0 & 0 \\ 1 & 2 & 2 & 2 & 2 & 0 & 0 & 2 & 0 & 0 & 2 & 1 & 1 & 1 & 1 & 2 \\ 2 & 0 & 1 & 1 & 1 & 2 & 1 & 2 & 1 & 1 & 2 & 2 & 1 & 2 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 2 & 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 2 & 2 & 0 & 1 & 0 & 0 & 2 & 2 & 0 & 1 & 0 & 1 \\ 1 & 2 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 2 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 & 2 & 0 & 2 & 2 & 2 & 2 & 0 & 2 & 0 & 2 & 1 & 2 & 0 \end{pmatrix}^{\top}.$$

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

This private key gives us a public key represented by the matrices $\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_{2n}$.

$$\mathbf{P}_1 = \begin{pmatrix} 0&0&2&0&2&1&0&2\\ 0&2&1&2&2&2&0&1\\ 2&1&0&1&0&0&0&1\\ 0&2&1&1&2&2&1&2\\ 2&2&0&2&0&2&2&0\\ 1&2&0&2&2&0&0&2\\ 0&0&0&1&2&0&0&0\\ 2&1&1&2&0&2&0&0 \end{pmatrix}, \mathbf{P}_2 = \begin{pmatrix} 2&1&0&1&1&0&2&0\\ 1&0&2&0&0&0&0&2\\ 0&2&0&2&0&0&2&1\\ 1&0&2&1&1&0&0&1\\ 1&0&0&1&1&0&0&1\\ 0&0&0&0&0&2&1&0\\ 2&0&2&0&0&1&0&1\\ 0&2&1&1&1&0&1&2 \end{pmatrix}, \mathbf{P}_3 = \begin{pmatrix} 1&0&0&0&2&1&1&1\\ 0&2&1&2&1&1&2&1\\ 0&1&1&2&0&1&1&2\\ 0&2&2&1&2&2&1&1\\ 2&1&0&2&2&0&0&2\\ 1&1&1&2&0&0&1&0\\ 1&2&1&1&0&1&2&2\\ 1&1&2&1&2&0&2&0 \end{pmatrix}, \mathbf{P}_4 = \begin{pmatrix} 0&1&2&2&2&1&1&0\\ 1&0&0&2&0&2&0&2\\ 2&0&1&0&2&0&2&2\\ 2&2&0&0&2&2&2&2\\ 2&0&2&2&0&2&1&0\\ 1&2&0&2&2&0&1&0\\ 1&0&2&2&1&1&1&0\\ 0&2&2&2&0&0&0&1 \end{pmatrix},$$

$$\mathbf{P}_5 = \begin{pmatrix} 2&0&1&1&1&0&2&1\\ 0&1&2&0&0&0&2&1\\ 1&2&2&0&0&0&1&0\\ 1&0&0&1&2&1&0&0\\ 1&0&0&2&1&1&2&1\\ 0&0&0&1&1&0&0&1\\ 2&2&1&0&2&0&0&2\\ 1&1&0&0&1&1&2&0 \end{pmatrix}, \mathbf{P}_6 = \begin{pmatrix} 0&1&1&0&0&2&1&1\\ 1&2&2&2&2&2&1&2\\ 1&2&0&1&1&2&0&2\\ 0&2&1&0&2&1&2&0\\ 0&2&1&2&0&2&0&1\\ 2&2&2&1&2&2&1&2\\ 1&1&0&2&0&1&2&0\\ 1&2&2&0&1&2&0&2 \end{pmatrix}, \mathbf{P}_7 = \begin{pmatrix} 0&1&2&2&1&2&2&0\\ 1&1&2&2&0&1&0&0\\ 2&2&0&1&2&1&1&0\\ 2&2&1&1&0&1&1&2\\ 1&0&2&0&0&2&2&2\\ 2&1&1&1&2&0&0&1\\ 2&0&1&1&2&0&0&0\\ 0&0&0&2&2&1&0&0 \end{pmatrix}, \mathbf{P}_8 = \begin{pmatrix} 1&2&0&0&0&2&0&2\\ 2&1&0&2&1&1&0&1\\ 0&0&2&1&0&1&1&1\\ 0&2&1&0&1&0&0&1\\ 0&1&0&1&1&1&0&1\\ 2&1&1&0&1&2&2&1\\ 0&0&1&0&0&2&2&1\\ 2&1&1&1&1&1&1&0 \end{pmatrix},$$

$$\mathbf{P}_9 = \begin{pmatrix} 2&0&2&1&2&0&1&0\\ 0&0&0&1&1&2&1&2\\ 2&0&1&2&1&2&2&2\\ 1&1&2&1&2&2&2&0\\ 2&1&1&2&0&1&0&1\\ 0&2&2&2&1&2&2&1\\ 1&1&2&2&0&2&0&2\\ 0&2&2&0&1&1&2&2 \end{pmatrix}, \mathbf{P}_{10} = \begin{pmatrix} 0&2&0&0&0&0&1&1\\ 2&2&2&2&0&1&1&1\\ 0&2&1&2&2&0&1&0\\ 0&2&2&1&1&1&0&1\\ 0&0&2&1&1&0&1&1\\ 0&1&0&1&0&1&1&0\\ 1&1&1&0&1&1&2&0\\ 1&1&0&1&1&0&0&1 \end{pmatrix}, \mathbf{P}_{11} = \begin{pmatrix} 0&0&0&2&2&2&0&1\\ 0&0&2&0&1&2&0&1\\ 0&2&2&0&1&2&2&1\\ 2&0&0&0&2&2&1&0\\ 2&1&1&2&2&1&1&1\\ 2&2&2&2&1&0&0&1\\ 0&0&2&1&1&0&1&2\\ 1&1&1&0&1&1&2&2 \end{pmatrix}, \mathbf{P}_{12} = \begin{pmatrix} 0&2&1&0&2&1&1&0\\ 2&1&0&1&2&1&1&2\\ 1&0&0&1&0&0&1&2\\ 0&1&1&2&2&2&1&1\\ 2&2&0&2&2&0&2&2\\ 1&1&0&2&0&0&2&2\\ 1&1&1&1&2&2&0&2\\ 0&2&2&1&2&2&2&0 \end{pmatrix},$$

$$\mathbf{P}_{13} = \begin{pmatrix} 2&0&0&0&2&1&1&0\\ 0&1&1&2&0&1&1&0\\ 0&1&1&0&2&1&2&1\\ 0&2&0&2&1&1&2&2\\ 2&0&2&1&2&0&2&2\\ 1&1&1&1&0&2&1&2\\ 1&1&2&2&2&1&1&0\\ 0&0&1&2&2&2&0&0 \end{pmatrix}, \mathbf{P}_{14} = \begin{pmatrix} 1&0&0&1&0&2&1&0\\ 0&0&2&2&0&1&0&0\\ 0&2&0&2&1&0&1&1\\ 1&2&2&2&0&1&0&2\\ 0&0&1&0&0&0&0&2\\ 2&1&0&1&0&2&0&1\\ 1&0&1&0&0&0&1&2\\ 0&0&1&2&2&1&2&0 \end{pmatrix}, \mathbf{P}_{15} = \begin{pmatrix} 1&0&2&2&2&1&1&0\\ 0&2&2&0&2&1&0&2\\ 2&2&2&1&2&1&2&0\\ 2&0&1&1&2&1&2&0\\ 2&2&2&2&0&0&2&0\\ 1&1&1&1&0&2&0&2\\ 1&0&2&2&2&0&1&2\\ 0&2&0&0&0&2&2&0 \end{pmatrix}, \mathbf{P}_{16} = \begin{pmatrix} 1&1&0&0&1&1&0&0\\ 1&1&2&1&0&1&0&1\\ 0&2&1&2&1&0&1&2\\ 0&1&2&2&0&2&1&0\\ 1&0&1&0&1&0&0&0\\ 1&1&0&2&0&2&1&0\\ 0&0&1&1&0&1&2&1\\ 0&1&2&0&0&0&1&1 \end{pmatrix}.$$

Recovering T: The first and harder step to recover an equivalent linear transformation T is to solve the MinRank problem associated with the public matrices $\mathbf{P}_1, \ldots, \mathbf{P}_{16}$ and $r+1$, with $r = \lceil \log_q D \rceil = 2$. Using the minors modeling, we construct a degree 4 polynomial system in $2n$ variables. We can fix the two first coordinates of the vector $\mathbf{u}'' = (u'_0, u'_1, \ldots, u'_7)$ as 1 and 0 respectively. A solution for this system is

$$\mathbf{u}' = (1, 0, b^{5854}, b^{4879}, b^{2843}, b^{2676}, b^{6279}, b^{1845}, b^{6102}, b^{5619}, b^{5448}, b^{6022}, b^{1721}, b^{2632}, b^{3738}, b^{6170}).$$

Next we compute

$$\mathbf{K}' = \ker\left(\sum_{i=0}^{2n-1} u'_i \mathbf{P}_{i+1}\right) = \left(\begin{pmatrix} 1&0&0&0&0&b^{6158}&b^{1567}&b^{6415}\\ 0&1&0&0&0&b^{3943}&b^{4591}&b^{95}\\ 0&0&1&0&0&b^{4461}&b^{4216}&b^{3027}\\ 0&0&0&1&0&b^{3577}&b^{5899}&b^{1096}\\ 0&0&0&0&1&b^{6554}&b^{4266}&b^{907} \end{pmatrix}\right),$$

and by solving the linear system

$$\mathbf{K}' \left(\sum_{i=0}^{2n-1} x_i \mathbf{P}_{i+1}\right) = \mathbf{0}_{(n-r)\times n},$$

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

we get another solution

$$\mathbf{v}' := (b^{1519}, b^{4750}, b^{4454}, b^{3326}, b^{2077}, b^{4519}, b^{3525}, b^{1978}, b^{5511}, b^{315}, b^{715}, b^{4722}, b^{5003}, b^{1895}, b^{2665}, b^{4505}).$$

Once we have two solution for the MinRank problem we compute

$$\mathbf{T}''^{-1} = \mathbf{U}''\mathbf{M}_{16}^{-1},$$

with $\mathbf{U}'' := [\mathbf{u}'|\cdots|\mathbf{u}'^{q^{n-1}}|\mathbf{v}'|\cdots|\mathbf{v}'^{q^{n-1}}]$, invert the output matrix to obtain $\mathbf{T}'' = [\mathbf{T}_1''|\mathbf{T}_2'']$, with

$$\mathbf{T}_1'' = \begin{pmatrix} 2&0&2&1&1&0&1&0\\ 1&2&2&2&0&2&0&0\\ 1&1&0&2&0&0&0&2\\ 2&2&1&2&0&1&0&0\\ 0&2&2&0&2&0&0&0\\ 2&1&0&1&1&2&0&0\\ 2&1&1&1&1&1&0&0\\ 2&2&1&0&2&1&0&1\\ 0&0&2&0&0&1&2&0\\ 1&2&2&0&0&0&0&1\\ 0&1&1&2&2&2&2&2\\ 2&2&0&1&0&1&2&1\\ 1&0&0&1&1&0&0&1\\ 1&2&2&1&1&2&1&0\\ 1&0&0&1&1&0&1&0\\ 0&0&2&0&2&2&1&1 \end{pmatrix}, \mathbf{T}_2'' = \begin{pmatrix} 1&1&0&0&2&1&1&1\\ 1&0&0&1&0&2&1&1\\ 2&1&1&2&2&1&1&0\\ 1&2&2&1&2&2&0&0\\ 0&1&1&2&0&1&2&1\\ 2&0&2&1&2&0&1&2\\ 2&0&2&2&2&0&2&1\\ 2&2&1&0&2&2&0&1\\ 2&1&0&1&2&1&1&0\\ 0&1&2&1&0&2&1&1\\ 2&2&2&1&1&0&1&2\\ 0&0&0&0&1&2&2&0\\ 0&2&0&0&2&1&1&0\\ 0&1&2&0&1&1&2&1\\ 1&2&0&2&0&1&2&1\\ 1&2&0&0&0&1&1&0 \end{pmatrix}$$

Recovering S: To find $\mathbf{W}'' := \mathbf{S}''\mathbf{M}_n = [\mathbf{w}''|\mathbf{w}''^q|\cdots|\mathbf{w}''^{q^{n-1}}]$, we find its first column $\mathbf{w}''$, which satisfy $\mathrm{Frob}_{j+1}(\mathbf{K}')\mathbf{w}'' = \mathbf{0}$, for $j = n-r, \ldots, n-1 = 7, 8$. By solving the overdetermined system

$$\begin{pmatrix} \mathbf{K}' \\ \mathrm{Frob}_7(\mathbf{K}') \end{pmatrix} \mathbf{w}'' = \begin{pmatrix} 1&0&0&0&0&b^{6158}&b^{1567}&b^{6415}\\ 0&1&0&0&0&b^{3943}&b^{4591}&b^{95}\\ 0&0&1&0&0&b^{4461}&b^{4216}&b^{3027}\\ 0&0&0&1&0&b^{3577}&b^{5899}&b^{1096}\\ 0&0&0&0&1&b^{6554}&b^{4266}&b^{907}\\ 1&0&0&0&0&b^{6426}&b^{2709}&b^{4325}\\ 0&1&0&0&0&b^{3501}&b^{3717}&b^{4405}\\ 0&0&1&0&0&b^{1487}&b^{3592}&b^{1009}\\ 0&0&0&1&0&b^{3379}&b^{4153}&b^{2552}\\ 0&0&0&0&1&b^{6558}&b^{1422}&b^{2489} \end{pmatrix} \mathbf{w}'' = \mathbf{0},$$

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

we obtain $\mathbf{w}'' = (b^{929}, b^{2174}, b^{2323}, b^{4231}, b^{3677}, b^{6313}, b^{2372}, b^{3245})$. We then compute

$$\mathbf{W}'' = \begin{pmatrix} b^{929} & b^{2787} & b^{1801} & b^{5403} & b^{3089} & b^{2707} & b^{1561} & b^{4683} \\ b^{2174} & b^{6522} & b^{6446} & b^{6218} & b^{5534} & b^{3482} & b^{3886} & b^{5098} \\ b^{2323} & b^{409} & b^{1227} & b^{3681} & b^{4483} & b^{329} & b^{987} & b^{2961} \\ b^{4231} & b^{6133} & b^{5279} & b^{2717} & b^{1591} & b^{4773} & b^{1199} & b^{3597} \\ b^{3677} & b^{4471} & b^{293} & b^{879} & b^{2637} & b^{1351} & b^{4053} & b^{5599} \\ b^{6313} & b^{5819} & b^{4337} & b^{6451} & b^{6233} & b^{5579} & b^{3617} & b^{4291} \\ b^{2372} & b^{556} & b^{1668} & b^{5004} & b^{1892} & b^{5676} & b^{3908} & b^{5164} \\ b^{3245} & b^{3175} & b^{2965} & b^{2335} & b^{445} & b^{1335} & b^{4005} & b^{5455} \end{pmatrix},$$

and

$$\mathbf{S}'' = \mathbf{W}''\mathbf{M}_8^{-1} = \begin{pmatrix} 2 & 2 & 2 & 1 & 2 & 0 & 0 & 2 \\ 1 & 2 & 1 & 1 & 2 & 0 & 1 & 2 \\ 2 & 1 & 0 & 2 & 0 & 2 & 1 & 0 \\ 2 & 2 & 1 & 1 & 2 & 1 & 2 & 2 \\ 0 & 2 & 1 & 2 & 0 & 0 & 0 & 2 \\ 1 & 0 & 1 & 0 & 1 & 1 & 1 & 2 \\ 1 & 0 & 2 & 0 & 1 & 2 & 2 & 0 \\ 0 & 1 & 1 & 0 & 2 & 2 & 0 & 1 \end{pmatrix}$$

Recovering core polynomials: To find our equivalent core polynomials H' and $\tilde{H}'$ we calculate $\mathbf{H}' = \mathbf{W}''^{-1}\left(\sum_{i=0}^{7} u'_i \mathbf{P}_{i+1}\right)\mathbf{W}''^{-t}$ as well as the value of $\tilde{\mathbf{H}}' = \mathbf{W}''^{-1}\left(\sum_{i=0}^{7} v'_i \mathbf{P}_{i+1}\right)\mathbf{W}''^{-t}$ and obtain

$$\mathbf{H}' = \begin{pmatrix} b^{2287} & b^{992} & b^{5159} & b^{4953} & b^{4144} & b^{6518} & b^{3920} & b^{4127} \\ b^{992} & b^{5165} & b^{5229} & b^{5023} & b^{4214} & b^{28} & b^{3990} & b^{4197} \\ b^{5159} & b^{5229} & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{4953} & b^{5023} & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{4144} & b^{4214} & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{6518} & b^{28} & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{3920} & b^{3990} & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{4127} & b^{4197} & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \tilde{\mathbf{H}}' = \begin{pmatrix} b^{87} & b^{1874} & b^{3075} & b^{2869} & b^{2060} & b^{4434} & b^{1836} & b^{2043} \\ b^{1874} & b^{6189} & b^{1832} & b^{1626} & b^{817} & b^{3191} & b^{593} & b^{800} \\ b^{3075} & b^{1832} & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{2869} & b^{1626} & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{2060} & b^{817} & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{4434} & b^{3191} & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{1836} & b^{593} & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{2043} & b^{800} & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Recovering the low degree polynomial: Once the core polynomials $\mathbf{H}' = [h_{ij}]$, $\tilde{\mathbf{H}}' = [\tilde{h}_{ij}]$ are recovered, our target is to build the low degree polynomial Ψ'' fundamental for the attacker to be able decrypt. So, we solve the following overdetermined systems

$$\begin{bmatrix} h_{1,r+1} & h_{1,r+2} & \cdots & h_{1,n-1} & h_{1,n} \\ \tilde{h}_{1,r+1} & \tilde{h}_{1,r+2} & \cdots & \tilde{h}_{1,n-1} & \tilde{h}_{1,n} \end{bmatrix}^\top \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} b^{5159} & b^{4953} & b^{4144} & b^{6518} & b^{3920} & b^{4127} \\ b^{3075} & b^{2869} & b^{2060} & b^{4434} & b^{1836} & b^{2043} \end{bmatrix}^\top \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \mathbf{0},$$

$$\begin{bmatrix} h_{2,r+1} & h_{2,r+2} & \cdots & h_{2,n-1} & h_{2,n} \\ \tilde{h}_{2,r+1} & \tilde{h}_{2,r+2} & \cdots & \tilde{h}_{2,n-1} & \tilde{h}_{2,n} \end{bmatrix}^\top \begin{bmatrix} y_0 \\ y_1 \end{bmatrix} = \begin{bmatrix} b^{5229} & b^{5023} & b^{4214} & b^{28} & b^{3990} & b^{4197} \\ b^{1832} & b^{1626} & b^{817} & b^{3191} & b^{593} & b^{800} \end{bmatrix}^\top \begin{bmatrix} y_0 \\ y_1 \end{bmatrix} = \mathbf{0},$$

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

and we obtain the solutions $[x_0, x_1]^\top = [b^{1418},\ b^{222}]^\top$ and $[y_0, y_1]^\top = [b^{2162},\ b^{2279}]^\top$. Then, we compute $b^{1418}\mathbf{H}' + b^{222}\breve{\mathbf{H}}'$ and $b^{2162}\mathbf{H}' + b^{2279}\tilde{\mathbf{H}}'$ obtaining respectively

$$\begin{pmatrix} b^{106} & b^{6092} & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{6092} & b^{3643} & b^{4437} & b^{4231} & b^{3422} & b^{5796} & b^{3198} & b^{3405} \\ 0 & b^{4437} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & b^{4231} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & b^{3422} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & b^{5796} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & b^{3198} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & b^{3405} & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} b^{1294} & b^{536} & b^{3144} & b^{2938} & b^{2129} & b^{4403} & b^{1905} & b^{2112} \\ b^{536} & b^{844} & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{3144} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{2938} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{2129} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{4503} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{1905} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ b^{2112} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Finally, we form the system

$$\begin{bmatrix} b^{4437} & b^{4231} & b^{3422} & b^{5796} & b^{3198} & b^{3405} \\ b^{3144} & b^{2938} & b^{2129} & b^{4403} & b^{1905} & b^{2112} \end{bmatrix}^\top \begin{bmatrix} z_0 \\ z_1 \end{bmatrix} = \mathbf{0},$$

we a solution $[z_0,\ z_1]^\top = [b^{1024},\ b^{5597}]^\top$, and we use it to compute our low degree polynomial,

$$\begin{aligned} \Psi'' &= b^{1024} X (b^{1418} H' + b^{222} \tilde{H}') + X^q (b^{2162} H' + b^{2279} \tilde{H}') \\ &= b^{6441} X^9 + b^{2097} X^7 + b^{852} X^5 + b^{1130} X^3 \end{aligned}$$

A.2 Low rank matrix forms

$$\begin{pmatrix} * & * & * & & * & & & \\ * & * & * & \dots & * & * & \dots & * \\ * & * & * & & * & & & \\ \vdots & & & \ddots & & & & \\ * & * & * & & * & & & \\ * & & & & & & & \\ \vdots & & & & & & & \\ * & & & & & & & \end{pmatrix}$$

Case $s = 0$.

$$\begin{pmatrix} * & * & * & & * & * & \dots & * \\ * & * & * & \dots & * & & & \\ * & * & * & & * & & & \\ \vdots & & & \ddots & & & & \\ * & * & * & & * & & & \\ * & & & & & & & \\ \vdots & & & & & & & \\ * & & & & & & & \end{pmatrix}$$

Case $s = 1$.

A.3 Comparison to Previous MinRank Analysis

It has been noted in [26] and [32], for example, that we may consider ZHFE to be a high degree instance of multi-HFE with two branches, i.e. $(X_1, X_2) \mapsto (F_1(X_1), F_2(X_2))$. This intuition is, however, mistaken. If we regard ZHFE as an instance of multi-HFE with $N = 2$, we must impose the relation $X_1 = X_2$, which

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

considerably changes the rank analysis. This fact is missing from the discussion of the KS-attack complexity in both [26], before the low Q-rank property was discovered and in [32] after the low Q-rank property of ZHFE was announced in [23].

