NIST Technical Note 2151

Challenge Design and Lessons Learned from the 2018 Differential Privacy Challenges

Diane Ridgeway
Mary F. Theofanos
Terese W. Manley
Christine Task

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2151

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

NIST Technical Note 2151

Challenge Design and Lessons Learned from the 2018 Differential Privacy Challenges

Diane Ridgeway
Mary F. Theofanos
Information Technology Laboratory
Information Access Division

Terese W. Manley
Communications Technology Laboratory
Public Safety Communications Research Division

Christine Task
Knexus Research Corporation
Oxon Hill, Maryland

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2151

April 2021

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 Technical Note 2151
Natl. Inst. Stand. Technol. Tech. Note 2151, 55 pages (April 2021)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2151

The Information Technology Laboratory (ITL) at the National Institute of Standards and Technology (NIST) promotes the U.S. economy and public welfare by providing technical leadership for the Nation's measurement and standards infrastructure. ITL develops tests, test methods, reference data, proof of concept implementations, and technical analyses to advance the development and productive use of information technology. ITL responsibilities include the development of management, administrative, technical, and physical standards and guidelines for the cost-effective security and privacy of other than national security-related information in federal information systems.

The Public Safety Communications Research (PSCR) Division is the primary federal laboratory conducting research, development, testing, and evaluation for public safety communications technologies. It is housed within the Communications Technology Laboratory (CTL) at the National Institute of Standards and Technology (NIST). It addresses the research and development (R&D) necessary for critical features identified by public safety entities beyond the current generation of broadband technology. PSCR conducts internal research, prize challenges and sponsors federal grants across key public safety technology areas, otherwise known as research portfolios including applied analytics for multi-modal real time data in conjunction with the Information Technology Laboratory.

Executive Summary

Datasets available to researchers and the public have proliferated in the past 10 years. These datasets have been analyzed using various statistical and machine learning methods, resulting in many useful insights, which have in turn helped to shape public policy and impacted other large-scale decision-making processes. However, certain risks have been associated with the release of many of these datasets as they may contain potentially sensitive information about individuals. The National Institute of Standards and Technology (NIST) Technical Note 1917 Public Safety Analytics Research and Development (R&D) Roadmap specifically notes that "monitoring proprietary or individual citizen data may raise privacy concerns" and recognizes that the assurance of data privacy is a critical condition in the development of public safety analytic capabilities. [1]

The public safety community's move to provide transparency through open data initiatives and the rise of advanced analytics warrants consideration for the processing procedures and techniques that de-identify data; and necessitate the use of tested, validated, high-speed algorithms that ensure the protection of both public safety personnel and the communities they serve.

De-identification which is also referred to as anonymization in Europe, is a set of approaches that strips personal information from a dataset. This term encompasses a broad and diverse range of techniques for mitigating the risk of linkage attacks and other misuses of datasets that contain personally identifiable information (PII). However, there is a utility vs. privacy tradeoff, in that a greater level of difficulty for carrying out a linkage attack will most likely imply a reduced utility for analysis and research purposes when it comes to de-identified datasets. Popular de-identification techniques, such as field suppression (and other field-specific perturbations) and guaranteeing *k*-anonymity, which preserve the privacy of the dataset, often must sacrifice too great a level of utility in order to prevent linkage attacks and other potentially damaging uses of the datasets. In addition, it is difficult or most often impossible to quantify the amount of privacy that is lost with these techniques.

A growing body of academic research in the field of differential privacy claims strict mathematical guarantees of data privacy, but with a potentially greater loss of dataset utility. Introduced by Cynthia Dwork in 2006, differential privacy (DP) is a mathematical theory and set of computational techniques that provide a method of de-identifying datasets—under the restriction of a quantifiable level of privacy loss. [2] Algorithms that satisfy the DP guarantee provide privacy protection that is robust against re-identification attacks, independent of an attacker's background knowledge. They use randomized mechanisms and provide a tunable trade-off between utility and privacy.

Informally, DP is a technique that serves to protects privacy no matter what third-party data is available; it strictly limits what is possible to learn about any one individual in the dataset. More formally, epsilon or ε-differential privacy considers the output probability distribution of a randomized data privatization process and bounds the amount that probability distribution can shift when one individual's data is added or removed.

To address the growing need for privacy techniques that can support high risk and demographic-rich data, like that found in the public safety sector, Applied Analytics Portfolio of the National Institute of Standards and Technology (NIST) Public Safety Communications Research (PSCR) Division partnered with the Information Technology Laboratory Information Access Division to establish a project to test, evaluate and strengthen research in DP. This effort lead to the creation of a series of prize challenges, or head-to-head competitions, and making the open source algorithms available for public safety use. This publication describes these efforts, focusing primarily on the design and results of PSCR's multi-phased innovation challenge which awarded a cumulative of $190K in cash prizes and makes recommendations for conducting future challenges in DP.

Beginning in 2017, PSCR's Open Innovation Office leveraged a contracted subject matter expert in DP, Knexus Research Corporation, and contracted challenge implementers, HeroX and Topcoder, through the National Aeronautics and Space Administration (NASA) Center of Excellence for Collaborative Innovation (CoECI) contract vehicle to aid in the design and to implement the 2018 Differential Privacy Challenges. A team of 10 experts drawn from academia, industry, and government were recruited to validate challenge design, review submissions for adherence to DP, and make recommendations to the PSCR judge panel to the Division Chief of PSCR who served as the NIST appointed judge and made final decisions on prize awards.

The challenge was split into two distinct phases. The first, a conceptual phase, titled 2017 Unlinkable Data Challenge, elicited new ideas on DP methods through a white paper contest and leveraged a two-fold approach for evaluation and awards: a manual technical review by experts, and a "People's Choice" award.