Although our complexity analysis is quite similar to the analysis of the multi-HFE attack of [2], there are, however, a few important distinctions that arise and elucidate the disparity between the complexity reported in [32] and our derived complexity. First, in multi-HFE, with N branches, the number of variables over the extension field required to express the quadratic function is N; thus the dimension of matrices required to construct a matrix representation for the central map is Nn, see [2, Proposition 5]. In ZHFE, a single variable is required over the extension field, and thus dimension n matrices are all that is required. Another distinction is that the rank bound for multi-HFE is due to a simultaneous degree bound in each of N variables over the extension field, producing a rank bound on the dimension Nn matrices of NR, where R is the rank, see [2, Lemma 3]. In ZHFE, the rank bound is due to the degree bound on Ψ, and only applies to a single variable; thus the rank bound is merely $R = r + 1$ where $r = \lceil log_q(D) \rceil$. Moreover, the minrank instance involves twice as many matrices in relation to the dimension of the matrices when compared with the minrank instances arising in multi-HFE. A final important distinction is that after the simultaneous MinRank is solved, an extra step, the derivation of an equivalent Ψ map, is required to recover a full private key.

These distinctions lead to vastly different complexity estimates on the KS-attack with minors modeling for ZHFE. In [32], the complexity of the KS-attack is reported as $\mathcal{O}(n^{2(R+1)\omega})$ citing the complexity estimate in [2]. Indeed, the complexity would be $\mathcal{O}(n^{(2R+1)\omega})$ for the KS-attack on a multi-HFE instance with Q-rank R according to [2, Proposition 13]. We are uncertain where the extra power of ω enters the analysis of [32]. We note that in [32] they claim that with an unrealistic linear algebra constant of $\omega = 2$ they obtain from this formula a complexity of 2^{138} for the KS-attack; however, computing $n^{2(R+1)\omega} = 55^{2(4+1)(2)} \approx 2^{115}$, whereas using the more realistic value $\omega = 2.3766$ we obtain 2^{138} as reported. This is apparently a minor editing mistake.

The reality is that the analysis in [2, Proposition 13] is related but not directly applicable to ZHFE since ZHFE does *not* correspond to multi-HFE with $N = 2$. Using an analysis analogous to the techniques in [2, Section 7], we derive above, using rank $R = r + 1$, an estimate of $\mathcal{O}(n^{(r+2)\omega})$. Using the proposed parameters $q = 7$, $n = 55$, and $D = 105$ which imply $r = 3$, we obtain an attack complexity of 2^{64}. We thus conclude that ZHFE is broken.

Cabarcas, Daniel; Smith-Tone, Daniel; Verbel, Javier.
"Key Recovery Attack for ZHFE."
Paper presented at PQCrypto 2017: The Eighth International Conference on Post-Quantum Cryptography, Utrecht, Netherlands. June 26, 2017 - June 28, 2017.

Security Analysis and Key Modification for $ZHFE$

Ray Perlner[1] and Daniel Smith-Tone[1,2]

[1]National Institute of Standards and Technology,
Gaithersburg, Maryland, USA
[2]Department of Mathematics, University of Louisville,
Louisville, Kentucky, USA

ray.perlner@nist.gov, daniel.smith@nist.gov

Abstract. $ZHFE$, designed by Porras et al., is one of the few promising candidates for a multivariate public-key encryption algorithm. In this article we extend and expound upon the existing security analysis on this scheme. We prove security against differential adversaries, complementing a more accurate and robust discussion of resistance to rank and algebraic attacks. We further suggest a modification, $ZHFE^-$, a multivariate encryption scheme which retains the security and performance properties of $ZHFE$ while optimizing key size in this theoretical framework.

Key words: Multivariate Cryptography, HFE, $ZHFE$, Discrete Differential, MinRank, Q-rank

1 Introduction

Since the late 1990s, a large international community has emerged to face the challenge of developing cryptographic constructions which resist attacks from quantum computers. The birth of this new discipline is due primarily to the discovery by Peter Shor in the mid 90s, see [1], of algorithms for factoring and computing discrete logarithms in polynomial time on a quantum computing device. The term post-quantum cryptography was coined to refer to this developing field and to emphasize the fact that information security in a quantum computing world is a fundamentally new science.

Today, we face mounting evidence that quantum computing is not a physical impossibility but merely a colossal engineering challenge. With the specter of the death of classical asymmetric cryptography looming on the horizon, it is more important than ever that we develop systems for authentication, confidentiality and key exchange which are secure in the quantum paradigm. We thus are forced to turn to problems of greater difficulty than the classical number theoretic constructs.

Systems of polynomial equations have been studied for thousands of years and have fueled the development of several branches of mathematics from classical

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

to modern times. Multivariate Public Key Cryptography(MPKC) has emerged from the serious investigation of computational algebraic geometry that reached maturity in the latter half of the last century. Today, we see MPKC as one of a few serious candidates for security in the post-quantum world.

A fundamental problem on which the security of any multivariate cryptosystem rests is the problem of solving systems of quadratic equations over finite fields. This problem is known to be NP-hard, and copious empirical evidence indicates that the problem is hard even in the average case. There is no known significant reduction of the complexity of this problem in the quantum model of computing, and, indeed, if this problem is discovered to be solvable in the quantum model, we can solve all NP problems and the task of securing information might be hopeless in principle. We thus reasonably suspect that MPKC will survive the transition into the quantum world.

Though multivariate cryptosystems almost always suffer from rather large key sizes, the key sizes are rarely so large that they are impractical and these systems can often be quite attractive in certain other aspects of performance. Some systems are very fast, having speeds orders of magnitude faster than RSA, [2–4]. Some schemes combine speed with power efficiency and small signature sizes, [5, 6]. Perhaps most importantly, it is generally simple to parameterize multivariate systems in such a way that vastly different properties are derived foiling various attack methodologies.

One great difficulty historically for MPKC is encryption. Though there are several viable options for digital signatures, see [5–8], there is a general absence of long-lived encryption systems. In the last couple of years, a couple of new encryption techniques have been proposed, see [9–11]. These systems are based on the simple idea, proposed by Ding, that the structure of a system of equations can retain injectivity without an extremely restrictive structure if the codomain is of much larger dimension than the domain.

In [12], however, a new and unexpected attack was presented on the *ABC* simple matrix encryption scheme of [9]. This attack is notable in that the complexity is far less asymptotically than predicted by the analysis in [9], though it does not break the scheme outright. This begs the question of the tightness of the security analyses in [10, 11] and the extent to which we can trust in the security of such young schemes in a field which has no significant success history in encryption.

Furthermore, one might ask whether there is some middleground on the ratio of the dimension of the codomain to that of the domain for these multivariate encryption schemes. Even if one concurs that relaxing the relationship between the dimensions of the domain and codomain enhance the security of injective maps, it remains unclear that the disparity should be so large as in the proposed schemes in which there are at least twice as many equations as variables.

In this article we extend and expound upon the security analysis in [11], incorporating some of the theoretical models of assurance presented in [13–15]. We prove security against differential adversaries complementing the discussion of resistance to algebraic attacks provided in [11]. We further elucidate the rank

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

structure of $ZHFE$ and specifically note some necessary, but trivial, key restrictions for security which were apparently overlooked in [11]. We further suggest a modification, $ZHFE^-$, a multivariate encryption scheme which retains the security and performance properties of $ZHFE$ while optimizing key size in this theoretical framework.

The paper is organized as follows. The next section introduces the notion of big field schemes and presents the prototypical such cryptosystem, HFE. In the following section, we define the Q-rank of a multivariate system of equations and discuss the central nature of this concept in the field. The subsequent section presents the $ZHFE$ encryption scheme and calculates some of its inherent parameters. Next we present a thorough security analysis of $ZHFE$, complementing and expanding the analysis provided in [11] and offering security assurance against a differential adversary as well as discussing parameters securing $ZHFE$ against rank and algebraic attacks. Subsequently, we present and analyze $ZHFE^-$, a new multivariate encryption scheme based on $ZHFE$ and the minus modifier. Finally, we note parameter choices for $ZHFE^-$ and discuss the role that the new methodology for multivariate encryption fills in the literature.

2 *HFE*

Several multivariate cryptosystems belong to a family collectively known as "big field" schemes. Such schemes are constructed using two ideas. The first is an equivalence between functions on a degree n extension k of a finite field $\mathbb{F}_q$ and functions on an n-dimensional $\mathbb{F}_q$-vector space. The second is an isomorphism of polynomials which allows one to hide structure in a function.

To see the equivalence, notice that a vector space isomorphism between k and an n-dimensional vector space over $\mathbb{F}_q$ extends to a vector space isomorphism between the space of univariate functions from k to itself and the space of multivariate n-dimensional vector-valued polynomial functions from $\mathbb{F}_q^n$ to itself. (Specifically, given an isomorphism $\phi : \mathbb{F}_q^n \to k$ and a function $f : k \to k$, the function $\phi^{-1} \circ f \circ \phi$ is such a function from $\mathbb{F}_q^n$ to itself; furthermore, this identification is a 1-1 correspondence.)

The second idea, the isomorphism of polynomials, is defined in the following manner.

Definition 1 *Two vector valued multivariate polynomials f and g are said to be* isomorpic *if there exist two affine maps T, U such that $g = T \circ f \circ U$.*

Together these ideas allow us to build an isomorphic copy of a structured univariate map with domain k while hiding the structure. The construction is sometimes called the butterfly construction because of the shape of its defining commutative diagram. Specifically, $P = T \circ \phi^{-1} \circ f \circ \phi \circ U$ produces a perturbed vector-valued version of the structured univariate polynomial f.

The Hidden Field Equations (HFE) scheme was first presented by Patarin in [16] as a method of avoiding his linearization equations attack which broke

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

the C^* scheme of Matsumoto and Imai, see [17] and [18]. The basic idea of the system is to use the butterfly construction to hide the structure of a low degree polynomial that can be inverted efficiently over k via the Berlekamp algorithm [19], for example.

More specifically, we select an effectively invertible "quadratic" map $f : k \to k$, quadratic in the sense that every monomial of f is a product of a constant and two Frobenius multiples of x. Explicitly any such "core" map f has the form:

$$f(x) = \sum_{\substack{i \leq j \\ q^i + q^j \leq D}} \alpha_{i,j} x^{q^i + q^j} + \sum_{\substack{i \\ q^i \leq D}} \beta_i x^{q^i} + \gamma.$$

The bound D on the degree of the polynomial is required to be quite low for efficient inversion.

The HFE scheme was designed to be used as an encryption or a signature scheme. To generate a signature (or to decrypt), one computes, successively, $v = T^{-1}y$, $u = f^{-1}(v)$ and $x = U^{-1}u$. The vector x is the signature (or the plaintext). For verification (or encryption), one simply evaluates the public polynomials, P, at x. If $P(x)$ which is equal to $T \circ f \circ U(x)$ is equal to y, the signature is authenticated (or the ciphertext is y).

3 Q-Rank

The defining charactersitic of HFE, the degree bound, which is necessary for the effective inversion of the central map, ensures that the scheme has low rank as a quadratic form over k, as described below. This property assures that the central map of HFE is vulnerable to Kipnis-Shamir modeling, see [20, 21].

Recall that any quadratic map $f : k \to k$ can be written

$$f(x) = \sum_{0 \leq i,j < n} \alpha_{ij} x^{q^i + q^j}.$$

We can equivalently express f as a vector function over the 1-dimensional k-algebra $\psi : k \to k^n$ where

$$\alpha \overset{\psi}{\mapsto} \left[\alpha \; \alpha^q \; \ldots \alpha^{q^{n-1}}\right]^T,$$

in the form $f(X) = X^T[\alpha_{ij}]X$ where $X = [x \; x^q \; \ldots \; x^{q^{n-1}}]^T$.

Any quadratic form over k can be expressed as a symmetric matrix, and over characteristic $p \neq 2$ a change of basis can be performed which transforms this matrix into an equivalent diagonal form. The rank of this matrix is the rank of the quadratic form. We call this rank the Q-rank of f, that is the rank of f as a quadratic function.

We note here that Q-rank is invariant under polynomial isomorphism, thus the Q-rank of a central map of a cryptosystem is the same as the Q-rank of the public key, unless, of course, the minus or projection modifiers are utilized. We

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

also note that the Q-rank is explicitly exploited in the attacks of [20, 21] and plays a central role in the derivation of degree of regularity bounds for several prominent cryptosystems, see [22–24]. Further, there seems to be a complicated relationship between the Q-rank of a field map and the presence of differential symmetric or invariant relations, see, for example [15]. Consequently, Q-rank seems to be emerging as a central concept in multivariate cryptography and in computational algebra.

4 *ZHFE*

$ZHFE$ was introduced in [11]. The idea is to construct an encryption scheme with a high Q-rank central map preventing attacks such as [21] exploiting this weakness. The scheme is notable among "big field" schemes which typically require some low Q-rank map for efficient inversion. Low Q-rank is in fact required for inversion in this setting as well, however, the system attempts to hide the low Q-rank structure in the public key.

The construction concatenates two high degree quadratic maps (with special structure) to form the central map. Specifically, the two general form quadratic maps f_0 and f_1 are derived by constructing a low degree (maximum degree D) cubic map

$$\Psi(x) = x\left[L_{00}f_0(x) + L_{01}f_1\right] + x^q\left[L_{10}f_0 + L_{11}f_1\right], \tag{1}$$

where L_{ij} is a linear map and the square brackets indicate multiplication over k.

To solve for f_0 and f_1 it suffices to set coefficients for the linear maps and for Ψ to recover a system of linear equations in the unknown coefficients of f_0 and f_1. In the homogeneous case, there are collectively $n^2 + n$ coefficients of f_0 and f_1 in k. Due to its low degree and the requirement that it satisfy (1), Ψ is constrained to be of the form

$$\Psi(x) = \sum_{i=0}^{1} \sum_{\substack{i\le j\le k \\ q^i+q^j+q^k\le D}} \alpha_{i,j,k}x^{q^i+q^j+q^k} + \sum_{i=0}^{1} \sum_{\substack{i\le j \\ q^i+q^j\le D}} \beta_{i,j}x^{q^i+q^j} + \sum_{i=0}^{1}\gamma_i x^{q^i}. \tag{2}$$

A cubic of the form (2) has n^2 coefficients over k, and thus for *any fixed choice* of Ψ and L_{ij} there are n^2 constraints on a linear system of dimension $n^2 + n$. Thus with probability roughly $1 - q^{-n}$, there is an n-dimensional space of coefficients for the maps f_0 and f_1.

Once, constructed, the central map $(y_0, y_1) = (f_0(x), f_1(x))$ can be inverted by using Berlekamp's algorithm to solve the low degree polynomial equation:

$$\Psi(x) - x\left[L_{00}y_0 + L_{01}y_1\right] - x^q\left[L_{10}y_0 + L_{11}y_1\right] = 0.$$

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

5 Analysis of $ZHFE$

A few avenues of attack have evolved along with the development of multivariate cryptosystems relying on a hidden large algebra structure. These attacks can be characterized as differential, see [25, 12], as minrank, see [20, 21], or as algebraic, see [26]. We analyze the security of $ZHFE$ against each of these attack models.

5.1 Algebraic

Algebraic attacks attempt to decrypt a given ciphertext y by solving the system of equations $P(x) = y$ directly. The term "algebraic" refers to the fact that these are generic algorithms for solving arbitrary systems of polynomial equations.

While these attacks are not structural, in the sense of being defined based on the structure of the system of equations, the algorithms employed can naturally take advantage of certain properties of the systems. In practice, the complexity of algorithms for solving these systems of equations is closely connected to the degree of regularity of the system.

The degree of regularity of a system of equations is the degree at which the first nontrivial degree fall occurs. Specifically, consider a generating set of an ideal $I = \langle g_1, \ldots, g_m \rangle \in \mathbb{F}_q[x_1, \ldots, x_n]$. We may generate elements of I by selecting polynomials $p_i \in \mathbb{F}_q[x_1, \ldots, x_n]$ and computing

$$\sum_{i=1}^{m} p_i g_i.$$

A degree fall occurs when the degree of this sum is less than the maximum degree of $p_i g_i$. Clearly some degree falls are due to trivial syzygies such as $-g_j g_i + g_i g_j = 0$ and $(g_i^{q-1} - 1)g_i = 0$. The smallest degree, $max_i p_i g_i$ such that the above sum has a nontrivial degree fall is the degree of regularity.

A great deal of literature is devoted to finding bounds for the degree of regularity of quadratic systems, see [22–24, 27]. In practice one can find a lower bound for the degree of regularity by studying toy examples of schemes and seeing how the degree of regularity changes as the parameters change.

Such an analysis for $ZHFE$ is quite straight forward. As mentioned in [11] the degree of regularity for toy $ZHFE$ systems matches exactly the degree of regularity for random systems of equations of the same size, at least for relatively small instances. Considering the connection between Q-rank and the degree of regularity as derived in [22–24, 27], we conclude that a thorough Q-rank analysis of $ZHFE$ will verify the security of the scheme against algebraic attacks. We perform this analysis in Section 5.4.

5.2 Differential Symmetric

As shown in [25], symmetric relations involving the discrete differential of a central map can induce a symmetry in the public key of a multivariate cryptosystem.

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

In certain circumstances, these relations can reveal properties of the extension field structure, and weaken the public key. Indeed one can easily turn the attack on SFLASH of [25], which converts an instance of C^{*-} into a compatible instance of C^*, into a direct key-recovery attack utilizing the derived representation of elements of the extension field.

As shown in [13] the maps inducing a linear differential symmetry for C^* schemes are precisely those corresponding to multiplication by an element of the extension field. Thus one may rightfully expect that nontrivial symmetric relations on the differential of a central map are uncommon. It is shown, however, in [13] and [15] that nontrivial symmetries can and often do exist even for cases as general as HFE.

As a specific example of the phenomenon of differential symmetries for general polynomials, consider the map $f(x) = x^{q^3+q^2} + x^{q^2+1}$ over a degree 6 extension of the characteristic 2 field $\mathbb{F}_q$. One can easily verify that the general linear symmetry structure, defined as

$$Df(La, x) + Df(a, Lx) = \Lambda_L Df(a, x),$$

is satisfied by the selection

$$Lx = \alpha x^{q^4} + \alpha x^q + \beta x \text{ and } \Lambda_L x = 0,$$

where $\alpha^{q^3} = \alpha$ and $\beta^q = \beta$. Thus there is a 4-dimensional $\mathbb{F}_q$-subspace of linear maps L satisfying the above differential symmetric relation for some choice of Λ_L, while the space of all $\mathbb{F}_q$-linear maps from the extension to itself is only of dimension 36. Consequently, a hypothetical cryptosystem based on this map would be vulnerable to an attack removing the minus modifier, similar to [25], among other weaknesses. Quite specifically, the distillation procedure described in [25] is effective in this instance. We note that this scenario is by no means limited to toy examples such as this one or even instances with Q-rank one; thus, the verification of the absence of differential symmetries is an important task for any multivariate cryptosystem, particularly those including the minus modifier.

In analyzing the differential symmetric properties of $ZHFE$, we may directly analyze the public key or we may study the differential of the Ψ map. We consider both interlinked cases explicitly.

The public key P consists of $2n$ polynomials. The defining characteristic of these polynomials is that $P = T(f_0||f_1)U$. Thus P does not behave like a random system. There exists a low degree cubic map Ψ such that

$$\begin{aligned}\Psi(Ux) =&(Ux)(L_{00}(T^{-1})_1 P(x) + L_{01}(T^{-1})_2 P(x)) \\ &+ (Ux)^q (L_{10}(T^{-1})_1 P(x) + L_{11}(T^{-1})_2 P(x)).\end{aligned} \tag{3}$$

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

We note that $(T^{-1})_i P(x) = f_i(Ux)$. We may now implicitly differentiate this equation obtaining

$$\begin{aligned} D\Psi(Ua, Ux) =&(Ua)(L_{00}f_0(Ux) + L_{01}f_1(Ux)) \\ &+ (Ua)^q(L_{10}f_0(Ux) + L_{11}f_1(Ux)) \\ &+ (Ux)(L_{00}Df_0(Ua, Ux) + L_{01}Df_1(Ua, Ux)) \\ &+ (Ux)^q(L_{10}Df_0(Ua, Ux) + L_{11}Df_1(Ua, Ux)). \end{aligned} \tag{4}$$

The above is a biquadratic relation in a and x, and as such doesn't immediately reveal a computational way to recover information about the hidden structure of P. To convert this relation into a form in which we can apply linear algebra techniques we require a second differential. For more information on a more general theory of discrete differential equations, see [28].

Since the differential is symmetric, we get the same answer whether we differentiate with respect to a or to x.

$$\begin{aligned} D^2\Psi(Ua, Ub, Ux) =&(Ua)(L_{00}Df_0(Ub, Ux) + L_{01}Df_1(Ub, Ux)) \\ &+ (Ua)^q(L_{10}Df_0(Ub, Ux) + L_{11}Df_1(Ub, Ux)) \\ &+ (Ub)(L_{00}Df_0(Ua, Ux) + L_{01}Df_1(Ua, Ux)) \\ &+ (Ub)^q(L_{10}Df_0(Ua, Ux) + L_{11}Df_1(Ua, Ux)) \\ &+ (Ux)(L_{00}Df_0(Ua, Ub) + L_{01}Df_1(Ua, Ub)) \\ &+ (Ux)^q(L_{10}Df_0(Ua, Ub) + L_{11}Df_1(Ua, Ub)). \end{aligned} \tag{5}$$

Now, due to the fact that Ψ is cubic with a small degree bound, $D^2\Psi$ is a cubic form of low rank. In fact, the existence of linear maps U and $L_{ij}(T^{-1})_j$ such that equations (3) and (5) hold while $D^2\Psi$ has low cubic rank is the defining characteristic of $ZHFE$.

In spite of the existence of this structure, it is unclear how to proceed. One might consider a cubic version of the rank attack from [29], however, the selection of the maps $L_{ij}(T^{-1})_j$ corresponds to solving a minrank problem on a 3-tensor, $D^2\Psi$. Though there is a possibility that the instances of the 3-tensor rank problem arising from this differential equation may lie in a class which are easy to solve, the general 3-tensor rank problem is known to be NP-hard and there does not seem to be any evidence that these instances are any more structured than arbitrary instances of the same rank.

5.3 Differential Invariant

As exemplified in [12] and [30], invariant relations on the differential of a public key can be exploited in key recovery. Although we may analyze the differential invariant structure of the public key of $ZHFE$ directly, there is not in general any nontrivial invariant due to the fact that the structure of $ZHFE$ is hidden in the cubic Ψ map. A couple of generalizations of differential invariants of quadratic functions are derived for higher q-degree functions in [28]. The most relaxed generalization for cubics is given in the following definition.

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

Definition 2 *A* differential invariant *of a cubic function f is a pair of subspaces $V_1, V_2 \subseteq k$ for which there exists a subspace W with $dim(W) \leq mindim(V_i)$ such that for all $A \in span D^2 f_i$, we have $D^2 f(a, b, x) = 0$ for all $a \in V_1$, $b \in V_2$ and $x \in W^{\perp}$.*

In the quadratic case, a differential invariant could be seen as a subspace of k on which Df simultaneously acts in every coordinate the same way, that is, always sending that subspace to the same space of linear forms of no larger dimension. In the cubic case we can realize a differential invariant as a subspace V_1 of k and a subspace (defined by V_2) of induced bilinear forms from $D^2 f$ each element of which maps V_1 to the same space of linear forms, W, of no larger dimension. The minimum condition on the dimension of W is due to the symmetry of $D^2 f$; we could equivalently consider the subspace V_2 of k and the subspace of bilinear forms from $D^2 f$ induced from V_1.

It is straightforward to show that the Ψ map of $ZHFE$ has no differential invariant structure. Following the technique of [15], without loss of generality, due to the symmetry, we let $\hat{a} \in V_1$, $\hat{b}, \hat{x} \in V_2$, and let S be a surjective linear map from V_2 to W. The existence of a differential invariant implies the equation

$$\begin{aligned} 0 &= D^2\Psi(\hat{a}, \hat{b}, S\hat{x}) \\ &= \sum_{\substack{0 \leq i,j,l < n \\ q^i + q^j + q^l \leq D}} \alpha_{ijl} \hat{a}^{q^i} \hat{b}^{q^j} (S\hat{x})^{q^l}. \end{aligned} \tag{6}$$

Since by symmetry D is much smaller than $dim(V_1)$ or $dim(V_2)$, (6) is already reduced modulo the minimal polynomial $\mathcal{M}_{V_1}(a)$ of V_1 as an element in $k[a]$ and modulo the minimal polynomial $\mathcal{M}_{V_2}(b)$ of V_2 as an element in $k[b]$. Thus the collection $\{\hat{a}, \hat{a}^q, \ldots, \hat{a}^{q^{d_1}}, \hat{b}, \ldots, \hat{b}^{q^{d_2}}\}$ is independent in $k[a, b]/\langle \mathcal{M}_{V_1}(a), \mathcal{M}_{V_2}(b) \rangle$. Therefore, we obtain the equations

$$\sum_{\substack{0 \leq i,j < n \\ q^i + q^j + q^l \leq D}}^{0 \leq l < n} \alpha_{ijl} (S\hat{x})^{q^l} = 0.$$

We then obtain the analogous result of [15]; statistically, S must be the zero map on V_2, contradicting the nontriviallity of the differential invariant. Furthermore, we also obtain the result that if any power of q is unique there is no nontrivial differential invariant.

5.4 Q-rank

A further attack vector for $ZHFE$ is to perform a minrank attack using the Kipnis-Shamir methodology of [20] and the improved version in [21]. The attack searches for a low rank k-linear combination of the differentials of the public key. The general minrank problem is known to be NP-complete, see [31] but in practice the complexity depends on the lowest rank map in the space.

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

It was shown in [21] that the smallest such rank is equal to the smallest Q-rank of the image of the public key under any full rank $\mathbb{F}_q$-linear map. Notice that for (1) to hold we must have that the $x^{q^i+q^j}$ term in $L_{00}f_0 + L_{01}f_1$ to have coefficient 0 for $q^i + q^j + 1 > D$ and $i, j \neq 1$. This restriction induces a relation on the quadratic representations of $L_{00}f_0$ and $L_{01}f_1$. Specifically, if

$$L_{00}f_0(x) + L_{01}f_1(x) = \begin{bmatrix} x \\ x^q \\ \vdots \\ x^{q^{n-1}} \end{bmatrix}^T \begin{bmatrix} \alpha_{00} & \frac{\alpha_{01}}{2} & \cdots & \frac{\alpha_{0(n-1)}}{2} \\ \frac{\alpha_{01}}{2} & \alpha_{11} & \cdots & \frac{\alpha_{1(n-1)}}{2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\alpha_{0(n-1)}}{2} & \frac{\alpha_{1(n-1)}}{2} & \cdots & \alpha_{(n-1)(n-1)} \end{bmatrix} \begin{bmatrix} x \\ x^q \\ \vdots \\ x^{q^{n-1}} \end{bmatrix},$$

then $\alpha_{ij} = 0$ for $q^i + q^j > D$ and $i, j \neq 1$. Thus $L_{00}f_0 + L_{01}f_1$ has the form

$$\left[\begin{array}{ccccc|ccc} \alpha_{00} & \frac{\alpha_{01}}{2} & \frac{\alpha_{02}}{2} & \cdots & \frac{\alpha_{0D}}{2} & 0 & \cdots & 0 \\ \frac{\alpha_{01}}{2} & \alpha_{11} & \frac{\alpha_{12}}{2} & \cdots & \frac{\alpha_{1D}}{2} & \frac{\alpha_{1(D+1)}}{2} & \cdots & \frac{\alpha_{1(n-1)}}{2} \\ \alpha_{02} & \frac{\alpha_{12}}{2} & \alpha_{22} & \cdots & \frac{\alpha_{2D}}{2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \frac{\alpha_{0,D}}{2} & \frac{\alpha_{1D}}{2} & \frac{\alpha_{2D}}{2} & \cdots & \alpha_{DD} & 0 & \cdots & 0 \\ \hline 0 & \frac{\alpha_{1(D+1)}}{2} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \frac{\alpha_{1(n-1)}}{2} & 0 & \cdots & 0 & 0 & \cdots & 0 \end{array}\right],$$

and has rank no more than $\lceil log_q(D) \rceil + 2$. Hence, if L_{ij} are nonsingular, the $Q - rank$ of $f_0||f_1$ is bounded by $\lceil log_q(D) \rceil + 2$.

In spite of the alarming relation derived above, Q-rank does not appear to be a weakness for ZHFE when one selects L_{ij} to have reasonable corank. One can check that for small r, insisting that L_{ij} have corank r increases the possible Q-rank of $f_0||f_1$ by $2r$. Also, having L_{ij} with even moderately large corank doesn't produce a non-negligible probability of decryption ambiguity due to the zero expectation of the dimension of the intersection of the kernels of L_{ij}. Furthermore, recall that we have at least n degrees of freedom over k in selecting f_0 and f_1 for *any* choice of L_{ij}. Thus the Kipnis-Shamir attack, which is exponential in the Q-rank of the scheme, is trivially thwarted with simple parameter restrictions, though we note that the lack of such restriction on the rank of L_{ij} in [11] is apparently an oversight.

5.5 Equivalent Keys

In [32], the question of the number of equivalent keys for multivariate cryptosystems is explored. This question is quite relevant for $ZHFE$, as well, since there can clearly be multiple private keys allowing one to decrypt a public key. The danger in this vein would be if there is insufficient entropy in public keys due to massive redundancy in private keys.

To analyze the number of equivalent keys, we first determine the number of possible pairs f_0, f_1 satisfying (1) for a fixed Ψ and L_{ij}. As mentioned in Section

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

4, a map of the form Ψ has n^2 coefficients over k, and due to the degree bound only s of these can be nonzero. Thus with L_{ij} fixed, we have n^2+n unknown coefficients for f_0 and f_1 over k, and so we have $n^2+n-(n^2-s)=n+s$ degrees of freedom in choosing the pair f_0, f_1 for a fixed private key.

Next we consider the same relation with f_0, f_1 fixed. For specificity, let $f_i(x)=\sum_{0\le v\le w<n}\alpha_{ivw}x^{q^v+q^w}$. Given the existence of L_{ij} and Ψ, we have the relation

$$\begin{aligned}\Psi(x)=&\sum_{t=0}^{1}\sum_{i=0}^{n-1}\sum_{0\le v\le w<n} l_{0ti}\alpha_{tvw}^{q^i}x^{q^{v+i}+q^{w+i}+1}\\ &+\sum_{t=0}^{1}\sum_{i=0}^{n-1}\sum_{0\le v\le w<n} l_{1ti}\alpha_{tvw}^{q^i}x^{q^{v+i}+q^{w+i}+q},\end{aligned}\tag{7}$$

where l_{ijl} are the unknown coefficients of the linearized polynomial form of L_{ij}. There are implicitly n^2-s linear relations on the $4n$ unknown coefficients of L_{ij}, as well as the rank restrictions on these maps; thus, for $n>4$ we expect an unique solution, and thus an unique Ψ as well.

Given a public key, there is a fixed relationship $P=T(f_0||f_1)U$. We note that different choices of T can be accommodated by different choices of L_{ij} by (3). In contrast, statistically there is only one selection of U which maintains the structure of the key. Thus $M(f_0||f_1)$, $L_{ij}(M^{-1})_i$, Ψ form distinct equivalent private keys for all invertible M. One can see this result as indicating that the security of $ZHFE$ is more closely related to the IP1S problem than the IP problem.

We therefore have roughly q^{4n^2} equivalent private keys for any given public key. Since there are q^{5n^2+sn} possible choices of private keys, there are on the order of q^{n^2+sn} nonequivalent public keys. Consequently, there is sufficient entropy in public keys.

6 $ZHFE$ Key Modification, $ZHFE^-$

6.1 Design

As mentioned in the previous section, there are many degrees of freedom in selecting f_0 and f_1, even when Ψ and L_{ij} for $(i,j)\in\{0,1\}^2$ are fixed. These facts naturally lead to the question of whether it is possible to develop a "minus" modification of $ZHFE$ preserving the essential injectivity of the original scheme.

Analogous to the analysis in the last section, we compute the degrees of freedom in selecting f_0 and f_1 when the L_{ij} for $(i,j)\in\{0,1\}^2$ are fixed and when the degree bound for Ψ is fixed. Because we are decreasing the dimension of f_0 or f_1 or both, we compute over $\mathbb{F}_q$.

Recall from section 5 that there are n^2 possible nonzero coefficients of a cubic polynomial of the form of Ψ over k, and that with only the degree bound restriction, n^2-s of these must be zero. Expressing this fact over $\mathbb{F}_q$, we see

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

that there are $n^3 - sn$ linear constraints. Considering the maps $L_{i,j}$ to be of corank c, we require an additional $2cn - 2n$ relations to be satisfied, for a total of $n^3 - sn + 2cn - 2n$ linear constraints. Allow the total combined output dimension of f_0 and f_1 over $\mathbb{F}_q$ to be $n + t$. Since there are $\binom{n}{2} + n = \binom{n+1}{2}$ homogeneous quadratic monomials in each coordinate, there are $(n+t)\binom{n+1}{2}$ coefficients in our linear system.

$$(n+t)\binom{n+1}{2} \geq n^3 - sn + 2cn - 2n$$
$$(n+1)t \geq n^2 - n - 2s + 4c - 4.$$

For realistic values of s, it is possible to get t as low as $n-2$, and $n-1$ is always possible. Thus we consider removing two public equations. For symmetry and simplicity, we choose to remove one coordinate from each of f_0 and f_1, making them both maps from $\mathbb{F}_q^n$ to $\mathbb{F}_q^{n-1}$.

Remark 1 *This technique makes $ZHFE^-$ much more similar to small field schemes. The central map is no longer defined as a pair of maps over the extension field.*

Generation of the central map proceeds exactly as in $ZHFE$, with the exception that the linear maps L_{ij} are now representable as $n \times (n-1)$ matrices with entries in $\mathbb{F}_q$. As with $ZHFE$ we identify the image of L_{ij} with k to obtain relation (1).

Inversion of the central map proceeds exactly as with $ZHFE$. Now since both f_0 and f_1 map into a smaller space, there is a possibility of decryption failure beyond that of $ZHFE$. Under the heuristic that f_0 and f_1 are random quadratic maps from $\mathbb{F}_q^n$ to $\mathbb{F}_q^{n-1}$, one computes the probability that $f_0(y)||f_1(y) = f_0(x)||f_1(y)$ for a fixed x to be q^{2-2n}. While f_0 and f_1 are not random, we expect this quantity to be correct, and therefore the probability of decryption failure is increased by q^{2-2n}. Assuming parameters similar to $ZFHE$, this probability is roughly 2^{-300}, which is well within reason.

6.2 Analysis

The differential analysis from the previous section carries over nearly verbatim to the case of $ZHFE^-$. In particular, the 3-tensor structure of the differential remains essentially the same, though over a slightly diminished space. We therefore conclude that $ZHFE^-$ is as secure as $ZHFE$ against a differential symmetric or invariant attack.

Further, the degree of regularity of a subset of a system of relations is bounded below, as noted in [22], by the degree of regularity of the entire system. Thus, in comparison with any full rank $ZHFE$ scheme of the same Q-rank, the degree of regularity is at least as high, and so once again the resistance to algebraic attacks and attacks in the Kipnis-Shamir model is reduced to Q-rank analysis.

Unlike the differential security criteria, Q-rank is not monotone with respect to the composition of projections, a fact which can be seen by observing that

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

$g(x) \in k[x]$, where k is an even degree n extension of $\mathbb{F}_q$, defined by $g(x) = x^{2q^{n/2}} + x^2$ clearly has Q-rank 2, whereas the composition with the projection $\pi(x) = x^{q^{n/2}} - x$ produces

$$\begin{aligned}\pi(g(x)) &= (x^{2q^{n/2}} + x^2)^{q^{n/2}} - (x^{2q^{n/2}} + x^2) \\ &= x^{2q^n} + x^{2q^{n/2}} - x^{2q^{n/2}} - x^2 = 0.\end{aligned}$$

This strange result is due to the fact that $g(x)$ maps into a subfield L of k of degree $n/2$ over $\mathbb{F}_q$, and π is the minimal polynomial of L. To verify that this phenomenon does not preclude the use of the minus modifier, we find a bound on the reduction of Q-rank for $ZFHE^-$.

First, we note that all options for removing two equations are equivalent with respect to Q-rank. Therefore our specification that the dimension of each f_i for $i \in \{0, 1\}$ is reduced by one suffices for Q-rank analysis. In this case, the minus modifier projects f_i onto a hyperplane. There is a basis in which this codimension one projection is given by $\pi(x) = x^q - x$. Since Q-rank is invariant under isomorphism, we may take $\tilde{f}_i$ isomorphic to f_i with respect to this basis.

Relative to this basis we may view the operation of projection on the associated matrices to be raising each element to the power q, shifting one unit down and to the right, and subtracting the original, thusly:

$$\pi \begin{bmatrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{n,1} & \alpha_{n,2} & \cdots & \alpha_{n,n} \end{bmatrix} = \begin{bmatrix} \alpha_{n,n}^q - \alpha_{11} & \alpha_{n,1}^q - \alpha_{12} & \cdots & \alpha_{n,n-1}^q - \alpha_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{n-1,n}^q - \alpha_{n,1} & \alpha_{n-1,1}^q - \alpha_{n,2} & \cdots & \alpha_{n-1,n-1}^q - \alpha_{n,n} \end{bmatrix}.$$

We are assured that this operation does not reduce the rank by more than one and thus the Q-rank of the public key is reduced by at most two. Since we can control the Q-rank via selection of L_{ij}, we conclude that $ZHFE^-$ is secure against the Kipnis-Shamir minrank attack.

6.3 Suggested Parameters

In this section we propose practical parameters for a realistic implementation of $ZHFE^-$. Since the most costly operations, encryption and decryption, utilize algorithms identical to those of $ZHFE$, and due to the tightness between the security analyses of the two schemes, we recommend parameters similar to those of the original scheme.

In an earlier version of this manuscript, we suggested as a parameter set $(q, n, D, r, c) = (7, 55, 105, 2, 6)$, where q is the size of the base field, n is the degree of the extension k over $\mathbb{F}_q$, D is the degree bound for Ψ (in this case $105 = 2{*}7^2 + 7$), r is the number of equations removed, and c is the corank of the parameters L_{ij}, having non-intersecting kernels. In discussions with the authors of [33], it became apparent that we overlooked the added restrictions from insisting on corank 6 matrices L_{ij}. Furthermore, we may have been overcautious about the

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

risk of the Q-rank property of $ZHFE$. Any linear system derived from the Q-rank property is inherently overdefined, and so we dare to be more aggressive. Based in part on their analysis, we propose new parameters for our scheme:

$$108 - ZHFE^{-}: \quad (q, n, D, r, c) = (7, 55, 393, 2, 3).$$

The experiments of the authors of [33] support the viability of these parameters while retaining the significant advance in key generation efficiency even in the minus case.

These parameters correspond to a public key Q-rank of approximately 6, and a degree of regularity of 9 (est.). Given the overdefined nature of the Q-rank attacks and the above analysis verifying resistance to all other known attacks, we conclude that these parameters achieve a security level greater than 80 bits. The performance and security data are essentially the same as the original scheme with L_{ij} of the same moderate corank, 3.

The main differences between $ZHFE^{-}$ and its progenitor with the same parameters is key size and encryption time. Since a plaintext is in $\mathbb{F}_7^{55}$, its length is 165 bits. The ciphertext lies in $\mathbb{F}_7^{2*55-2}$ and is thus 324 bits in length. Thus the public key size is determined by the storage requirements of 108 equations in 55 variables over $\mathbb{F}_7$. This quantity is roughly $63.1K$. In comparison, the public key size of $110 - ZHFE(7, 55, 105, 6)$ is $64.3K$, which is about 2% larger. Finally, since $ZHFE^{-}$ has about 2% fewer public equations than $ZHFE$, encryption is about 2% faster.

7 Conclusion

For many years, multivariate cryptography has had effective tools for building secure and efficient post-quantum signature schemes, but has had much less success for encryption. New schemes such as $ZHFE$ and ABC are promising candidates to fill that gap. Nonetheless, being trapdoor constructions, these schemes can only be trusted after a detailed security analysis.

This work provides much of the security analysis needed to establish trust in the $ZHFE$ construction. In addition to the existing analysis of the difficulty of applying direct algebraic attack to $ZHFE$, we analyze the scheme's security against differential attacks, specify parameters precluding rank attacks, and verify resistance to IP-based equivalent-key attacks. This analysis serves to elucidate the structure of the $ZHFE$ public key, but does not break the cryptosystem, reinforcing the likelihood that the scheme is indeed secure.

The elucidation of the structure of $ZHFE$ also allows us to propose the modified scheme $ZHFE^{-}$. $ZHFE^{-}$ modifies the core map of $ZHFE$ and thereby reduces its key size, while still remaining secure with respect to the attacks analyzed above. While the reduction in key size is relatively small, it opens up the possibility of using Ding's idea of constructing an injective multivariate encryption map whose codomain is much larger than its domain, without requring the dimension of the codomain to exceed that of the domain by a factor of two or more, as do all existing schemes that use this approach.