Prize	Team Name	Affiliation	Entry Title	Summary
Grand Prize $15,000 + People's Choice $5,000	**Georgia Tech Privacy Team**	Georgia Institute of Technology: Atlanta, GA	Private Synthetic Data Generation via GANs	Used a Differentially Private Generative Adversarial Network {DP-GAN} to generate private synthetic data for analysis tasks.
Runner Up $10,000 + People's Choice $5,000	**DPSyn**	Purdue University; West Lafayette, IN	DPSyn: Differentially Private Data Synthesizer	Our approach is to generate a synthetic dataset that approximates many randomly-chosen marginal distributions of the input dataset.
Honorable Mention $5,000	**WesTeam**	Westat; Atlanta, GA	>>>Advancing Models in Differential Privacy>>>	Generated a synthetic dataset that approximates many randomly-chosen marginal distributions of the input dataset.

Figure 1 – 2017 Unlinkable Data Challenge Concept Paper Winners

The second phase of the challenge, titled 2018 Differential Privacy Synthetic Data Challenge, took an empirical approach to aggressively advance concepts for generating differentially private synthetic data via a series of three coding competitions or matches. These competitions introduced increasingly complex metrics to score Docker container solutions at three levels of ε

via a leaderboard during a provision phase, and final submissions against withheld data at three or more levels of ε during a sequestered phase. The stochastic nature of DP was addressed by a manual validation of DP, one voluntary offered in the provisional phase in exchange for a score boost for the contestants; and one mandatory during the sequestered stage. Additional details on challenge design including marketing, data, metrics, and scoring are detailed in the document.

The five final winning solutions of the Synthetic Data Challenge fell into three basic categories of approaches:

- *Marginals* - determined the probability distribution of the variables contained in the ground truth data by using the marginal distribution of subsets of the random variables.
- *Probabilistic Graphical Models (PGM)* - constructed of interpretable models that use a graph structure to record patterns of variable correlations; these graphs are learned automatically from data and then manipulated by reasoning algorithms to generate synthetic data.
- *Generalized Adversarial Networks (GAN)* - utilized a generator and a discriminator which are trained under adversarial learning approach to estimate the potential distribution of original data samples and generate new synthetic data samples from that distribution.

Challenge results are expanded in the document and include descriptions on the effectiveness of outreach and the focus, participation, and scoring methodology for each match.

Match 1			Match 2			Match 3		
Prize	Team Name	Points	Prize	Team Name	Points	Prize	Team Name	Points
1st Place $10,000 + Progressive Prize $1,000	jonathanps (Marginal)	781,953	1st Place $15,000 + Progressive Prize $1,000	jonathanps (Marginal)	748,427	1st Place $25,000 + Progressive Prize $1,000	rmckenna (PGM)	902,307
2nd Place $7,000	ninghui (Marginal)	736,780	2nd Place $10,000 + Progressive Prize $1,000	ninghui (Marginal)	705,843	2nd Place $15,000 + Progressive Prize $1,000	ninghui (Marginal)	870,097
3rd Place $5,000 + Progressive Prize $1,000	rmckenna (PGM)	664,623	3rd Place $5,000	privbayes (PGM)	641,671	3rd Place $10,000 + Progressive Prize $1,000	privbayes (PGM)	823,513
4th Place $2,000	manisrivatava (GAN)	93,955	4th Place $3,000 + Progressive Prize $1,000	rmckenna (PGM)	639,887	4th Place $5,000 + Progressive Prize $1,000	gardn999 (Marginal)	768,802
5th Place $1,000	privbayes (PGM)	82,414	5th Place $2,000 + Progressive Prize $1,000	gardn999 (Marginal)	604,066	5th Place $3,000	manisrivatava (GAN)	541,494
Progressive Prize $1,000	brettbj							
Progressive Prize $1,001	eceva							

Figure 2 - Winners of the Synthetic Data Challenge by Match

Overall the strategy used for this challenge, which led contestants to evaluate, improve, and document their solutions from a conceptual phase through an increasingly difficult sequence of empirical matches, was well-suited to the problem and was a generally successful approach for moving solutions from theory to practice. Scores improved over the course of the challenge, and the top-ranked, final-winning solutions produced high-quality results in a difficult, real-world use case. Unanticipated technical outcomes included the higher performance of simpler marginal based approaches against more recent GAN models, high performance despite unspecified workload, and good performance at lower levels of ε.

The challenge garnered global interest and U.S.-led teams comprised of international partners. However, due to the limited maturity of DP technologies and few production implementations, challenge implementors overestimated the number of potential challenge participants. This challenge relied heavily on a limited number of NIST-externally recruited subject matter experts and revealed critical lessons learned in regard to recruitment, scheduling, and collaboration, as well as, data selection and test harness design. These are further detailed in the document.

The NIST PSCR Differential Privacy Synthetic Data challenge results convincingly established that DP theory can be applied with current technology, and that the high ranked final solutions provide very meaningful insight as to how it can be done. The set of successful techniques far exceeded the success anticipated by the DP academic community at the outset of the challenge. The challenge effectively garnered global participation and support from the small circle of researchers accelerating advancement in the field, as well as, expanding the acumen of DP to public safety data owners and technologists.

The challenge's winning solutions not only sparked interest in privacy and statistics circles which resulted in invited talks at conferences, but they also caught the attention of Fortune 500 companies looking to leverage demographic-rich data in an era of growing privacy restrictions. While further work and investment will be required in the areas of automated algorithm tuning, software and computer engineering, and user-interface development to create a commercial application that can be used to produce synthetic data, the PSCR challenge served as a necessary bridge over the wide development gap.

Future research by the competing teams, as well as new researchers, and collaborations with the public safety and commercial sector can continue to build and improve further on these solutions. New techniques may also use these challenge results as a benchmarking tool. DP technology solutions are improving rapidly and have promise to provide levels of privitization that would allow broader use of data by public safety, government, academia and industry.

Purpose

The purpose of this publication is to document the use of a prize challenge as a means for driving innovation in the developing field of differential privacy and to describe the process, considerations, outcomes, and lessons learned.