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

References

1. Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J. Sci. Stat. Comp. **26, 1484** (1997)
2. Chen, A.I.T., Chen, M.S., Chen, T.R., Cheng, C.M., Ding, J., Kuo, E.L.H., Lee, F.Y.S., Yang, B.Y.: Sse implementation of multivariate pkcs on modern x86 cpus. CHES 2009, LNCS, Springer, IACR **5747** (2009) 33–48
3. Chen, A.I.T., Chen, C.H.O., Chen, M.S., Cheng, C.M., Yang, B.Y.: Practical-sized instances of multivariate pkcs: Rainbow, tts, and ℓic-derivatives. Post-Quantum Crypto, LNCS **5299** (2008) 95–106
4. Yang, B.Y., Cheng, C.M., Chen, B.R., Chen, J.M.: Implementing minimized multivariate public-key cryptosystems on low-resource embedded systems. 3rd Security of Pervasive Computing Conference, LNCS **3934** (2006) 73–88
5. Ding, J., Schmidt, D.: Rainbow, a new multivariable polynomial signature scheme. ACNS 2005, LNCS **3531** (2005) 164–175
6. Chen, M.S., Yang, B.Y., Smith-Tone, D.: Pflash - secure asymmetric signatures on smart cards. Lightweight Cryptography Workshop 2015 (2015) http://csrc.nist.gov/groups/ST/lwc-workshop2015/papers/session3-smith-tone-paper.pdf.
7. Kipnis, A., Patarin, J., Goubin, L.: Unbalanced oil and vinegar signature schemes. EUROCRYPT 1999. LNCS **1592** (1999) 206–222
8. Patarin, J., Courtois, N., Goubin, L.: Quartz, 128-bit long digital signatures. In Naccache, D., ed.: CT-RSA. Volume 2020 of Lecture Notes in Computer Science., Springer (2001) 282–297
9. Tao, C., Diene, A., Tang, S., Ding, J.: Simple matrix scheme for encryption. [34] 231–242
10. Ding, J., Petzoldt, A., Wang, L.: The cubic simple matrix encryption scheme. [35] 76–87
11. Porras, J., Baena, J., Ding, J.: Zhfe, a new multivariate public key encryption scheme. [35] 229–245
12. Moody, D., Perlner, R.A., Smith-Tone, D.: An asymptotically optimal structural attack on the ABC multivariate encryption scheme. [35] 180–196
13. Smith-Tone, D.: On the differential security of multivariate public key cryptosystems. In Yang, B.Y., ed.: PQCrypto. Volume 7071 of Lecture Notes in Computer Science., Springer (2011) 130–142
14. Perlner, R.A., Smith-Tone, D.: A classification of differential invariants for multivariate post-quantum cryptosystems. [34] 165–173
15. Daniels, T., Smith-Tone, D.: Differential properties of the HFE cryptosystem. [35] 59–75
16. Patarin, J.: Hidden Fields Equations (HFE) and Isomorphisms of Polynomials (IP): Two New Families of Asymmetric Algorithms. In: EUROCRYPT. (1996) 33–48
17. Patarin, J.: Cryptoanalysis of the Matsumoto and Imai Public Key Scheme of Eurocrypt'88. In Coppersmith, D., ed.: CRYPTO. Volume 963 of Lecture Notes in Computer Science., Springer (1995) 248–261
18. Matsumoto, T., Imai, H.: Public Quadratic Polynominal-Tuples for Efficient Signature-Verification and Message-Encryption. In: EUROCRYPT. (1988) 419–453
19. Berlekamp, E.R.: Factoring polynomials over large finite fields. Mathematics of Computation **24** (1970) pp. 713–735

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

20. Kipnis, A., Shamir, A.: Cryptanalysis of the hfe public key cryptosystem by relinearization. Advances in Cryptology - CRYPTO 1999, Springer **1666** (1999) 788
21. Bettale, L., Faugère, J., Perret, L.: Cryptanalysis of hfe, multi-hfe and variants for odd and even characteristic. Des. Codes Cryptography **69** (2013) 1–52
22. Dubois, V., Gama, N.: The degree of regularity of HFE systems. In Abe, M., ed.: Advances in Cryptology - ASIACRYPT 2010 - 16th International Conference on the Theory and Application of Cryptology and Information Security, Singapore, December 5-9, 2010. Proceedings. Volume 6477 of Lecture Notes in Computer Science., Springer (2010) 557–576
23. Ding, J., Hodges, T.J.: Inverting HFE systems is quasi-polynomial for all fields. In Rogaway, P., ed.: Advances in Cryptology - CRYPTO 2011 - 31st Annual Cryptology Conference, Santa Barbara, CA, USA, August 14-18, 2011. Proceedings. Volume 6841 of Lecture Notes in Computer Science., Springer (2011) 724–742
24. Ding, J., Yang, B.Y.: Degree of regularity for hfev and hfev-. [34] 52–66
25. Dubois, V., Fouque, P.A., Shamir, A., Stern, J.: Practical Cryptanalysis of SFLASH. In Menezes, A., ed.: CRYPTO. Volume 4622 of Lecture Notes in Computer Science., Springer (2007) 1–12
26. Faugere, J.C.: Algebraic cryptanalysis of hidden field equations (hfe) using grobner bases. CRYPTO 2003, LNCS **2729** (2003) 44–60
27. Ding, J., Kleinjung, T.: Degree of regularity for HFE-. IACR Cryptology ePrint Archive **2011** (2011) 570
28. Smith-Tone, D.: Discrete geometric foundations for multivariate public key cryptography. (In Submission)
29. Goubin, L., Courtois, N.: Cryptanalysis of the ttm cryptosystem. In Okamoto, T., ed.: ASIACRYPT. Volume 1976 of Lecture Notes in Computer Science., Springer (2000) 44–57
30. Faugère, J., Gligoroski, D., Perret, L., Samardjiska, S., Thomae, E.: A polynomial-time key-recovery attack on MQQ cryptosystems. In Katz, J., ed.: Public-Key Cryptography - PKC 2015 - 18th IACR International Conference on Practice and Theory in Public-Key Cryptography, Gaithersburg, MD, USA, March 30 - April 1, 2015, Proceedings. Volume 9020 of Lecture Notes in Computer Science., Springer (2015) 150–174
31. Buss, J.F., Frandsen, G.S., Shallit, J.O.: The computational complexity of some problems of linear algebra. Journal of Computer and System Sciences **58** (1999) 572 – 596
32. Wolf, C., Preneel, B.: Equivalent keys in multivariate quadratic public key systems. J. Mathematical Cryptology **4** (2011) 375–415
33. Baena, J., Cabarcas, D., Escudero, D., Porras-Barrera, J., Verbel, J.: Efficient zhfe key generation. In: Post-Quantum Cryptography - 7th International Conference, PQCrypto 2016, Fukuoka, Japan, February 24-26, 2016. Proceedings. (2016)
34. Gaborit, P., ed.: Post-Quantum Cryptography - 5th International Workshop, PQCrypto 2013, Limoges, France, June 4-7, 2013. Proceedings. In Gaborit, P., ed.: PQCrypto. Volume 7932 of Lecture Notes in Computer Science., Springer (2013)
35. Mosca, M., ed.: Post-Quantum Cryptography - 6th International Workshop, PQCrypto 2014, Waterloo, ON, Canada, October 1-3, 2014. Proceedings. Volume 8772 of Lecture Notes in Computer Science., Springer (2014)

Perlner, Ray; Smith-Tone, Daniel.
"Security Analysis and Key Modification for ZHFE."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

On the Differential Security of the $HFEv^-$ Signature Primitive

Ryann Cartor[1], Ryan Gipson[1], Daniel Smith-Tone[1,2], and Jeremy Vates[1]

[1]Department of Mathematics, University of Louisville,
Louisville, Kentucky, USA
[2]National Institute of Standards and Technology,
Gaithersburg, Maryland, USA

ryann.cartor@louisville.edu, ryan.gipson@louisville.edu,
jeremy.vates@louisville.edu, daniel.smith@nist.gov

Abstract. Multivariate Public Key Cryptography (MPKC) is one of the most attractive post-quantum options for digital signatures in a wide array of applications. The history of multivariate signature schemes is tumultuous, however, and solid security arguments are required to inspire faith in the schemes and to verify their security against yet undiscovered attacks. The effectiveness of "differential attacks" on various field-based systems has prompted the investigation of the resistance of schemes against differential adversaries. Due to its prominence in the area and the recent optimization of its parameters, we prove the security of $HFEv^-$ against differential adversaries. We investigate the newly suggested parameters and conclude that the proposed scheme is secure against all known attacks and against any differential adversary.

Key words: Multivariate Cryptography, HFEv-, Discrete Differential, MinRank, Q-rank

1 Introduction and Outline

In the mid 1990s, Peter Shor discovered a way to efficiently implement quantum period finding algorithms on structures of exponential size and showed how the modern world as we know it will change forever once the behemoth engineering challenge of constructing a large scale quantum computing device is overcome. His polynomial time quantum Fourier transforms for smooth integers can be employed to factor integers, to compute discrete logarithms and is powerful enough to efficiently solve hidden subgroup problems for well behaved (usually Abelian) groups. Given the ubiquity of these problems in deployed technologies, our e-society is confronted with the possibility that its public key infrastructure is terminally ill.

It is not known how far this computational cancer may spread, how pervasive exponential quantum speed-ups will prove to be nor how fundamentally wide the gap between feasibility in the classical and quantum world are. Thus we

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

face the task in a rapidly maturing twenty-first century, with ever expanding interconnectivity, of securing open channel communication between unknown future devices, against machines with unknown capabilities, with an unknown date of inception.

Charged with this challenge is a growing international community of experts in quantum-resistant cryptography. The world-wide effort has spawned international standardization efforts including the European Union Horizon 2020 Project, "Post-Quantum Cryptography for Long-Term Security" PQCRYPTO ICT-645622 [1], ETSI's Quantum Safe Cryptography Specification Group [2], and NIST's Post-Quantum Cryptography Workgroup [3]. The dedication of these resources is evidence that the field of post-quantum cryptography is evolving into a state in which we can identify practical technologies with confidence that they will remain secure in a quantum computing world.

One of a few reasonable candidates for post-quantum security is multivariate cryptography. We already rely heavily on the difficulty of inverting nonlinear systems of equations in symmetric cryptography, and we quite reasonably suspect that that security will remain in the quantum paradigm. Multivariate Public Key Cryptography (MPKC) has the added challenge of resisting quantum attack in the asymmetric setting.

While it is difficult to be assured of a cryptosystem's post-quantum security in light of the continual evolution of the relatively young field of quantum algorithms, it is reasonable to start by developing schemes which resist classical attack and for which there is no known significant weakness in the quantum realm. Furthermore, the establishment of security metrics provides insight that educates us about the possibilities for attacks and the correct strategies for the development of cryptosystems.

In this vein, some classification metrics are introduced in [4–6] which can be utilized to rule out certain classes of attacks. While not reduction theoretic proof, reducing the task of breaking the scheme to a known (or often suspected) hard problem, these metrics can be used to prove that certain classes of attacks fail or to illustrate specific computational challenges which an adversary must face to effect an attack.

Many attacks on multivariate public key cryptosystems can be viewed as differential attacks, in that they utilize some symmetric relation or some invariant property of the public polynomials. These attacks have proved effective in application to several cryptosystems. For instance, the attack on SFLASH, see [7], is an attack utilizing differential symmetry, the attack of Kipnis and Shamir [8] on the oil-and-vinegar scheme is actually an attack exploiting a differential invariant, the attack on the ABC matrix encryption scheme of [9] utilizes a subspace differential invariant; even Patarin's initial attack on C^* [10] can be viewed as an exploitation of a trivial differential symmetry, see [5].

As is demonstrated in [4, 6, 11], many general polynomial schemes can have nontrivial linear differential symmetries. Specifically, in [6], systems of linear equations are presented which can have solution spaces large enough to guarantee the existence of nontrivial linear differential symmetries, while in both [4] and

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Ŝignature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

[11] explicit constructions of maps with nontrivial symmetries are provided. The existence of such symmetries in abundance is the basis of attacks removing the minus modifier as in [7], and depending on the structure of the maps inducing the symmetry, may even provide a direct key recovery attack. Furthermore, the attack of [9] on the ABC simple matrix scheme teaches us that differential invariant techniques are a current concern as well. These facts along with the ubiquity of differential attacks in the literature are evidence that the program developed in [4–6] to verify security against differential adversaries is a necessary component of any theory of security for practical and desirable multivariate cryptosystems.

This challenge leads us to an investigation of the $HFEv$ and $HFEv^-$ cryptosystems, see [12], and a characterization of their differential properties. Results similar to those of [4–6] will allow us to make conclusions about the differential security of $HFEv$, and provide a platform for deriving such results for $HFEv^-$.

Specifically, we reduce the task of verifying trivial differential symmetric structure for a polynomial f to the task of verifying that the solution space of a large system of linear equations related to f has a special form. We elucidate the structure of these equations in the case of the central map of $HFEv$ and provide an algorithm for generating keys which provably have trivial differential symmetric structure. In conjunction with our later results on differential invariants, the proof of concept algorithm verifies that information theoretic security against differential adversaries, as defined in [6], is possible with an instantaneous addition to key generation while maintaining sufficient entropy in the key space to avoid "guess-then-IP" attacks. We then extend these methods to the case of $HFEv^-$, deriving the same conclusion.

Expanding on the methods of [6], we prove the following.

Theorem 1 *Let k be a degree n extension of the finite field $\mathbb{F}_q$. Let f be an $HFEv$ central maps. With high probability, f has no nontrivial differential invariant structure.*

With a minimal augmentation of this method we extend this result to the case of $HFEv^-$.

Theorem 2 *Let f be an $HFEv$ central map and let π be a linear projection. With high probability, $\pi \circ f$ has no nontrivial differential invariant structure.*

Thus, with proper parameter selection, $HFEv^-$ is provably secure against differential adversaries. Together with the existant literature on resistance to algebraic and rank attacks, this security argument provides significant theoretical support for the security of aggressive $HFEv^-$ parameters, such as those presented in [13].

The paper is organized as follows. First, we recall big field constructions in multivariate public key cryptography. Next we review the HFE scheme from [14] and the $HFEv^-$ scheme from [12]. In the following section, we provide criteria for the nonexistence of a differential symmetric relation on the private key of both $HFEv$ and $HFEv^-$ and discuss an efficient addition to key generation

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

that allows provably secure keys to be generated automatically. We next review the notion of a differential invariant and a method of classifying differential invariants. We continue, analyzing the differential invariant structure of $HFEv$ and $HFEv^-$, deriving bounds on the probability of differential invariants in the general case. Next, we review the Q-rank and degree of regularity of $HFEv^-$, and discuss resistance to attacks exploiting equivalent keys. Finally, we conclude, discussing the impact of these results on the $HFEv^-$ pedigree.

2 Big Field Signature Schemes

At Eurocrypt '88, Matsumoto and Imai introduced the first massively multivariate cryptosystem which we now call C^*, in [15]. This contribution was based on a fundamentally new idea for developing a trapdoor one-way function. Specifically, they used finite extensions of Galois fields to obtain two representations of the same function: one, a vector-valued function over the base field; the other, an univariate function over the extension field.

One benefit of using this "big field" structure, is that Frobenius operations in extensions of conveniently sized Galois fields can be modeled as permutations of elements in the small field while computations in the small field can be cleverly coded to utilize current architectures optimally. Thus, one can compute a variety of exponential maps and products with great efficiency and obfuscate a simple structure by perturbing the vector representation.

Typically, a big field scheme is built using what is sometimes called the butterfly construction. Given a finite field $\mathbb{F}_q$, a degree n extension $\mathbb{K}$, and an $\mathbb{F}_q$-vector space isomorphism $\phi : \mathbb{F}_q^n \to \mathbb{K}$, one can find an $\mathbb{F}_q$-vector representation of the function $f : \mathbb{K} \to \mathbb{K}$. To hide the choice of basis for the input and output of f, we may compose two affine transformations $T, U : \mathbb{F}_q^n \to \mathbb{F}_q^n$. The resulting composition $P = T \circ \phi^{-q} \circ f \circ \phi \circ U$ is then the public key. The construction is summarized in the figure below:

$$\begin{array}{ccccccc} & & \mathbb{K} & \xrightarrow{\quad f \quad} & \mathbb{K} & & \\ & & \phi \downarrow & & \downarrow \phi^{-1} & & \\ \mathbb{F}_q^n & \xrightarrow{U} & \mathbb{F}_q^n & \xrightarrow{F} & \mathbb{F}_q^n & \xrightarrow{T} & \mathbb{F}_q^n \end{array}$$

2.1 *HFE*

The Hidden Field Equations (HFE) scheme was first presented by Patarin in [14] as a method of avoiding his linearization equations attack which broke the C^* scheme of Matsumoto and Imai, see [10] and [15]. The basic idea of the system is to use the butterfly construction to hide the structure of a low degree polynomial that can be inverted efficiently over $\mathbb{K}$ via the Berlekamp algorithm [16], for example.

More specifically, we select an effectively invertible "quadratic" map $f : \mathbb{K} \to \mathbb{K}$, quadratic in the sense that every monomial of f is a product of a constant

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

and two Frobenius multiples of x. Explicitly any such "core" map f has the form:

$$f(x) = \sum_{\substack{i\leq j \\ q^i+q^j\leq D}} \alpha_{i,j}x^{q^i+q^j} + \sum_{\substack{i \\ q^i\leq D}} \beta_i x^{q^i} + \gamma.$$

The bound D on the degree of the polynomial is required to be quite low for efficient inversion.

One generates a signature by setting $y = h$, a hash digest, and computing, successively, $v = T^{-1}y$, $u = f^{-1}(v)$ and $x = U^{-1}u$. The vector x acts as the signature.

For verification, one simply evaluates the public polynomials, P, at x. If $P(x)$ which is equal to $T \circ f \circ U(x)$ is equal to y, the signature is authenticated. Otherwise, the signature is rejected.

2.2 $HFEv^-$

Taking the HFE construction one step further, we may apply the vinegar modifier, adding extra variables $\tilde{x}_1, \ldots \tilde{x}_v$ to be assigned random values upon inversion. The effect of adding vinegar variables is that new quadratic terms, formed from both products of vinegar variables and HFE variables and products among vinegar variables, increase the rank of the public key. The central map of the $HFEv$ scheme has the form:

$$f(\mathbf{x}) = \sum_{\substack{i\leq j \\ q^i+q^j\leq D}} \alpha_{i,j}x^{q^i+q^j} + \sum_{\substack{i \\ q^i\leq D}} \beta_i(\tilde{x}_1, \ldots, \tilde{x}_v)x^{q^i} + \gamma(\tilde{x}_1, \ldots, \tilde{x}_v),$$

where $\alpha_{i,j} \in \mathbb{K}$, $\beta_i : \mathbb{F}_q^v \to \mathbb{K}$ is linear, and $\gamma : \mathbb{F}_q^v \to \mathbb{K}$ is quadratic.

In contrast to HFE, f is a vector-valued function mapping $\mathbb{F}_q^{n+v}$ to $\mathbb{F}_q^n$. The work of [17, 18, 6] show that representations of such functions over $\mathbb{K}$ are quite valuable. Thus it is beneficial to employ an augmentation of f, adding $n - v$ additional vinegar variables, and say $\hat{y} = \{\tilde{x}_1, \ldots, \tilde{x}_v, \ldots, \tilde{x}_n\}$, where $\tilde{x}_{v+1} = \tilde{x}_{v+2} = \ldots = \tilde{x}_n = 0$. Thus, our core map becomes

$$f(\mathbf{x}) = \hat{f}\begin{pmatrix}\hat{x} \\ \hat{y}\end{pmatrix}.$$

which algebraically identifies f as a bivariate function over $\mathbb{K}$. We may now write f in the following form:

$$f(x,y) = \sum_{\substack{0\leq i\leq j<n \\ q^i+q^j\leq D}} \alpha_{ij}x^{q^i+q^j} + \sum_{\substack{0\leq i,j<n \\ q^i\leq D}} \beta_{ij}x^{q^i}y^{q^j} + \sum_{0\leq i\leq j<n} \gamma_{ij}y^{q^i+q^j}. \quad (1)$$

Here we see an obvious distinction among the types of monomials. We will label the monomials with α coefficients the "HFE monomials," those with β coefficients the "mixing monomials" and the monomials with γ coefficients the "vinegar monomials."

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

The $HFEv^-$ scheme uses the $HFEv$ primitive f above and augments the public key with the minus modifier. The minus modifier removes r of the public equations. This alteration is designed to destroy some of the information of the big field operations latent in the public key.

3 Differential Symmetry

The discrete differential of a field map $f : \mathbb{K} \to \mathbb{K}$ is given by:

$$Df(a, x) = f(a + x) - f(a) - f(x) + f(0).$$

It is simply a normalized difference operator with variable interval. In [7], the SFLASH signature scheme was broken by exploiting a symmetric relation of the differential of the public key. This relation was inherited from the core map of the scheme.

Definition 1 *A* general linear differential symmetry *is a relation of the form*

$$Df(Mx, a) + Df(x, Ma) = \Lambda_M Df(a, x),$$

where $M, \Lambda_M : \mathbb{K} \to \mathbb{K}$ *are* $\mathbb{F}_q$*-linear maps.*

A differential symmetry exists when linear maps may be applied to the discrete differential inputs in such a way that the effect can be factored out of the differential. Furthermore, we say that the symmetry is *linear* when the relation is linear in the unknown coefficients of the linear maps. It can be shown that any such linear symmetric relation implies the existence of a symmetry of the above form, hence the term "general."

While attacks similar to that of [7, 19] exploited some multiplicative relation on central maps of schemes with some algebraic structure over the base field, it was shown in [4] that general linear differential symmetries based on more complex relations exist, in general. Therefore, when analyzing the potential threat of a differential adversary, as defined in [6], it becomes necessary to classify the possible linear differential symmetries. If we succeed in characterizing parameters which provably eliminate nontrivial differential symmetric relations, we prove security against the entire class of differential symmetric attacks, even those utilizing relations not yet discovered.

To this end, we evaluate the security of $HFEv$ against such adversaries. We explicitly consider parameter restrictions which necessarily preclude the existence of any nontrivial differential symmetry.