Abstract

The push for open data has made a multitude of datasets available, enabling researchers to analyze publicly available information using various statistical and machine learning methods in support of policy development. An area of increasing interest that is being made available is public safety data, which can include both sensitive information and personally identifiable information (PII). Release of sensitive data and PII can lead to individual and organizational harm. However, the removal of PII alone is an insufficient approach to preventing linkage attacks -- the process of combining unrelated data to identify individuals and entities. A growing body of academic research in the field of differential privacy exists which claims strict mathematical guarantees of data privacy, but with a potentially greater loss of dataset utility. In 2017 National Institute of Standards and Technology (NIST) Public Safety Communications Research (PSCR) Division initiated efforts to test, evaluate and strengthen research in differential privacy and add to its growing body of knowledge by making available open source algorithms for public safety use. This publication describes the efforts focusing primarily on the design and results of PSCR's multi-phased innovation prize challenge and makes recommendations for conducting future challenges in differential privacy.

Key words

Challenge; Innovation Challenge; Deidentification; Differential Privacy; Synthetic Data; Data Privacy; Privacy; Public Safety; Public Safety Communications Research.

Acknowledgements

The authors thank all contributors to this publication, particularly the challenge participants and their advisors, who made this challenge a success and provided feedback for this report. We also thank Dr. Sergey Pogodin and David Lee for their contributions to the document. Finally, we thank the staff at HeroX and Topcoder, challenge implementors and platform providers, who managed and supported the infrastructure and outreach for the challenge.

Subject Matter Expert Panel Members and Advisors: Christine Task, Claire Bowen, Joshua Snoke, Joseph Near, Changchang Lui, Jason Suagee, Kirk Wolter, Om Thakkar, Erin Kenneally, Daniel Kifer, Quentin Brumme, Steve Cohen, Ashwin Machanavajjhala

Federal Advisors: Dereck Orr, Division Chief, NIST Communications Technology Laboratory Public Safety Communications Division; Ellen Ryan, Deputy Division Chief and Open Innovation Office Portfolio Manager, NIST Communications Technology Laboratory Public Safety Communications Research Division; and John Garofolo, Applied Analytics Portfolio Manager, NIST Information Technology Laboratory

Technical Reviewers: Gary Howarth, NIST Communications Technology Laboratory; Kaitlin Boeckl, NIST Information Technology Laboratory

Table of Contents

Table of Figures

1.1 Introduction

A rapid proliferation of datasets has been made available to researchers and the public within the last 10 years. These datasets have been analyzed using various statistical and machine learning methods, resulting in many useful insights, which have in turn helped to shape public policy and impacted other large-scale decision-making processes. However, certain risks have been associated with the release of many of these datasets as they may contain potentially sensitive information about individuals. National Institute of Standards and Technology (NIST) Technical Note 1917 Public Safety Analytics (PSA) Research and Development (R&D) Roadmap specifically notes that "monitoring proprietary or individual citizen data may raise privacy concerns" and recognizes that the assurance of data privacy is a critical condition in the development of public safety analytic capabilities. [1]

Nonetheless, some public safety datasets, for example, the Dallas Police's RMS Incidents within the City of Dallas Open Data Repository contains 602,097 entries of up to 100 fields and includes the personally identifiable information (PII) of officers, victims, and suspects. [3] Datasets such as this contain valuable information with potentially important research implications, including location data collected from mobile devices, which can be used for contingency planning for disaster scenarios; travel data, which can be used to identify safety risks within the industry; hospital and medical record data, which can assist researchers in tracking contagious diseases, such as virus outbreaks, the epidemiology of drug abuse, and other health epidemics; and patterns of violence in local communities. However, in most cases within the public safety sector, privacy concerns surrounding PII limit both the use of this data and the ability to share this data freely.

The availability of this type of public data is expected to rise in the near future. Due to the sensitive nature of information contained in these types of datasets, steps can be taken to remove PII prior to the datasets being made publicly available to analysts and researchers. However, achieving privacy is not as simple as redacting identifiers. Simply removing PII from these datasets is an insufficient approach because contextual information, particularly when combined with external databases, can allow for re-identification. It is well-known that auxiliary datasets can be used in combination with records in a redacted dataset to identify an individual. The process of combining auxiliary information with released data to identify individuals and entities is known as a *linkage attack*. [2]

Examples of linkage attacks utilizing released public safety data can be traced in literature back as early as 2001 when Ochoa, et al. were able to positively identify a significant percentage (35%) of the homicide victims contained in the Illinois Criminal Justice Information Authority database by linking to the social security death index. [4]

Public safety's move to provide transparency through open data initiatives and the rise of advanced analytics warrants consideration for the processing procedures and techniques that de-identify data; and necessitate the use of tested, validated, high-speed algorithms that ensure the protection of both public safety personnel and the communities they serve.

1.1. De-Identification and Differential Privacy

De-identification, which is also referred to as anonymization in Europe, is an approach that strips personal information from a dataset. This term encompasses a broad and diverse range of techniques for mitigating the risk of linkage attacks and other misuses of datasets that contain PII. However, there is a utility vs. privacy tradeoff, in that a greater level of difficulty for carrying out a linkage attack will most likely imply a reduced utility for analysis and research purposes when it comes to de-identified datasets. Popular de-identification techniques, such as field suppression (and other field-specific perturbations) and guaranteeing *k*-anonymity, which preserve the privacy of the dataset, often must sacrifice too great a level of utility in order to limit linkage attacks and other potentially damaging uses of the datasets. In addition, it is difficult or most often impossible to quantify the amount of privacy that is lost with these techniques.

A growing body of academic research in the field of differential privacy claims strict mathematical guarantees of data privacy, but with a potentially greater loss of dataset utility. Introduced by Cynthia Dwork in 2006, differential privacy (DP) is a mathematical theory, and set of computational techniques, that provide a method of de-identifying datasets—under the restriction of a quantifiable level of privacy loss. [2] DP analysis also known as mechanisms ($\mathcal{M}$) provide privacy protection that's robust against re-identification attacks, independent of an attacker's background knowledge.

The process of publicly releasing datasets while guaranteeing privacy through DP consists of three parts: a generative model is built which captures the distribution of the original sensitive data, perturbation steps are applied at various points to ensure the model satisfies DP, and then the privatized model is used to synthesize a new dataset consisting of synthetic individuals. Because the synthetic data satisfies DP, the synthetic data provably contains no real individuals, and thus, individuals cannot be re-identified. Informally, differential privacy is satisfied if given two databases (D_1, D_2) which differ by the data of a single individual, synthetic data output (O) reveals no distinguishable information about the individual from either database.

The strength and probability of this privacy guarantee is controlled by tuning the privacy loss budget, or privacy parameter ε. Lower levels of ε provide more indistinguishable results, thereby increasing each individual's privacy. [2]

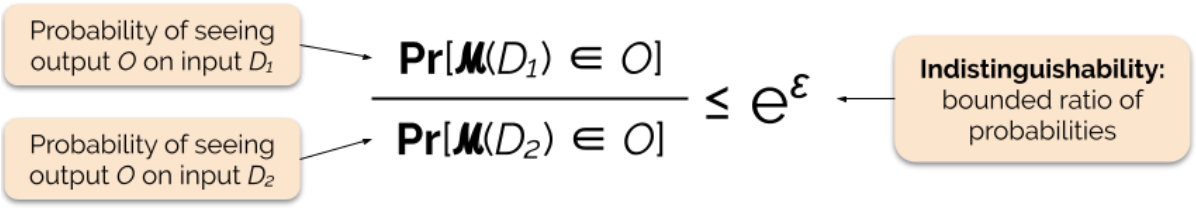


Figure 3 - Formal Definition of Differential Privacy [5]

This constraint ensures that re-identification attacks will not be feasible on the privatized results, which we can demonstrate with a contradiction argument. If a specific unique person is recognizable with certainty in the published results, then the probability is unbounded, and this violates the constraint of DP.

The bound A provides a formal measure of individual privacy—the larger A is the more distance that is permitted between probabilities $P1$ and $P2$, resulting in less overlap between possible realities and a weaker privacy guarantee. The definition of DP sets $A = e^{\epsilon}$ where the parameter ϵ is used to tune the privacy/utility trade-off or privacy budget, with small values of ϵ providing better privacy and requiring larger amounts of added noise. Very large values of ϵ weaken the constraint until it no longer ensures protection against re-identification; very small values of ϵ can require the probability distribution to be so wide (and the added random noise to be sufficiently large) that the published results no longer bear any resemblance to the true data and provide no utility for analysis.

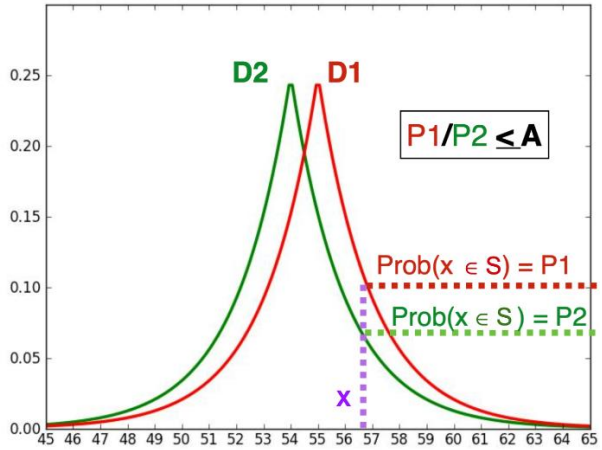


Figure 4 – Differential privacy definition illustration, showing overlap of output probability distributions from neighboring datasets D1, D2 and a specific sampled result S

Well-designed differentially private algorithms capture the desired information in the data while using techniques that are robust to small changes, producing results that naturally shift relatively little when single individual's records are added or removed, requiring relatively little randomization to privatize. This problem becomes more challenging as the complexity of the data, the size of the data space, and the number of features in the desired output increase. The problem of differentially private synthetic data, which requires retaining all information of potential interest in a possibly large and complex dataspace, however, has remained a notoriously difficult problem.

DP techniques may hold great promise in the field of data de-identification and provide a pathway for publicly releasing public safety datasets, but only if the utility of the de-identified datasets that they produce can be substantially improved.

1.2. Research Approach

Increasing demands for public safety data necessitate the ability to properly de-identify datasets with tested, validated, high-speed algorithms that ensure the protection of PII for both public safety personnel and the community. However, proving these techniques and refining the algorithms to the point where they can be applied in privacy-preserving data release pipelines requires accelerated innovation to make de-identification of privacy-sensitive datasets practical in time to meet the demand.

To address this need, in 2017 the Applied Analytics Portfolio of the National Institute of Standards and Technology (NIST) Public Safety Communications Research (PSCR) Division partnered with the Information Technology Laboratory Information Access Division to establish a project to test, evaluate and strengthen research in DP and add to its growing body of knowledge by exposing research by way of prize challenges, or head-to-head competitions, and making the open source algorithms available for public safety use.

Additionally, following the challenge, NIST worked with representatives from Los Alamos National Laboratory and RAND Corporation to evaluate the utility metrics used in the challenge for determining the accuracy of synthetic data, as well as other applicable metrics. Detailed information on the evaluation of techniques for differentially private synthetic data is covered by Bowen and Snoke. [6]

This publication describes the consideration and design of the NIST de-identification contest and the 2018 NIST Differential Privacy Synthetic Data Challenge, a two-phase innovation prize competition. It also identifies lessons learned and considerations for conducting future challenges in DP. The information is presented in the following manner:

- First, we outline the rationale for utilizing the prize challenge approach, and cover considerations taken into account in preparation for the challenge.
- Second, we describe our design, data preparation, and scoring methodologies for the conceptual and empirical contests and follow with the results of each respectively.
- Third, we provide details on participant approaches.
- Fourth, we highlight technical observations from the challenge.
- Fifth, we conclude with lessons learned, and future research directions in the public safety and DP space.