3.1 Linear Symmetry for HFEv

In our analysis, we will begin by considering the differential of our core map. From the perspective of our adversary, the discrete differential would be

$$D\hat{f}\left(\begin{bmatrix}\hat{a}\\ \hat{b}\end{bmatrix}, \begin{bmatrix}\hat{x}\\ \hat{y}\end{bmatrix}\right) = Df(a, b, x, y).$$

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

By the bilinearity of $D\hat{f}$ we see that Df is multi-affine; Df is affine in each of its inputs when the remaining inputs are fixed. Evaluating this differential we obtain

$$\begin{aligned} Df(a,b,x,y) = &\sum_{\substack{0\le i\le j<n \\ q^i+q^j\le D}} \alpha_{i,j}(x^{q^i}a^{q^j} + x^{q^j}a^{q^i}) \\ &+ \sum_{\substack{0\le i,j<n \\ q^i\le D}} \beta_{i,j}(x^{q^i}b^{q^j} + a^{q^i}y^{q^j}) \\ &+ \sum_{0\le i\le j<n} \gamma_{i,j}(y^{q^i}b^{q^j} + y^{q^j}b^{q^i}), \end{aligned} \tag{2}$$

noting that Df is a $\mathbb{K}$-bilinear form in $[a\ b]^T$ and $[x\ y]^T$. For ease of computation, we will choose the following representation for $\mathbb{K}$:

$$x \mapsto [x \ \ x^q \ \ x^{q^2} \ \ \ldots \ \ x^{q^{n-1}}]^T.$$

Similarly, we may map our oil-vinegar vector as

$$[x \ \ y] \mapsto [x \ \ x^q \ \ x^{q^2} \ \ \ldots \ \ x^{q^{n-1}} \ \ y \ \ y^q \ \ y^{q^2} \ \ \ldots \ \ y^{q^{n-1}}]^T,$$

and Df is thus represented by the $2n \times 2n$ matrix where the (i,j)th and (j,i)th entries in the upper left $n \times n$ block are the coefficients $\alpha_{i,j}$, and the (i,j)th entries in the upper right block and the (j,i)th entries in the lower left block are the coefficients $\beta_{i,j}$, while the (i,j)th and the (j,i)th entries in the lower right block are the coefficients $\gamma_{i,j}$.
Note, that any $\mathbb{F}_q$-linear map $M : \mathbb{K} \to \mathbb{K}$ can be represented by $Mx = \sum_{i=0}^{n-1} m_i x$. Thus, as demonstrated in [6], under our representation,

$$M = \begin{pmatrix} m_0 & m_1 & \cdots & m_{n-1} \\ m_{n-1}^q & m_0^q & \cdots & m_{n-2}^q \\ \vdots & \vdots & \ddots & \vdots \\ m_1^{q^{n-1}} & m_2^{q^{n-1}} & \cdots & m_0^{q^{n-1}} \end{pmatrix}.$$

However, when viewing an $\mathbb{F}_q$-linear map over our vector $\begin{bmatrix} \hat{x} \\ \hat{y} \end{bmatrix}$, we may consider the $2n \times 2n$ matrix

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Ŝignature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

$$\overline{M} = \begin{pmatrix} m_{00,0} & m_{00,1} & \cdots & m_{00,n-1} & m_{01,0} & m_{01,1} & \cdots & m_{01,n-1} \\ m_{00,n-1}^{q} & m_{00,0}^{q} & \cdots & m_{00,n-2}^{q} & m_{01,n-1}^{q} & m_{01,0}^{q} & \cdots & m_{01,n-2}^{q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{00,1}^{q^{n-1}} & m_{00,2}^{q^{n-1}} & \cdots & m_{00,0}^{q^{n-1}} & m_{01,1}^{q^{n-1}} & m_{01,2}^{q^{n-1}} & \cdots & m_{01,0}^{q^{n-1}} \\ m_{10,0} & m_{10,1} & \cdots & m_{10,n-1} & m_{11,0} & m_{11,1} & \cdots & m_{11,n-1} \\ m_{10,n-1}^{q} & m_{10,0}^{q} & \cdots & m_{10,n-2}^{q} & m_{11,n-1}^{q} & m_{11,0}^{q} & \cdots & m_{11,n-2}^{q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{10,1}^{q^{n-1}} & m_{10,2}^{q^{n-1}} & \cdots & m_{10,0}^{q^{n-1}} & m_{11,1}^{q^{n-1}} & m_{11,2}^{q^{n-1}} & \cdots & m_{11,0}^{q^{n-1}} \end{pmatrix}.$$

For computational reference, we will label each row and column *modulo*(n), i.e., each coordinate of the entry (i, j), will be represented by a residue class modulo n.

If we assume that f is vulnerable to a differential attack, then there exists a non-trivial linear mapping $\overline{M}$ such that the differential symmetry in (1) is satisfied. To compute such a symmetry inducing map requires the solution of $4n^2$ highly dependent but random equations in the $8n$ unknown coefficients of $\overline{M}$ and $\overline{A_M}$ over $\mathbb{K}$. Since trivial symmetries (such as multiplication by scalars) are exhibited by every map, we know that there exist nontrivial solutions. Even assuming unit time for $\mathbb{K}$-arithmetic operations, for realistic parameters this process is very inefficient; with the more realistic assumption of costly $\mathbb{K}$-arithmetic operations, this task is unsatisfactory in key generation.

To make the solution of such systems of equations more efficient, we derive the structure of the equations and develop a two step process for verifying trivial differential symmetric structure. The first step involves finding equations which only involve a subset of the variables. The existence of such equations is guaranteed by the degree bound of the HFE monomials. This information is then bootstrapped to eliminate many unknown coefficients of $\overline{M}$ resulting in a very small system of equations which can be solved explicitly.

We remark here that this methodology also suggests a method for estimating the probability of the existence of a differential symmetry for the $HFEv$ primitive. The existence of a nontrivial symmetry corresponds to systems for which the rank of the system of equations is less than $8n$. Under the heuristic that under row reduction these systems of equations behave like random $8n \times 8n$ matrices, we obtain a probability of roughly $1 - q^{-1}$ that the scheme has no nontrivial differential symmetry. We note that this heuristic is almost certainly false since trivial symmetries do exist. This quantity does represent a lower bound, however, and thus may offer support for larger base fields.

We begin by considering the entries of the matrix $\overline{M}^T Df + Df\overline{M}$. The contribution of any monomial $\alpha_{i,j}x^{q^i+q^j}$ to the ith row of $Df\overline{M}$ is given by

$$\left(\alpha_{i,j}m_{00,-j}^{j}\ \alpha_{i,j}m_{00,1-j}^{j}\ \cdots\ \alpha_{i,j}m_{00,-1-j}^{j}\ \alpha_{i,j}m_{01,-j}^{j}\ \alpha_{i,j}m_{01,1-j}^{j}\ \cdots\ \alpha_{i,j}m_{01,-1-j}^{j}\right)$$

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Ŝignature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

while the contribution to the jth row is

$$\left(\alpha_{i,j}m^i_{00,-i}\ \alpha_{i,j}m^i_{00,1-i}\cdots\ \alpha_{i,j}m^i_{00,-1-i}\ \alpha_{i,j}m^i_{01,-i}\ \alpha_{i,j}m^i_{01,1-i}\cdots\ \alpha_{i,j}m^i_{01,-1-i}\right).$$

By symmetry, the ith and and jth columns of $\overline{M}^T Df$ are the same as their respective rows.

It is clear that the rows and columns associated with coefficients of vinegar monomials as well as terms associated with mixing monomials may be represented similarly. However, it should be noted that those terms associated with mixing monomials will be multiplied by linear coefficients $m_{00,\cdot}$, $m_{01,\cdot}$, $m_{10,\cdot}$, and $m_{11,\cdot}$, while coefficients associated with vinegar variables are multiplied only by linear coefficients $m_{10,\cdot}$ and $m_{11,\cdot}$.

The above patterns can be extended to characterize the contribution to the ith row and jth row of monomials of the form $\beta_{i,j}x^{q^i}y^{q^j}$ and $\gamma_{i,j}y^{q^i+q^j}$, as well. We note, however, that γ coefficients interact with entries from the lower block matrices while β coefficients interact with coefficients from all block matrices.

Now that we have characterized the left side of (1), we will consider the entries of $\Lambda_{\overline{M}}Df$. For every monomial of f, say $\alpha_{i',j'}x^{q^i+q^j}$, $\beta_{r,s}x^{q^r}y^{q^s}$, or $\gamma_{u,v}y^{q^s+q^v}$, we have under the mapping of $\Lambda_{\overline{M}}$ terms of the form: $l_\ell\alpha^{q^\ell}_{i,j}x^{q^{i+\ell}+q^{j+\ell}}$, $l_\ell\beta^{q^{r+\ell}}_{r,s}x^{q^{s+\ell}}y^{q^j}$, and $l_\ell\gamma^{q^\ell}_{u,v}y^{q^{u+\ell}+q^{v+\ell}}$. Clearly, this results in every nonzero entry, say (r,s), of our Df matrix being raised to the power of q^ℓ and shifted along a forty-five degree angle to entry $(r+\ell,s+\ell)$. Thus, for each monomial in f there are two possible nonzero entries in the ith row, with possible overlap.

This discrete geometrical interpretation of the action of M and D on the coefficients of f is central to this analysis. A graphical representation of these relations is provided in Figure 1.

As in [6], the possibility of a differential symmetry can be determined by setting the matrix representation of $M^TDf+DfM$ equal to the matrix $\Lambda_M Df$. We will demonstrate an algorithm, given some specific constraints, that will help provide secure keys to be generated automatically.

Due to the structure of our M matrix, we need to work within each $m_{i,j}$ matrix independently. The following algorithm for $m_{0,0}$ extends very naturally to the other 3 matrices. For clarity, all m terms in description below are $m_{0,0}$ terms.

Let $\alpha_{i,j},\beta_{r,s},\gamma_{u,v}$ represent the coefficients of our monomials in our core map. Consider the ith row of $M^TDf+DfM$. For all w not occurring as a power of q of our HFE or mixing monomials in f, or difference of powers of q in an exponent of a monomial in f plus i, the (i,w) entry is $\alpha_{i,j}m^{q^j}_{w-j}=0$ (resp. $\beta_{i,j}m^{q^j}_{w-j}$). Consider the rth row. For all w not occuring as an exponent of q in a vinegar monomial or as a difference of powers of q in an exponent of a monomial in f plus s, the (r,w)th entry is $\beta_{r,s}m^{q^s}_{k-s}=0$. Hence, we can use those relations to look for non-zero entries of $m_{0,0}$.

After putting those relations into Algorithm 1, see Figure 3a, you can generate a set for every i and r, exponents that occur in your core map. Each set provides a list of indices of all possible non-zero m's. For each index not occuring

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

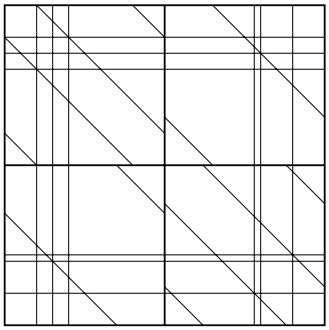


Fig. 1: Graphical representation of the equation $M^T Df + DfM = \Lambda_M Df$ for the $HFEv$ (actually, vC^*) polynomial $f(x) = \alpha_{i,j} x^{q^i+q^j} + \beta_{r,s} x^{q^r} y^{q^s} + \gamma_{u,v} y^{q^u+q^v}$. Horizontal and vertical lines represent nonzero entries in $M^T Df + DfM$ while diagonal lines represent nonzero entries in $\Lambda_M Df$. We may consider this diagram as a genus 4 surface containing straight lines.

in any such set, the corresponding coefficient m must equal zero due to the fact that there must be a coordinate in the equation $M^T Df + DfM = \Lambda_M Df$ setting a constant multiple of m to zero. Thus, the intersection off all sets generated produces a list of all possible non-zero entries for the sub-matrix $m_{0,0}$.

Once this list is obtained, the variables shown to have value zero are eliminated from the system of equations. After repeating a similar algorithm for each of the remaining three submatrices a significantly diminished system of equations is produced which is then solved explicitly.

After running this algorithm with realistic values satisfying the above constraints and matching the parameter sizes of [13] along with using mild restrictions on the powers of the mixing and vinegar monomials, the only non-zero value obtained is m_0.

We note that it is possible that these restrictions, especially the restriction for these experiments on the number of monomials, place a lower bound on the number of vinegar variables required to achieve such a structure. On the other hand, with numerous small-scale experiments without parameter restrictions and using the full number of monomials we found that structurally the only nonzero value for the matrix $m_{0,0}$ is the m_0 term.

Since we have only a single non-zero term, our $m_{0,0}$ matrix is a diagonal matrix. A similar analysis for each of the remaining submatrices reveals the same structure. Thus we find that the only possible structure for $\overline{M}$ under these constraints satisfying a differential symmetry for $HFEv$ is

$$\overline{M} = \left[\begin{array}{c|c} cI & dI \\ \hline dI & cI \end{array}\right].$$

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Ŝignature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

Furthermore, we can prove by way of Theorem 2 from [20], that the coefficients $c, d \in \mathbb{F}_q$.

We note that this map induces a trivial differential symmetry. To see this, note that the (nonpartial) differential of any bivariate function is bilinear in its vector inputs. Thus

$$\begin{aligned} Dg(\overline{M}[a\ b]^T, [x\ y]^T) &= Dg([ca+db\ da+cb]^T, [x\ y]^T) \\ &= Dg([ca+db\ cb+da]^T, [x\ y]^T) \\ &= Dg(c[a\ b]^T, [x\ y]^T) + Dg(d[b\ a]^T, [x\ y]^T) \\ &= cDg(a,b,x,y) + dDg(b,a,x,y) \\ &= (c+d)Dg(a,b,x,y). \end{aligned} \tag{3}$$

Consequently, for the parameters provided by Algorithm 1, $HFEv$ provably has no nontrivial differential symmetric structure.

It should be noted that the restrictions provided on the powers of q of the monomials of our f does lower the entropy of our key space and likely raise the number of required vinegar variables to a level which is either unsafe or undesirable. However, there is still plenty of entropy with these restrictions and we obtain provable security against the differential symmetric attack. The restrictions provided are just a base line for this technique and our experiments with small scale examples indicate that even when we insist that every possible monomial satisfying the HFE degree bound is required to have a nonzero coefficient, the generalized algorithm still outputs only the trivial solution. Thus we can achieve provable security with minimal loss of entropy.

3.2 $HFEv^-$

Now, the algorithm extends naturally to $HFEv^-$. Every non-zero entry from the system generated by $HFEv$ is also in that generated by $HFEv^-$, but with a few more, see Figure 2. We choose a basis in which an example minus projection is a polynomial of degree q^2. For every ith row, we also have for any w not a power of $\alpha+n$ or $\beta+n$ where $n<2$, the (i,w)th entry is $\alpha_{i,j}m_{w-j}^{q^j}=0$. For the sth row, for all w not being a power of $\beta+n$ or $r+n$ where $n<2$, the (s,w)th entry is $\beta_{r,s}m_{w-r}^{q^r}=0$. A visualization is provided in Figure 2.

Again, we can use these relations, along with the relations described in the $HFEv$ system, to create a list of sets of all non-zero areas on $m_{0,0}$ using Algorithm 2, see Figure **??**. Each of these sets contains indices which are possibly non-zero, thus entries not in that set are definitively equal to zero.

By taking the intersection of all the sets, you can find the final locations of non-zero entries for our sub matrix $m_{0,0}$. In doing so, with realistic values from [13], the only non-zero value obtained is m_0. This again gives us security against symmetrical attacks by having M being a block matrix consisting of diagonal matrices with an argument similar to [6].

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

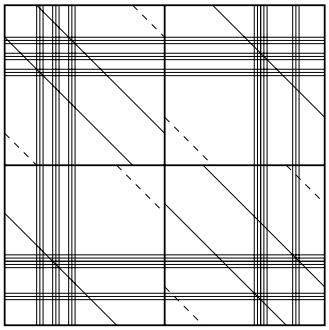


Fig. 2: Graphical representation of the equation $M^T Df + DfM = \Lambda_M Df$ for the $HFEv^-$ with the minus modifier given by the projection $\pi(x) = x^{q^2} + \rho x^q + \tau x$. Horizontal and vertical lines represent nonzero entries in $M^T Df + DfM$ while diagonal lines represent nonzero entries in $\Lambda_M Df$. We note that each triple of lines corresponds to a single monomial in the central map.

HFEvKeyCheck
Input: An $HFEv$ central map f, a flag flg
Output: Set of indices of coefficients m_i of submatrix m_{00} which are possibly nonzero in a linear map inducing differential symmetry for f.
01. **for** monomial $\alpha_{i,j} x^{q^i+q^j}$ in f
02. $S_i = \{\}$;
03 $S_j = \{\}$;
04. **for** monomial with powers r and s in f
05. $S_i = S_i \cup \{r-j, s-j, i-j+r-s, i-j+s-r\}$;
06. $S_j = S_j \cup \{r-i, s-i, j-i+r-s, j-i+s-r\}$;
07. **end for;**
08. **end for;**
09. **if** flg
10. **then**
11. **return** all S_i;
12. **else**
13. **return** $\bigcap S_i$;
14. **end if;**

(a) Algorithm 1: $HFEv$

HFEv-KeyCheck
Input: An $HFEv^-$ central map $\pi(f)$, the corank of π, r
Output: Set of indices of coefficients m_i of submatrix m_{00} which are possibly nonzero in a linear map inducing differential symmetry for $\pi(f)$.
01. **Call:** HFEvKeyCheck(f,1);
02. **for** all S_i
03 $T_i = \{\}$;
04. **for** j from 0 to $r-1$
05. $T_i = T_i \cup (j + S_i)$;
06. **end for;**
07. **end for;**
08. **return** $\bigcap T_i$;

(b) Algorithm 2: $HFEv^-$

Fig. 3: Algorithms 1 and 2

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

4 Differential Invariants

Definition 2 *Let $f : \mathbb{F}_q^n \to \mathbb{F}_q^m$ be a function. A* differential invariant *of f is a subspace $V \subseteq \mathbb{K}$ with the property that there is a subspace $W \subseteq \mathbb{K}$ such that $dim(W) \leq dim(V)$ and $\forall A \in Span_{\mathbb{F}_q}(Df_i)$, $AV \subseteq W$.*

Informally speaking, a function has a differential invariant if the image of a subspace under all differential coordinate forms lies in a fixed subspace of dimension no larger. This definition captures the notion of *simultaneous invariants*, subspaces which are simultaneously invariant subspaces of Df_i for all i, and detects when large subspaces are acted upon linearly.

If we assume the existence of a differential invariant V, we can define a corresponding subspace $V^\perp$ as the set of all elements $x \in \mathbb{K}$ such that the dot product $\langle x, Av \rangle = 0 \ \forall v \in V, \forall A \in Span(Df_i)$. We note that this is not the standard definition of an orthogonal complement. $V^\perp$ is not the set of everything orthogonal to V, but rather everything orthogonal to AV, which may or may not be in V. By definition, it is clear that V and $V^\perp$ satisfy the relation

$$dim(V) + dim(V^\perp) \geq n.$$

Assume there is a differential invariant $V \subseteq \mathbb{F}_q^n$, and choose linear maps $M : \mathbb{F}_q^n \to V$ and $M^\perp : \mathbb{F}_q^n \to V^\perp$. For any differential-coordinate-form, we have

$$[Df(M^\perp y, Mx)]_i = (M^\perp y)^T \left(Df_i(Mx)\right) \tag{4}$$

Since $M^\perp y$ is in $V^\perp$, and $Df_i Mx \in AV$, we must then have that

$$[Df(M^\perp y, Mx)]_i = (M^\perp a)^T \left(Df_i(Mx)\right) = 0 \tag{5}$$

Thus, as derived in [5],

$$\forall y, x \in \mathbb{F}_q^n, Df(M^\perp y, Mx) = 0 \quad \text{or equivalently,} \quad Df(M^\perp \mathbb{F}_q^n, M\mathbb{F}_q^n) = 0 \tag{6}$$

This relation restricts the structure of M and $M^\perp$, and provides a direct means of classifying the differential invariant structure of f.

We follow an analogous strategy to that of [6], adapted to the structure of the central $HFEv^-$ map f. First, we recall a result of [6].

Proposition 1. *([6]) If A, B are two $m \times n$ matrices, then $rank(A) = rank(B)$ if and only if there exist nonsingular matrices C, D, such that $A = CBD$.*

Without loss of generality we assume that $rank(M^\perp) \leq rank(M)$. If the ranks are equal, then we may apply the proposition and write $M^\perp = SMT$, with S and T nonsingular. If $rank(M^\perp) < rank(M)$, compose M with a singular matrix X so that $rank(XM) = rank(M^\perp)$, and then apply the above result so that $M^\perp = S(XM)T$. Then we can express $M^\perp = S'MT$, where S' is singular. Restating our differential result (6) in this manner, we have that if $M^\perp = SMT$, and $M : \mathbb{F}_q^{n+v} \to V$, then

$$\forall x, y \in \mathbb{F}_q^n, Df(SMTy, MTx) = 0. \tag{7}$$

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

4.1 Minimal Generators over Intermediate Subfield

For lack of a good reference, we prove the following statement about the structure of the coordinate ring of a subspace of an extension field over an intermediate extension.

Lemma 1 *Let $\mathbb{L}/\mathbb{K}/\mathbb{F}_q$ be a tower of finite extensions with $|\mathbb{L} : \mathbb{K}| = m$ and $|\mathbb{K} : \mathbb{F}_q| = n$. Let V be an $\mathbb{F}_q$-subspace of $\mathbb{L}$. Then $I(V)$ has m multivariate generators over $\mathbb{K}$ of the form*

$$\mathcal{M}_V^{(k)}(x_0, \ldots, x_{m-1}) = \sum_{\substack{0 \leq i < n \\ 0 \leq j < m}} a_{ijk} x_j^{q^i}.$$

Proof. Choose a basis $\{\overline{e_0} = \overline{1}, \overline{e_1}, \ldots, \overline{e_{mn-1}}\}$ for $\mathbb{L}$ over $\mathbb{K}$. Since V is an $\mathbb{F}_q$-subspace of $\mathbb{L}$, the minimal polynomial of V over $\mathbb{L}$, $\mathcal{M}_V(\overline{X}) = \sum_{i=0}^{mn-1} \overline{\alpha_i} \overline{X}^{q^i}$, is $\mathbb{F}_q$-linear. Note that the operations of addition and left multiplication by elements in $\mathbb{L}$ are $\mathbb{K}$-linear, whereas the Frobenius maps are merely $\mathbb{F}$-linear.

Now, since $\mathcal{M}_V(\overline{X})$ is linear it is additive, hence

$$\mathcal{M}_V(\overline{X}) = \mathcal{M}_V\left(\begin{bmatrix} x_0 \\ \vdots \\ x_{m-1} \end{bmatrix}\right) = \sum_{i=0}^{m-1} \mathcal{M}_V(x_i \overline{e_i}).$$

In each summand of $\mathcal{M}_V(x_j \overline{e_j})$, we have

$$(x_j \overline{e_j})^{q^i} = x_j^{q^i} \overline{e_j}^{q^i} = x_j^{q^i} \sum_{i=0}^{m-1} r_i \overline{e_i}$$

for some $r_0, \ldots, r_{m-1} \in \mathbb{K}$. As a vector over $\mathbb{K}$ this quantity is

$$\begin{bmatrix} r_0 x_j^{q^i} \\ \vdots \\ r_{m-1} x_j^{q^i} \end{bmatrix}.$$

Thus $\mathcal{M}_V(x_j \overline{e_j})$ is an m-dimensional vector of $\mathbb{K}$-linear combinations of $x_j, x_j^q, \ldots, x_j^{q^{n-1}}$. Thus $\mathcal{M}_V(\overline{X})$ is of the form

$$\mathcal{M}_V(\overline{X}) = \begin{bmatrix} \mathcal{M}_V^{(0)}(x_0, \ldots, x_{m-1}) \\ \vdots \\ \mathcal{M}_V^{(m-1)}(_0, \ldots, x_{m-1}) \end{bmatrix} = \begin{bmatrix} \sum_{\substack{0 \leq i < n \\ 0 \leq j < m}} a_{ij0} x_j^{q^i} \\ \vdots \\ \sum_{\substack{0 \leq i < n \\ 0 \leq j < m}} a_{ij(m-1)} x_j^{q^i} \end{bmatrix},$$

as required.