We acknowledge the distinction between registered participants and contestants, although we utilize the terms synonymously throughout the document. We note the difference in the results discussion as registered participants or registrants who completed online registration on the challenge platform but did not submit entries to the challenge contests; and contestants as those who provided submissions and actively participated in one or more phases of the challenge.

2. Prize Competition Approach and Challenge Management

The America COMPETES Reauthorization Act of 2010 (Pub. Law 111-358, title I, § 105(a), Jan. 4, 2011) authorizes the use of point solution, exposition, and participation prize competitions, or more commonly referred to as prize challenges, to rejuvenate investment focus on science, technology, engineering, and mathematics. [7] Since that time, prize challenges have offered many advantages for rapidly advancing innovation at NIST. A major benefit of prize challenges is the ability to expand the pool of problem solvers beyond the traditional candidates. Prize challenges generally extend the typical pool of stakeholders, or interested parties, to attract diverse talent from multiple disciplines, who come together to solve the problem. Other benefits include:

- The acceleration of the timeline; while traditional grants and contracting mechanisms may address the problem statement, prize challenges can be designed to reach a specific goal within a strict timeline.

- The number of participants is limitless and the outreach global; prize challenges encourage diverse groups to participate, resulting in many different solutions and more of them.
- Cost effectiveness and development of many solutions.
- Coalition and stimulation of the DP marketplace and focused research.
- Cultivation of a collaborative community interested in solving and progressing DP applications.

Over the last decade, NIST has explored the use of prize challenges to foster innovation and drive standards. For example, to encourage Internet of Things (IOT) stakeholders to collaborate with municipal leaders and develop smart cities, in 2014, NIST launched the Global City Teams Challenge (GCTC), an exposition prize challenge which fostered innovation acting as a matchmaker and incubator to form public-private partnerships creating opportunities for engagement and collaboration. The process and results informed foundation publications including IOT-Enabled Smart Cities Framework and Municipal Internet of Things Blueprint to assist city leader's decision-making. [8] Utilization of prize challenges within the NIST Information Technology Laboratory had been limited to the Intelligence Advanced Research Projects Activity (IARPA) sponsored Open Cross-Lingual Information Retrieval prize challenge which sought to develop "a cross-lingual information retrieval (CLIR) system to assist English speaking experts with data triaging." [9] However, the Public Safety Communications Research division of the Communications Technology Laboratory has taken a forward leaning approach, establishing a dedicated Open Innovation Office to drive rapid advancement for public safety communications and relevant supporting analytic needs.

PSCR's Open Innovation Office leveraged a contracted subject matter expert in DP, Knexus Research Corporation, and contracted challenge implementers, HeroX and TopCoder, through the National Aeronautics and Space Administration (NASA) Center of Excellence for Collaborative Innovation (CoECI) contract vehicle to aid in the design and to conduct the Differential Privacy Synthetic Data Challenge.

> "Ultimately, the ability of prizes to mobilize participants and capital, spread the burden of risk, and set a problem-solving agenda makes them a powerful instrument of change. They offer a valuable form of leverage to sponsors that use them as part of a well-designed strategy."
>
> Mckinsey & Company - 2009 [24]

A team of 10 experts drawn from academia, industry, and government were recruited to validate challenge design, review submissions for adherence to DP, and make recommendations to the PSCR judge panel for awards to PSCR. The Division Chief of PSCR served as the NIST appointed judge and made final decisions on prize awards.

2.1 Challenge Considerations and Assumptions

For the first national level contest in DP to be successful, we understood that the challenge would need to provide the necessary benchmarking tools, data, and metrics, along with the motivation, leaderboards, and incentive prizes to spur the research community to determine

whether effective practical solutions can be developed for the differentially private synthetic data problem. Considerations and assumptions that drove the challenge design included:

- *Concept maturity* - We understood that DP methods were nascent and experts in the field had limited access to outside training datasets or real world application platforms. The challenge would require a multi-phased approach, almost a bootcamp style, which would drive and motivate teams to repeatedly evaluate their algorithms on real world problems, use results to improve their algorithms, and then evaluate and improve again.

- *Data* – The data to be de-identified needed to be relevant to public safety needs. Per NIST Institutional Review Board guidelines, data could not contain PII or be obtained without consent. Sample data would need to be provided for development but must be independent of the final evaluation data. Therefore, the challenge required at least three datasets: 1) sample data for download and development, 2) provisional data for participant testing, and 3) sequestered data for final evaluation. Data should increase in difficulty as the phases of the contest progressed.

- *Benchmarking tools and metrics* - We recognized that not all algorithms are created equal. Sanitizing data using algorithms that satisfy DP will prevent re-identification, but poorly designed algorithms may add too much randomized "noise" to protect the data and become useless for analysis. Therefore, the contest would need to evaluate algorithm design as well as measurably determine the techniques that work well at preserving utility while protecting privacy. To ensure fairness and efficiency, and to promote the development of data-agnostic algorithms, multiple *non-redundant* metrics would be required for the challenge. Two specific concerns drove metrics selection:
    - ***Coverage*** - to address a breadth of use cases for the data
    - ***Discrimination*** – to distinguish between real and synthetic data

- *Motivation* – We understood that the contestants would consist of data science teams composed of students, academics, and industry professionals with already busy schedules. In order to ensure that the challenge was more exciting than stressful, and to help contestants stay actively engaged and striving to produce their best work throughout, it was important to have a well-designed set of incentives and milestones.

3. Challenge Design

In order to address the considerations and assumptions, the challenge was split into two distinct phases. The first, a conceptual phase, elicited new ideas on DP methods through a white paper. The second, an empirical phase, aggressively advanced the concepts into applied research via a coding competition. In this section, we discuss the challenge design for both phases, covering overarching issues applicable to both, then address details for each phase.

3.1 Team Eligibility

To be eligible for the cash prizes, each contestant or team of contestants were required to include an Official Representative who was age 18 or older at the time of entry and a U.S. citizen or permanent resident of the United States or its territories.

3.2 Milestones and Incentives

The conceptual phase contest known as the 2018 NIST "The Unlinkable Data Challenge: Advancing Methods in Differential Privacy" occurred over three months in the summer of 2018. This was followed by the empirical phase contest, known as the 2018 NIST "Differential Privacy Synthetic Data Challenge" which consisted of a sequence of three consecutive marathon matches that took place over eight months from October 2018 through May 2019. These milestones were designed to break down what may have seemed like an insurmountable problem into comprehensible and conquerable stages. Financial incentives were provided early and frequently to maintain momentum and participation.

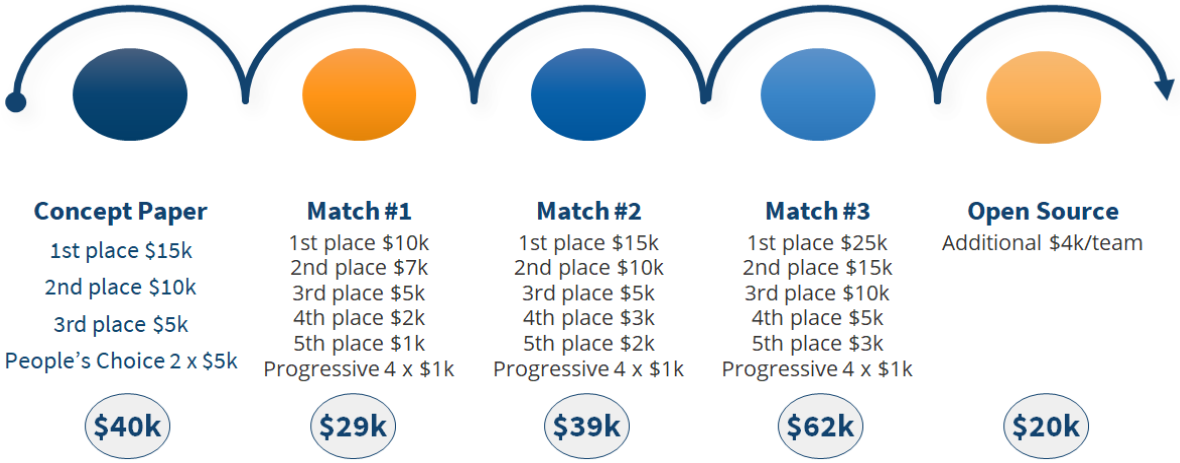


Figure 5 - Challenge Milestones and Incentives

3.3 Experts and Judges

NIST identified 10 DP subject matter experts to assist with the challenge. The selected experts were academic, government, and industry professionals with extensive backgrounds in statistics and mathematics, as well as experience in developing DP-based solutions. The subject matter experts participated in ad hoc design meetings and served as reviewers for contestant submissions. Judges were appointed by NIST and consisted of PSCR senior staff.

3.4 Outreach

Ongoing outreach, accessibility, and engagement throughout the challenge was critical to rapid advancement of the discipline within the allotted time frame of the project. We leveraged a variety of outreach approaches to recruit, educate, and provide feedback to participants.

3.4.1 Marketing Strategy

Official details about the challenge were posted to the Challenge.gov website, in addition to the NIST PSCR and specific challenge websites. A NIST press release[1] and NIST Tech Beat[2] articles were issued to increase visibility to DP while inviting interested parties to participate. Due to the complexity of the problem, the marketing strategy for the 2018 NIST Differential Privacy Synthetic Data Challenge specifically targeted solvers with experience in DP, in addition

[1] https://www.nist.gov/blogs/taking-measure/differential-privacy-qa-nists-mary-theofanos
[2] https://www.nist.gov/news-events/news/2018/05/help-keep-big-data-safe-entering-nists-unlinkable-data-challenge

to reaching out to the general data science challenge community. A two-prong approach for recruitment was applied:

- *Social Media* - HeroX developed and ran social media ad campaigns targeted to the DP community on Facebook and Twitter. Metrics from each ad set were continually tracked and revised throughout the duration of the campaign in order to maximize success.
- *Targeted Emails* - HeroX, NIST PSCR staff, and the challenge subject matter expert panel worked together to develop a targeted outreach list, focusing on experts in DP and academic leadership for relevant departments in prominent universities. As the marathon matches progressed, reminder emails were sent only to those that had opened a prior communication, and the frequency of emails decreased as interested users registered on the HeroX platform.

3.4.2 Registration

Registration was conducted through the HeroX platform for the conceptual phase and the TopCoder platform for the empirical phase. To provide schedule flexibility for participants, registration was intentionally fluid. For the conceptual phase, registration was open for three solid months from February 1, 2018 to contest launch on May 1, 2018. In the empirical phase, participants could join any of the three matches up to one week prior to the final scoring stages, to allow time to submit their entry to the public leaderboard and pre-screening. Additionally, contestants were not required to participate in all matches and could join, leave, and rejoin the challenge, as their schedules allowed.

3.4.3 Webinars

For the conceptual phase, and after the rules and guidelines were made publicly available, the Topcoder, HeroX, and NIST teams requested and collected questions from prospective contestants about the challenge. The NIST principal investigator responded to each question in a pre-recorded video which was made available on the HeroX competitor forum as a way to open the conversation amongst participants and clarify complex questions. For the empirical phase, the challenge team conducted an educational webinar for each marathon match. The goal of the webinar was to increase engagement and educate solvers who were not already familiar with the field of DP.

1. NIST DP #1 Webinar [10]
2. NIST DP #2 Webinar [11]
3. NIST DP #3 Webinar [12]

3.5 Conceptual Phase

The conceptual phase was designed to identify unique solvers in DP, expose the ideas that may have been evolving, but had yet to be documented, and encourage new community involvement. It also served to check current, relevant thoughts in the area of DP and set expectations for the challenge team on which techniques could be applied in the empirical phase. [13]

Conceptual Phase Goals

1. Describe public safety problem
2. Increase visibility on differential privacy
3. Invigorate data science community
4. Gather new techniques

Challenge rules, registration page, submission page, notifications, and an open online discussion forum were made available through both the HeroX and TopCoder platforms. The challenge page and accompanying marketing strategy invited individuals with an interest in DP to submit white papers proposing algorithms and solution features against the following problem statement. [14]

3.5.1 Problem Statement

Dedicated web pages for the challenge also offered a problem statement to spark contestants' interest. "The Unlinkable Data Challenge: Advancing Methods in Differential Privacy seeks a mechanism to enable the protection of PII while maintaining a dataset's utility for analysis."