We note that the minimal polynomials studied in [6] correspond to the special case of the above lemma in which $m = 1$. Given our characterization from Section 2.2 of the central map of $HFEv^-$ as a bivariate polynomial over $\mathbb{K}$, we are primarily interested in the $m = 2$ case of Lemma 1.

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

4.2 Invariant Analysis of $HFEv$

As in [6], we consider $Df(SMTa, MTx)$, where T is nonsingular, S is a possibly singular map which sends V into $V^\perp$ and $M : k \to k$ is a projection onto V. Without loss of generality we'll assume that M projects onto V. Then MT is another projection onto V. SMT is a projection onto $V^\perp$. An important distinction is that for this case, the a and x above are actually two dimensional vectors over k. Thus $dim(V) + dim(V^\perp) \geq n$.

Proof (of Theorem 1). Let us denote by $[\hat{x}\ \hat{y}]^T$ the quantity $MT[x\ y]^T$.

Suppose we have

$$f(x,y) = \sum_{\substack{0\leq i\leq j<n \\ q^i+q^j\leq D}} \alpha_{ij} x^{q^i+q^j} + \sum_{\substack{0\leq i,j<n \\ q^i\leq D}} \beta_{ij} x^{q^i} y^{q^j} + \sum_{0\leq i\leq j<n} \gamma_{ij} y^{q^i+q^j}.$$

Applying the differential (w.r.t. the vector $[x\ y]^T$) as described in Section 3.1, we obtain:

$$\begin{aligned} Df(a,b,x,y) = &\sum_{\substack{0\leq i\leq j<n \\ q^i+q^j\leq D}} \alpha_{ij} \left(a^{q^i} x^{q^j} + a^{q^j} x^{q^i} \right) \\ &+ \sum_{\substack{0\leq i,j<n \\ q^i\leq D}} \beta_{ij} \left(a^{q^i} y^{q^j} + x^{q^i} b^{q^j} \right) \\ &+ \sum_{0\leq i\leq j<n} \gamma_{ij} \left(b^{q^i} y^{q^j} + b^{q^j} y^{q^i} \right). \end{aligned} \tag{8}$$

Substituting $SMT[a\ b]^T$ and $MT[x\ y]^T$, we derive

$$Df(S[\hat{a}\ \hat{b}]^T, \hat{x}, \hat{y}) = Df(S_{11}\hat{a} + S_{12}\hat{b}, S_{21}\hat{a} + S_{22}\hat{b}, \hat{x}, \hat{y}).$$

For notational convenience let $\hat{a} = S_{11}\hat{a} + S_{12}\hat{b}$ and $\hat{\hat{b}} = S_{21}\hat{a} + S_{22}\hat{b}$. Plugging in these values in the previous equation we get

$$\begin{aligned} Df(\hat{\hat{a}}, \hat{\hat{b}}, \hat{x}, \hat{y}) = &\sum_{\substack{0\leq i\leq j<n \\ q^i+q^j\leq D}} \alpha_{ij} \left((\hat{\hat{a}})^{q^i} \hat{x}^{q^j} + (\hat{\hat{a}})^{q^j} \hat{x}^{q^i} \right) \\ &+ \sum_{\substack{0\leq i,j<n \\ q^i\leq D}} \beta_{ij} \left((\hat{\hat{a}})^{q^i} \hat{y}^{q^j} + \hat{x}^{q^i} (\hat{\hat{b}})^{q^j} \right) \\ &+ \sum_{0\leq i\leq j<n} \gamma_{ij} \left((\hat{\hat{b}})^{q^i} \hat{y}^{q^j} + (\hat{\hat{b}})^{q^j} \hat{y}^{q^i} \right). \end{aligned} \tag{9}$$

In contrast to the situation with HFE, these monomials are not necessarily independent. By Lemma 1, the generators of $I(V)$ have the form

$$\sum_{0\leq i<n} r_{ij} x^{q^i} + \sum_{0\leq i<n} s_{ij} y^{q^i} \text{ for } j \in \{1,2\},$$

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Ŝignature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

where $r_{ij}, s_{ij} \in \mathbb{K}$. Clearly, these expressions evaluate to zero on $(\hat{x}, \hat{y})$. Evaluating (9) modulo $I(V)$ (only on the variables $\hat{x}$ and $\hat{y}$), we obtain:

$$\begin{aligned} Df(\hat{\mathring{a}}, \hat{\mathring{b}}, \hat{x}, \hat{y}) = &\sum_{\substack{0 \le i < n \\ 0 \le j < d_x}} \left[\alpha'_{ij} (\hat{\mathring{a}})^{q^i} + \beta'_{ij} (\hat{\mathring{b}})^{q^i} \right] \hat{x}^{q^j} \\ &+ \sum_{\substack{0 \le i < n \\ 0 \le j < d_y}} \left[\gamma'_{ij} (\hat{\mathring{a}})^{q^i} + \delta'_{ij} (\hat{\mathring{b}})^{q^i} \right] \hat{y}^{q^j}, \end{aligned} \tag{10}$$

where d_x and d_y are the largest powers of $\hat{x}$ (resp. $\hat{y}$) occuring. After the reduction modulo $I(V)$, the remaining monomials $\hat{x}, \ldots, \hat{x}^{q^{d_x}}$ and $\hat{y}, \ldots, \hat{y}^{q^{d_y}}$ are independent. Thus, for $Df(\hat{\mathring{a}}, \hat{\mathring{b}}, \hat{x}, \hat{y}) = 0$, each polynomial expression multiplied by a single $\hat{x}^{q^j}$ or $\hat{y}^{q^j}$ must be identically zero, that is to say that for all $0 \le j \le d_x$

$$\sum_{0 \le i < n} \left[\alpha'_{ij} (\hat{\mathring{a}})^{q^i} + \beta'_{ij} (\hat{\mathring{b}})^{q^i} \right] = 0 \tag{11}$$

and for all $0 \le j \le d_y$

$$\sum_{0 \le i < n} \left[\gamma'_{ij} (\hat{\mathring{a}})^{q^i} + \delta'_{ij} (\hat{\mathring{b}})^{q^i} \right] = 0. \tag{12}$$

The left hand sides of (11) and (12) are $\mathbb{F}$-linear functions in $S[\hat{a}\ \hat{b}]^T$. Thus we can express each such equality over $\mathbb{F}$ as

$$LS \left[\hat{a}_0 \ \cdots \ \hat{a}_{n-1}\ \hat{b}_0 \ \cdots \ \hat{b}_{n-1} \right]^T = 0,$$

where L is an $n \times 2n$ matrix with entries in $\mathbb{F}$. We note specifically that the coefficients of L depend on V and the choices of coefficients in the central map f. For randomly chosen coefficients retaining the $HFEv$ structure, we expect an L derived from an equation of the form (11) or (12) to have high rank with very high probability, more than $1 - q^{-n}$. Thus the dimension of the intersections of the nullspaces of each L is zero with probability at least $1 - 2q^{-n}$.

Clearly, the condition for these equations to be satisfied is that S sends V to the intersection of the nullspaces of each such L. Thus S is with high probability the zero map on V and so $V^\perp = \{0\}$. This generates a contradiction, however, since $2n \le dim(V) + \dim(V^\perp) < 2n$. Thus, with probability greater than $1 - 2q^{-n}$, f has no nontrivial differential invariant structure.

4.3 $HFEv^-$

The situation for $HFEv^-$ is quite similar, but the probabilities are slightly different. Specifically one must note that since the condition of being a differential invariant is a condition on the span of the public differential forms, under projection this condition is weaker and easier to satisfy. For specificity, we consider

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

the removal of a single public equation, though, critically, a very similar though notationally messy analysis is easy to derive in the general case.

We may model the removal of a single equation as a projection of the form $\pi(x) = x^q + x$ applied after the central map.

Proof (of Theorem 2). Consider

$$\begin{aligned}\pi(f(x,y)) = &\sum_{\substack{0\le i\le j<n\\ q^i+q^j\le D}} \alpha_{ij}x^{q^i+q^j} + \sum_{\substack{0\le i,j<n\\ q^i\le D}} \beta_{ij}x^{q^i}y^{q^j} + \sum_{0\le i\le j<n} \gamma_{ij}y^{q^i+q^j}\\ &+ \sum_{\substack{0\le i\le j<n\\ q^i+q^j\le D}} \alpha^q_{ij}x^{q^{i+1}+q^{j+1}} + \sum_{\substack{0\le i,j<n\\ q^i\le D}} \beta^q_{ij}x^{q^{i+1}}y^{q^{j+1}} + \sum_{0\le i\le j<n} \gamma^q_{ij}y^{q^{i+1}+q^{j+1}}.\end{aligned} \tag{13}$$

Taking the differential, we obtain

$$\begin{aligned}D(\pi\circ f)(\hat{a},\hat{b},\hat{x},\hat{y}) = &\sum_{\substack{0\le i\le j<n\\ q^i+q^j\le D}} \alpha_{ij}\left((\hat{a})^{q^i}\hat{x}^{q^j} + (\hat{a})^{q^j}\hat{x}^{q^i}\right)\\ &+ \sum_{\substack{0\le i,j<n\\ q^i\le D}} \beta_{ij}\left((\hat{a})^{q^i}\hat{y}^{q^j} + \hat{x}^{q^i}(\hat{b})^{q^j}\right)\\ &+ \sum_{0\le i\le j<n} \gamma_{ij}\left((\hat{b})^{q^i}\hat{y}^{q^j} + (\hat{b})^{q^j}\hat{y}^{q^i}\right)\\ &+ \sum_{\substack{0\le i\le j<n\\ q^i+q^j\le D}} \alpha^q_{ij}\left((\hat{a})^{q^{i+1}}\hat{x}^{q^{j+1}} + (\hat{a})^{q^{j+1}}\hat{x}^{q^{i+1}}\right)\\ &+ \sum_{\substack{0\le i,j<n\\ q^i\le D}} \beta^q_{ij}\left((\hat{a})^{q^{i+1}}\hat{y}^{q^{j+1}} + \hat{x}^{q^{i+1}}(\hat{b})^{q^{j+1}}\right)\\ &+ \sum_{0\le i\le j<n} \gamma^q_{ij}\left((\hat{b})^{q^{i+1}}\hat{y}^{q^{j+1}} + (\hat{b})^{q^{j+1}}\hat{y}^{q^{i+1}}\right).\end{aligned} \tag{14}$$

Again, we may evaluate modulo $I(V)$ and collect the terms for the distinct powers of $\hat{x}$ and $\hat{y}$. By the independence of these monomials we obtain the relations

$$\begin{aligned}\sum_{0\le i<n}\left[\alpha''_{ij}(\hat{a})^{q^i} + \beta'_{ij}(\hat{\hat{b}})^{q^i}\right] = 0\\ \sum_{0\le i<n}\left[\gamma''_{ij}(\hat{a})^{q^i} + \delta'_{ij}(\hat{\hat{b}})^{q^i}\right] = 0.\end{aligned} \tag{15}$$

At this point, the analysis proceeds exactly as in the case of $HFEv$. We once again arrive at the conclusion that with high probability S is the zero map on V, contradicting the existence of a differential invariant. We note here that this analysis works for any projection, though the exact values of the α''_{ij} and γ''_{ij} depend on the specific projection and the structure of f.

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

5 Degree of Regularity, Q-rank and Parameters

Further considerations for the security of $HFEv^-$ are the degree of regularity, a quantity closely connected to the complexity of algebraic attacks, and the Q-rank of the public key. A careful analysis of each of these quantities reveals that they support the security of $HFEv^-$ against an algebraic attack such as [21] and against the Kipnis-Shamir methodology and its improvements, see [17, 18].

In [22], it is shown that an upper bound for the Q-rank of an $HFEv^-$ system is given by the sum of the Q-rank of the HFE component, the number of removed equations, and the Q-rank of the vinegar component. For Gui-96(96,5,6,6), here $q = 2$, $n = 96$, $D = 5$, $v = 6$ and $r = 6$, this quantity is roughly 15. Furthermore, in [13], experimental evidence in the form of analysis of toy variants is provided indicating that this estimate is tight. Thus the complexity of a Kipnis-Shamir style attack is roughly $O(n^3 q^{15n})$.

Also in [22], a formula for an upper bound on the degree of regularity for $HFEv^-$ systems is derived. Given the parameters of Gui-96(96,5,6,6), the degree of regularity is expected to be 9. Further, experiments are provided in [13] supporting the tightness of this approximation formula for toy schemes with n as large as 38. With this degree of regularity the expected complexity of inverting the system via Gröbner basis techniques is given by

$$\binom{96 - 6 + 9}{9}^{2.3766} \approx 2^{93}.$$

We note that an error in the approximation of the degree of regularity can easily change this estimate by a factor of a few thousand. Still, it seems clear that each of these avenues of attack is unviable.

Still another attack vector is to put the entropy of the key space to the test with techniques such as those mentioned in [23] for deriving equivalence classes of keys. With our most restrictive instance of the key verification algorithm in Section 3.2, we have a key space consisting of roughly q^{13n} central maps, roughly q^{6n} of which can be seen as equivalent keys as in [23]. Thus provable security against the differential adversary can be achieved with a key space of size far beyond the reach of the "guess-then-IP" strategy. =

6 Conclusion

$HFEv^-$ is rapidly approaching twenty years of age and stands as one of the oldest post-quantum signature schemes remaining secure. With the new parameters suggested in [13], $HFEv^-$ has metamorphosed from the very slow form of QUARTZ into a perfectly reasonable option for practical and secure quantum-resistant signatures.

Our analysis contributes to the confidence and optimism which $HFEv^-$ inspires. By elucidating the differential structure of the central map of $HFEv^-$, we have verified that a class of attacks which has proven very powerful against multivariate schemes in the past cannot be employed against $HFEv^-$. In conjunction

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

with the careful analysis of the degree of regularity and Q-rank of the scheme already present in the literature, we have succeeded in showing that $HFEv^-$ is secure against every type of attack known. If the future holds a successful attack against $HFEv^-$ it must be by way of a fundamentally new advance.

References

1. Lange, T., et al.: Post-quantum cryptography for long term security. Horizon2020 ICT-645622 (2015) http://cordis.europa.eu/project/rcn/194347_en.html.
2. Campagna, M., Chen, L., et al.: Quantum safe cryptography and security. ETSI White Paper No. 8 (2015) http://www.etsi.org/images/files/ETSIWhitePapers/QuantumSafeWhitepaper.pdf.
3. Moody, D., Chen, L., Liu, Y.K.: Nist pqc workgroup. Computer Security Resource Center (2015) http://csrc.nist.gov/groups/ST/crypto-research-projects/#PQC.
4. Smith-Tone, D.: On the differential security of multivariate public key cryptosystems. In Yang, B.Y., ed.: PQCrypto. Volume 7071 of Lecture Notes in Computer Science., Springer (2011) 130–142
5. Perlner, R.A., Smith-Tone, D.: A classification of differential invariants for multivariate post-quantum cryptosystems. [24] 165–173
6. Daniels, T., Smith-Tone, D.: Differential properties of the HFE cryptosystem. [25] 59–75
7. Dubois, V., Fouque, P.A., Shamir, A., Stern, J.: Practical Cryptanalysis of SFLASH. In Menezes, A., ed.: CRYPTO. Volume 4622 of Lecture Notes in Computer Science., Springer (2007) 1–12
8. Shamir, A., Kipnis, A.: Cryptanalysis of the oil & vinegar signature scheme. CRYPTO 1998. LNCS **1462** (1998) 257–266
9. Moody, D., Perlner, R.A., Smith-Tone, D.: An asymptotically optimal structural attack on the ABC multivariate encryption scheme. [25] 180–196
10. Patarin, J.: Cryptoanalysis of the Matsumoto and Imai Public Key Scheme of Eurocrypt'88. In Coppersmith, D., ed.: CRYPTO. Volume 963 of Lecture Notes in Computer Science., Springer (1995) 248–261
11. Perlner, R., Smith-Tone, D.: Security analysis and key modification for zhfe. In: Post-Quantum Cryptography - 7th International Conference, PQCrypto 2016, Fukuoka, Japan, February 24-26, 2016. Proceedings. (2016)
12. Patarin, J., Courtois, N., Goubin, L.: Quartz, 128-bit long digital signatures. In Naccache, D., ed.: CT-RSA. Volume 2020 of Lecture Notes in Computer Science., Springer (2001) 282–297
13. Petzoldt, A., Chen, M., Yang, B., Tao, C., Ding, J.: Design principles for hfev-based multivariate signature schemes. In Iwata, T., Cheon, J.H., eds.: Advances in Cryptology - ASIACRYPT 2015 - 21st International Conference on the Theory and Application of Cryptology and Information Security, Auckland, New Zealand, November 29 - December 3, 2015, Proceedings, Part I. Volume 9452 of Lecture Notes in Computer Science., Springer (2015) 311–334
14. Patarin, J.: Hidden Fields Equations (HFE) and Isomorphisms of Polynomials (IP): Two New Families of Asymmetric Algorithms. In: EUROCRYPT. (1996) 33–48
15. Matsumoto, T., Imai, H.: Public Quadratic Polynominal-Tuples for Efficient Signature-Verification and Message-Encryption. In: EUROCRYPT. (1988) 419–453

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

16. Berlekamp, E.R.: Factoring polynomials over large finite fields. Mathematics of Computation **24** (1970) pp. 713–735
17. Kipnis, A., Shamir, A.: Cryptanalysis of the hfe public key cryptosystem by relinearization. Advances in Cryptology - CRYPTO 1999, Springer **1666** (1999) 788
18. Bettale, L., Faugère, J., Perret, L.: Cryptanalysis of hfe, multi-hfe and variants for odd and even characteristic. Des. Codes Cryptography **69** (2013) 1–52
19. Fouque, P.A., Macario-Rat, G., Perret, L., Stern, J.: Total break of the ℓic- signature scheme. PKC 2008, LNCS **4939** (2008) 1–17
20. Smith-Tone, D.: Properties of the discrete differential with cryptographic applications. In Sendrier, N., ed.: PQCrypto. Volume 6061 of Lecture Notes in Computer Science., Springer (2010) 1–12
21. Faugere, J.C.: Algebraic cryptanalysis of hidden field equations (hfe) using grobner bases. CRYPTO 2003, LNCS **2729** (2003) 44–60
22. Ding, J., Yang, B.Y.: Degree of regularity for hfev and hfev-. [24] 52–66
23. Wolf, C., Preneel, B.: Equivalent keys in multivariate quadratic public key systems. J. Mathematical Cryptology **4** (2011) 375–415
24. Gaborit, P., ed.: Post-Quantum Cryptography - 5th International Workshop, PQCrypto 2013, Limoges, France, June 4-7, 2013. Proceedings. In Gaborit, P., ed.: PQCrypto. Volume 7932 of Lecture Notes in Computer Science., Springer (2013)
25. Mosca, M., ed.: Post-Quantum Cryptography - 6th International Workshop, PQCrypto 2014, Waterloo, ON, Canada, October 1-3, 2014. Proceedings. Volume 8772 of Lecture Notes in Computer Science., Springer (2014)

Cartor, Ryann; Gipson, Ryan; Smith-Tone, Daniel; Vates, Jeremy.
"On the Differential Security of the HFEvû-Signature Primitive."
Paper presented at PQCrypto 2016: The Seventh International Conference on Post-Quantum Cryptography, Fukuoka, Japan. February 24, 2016 - February 26, 2016.

Sub Topic: Reaction Kinetics

Spring Technical Meeting
Eastern States Section of the Combustion Institute
March 4-7, 2018
State College, Pennsylvania

Kinetics of H Atom Addition to Cyclopentene

Jeffrey A. Manion[*] and Iftikhar A. Awan

[1]*Chemical Sciences Division*
National Institute of Standards and Technology
Gaithersburg, MD, 20899-8320
[*]*Corresponding author: jeffrey.manion@nist.gov*

Abstract: To provide benchmark information needed to develop kinetic models of the combustion and pyrolysis of hydrocarbon ring structures, we have used the single pulse shock tube technique to study the kinetics of H atom addition to cyclopentene at 863 K to 1167 K and pressures of 160 kPa to 370 kPa. Addition of H to the pi bond leads to the cyclopentyl radical, which rapidly decomposes to ethene and allyl radical. Rate constants for the overall process were determined relative to a reference reaction via post-shock GC/FID/MS monitoring of products. A Transition-State-Theory/Rice-Ramsberger-Kassel-Marcus (TST/RRKM) model has been applied in conjunction with evaluated literature data to convert the primary measurements to a high pressure limiting rate expression for H addition. Results are compared with related systems. Near 1000 K, our data require a minimum value of 1.5 for branching between beta C-C and C-H bond scission in cyclopentyl radicals to maintain established trends in H addition rates. This minimum is consistent with a branching value of about 3 that we determined in previous experiments, but conflicts with much smaller values derived by current computations and those used in recent kinetics models to describe jet-stirred reactor studies of cyclopentane combustion.

Keywords: Kinetics, Fuels, Cyclopentyl, Shock Tube, RRKM

1. Introduction

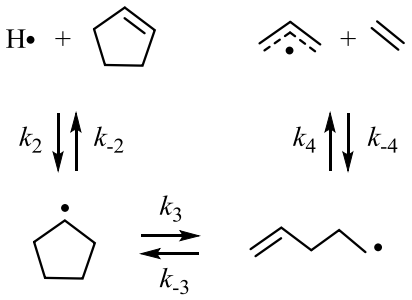
Cyclopentene (CPE) is a combustion intermediate and component of some fuels. It is the smallest hydrocarbon having an unsaturated C5 ring and is a prototypical species useful for predicting the behavior of larger analogs. Such ring structures may be present in liquid fuels or may arise from the short chain radicals and olefinic species formed during the breakup of larger fuels. The reaction of these species with small radicals leads to polycyclic aromatic hydrocarbons (PAH's), finally terminating with particulate soot,[1] which is an environmental and regulatory concern. Our interest is related to the development of reliable kinetic models for the combustion of cyclic fuels and their conversion to PAH. Benchmark information on the base ring structure is expected to provide useful information needed to reliably extend the models to larger systems

We report herein the use the shock tube technique to determine rate constants for the addition of H atoms to cyclopentene at temperatures near 1000 K. We also use a TST/RRKM model to explore the relationship of this process to the subsequent decomposition of the cyclopentyl adduct radical. Addition of H to cyclopentene forms the cyclopentyl radical, which at higher temperatures undergoes rapid ring opening and decomposition to ethene and allyl radical.[2-5] Ring opening competes with the re-ejection of a hydrogen atom, the kinetics of which is related to H addition through detailed balance. The overall decomposition is:

cyclopentene (CPE) + H → ethene + allyl (R1)

Awan, Iftikhar; Manion, Jeffrey.
"Kinetics of H Atom Addition to Cyclopentene."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

Sub Topic: Reaction Kinetics



Scheme 1: Decomposition of cyclopentene (CPE) induced by addition of H atoms.

As depicted in Scheme 1, The process consists of three elementary reactions, R2 to R4:

cyclopentene (CPE) + H ⇄ cyclopentyl (R2)

cyclopentyl ⇄ pent-4-en-1-yl (R3)

pent-4-en-1-yl ⇄ ethene + allyl (R4)

Although the mechanism is established, there is conflicting information about the various rate constants, and it is the balance of the various elementary reactions that determines the overall behavior under conditions of interest.

In experimental work reported in 2011 we measured the branching between cyclopentene and ethene products in the decomposition of cyclopentyl radical under conditions similar to the present experiments and developed a TST/RRKM model of R2 to R4.[2] Our measured ratio for CC/CH beta scission in cyclopentyl (k_3/k_{-2}) was about an order of magnitude larger than values derived by computation.[5] In the intervening years a number of relevant studies have been reported. Pertinent to k_{-3} (Scheme 1), Wang *et al.*[6] in 2015 derived structure activity rate estimation rules for ring closure in C_4 to C_8 alkenyl radicals. Herbinet *et al.*[1] studied and modeled cyclopentene pyrolysis and formation of the first aromatic ring in a jet-stirred reactor (JSR) in 2016. Finally, in 2017 Al Rashidi *et al.*[7] reported a study of cyclopentane combustion in a JSR and developed a detailed kinetic model to describe the results. They found unusual inhibition behavior and their model showed that the relative and absolute rate constants for cyclopentyl radical decomposition were critical parameters. They were unable to model their results using the rate constants and k_3/k_{-2} branching ratios we reported in 2011.

In the present work, we use thermal precursors to create H atoms in the presence of a large excess of cyclopentene and an additional reference compound. The stable ethene product of R1 is quantitated in post-shock GC analyses and compared with the amount of product formed by attack of H on the reference species This directly yields relative rate constants, which are then converted to absolute values of R1 based on the reference rate. We combine our data with computations and various literature data to create a detailed kinetic model that relates the results to the elementary reactions of Scheme 1. Our data and analysis lead to reliable high pressure rate expression for k_2 and an independent method of estimating k_3/k_{-2}.

2. Methods / Experimental

Experiments are carried out in a heated single pulse shock tube having an ≈ 500 μs reaction time and described in earlier publications.[8] In the current studies, reactants are highly diluted in

Awan, Iftikhar; Manion, Jeffrey.
"Kinetics of H Atom Addition to Cyclopentene."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

a bath gas of argon and we use small quantities of a precursor [≈50 μL/L (ppm)] to create a limited number of H atoms in the presence of much larger quantities (≈10,000 μL/L) of cyclopentene and 1,3,5-trimethylbenzene (135TMB). The latter compound serves as both a rate reference for H atom reactions and a radical scavenger. A Hewlett Packard 6890 N gas chromatograph equipped with FID and MS detection is used to measure the post-shock products. Including systematic errors, analytical uncertainties (2σ) for the main products are estimated as 6 %.

Generation of H atoms. We have used three different methods to thermally produce H atoms. On the ≈ 500 μs time scale of our experiments, thermolysis of hexamethylethane (HME) is an effective source of H at temperatures >1000 K:

hexamethylethane → 2 tert-butyl (R5)

tert-butyl → isobutene + H (R6)

To probe lower temperatures, a few experiments were conducted with mixtures containing small amounts (≈25 μL/L) of tert-butyl peroxide (tBPO) and a large amount of H_2 (≈20 %), which leads to H atoms via the following sequence:

tert-butyl peroxide → 2 $(CH_3)_3C\text{-}O$ → 2 acetone + 2 CH_3 (R7)

$CH_3 + H_2 \rightarrow CH_4 + H$ (R8)

Although useful as a confirmatory technique, we found this method to be of limited value due to the loss of methyl radicals to side reactions. The decomposition of 2-iodopropane proved to be a more effective source of H atoms at temperatures < 1000 K. 2-Iodopropane undergoes both molecular elimination of HI and fission of the weak C-I bond. The latter process generates an isopropyl radical that rapidly ejects H:

2-iodopropane → propene + HI (R9)

2-iodopropane → *i*-C_3H_7 + I (R10)

i-C_3H_7 → propene + H (R11)

We determine rate constants for the reaction of H with CPE relative to known values for the reaction of H with 135TMB. Attack of H on 135TMB leads either to H_2 and dimethylbenzyl radical (DMB) or displacement of methyl to give *m*-xylene:

H + 135TMB → H_2 + dimethylbenzyl (DMB) (R12)

H + 135TMB → CH_3 + *m*-xylene (R13)

Near 1000 K the branching is about 2:1 in favor of abstraction.[9] DMB does not readily propagate radical chains and inhibits secondary chemistry through recombination reactions. Formation of *m*-xylene through R13 is a highly specific rate reference for reactions of H atoms. For R13 we use the rate parameters of Tsang et al.[9], k(H + 135TMB → *m*-xylene + CH_3) = 6.7×10^{13} exp(−3255/T) cm^3 mol^{-1} s^{-1}. At our temperatures the uncertainty in k_{13} is about a factor of 1.5.[9] The existing database of rate constants measured relative to R13 provides valuable comparisons and a consistent scaling of absolute values.

3. Results and Discussion

Matheu et al.[10] used a mechanism generating computercode to identify 70 possible product channels in the H + cyclopentene addition reaction, but concluded that the reactions of Scheme 1 are the only important unimolecular processes. Our experimental product data support this

Awan, Iftikhar; Manion, Jeffrey.
"Kinetics of H Atom Addition to Cyclopentene."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

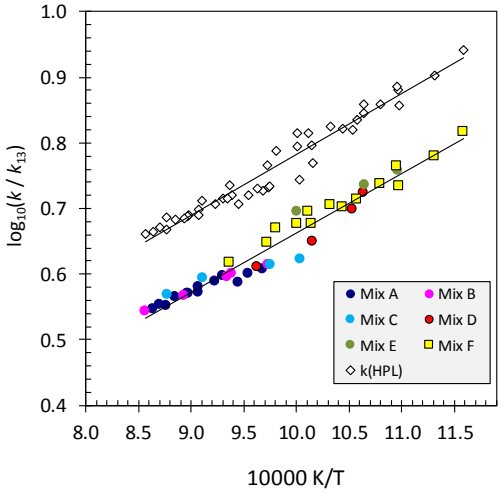


Figure 1. Experimental rate constant ratios for k_1/k_{13} (filled markers) and the corresponding high pressure limit values for $k_{2(\infty)}/k_{13}$ (unfilled markers) derived after applying an experimentally based correction for branching in the decomposition of the cyclopentyl radical (see text).

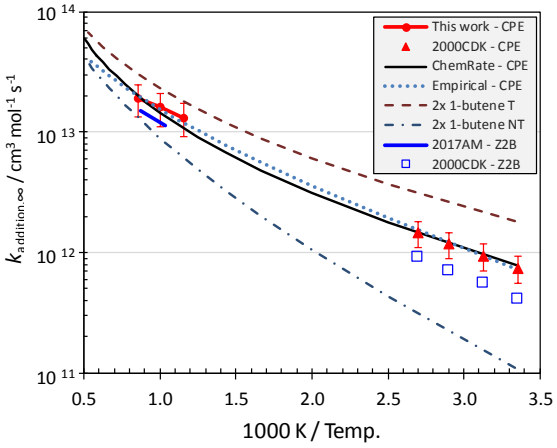


Figure 2. High pressure rate constants for addition of H to CPE and selected alkenes.; Z2B = (*Z*)-2-butene; T = terminal site; NT = nonterminal site. Dotted and solid lines indicate respective empirical and TST fits to the CPE data. Indicated uncertainties are 2σ. Data for 1-butene are multiplied by 2 to normalize the number of sites relative to CPE. Data sources: CPE[11]; 1-butene[8]; Z2B[11, 12].

conclusion. Thus, ethene formation is directly related to the overall reaction CPE + H → cyclopentyl → ethene + allyl (R1). The rate of this reaction relative to R13 is given by the molar yields of ethene and *m*-xylene normalized by the ratio of the reactants, k_1/k_{13} = [ethene][135TMB]/[*m*-xylene][CPE]. These data are plotted in Figure 1 and give:

$$\frac{k_1}{k_{13}} = 10^{-0.196\pm0.062}\exp[(1995 \pm 60)\ \mathrm{K}/T);\ 863\ \mathrm{K}\ \text{to}\ 1167\ \mathrm{K}$$

The given uncertainties are 2σ and represent precision only. Including systematic errors, the overall uncertainty (2σ) in the relative rate is estimated as about ± 15 %. Using k_{13} we derive:

$$k_1 = 4.27 \times 10^{13}\exp(-1260\ \mathrm{K}/T)\ \mathrm{cm^3\ mol^{-1}s^{-1}};\ 863\ \mathrm{K}\ \text{to}\ 1167\ \mathrm{K}$$

The uncertainty (2σ) in the absolute rate constant is about a factor of 1.5 and is due mainly to the uncertainty in the rate of the reference reaction. Because addition of H to cyclopentene, R2, is partially reversible, k_1 represents the minimum possible value of k_2. Assuming all H additions yield either ethene or cyclopentene as a stable olefin, then $k_2(\infty) = k_1(\gamma+1)/\gamma$, where γ is the molar product ratio, [ethene]/[CPE], in the decomposition of the cyclopentyl radicals formed by R2.

As reported in 2011,[9] we have previously measured values of γ by creating cyclopentyl radicals under dilute conditions and directly observing the ethene/CPE product ratio. G3MP2B3 computations were used to compute energies and other properties and an RRKM master equation analysis of the system depicted in Scheme 1 was carried out with ChemRate software[13]. The initial theoretical model was minimally tuned to match available data. Presently, we have slightly updated the 2011 model to better match additional literature information and then used it to derive values of γ for the chemically activated cyclopentyl radicals formed in the current experiment. Computed γ values are about 3 and little changed from the thermal values under our conditions. They lead to values of $k_{2(\infty)}$ that are about 30 % larger than k_1 (Figure 1). For $k_{2(\infty)}$, we find:

Awan, Iftikhar; Manion, Jeffrey.
"Kinetics of H Atom Addition to Cyclopentene."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

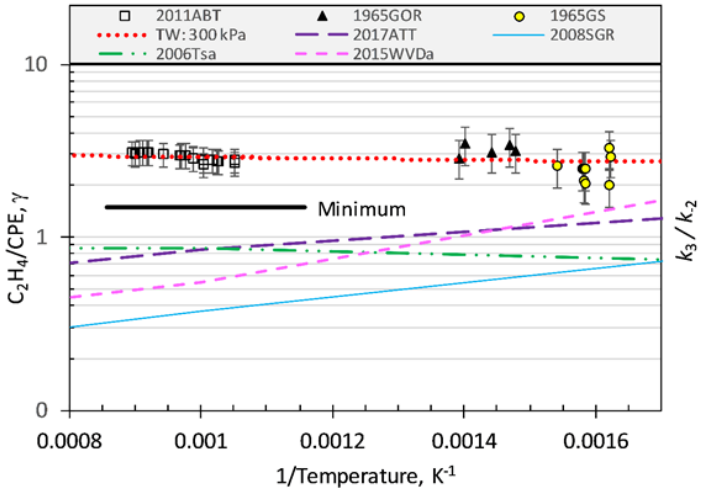


Figure 3. Branching in cyclopentyl decomposition. The symbols and dotted line show ethene/cyclopentene product ratios (γ, left axis) from experiment and our model at 300 kPa, respectively; the other lines are k_3/k_{-2} values (right axis) at the high pressure limit, which are approximately equivalent. The indicated minimum is that required by our current experiments to maintain consistency in H atom addition rates. Product ratios of 1965GOR[3] 1965GS[4] are plotted only at their higher temperatures where they relate to k_3/k_{-2}. Values from 1965GS are corrected for cyclopentene formed by disproportionation of cyclopentyl using k_{dis}/k_{rec} = 1 as reported in that work. Data sources: TW, this work; 2011ABT[2]; 2006Tsa[14]; 2017ATT[7]; 2015WVDa[6]; 2008SGR[5].

$$k_{2(\infty)} = 5.37 \times 10^{13} \exp(-1213\ \mathrm{K}/T)\ \mathrm{cm^3\ mol^{-1}s^{-1}};\ 863\ \mathrm{K}\ \text{to}\ 1167\ \mathrm{K}$$

Our TST/RRKM model has been tuned to fit the present data as well as those of Clarke *et al*. at 298 K to 370 K, and results in $k_{2(\infty)} = 9.09\text{x}10^7 T^{1.783}\exp(-324\ \mathrm{K}/T)$ between 298 K and 2000 K. Comparisons with pertinent data are presented in Figure 2. Data near 1000 K are from this laboratory and are all derived relative to k_{13}. The addition rate of H to cyclopentene lies between those for terminal and nonterminal addition of H to 1-butene and is slightly faster than the analogous reaction with (Z)-2-butene, a noncyclic analog.

We have derived $k_{2(\infty)}$ using γ values of about 3. At our temperatures, values of γ less than 1.1 would require, on a per site basis, that the high pressure rate constant for H atom addition to cyclopentene exceed that for terminal addition of H to 1-butene. Such a result would be strongly at odds with known trends in rates of H atom addition to olefins. To maintain self-consistency in H atom addition rates, the reasonable lower limit of γ is about 1.5. This derived minimum is consistent with our measured γ values (about 3), but is obtained in an independent manner. It is at odds, however, with γ values derived from computations and used in current kinetics models of cyclopentane combustion, all of which are in the range of 0.3 to 0.8 near 1000 K.

Figure 3 summarizes experimental values of γ and theoretical values for the closely related branching ratio k_3/k_{-2}. The two are not strictly identical because of the slight reversibility of ring opening (about 5%), but all current models agree that the difference is minimal. Note that extrapolation of the experimental γ values to high pressure limits would slightly increase differences with the k_3/k_{-2} values. The experimental product ratios exhibit only a weak temperature dependence between 600 K and 1100 K. These values are derived from GC measurements having high precision, so significant errors would have to be systematic and related to the mechanism assumed in the data analyses. Both Gordon[3] and Gunning and Stock[4] conclude that in their

Awan, Iftikhar; Manion, Jeffrey.
"Kinetics of H Atom Addition to Cyclopentene."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

studies cyclopentene is formed at lower temperatures primarily by disproportionation reactions of cyclopentyl. The data from these works plotted in Figure 3 are from their higher temperatures where C-H bond scission is believed to be the dominant source of cyclopentene, conditions where ethene/cyclopentene ratios should be closely related to k_3/k_{-2}. Notice that contributions from disproportionation would tend to make the derived ratios too small and would further increase the discrepancy with the values from theory or used in the modeling work.

In their experimental and modeling study of cyclopentane combustion in a JSR, Al Rashidi *et al.*[7] showed that the relative and absolute rate constants for cyclopentyl radical decomposition were critical parameters in their detailed kinetic model. They were unable, however, to reproduce the observed reactivity and inhibition behavior without using branching ratios much smaller than we reported in 2011.[2] Nonetheless, the present data and analysis appears to independently confirm the 2011 results. The reasons for this apparent disagreement remain to be explained.

4. Conclusions

Shock tube methods have been used to investigate the addition of H atoms to cyclopentene at temperatures of (863 to 1167) K and pressures of (160 to 370) kPa. Addition of H to the double bond leads to a cyclopentyl radical that rapidly ring opens and decomposes to ethene and allyl radical. We have determined rate constants for the overall process and related these to the high pressure limiting rate constant for H atom addition. The data allow the relative and absolute rates of CC and CH β-scission in cyclopentyl radical to be deduced. Near 1000 K, ratios of CC to CH scission smaller than about 1.5 are shown to result in inconsistencies in known relative rates of H atom addition to olefins. The present data are consistent with branching values near 3 as determined in our previous work, but do not agree with reported much smaller values derived from theoretical studies or those seemingly required in detailed kinetic models of cyclopentane combustion to explain the observed global reactivity behavior.

6. References

[1] O. Herbinet, A. Rodriguez, B. Husson, F. Battin-Leclerc, Z.D. Wang, Z.J. Cheng, F. Qi, J. Phys. Chem. A 120 (2016) 668-682.
[2] I.A. Awan, D.R. Burgess, Jr., W. Tsang, J.A. Manion, Proc. Combust. Inst. 33 (2011) 341-349.
[3] A.S. Gordon, Can. J. Chem. 43 (1965) 570-581.
[4] H.E. Gunning, R.L. Stock, Can. J. Chem.-Rev. Can. Chim. 42 (1964) 357-370.
[5] B. Sirjean, P.A. Glaude, M.F. Ruiz-Lopez, R. Fournet, J. Phys. Chem. A 112 (2008) 11598-11610.
[6] K. Wang, S.M. Villano, A.M. Dean, J. Phys. Chem. A 119 (2015) 7205-7221.
[7] M.J. Al Rashidi, S. Thion, C. Togbe, G. Dayma, M. Mehl, P. Dagaut, W.J. Pitz, J. Zador, S.M. Sarathy, Proc. Combust. Inst. 36 (2017) 469-477.
[8] J.A. Manion, I.A. Awan, J. Phys. Chem. A 119 (2015) 429−441.
[9] W. Tsang, J.P. Cui, J.A. Walker, Single Pulse Shock Tube Studies on the Reactions of Hydrogen Atoms with Unsaturated Compounds, Proceedings of the 17th International Symposium on Shock Waves & Shock Tubes, Bethlehem, PA, AIP Conference Proceedings 208, American Institute of Physics, Melville, NY, 1990, pp. 63 - 73, DOI: 10.1063/1061.39497.
[10] D.M. Matheu, W.H. Green, J.M. Grenda, Int. J. Chem. Kinet. 35 (2003) 95-119.
[11] J.S. Clarke, N.M. Donahue, J.H. Kroll, H.A. Rypkema, J.G. Anderson, J. Phys. Chem. A 104 (2000) 5254-5264.
[12] I.A. Awan, J.A. Manion, unpublished work, 2018.
[13] V. Mokrushin, V. Bedanov, W. Tsang, M. Zachariah, V.D. Knyazev, W.S. McGivern, ChemRate, National Institute of Standards and Technology, Gaithersburg, Maryland, 1996-2011.
[14] W. Tsang, J. Phys. Chem. A 110 (2006) 8501-8509.

Awan, Iftikhar; Manion, Jeffrey.
"Kinetics of H Atom Addition to Cyclopentene."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

Spring Technical Meeting
Eastern States Section of the Combustion Institute
March 4-7, 2018
State College, Pennsylvania

Evaluated Rate Constants for *i*-Butane + H and CH_3: Shock Tube Experiments with Bayesian Model Optimization

Laura A. Mertens,[1,*] *Ifthikar A. Awan*[1] *and Jeffrey A. Manion*[1]
[1]*Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD, 20899-8320*
[*]*Corresponding Author Email: laura.mertens@nist.gov*

Abstract: The reactions of *i*-butane with CH_3 and H were investigated with shock tube experiments (870 K to 1130 K and 140 kPa to 360 kPa). Propene and *i*-butene, measured with GC/FID and MS, were quantified as characteristic of H-abstraction from the primary and tertiary carbons, respectively. A comprehensive Cantera[1] kinetics model based on JetSurF 2.0[2] was optimized to these experimental measurements and literature data – including early experiments at low temperatures, measurements of the total rate constant, and measurement from ethane for the rate constant for primary carbons – using the Method of Uncertainty Minimization using Polynomial Chaos Expansions (MUM-PCE) – pioneered by David Sheen and Hai Wang.[3] For both H and CH_3, the optimization increased the rate H-abstraction from the tertiary carbon relative to the primary carbon. We combined our primary and tertiary rate constants with previous results from our group on *n*-butane[4] to get site-specific rate constants for the reaction of H and CH_3 with a generic primary, secondary and tertiary carbon.
***Key Words**: kinetics, pyrolysis, combustion, modeling, H-abstraction*

1. Introduction

Combustion chemistry is driven by chain reactions propagated by free radicals, like H and methyl (CH_3).[5] H and CH_3 radicals readily react with hydrocarbons, typically abstracting a H to form H_2 or CH_4, respectively. The rate of this reaction is highly dependent on the structure of the carbon center from which the H is abstracted. Due to the relative strengths of their C-H bonds, tertiary carbons react faster with H and CH_3 than secondary carbons, which react faster than primary carbons. The global structure and the mass of the alkane do not strongly affect the rate of hydrogen abstraction, so rate constants for a large hydrocarbon fuel are often calculated from the number of its primary, secondary and tertiary carbons.[6]

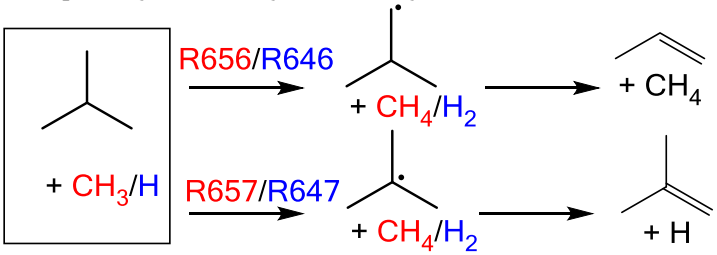


Figure 1. Reactions of CH_3 (red, reactions R656 and R657) and H (blue, reactions R646 and R647) with *i*-butane.