3.5.2 Milestones

- Pre-Registration: February 1, 2018 – May 1, 2018
- Submission Period: May 1, 2018 – August 2, 2018
- HeroX Eligibility Screening: August 3, 2018 – August 6, 2018
- NIST Evaluation and Judging: August 7, 2018 – September 10, 2018
- People's Choice Award Voting: August 14, 2018 – August 28, 2018
- Winner Announcement: September 12, 2018

3.5.3 Judging Procedures

A two-fold approach was utilized for evaluation and awards: a manual technical review by experts, and a peer review, public choice award.

The manual technical review process consisted of three levels. First, submissions were reviewed by the HeroX internal panel to ensure eligibility requirements were met. Next, the NIST selected panel of subject matter experts (SMEs) evaluated submissions for adherence to DP theory and provided comments and finalist recommendations. Comments and recommendations were then passed to the NIST appointed official judge for final ranking and award.

Voting for the People's Choice Awards was held on the HeroX platform. The four finalist submissions were posted on the challenge webpage for a period of two weeks. Visitors to the page were able to review the finalist entries and register to vote for their favorite submission. Votes were tallied and exposed in real time. Additionally, registered voters were designated as followers of the challenge and sent updates on results and future events.

3.5.4 Results

In this section we summarize the results of the marketing strategy and participation in the conceptual phase and highlight the winning team submissions.

3.5.4.1 Response and Participation

The Unlinkable Data Challenge garnered wide attention drawing 32 registered teams and a total of 144 registered competitors from 104 countries. The challenge community reached 610 participants which included individuals who voted in the People's Choice effort. Page views indicated a broader interest in the contest reaching 103 countries with spikes in views occurring pre-contest launch and during the People's Choice voting period. The targeted outreach effort recruited the most registrants (48), including the Honorable Mention award winner and was followed by Facebook advertising (10) and Twitter (2).

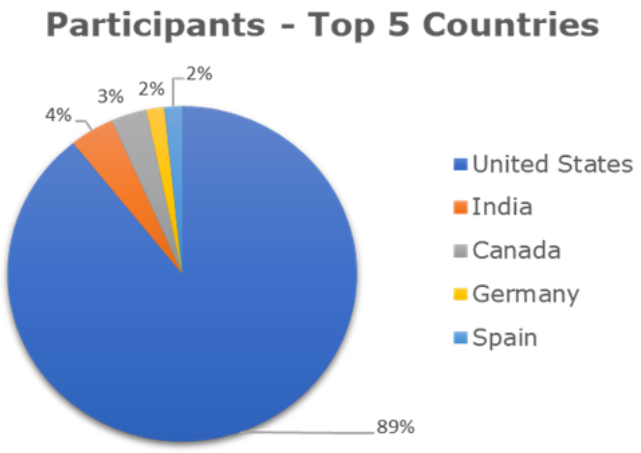


Figure 6 - Unlinkable Data Challenge Registered Competitors

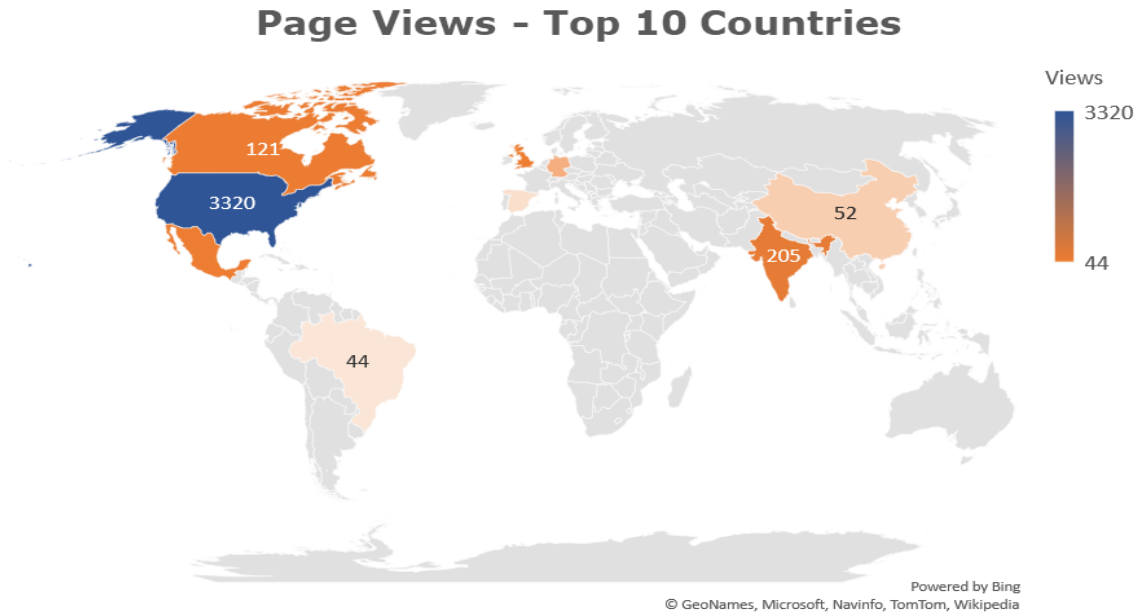


Figure 7 - Unlinkable Data Challenge Global Following

3.5.4.2 Awards

Less than half of the teams who registered submitted papers for the challenge. Only four of the 11 papers submitted for the conceptual contest met the minimum eligibility requirements for advancement. Eligibility requirements may have restricted registrations and submissions, however, the complexity of the problem and limited understanding of DP theory were deemed leading factors by the HeroX team.