Here, we found relative rate constants for the reaction of *i*-butane with H and CH_3 radicals with shock tube experiments. The shock tube data was combined with literature data to find to

Awan, Iftikhar; Manion, Jeffrey; Mertens, Laura.
"Evaluated Rate Constants for i-Butane + H and CH3: Shock Tube Experiments with Bayesian Model Optimization."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

find rates for abstraction at both the primary and tertiary carbons using MUM-PCE. Combining primary and tertiary rate constants from this work with previous work on *n*-butane,[4] we propose self-consistent rate constants for H-abstraction from primary, secondary and tertiary carbon centers by H and CH_3.

2. Methods

All experiments were performed in the NIST shock tube reactor which has been described elsewhere.[4] H and CH_3 radicals were produced from pyrolysis of C_2H_5I and *di-tert*-butylperoxide (*dt*BPO), respectively. After the H or CH_3 reacted with an excess of *i*-butane, the products – especially propene and *i*-butene, produced from H-abstraction from the primary carbon and secondary carbons, respectively– were measured with gas chromatography with flame ionization and mass spectrometric detection. The ratio of [propene]/[*i*-butene] approximately equals the branching ratio of k_{656}/k_{657} for experiments with *dt*BPO and of k_{646}/k_{647} for experiments with C_2H_5I. An excess of toluene was added to scavenge free radicals. The shock temperature and pressure were (868 to 1131) K and (130 to 360) kPa.

Table 1. Rate Constants for the Main Reactions of the Prior Model. Parameters are for $k = A \times T^n \times \exp(-E_a/RT)$. A is in units of mol, s, cm. E_a/R is in units of K.

Reaction number	Reaction		$\log_{10}A$	n	E_a/R	Reference
645	$CH_3 + iC_3H_7 \leftrightarrow iC_4H_{10}$	k_∞	15.15	-0.68	0	JetSurF 2.0[2]
		k_0	61.62	-13.33	1964.3	
646	$H + iC_4H_{10} \leftrightarrow H_2 + iC_4H_9$	k	6.257	2.54	3400.	Tsang and JetSurF 2.0[2, 7]
647	$H + iC_4H_{10} \leftrightarrow H_2 + tC_4H_9$	k	5.780	2.4	1300.	
656	$CH_3 + iC_4H_{10} \leftrightarrow CH_4 + iC_4H_9$	k	0.132	3.65	3600.	
657	$CH_3 + iC_4H_{10} \leftrightarrow CH_4 + tC_4H_9$	k	-0.044	3.46	2314	
673	$Toluene + H \leftrightarrow Benzyl + H_2$	k	14.17	0	4500.	Sheen *et al*.[8]
674	$Toluene + H \leftrightarrow Benzene + CH_3$	k	6.230	2.2	2000.	Sheen *et al*.[8]
676	$Toluene + CH_3 \leftrightarrow Benzyl + CH_4$	k	11.50	0	4780.6	JetSurF 2.0[2, 9]
805	$Benzyl + CH_3 \leftrightarrow Ethylbenzene$	k	13.08	0	111.2	Brand *et al*.[10]

We created a chemical model with the Cantera software package[1] based off the JetSurF 2.0 chemical mechanism.[2] This mechanism was modified by David Sheen to include compounds commonly used for shock tube experiments at NIST[11], [4, 8] and C_2H_5I decomposition from Bentz *et al.* Rate constants for the decomposition of *i*-butyl radicals were taken from experiments in our laboratory that measured that *i*-C_4H_9 is dominated by C-C scission. Under 1050 K, the kinetics model predicts that >96 % of propene is from R646 and R656 – abstraction from the primary carbon. For the entire temperature range of the experiments, ≥97 % of *i*-butane is from abstraction from the tertiary carbon, R647 and R657.

The Cantera kinetics was optimized to the experimental data using Method of Uncertainty Minimization using Polynomial Chaos Expansions (MUM-PCE) with the mumpce 0.1 software package published by David Sheen.[3, 12] First we performed a sensitivity analysis to find what reactions – called "active parameters" – are most important to predicting the outcome of the shock experiment (Table 1). With Bayes theorem, MUM-PCE then optimized the active parameters' rate constants to better match experimental data from the shock tube and relevant

Awan, Iftikhar; Manion, Jeffrey; Mertens, Laura.
"Evaluated Rate Constants for i-Butane + H and CH3: Shock Tube Experiments with Bayesian Model Optimization."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

literature data.[13] The literature data included measurements of raw experimental data, total rate constants ($k_p + k_s$), relative rate constants, and primary rate constants from ethane.

3. Results and Discussion

Product distributions are shown for a typical trial on Figure 2 for all species with abundances of 1 % or more of acetone. Besides acetone, which is produced from *dt*PBO decomposition, the two most abundant species are methane and ethane, both from CH_3 radical chemistry: methane from R656 and R657 and ethane from CH_3 self-recombination. As the main products of CH_3 reaction with *i*-butane, propene and *i*-butene were also detected in significant amounts. We also measured aromatic compounds from the reactions of CH_3 with toluene, including ethylbenzene (R805), benzene (R674, H reactant produced from R657) and bibenzyl. All products except acetone increase with temperature (Figure 2), and this increase accelerates above 1050 K. This correlates with the decomposition of *i*-butane (R645), which starts to decompose at 1050 K, producing additional CH_3 radicals.

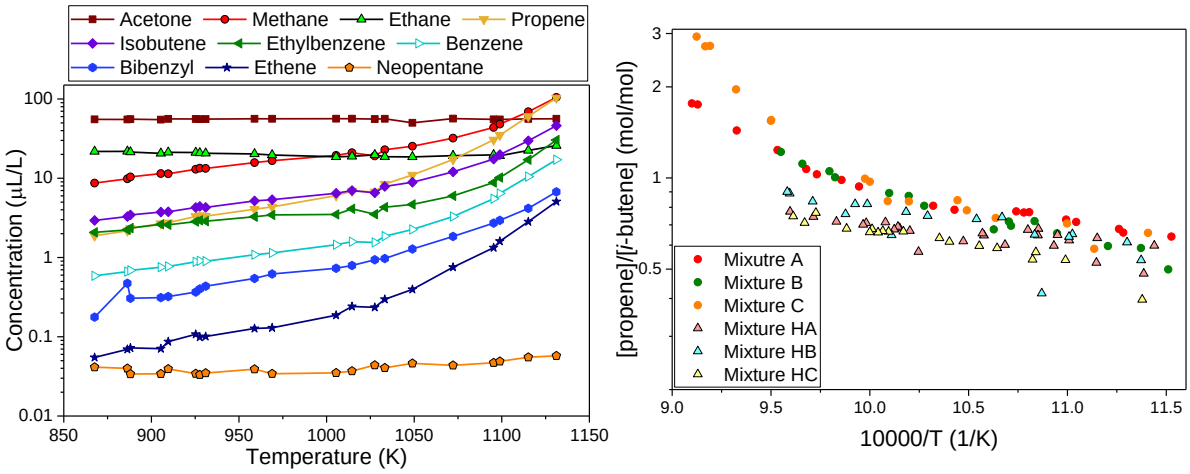


Figure 2. Left: Products a mixture with *dt*BPO. Right: [propene]/[i-butene] for all mixtures. Mixtures A, B and C have *dt*BPO and mixtures HA, HB and HC

As with mixtures *dt*BPO, mixtures with C_2H_5I had significant amounts of *i*-butene and propene from R646 and R647, respectively. Since CH_3 radicals were produced as co-products with propene from R646, we observed methane at concentrations about equal to those of propene. Ethane concentration, due to the lower CH_3 concentrations for mixtures with C_2H_5I – about an order of magnitude less than methane concentrations. We, again, measured significant amounts of benzene, ethylbenzene, and bibenzyl from toluene chemistry.

Assuming no side chemistry, the ratio of [propene]/[*i*-butene] is equal to the branching ratio of k_{646}/k_{647} for H-dominated mixtures with C_2H_5I and k_{656}/k_{657} CH_3-dominated mixtures with *dt*BPO. Figure 1 shows that for both H and CH_3 the ratio is about equal to 1 at temperatures between 868 and 1052K. This ratio increases with temperature, as tertiary abstraction is more favorable at lower temperatures due to the lower C-H bond strengths (and therefor lower activation energy for reaction R646 and R647) for tertiary carbons.[7] The branching ratio is higher for CH_3 than H, meaning that H is more selective for the tertiary carbon than CH_3.

The data from these shock tube experiments (including measurements of the ratio, propene, ethene, benzene, methane and ethane). We used MUM-PCE to find the sensitivity of the prior model – or more specifically, the model's predictions for all experiments used for the

Awan, Iftikhar; Manion, Jeffrey; Mertens, Laura.
"Evaluated Rate Constants for i-Butane + H and CH3: Shock Tube Experiments with Bayesian Model Optimization."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

optimization – to changing each of its 1487 rate constants. This analysis lead to the choice of nine rate constants to optimize – which we call active parameters (Table 1). For most of these nine active parameters (Table 1), we only optimized the A-factor (which is the same as multiplying the rate constant by a temperature- and pressure-independent constant). For reactions R646, R647, R656, and R657, we optimized both the A-factor and the activation energy, E_a, as these reactions are the primary focus of this work and are the most well-constrained by our experiment and the chosen literature data.

Table 2. Posterior Active Parameters with Factorial Uncertainties (f_A for A and f_{Ea} for E_a). A is in units of mol, s, cm. E_a/R is in units of K.

Reaction number		$\log_{10}A$	f_A	n	E_a/R	f_{Ea}
645	k_∞	15.57	1.32	-0.68	0	-
	k_0	61.62	-	-13.33	1964.3	-
646	k	6.17	1.55	2.54	3390	1.09
647	k	5.88	1.45	2.4	1409	1.18
656	k	0.233	1.67	3.65	3819	1.14
657	k	0.191	1.41	3.46	2362	1.08
673	k	14.23	1.42	0	4500	-
674	k	6.15	1.35	2.2	2000	-
676	k	11.54	1.45	0	4780.6	-
805	k	12.83	1.61	0	111.2	-

We first optimized all active parameters on Table 1 to only our data We updated the kinetics model with the optimized rate constants for R645, R673, R673, R676 and R805, and reoptimized R646, R647, R656 and R657 to (1) the literature data and (2) our rate constants for R646, R647, R656 and R657 found from the first optimization. This two-step optimization was done so that our data would have approximately equal weight to the other experiments.

The posterior model also well describes our experimental data for most products and most experimental conditions, to within 35 %. The optimization decreased this relative rate constant to match the experimental ratios. From 950 -1200 K the ratio of k_{656}/k_{657} was lowered by (61 to 62) %, mostly due to a 72 % increase in the A-factor for CH_3 reaction with the tertiary carbon of *i*-butane (R657). The optimization also increased the relative amount of tertiary abstraction for H reaction with *i*-butane (26 to 28) % from 950 K to 1100 K), by increasing the A-factor of R647 by 25 % and decreasing the A-factor of R646 by 18 %.

The relative rate constants for the posterior model are (again, on a per-H basis):

$$\text{H} + i\text{-butane: } k_p/k_t = 10^{(-0.660 \pm 0.250)} \times T^{0.14} \times \exp(-(1982 \pm 397)/T)$$

$$\text{CH}_3 + i\text{-butane: } k_p/k_t = 10^{(-0.912 \pm 0.268)} \times T^{0.19} \times \exp(-(1457 \pm 567)/T)$$

and are applicable to temperatures from (270 to 1327) K. All uncertainties are 2σ.

By combining k_p and k_t from this work with k_s/k_p (k_s denotes the rate constant for H-abstraction from a secondary carbon) from previous work by Manion *et al.*[4] on *n*-butane, we provide a full set of self-consistent rate constants for H and CH_3 H-abstraction from primary, secondary and tertiary carbons. For both CH_3 and H reaction with a generic hydrocarbon, we use H-abstraction rate constants for primary carbons from this work, and find the secondary rate constant by multiplying the relative rate constants by Manion *et al.*[4] for k_s/k_p by our value of k_p. While Manion *et al.* did find rate constants for k_p with MUM-PCE for both H and CH_3, the values of k_p from this work are based on a larger data set.

Awan, Iftikhar; Manion, Jeffrey; Mertens, Laura.
"Evaluated Rate Constants for i-Butane + H and CH3: Shock Tube Experiments with Bayesian Model Optimization."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

We found the following rate constants for H reaction with a generic hydrocarbon on a per-H basis, (270 to 1327) K:

$$k_p = 10^{(5.22 \pm 0.19)} \times n^{2.54} \times \exp(-(3391 \pm 305)\ \mathrm{K} / T)\ \mathrm{cm^3\ mol^{-1}\ s^{-1}}$$
$$k_s = 10^{(5.64 \pm 0.20)} \times n^{2.4} \times \exp(-(2145 \pm 336)\ \mathrm{K} / T)\ \mathrm{cm^3\ mol^{-1}\ s^{-1}}$$
$$k_t = 10^{(5.88 \pm 0.16)} \times n^{2.4} \times \exp(-(1409 \pm 253)\ \mathrm{K} / T)\ \mathrm{cm^3\ mol^{-1}\ s^{-1}}$$

These values are plotted on Figure 3 and compared to evaluated rate constants by Tsang.[7, 14] The rate constants are only slightly updated from the prior model.

We found the following rate constants for CH_3 reaction with a generic hydrocarbon on a per-H basis, (270 to 1327) K:

$$k_p = 10^{(-0.722 \pm 0.223)} \times n^{3.65} \times \exp(-(3819 \pm 535)\ \mathrm{K} / T)\ \mathrm{cm^3\ mol^{-1}\ s^{-1}}$$
$$k_s = 10^{(0.095 \pm 0.231)} \times n^{3.46} \times \exp(-(3193 \pm 547)\ \mathrm{K} / T)\ \mathrm{cm^3\ mol^{-1}\ s^{-1}}$$
$$k_t = 10^{(0.191 \pm 0.149)} \times n^{3.46} \times \exp(-(2362 \pm 189)\ \mathrm{K} / T)\ \mathrm{cm^3\ mol^{-1}\ s^{-1}}$$

These values are plotted and compared to evaluated rate constants from Tsang.[7, 15] While the primary rate constant agrees well with Tsang, our rate constants for abstraction from a secondary carbon and tertiary carbons are slightly lower and slightly higher, respectively, than Tsang's rate constants.

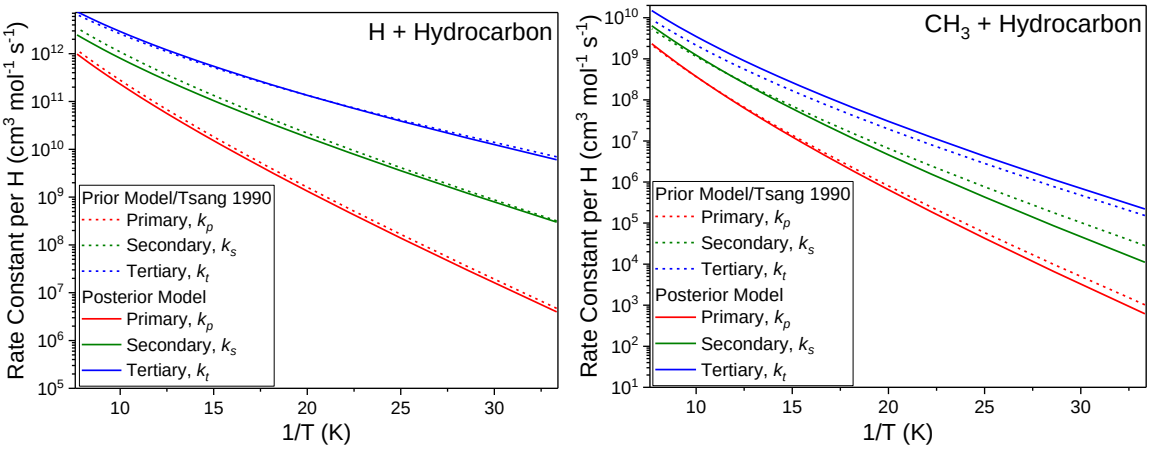


Figure 3. Our evaluated rate constants for H and CH_3 + a generic primary, secondary and tertiary carbon on a per-H basis compared to previous values by Tsang.[7, 15]

As expected due to the relative bond strengths, the activation energy for H-abstraction decreases from primary to secondary to tertiary for reaction with both H and CH_3. The A-factors increase from primary to secondary to tertiary, which ensures that the rate constants will not cross if they are extrapolated to higher temperatures. The rate constant for H and CH_3 reaction with any hydrocarbon can be found with the following formulate:

$$k_{total} = k_p \times N_{Hp} + k_s \times N_{Hs} + k_t \times N_{Ht}$$

where N_{Hp}, N_{Hs}, and N_{Ht} are the number of H's attached to primary, secondary and tertiary carbons, respectively.

4. Conclusions

We successfully used the MUM-PCE software package[12] to provide evaluated rate constants for H and CH_3 reaction with *i*-butane. The MUM-PCE method optimized a Cantera kinetics model to the experimental data we found using the NIST shock tube as well as to a range of literature data. The resulting rate constants were combined with previous work from our

Awan, Iftikhar; Manion, Jeffrey; Mertens, Laura.
"Evaluated Rate Constants for i-Butane + H and CH3: Shock Tube Experiments with Bayesian Model Optimization."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.

laboratory on *n*-butane[4] to give self-consistent rate constants for H and CH_3 abstraction of a generic H on primary, secondary and tertiary carbons.

9. Acknowledgments
David Sheen for providing technical support for Cantera and MUM-PCE. The National Academy of Science's NRC Research Associateship Program for financial support (LAM).

10. References
[1] D.G. Goodwin, H.K. Moffat, and R.L. Speth, *Cantera.* 2017.
[2] Wang, H., et al., *JetSurF version 2.0*. 2010.
[3] D.A. Sheen and H. Wang, The method of uncertainty quantification and minimization using polynomial chaos expansions, Combust. Flame 158 (2011) 2358-2374.
[4] J.A. Manion, D.A. Sheen, and I.A. Awan, Evaluated Kinetics of the Reactions of H and CH3 with *n*-Alkanes: Experiments with *n*-Butane and a Combustion Model Reaction Network Analysis, J. Phys. Chem. A 119 (2015) 7637-7658.
[5] H.J. Curran, Rate constant estimation for C-1 to C-4 alkyl and alkoxyl radical decomposition, International Journal of Chemical Kinetics 38 (2006) 250-275.
[6] N. Cohen, Are reaction rate coefficients additive? Revised transition state theory calculations for OH + alkane reactions, International Journal of Chemical Kinetics 23 (1991) 397-417.
[7] W. Tsang, Chemical Kinetic Data Base for Combustion Chemistry Part 4. Isobutane, Journal of Physical and Chemical Reference Data 19 (1990) 1-68.
[8] D.A. Sheen, C.M. Rosado-Reyes, and W. Tsang, Kinetics of H atom attack on unsaturated hydrocarbons using spectral uncertainty propagation and minimization techniques. Proceedings of the Combustion Institute, 34 (2013) 527-536.
[9] J. A. Kerr and M. J. Parsonage, Evaluated Kinetic Data on Gas Phase Hydrogen Transfer Reactions of Methyl Radicals, Berichte der Bunsengesellschaft für physikalische Chemie, 80 (1976) 825-825.
[10] U. Brand, *et al.*, Carbon-carbon and carbon-hydrogen bond splits of laser-excited aromatic molecules. 1. Specific and thermally averaged rate constants, J. Phys. Chem. 94 (1990) 6305-6316.
[11] D.A. Sheen. Personal correspondence.
[12] D.A. Sheen. mumpce 0.1 documentation, 2017 April 6, Available from: https://davidasheen.github.io/mumpce_py/#
[13] J. A. Manion, *et al.*, NIST Chemical Kinetics Database, NIST Standard Reference Database 17, Version 7.0 (Web Version), Release 1.6.8, Data version 2015.12, National Institute of Standards and Technology, Gaithersburg, Maryland, 20899-8320. http://kinetics.nist.gov/
[14] W. Tsang, Chemical kinetic data-base for combution chemistry. 3. Propane. Journal of Physical and Chemical Reference Data, 17 (1988) 887-952.
[15] W. Tsang, Chemical Kinetic Data Base for Combustion Chemistry 5. Propene. Journal of Physical and Chemical Reference Data, 20 (1991) 221-273

Awan, Iftikhar; Manion, Jeffrey; Mertens, Laura.
"Evaluated Rate Constants for i-Butane + H and CH3: Shock Tube Experiments with Bayesian Model Optimization."
Paper presented at Combustion Institute Eastern States Spring Meeting, Station College, PA, United States. March 4, 2018 - March 7, 2018.