Prize	Team Name	Affiliation	Entry Title	Summary
Grand Prize $15,000 + People's Choice $5,000	**Georgia Tech Privacy Team**	Georgia Institute of Technology: Atlanta, GA	Private Synthetic Data Generation via GANs	Used a Differentially Private Generative Adversarial Network {DP-GAN} to generate private synthetic data for analysis tasks.
Runner Up $10,000 + People's Choice $5,000	**DPSyn**	Purdue University; West Lafayette, IN	DPSyn: Differentially Private Data Synthesizer	Our approach is to generate a synthetic dataset that approximates many randomly-chosen marginal distributions of the input dataset.
Honorable Mention $5,000	**WesTeam**	Westat; Atlanta, GA	>>>Advancing Models in Differential Privacy>>>	Generated a synthetic dataset that approximates many randomly-chosen marginal distributions of the input dataset.

Figure 8 - Unlinkable Data Challenge Winners

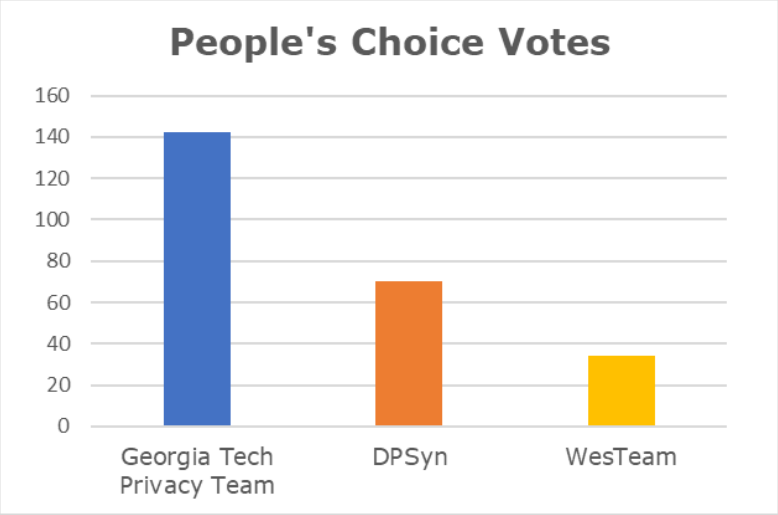


Figure 9 - Unlinkable Data Challenge People's Choice Award Winners

3.6 Empirical Phase

The empirical phase was designed to allow participants to apply and test the new techniques described in the conceptual phase, and for the NIST challenge team to develop and provide a means to evaluate and measure the techniques. Benchmarks and feedback were incorporated into the series of matches to rapidly drive iterative improvement of the techniques. The challenge approach was intended to provide the motivation and tools necessary for teams to steadily refine their solutions through repeated evaluation of their algorithms on real-world datasets.

3.6.1 Challenge Overview

The design of the empirical phase, or the 2018 NIST Differential Privacy Synthetic Data Challenge, consisted of three marathon matches, hosted on the TopCoder platform, lasting eight weeks each. Challenge participants were invited to develop an ε and δ DP algorithm at varying levels of ε for each match. The marathon matches were run as two-stage, head-to-head algorithm competitions. The initial five weeks consisted of a provisional stage, where contestants focused on submitting the synthetic data output by their algorithms for provisional scoring on a public leaderboard. This was followed by an invite-only, three weeklong sequestered stage, where teams submitted executable solutions, source code, and full documentation for review and final scoring.

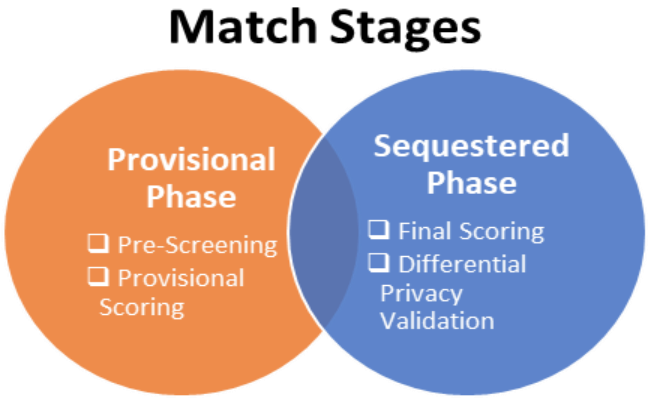


Figure 10 -Synthetic Data Challenge Match Stages

Each of the three matches began with a requirement for teams to submit correctly formatted synthetic datasets in order to earn a provisional leaderboard score. Scores could be boosted by submitting code and documentation for a pre-screening review by the SME panel mid-way through the provisional stage. At the end of the five-week provisional stage, all teams with a pre-screened score on the leaderboard were invited to the sequestered round of the match. The sequestered stage required more complete, stable solutions. Teams that advanced were required to submit code that accepted standardized δ and ε input with no hardcoded data schemas (schema given as input), and thorough code documentation aligned with the algorithm documentation. Each solution would then undergo a source code review by multiple DP SMEs, and their Docker containers would run on the Topcoder platform using the sequestered data to generate final scores. If the solution encountered problems in this process, the teams would be informed via a TopCoder forum post and allowed to fix and resubmit their code.

Standings

Handle	Score	Rank
rmckenna	934788.86	1
ninghui	930228.00	2
privbayes	839445.10	3
gardn999	805170.88	4
manisrivastava	652890.07	5
rachelcummings	407420.08	6

Figure 11 - Synthetic Data Challenge Leaderboard Snapshot from Match 3

In each match, teams received cash prize awards designed to encourage continued participation in the following match which would challenge teams with increasing difficulty. In the final match, the final five (5) winners had the option to receive a bonus cash prize award for posting their full source code to a publicly available website [15], [16]. Sharing of source code was intended to expand the knowledge base, spark collaborations, and accelerate development of production-level solutions that could be adopted by public safety.

Throughout the empirical phase, challenge rules, registration page, notifications, and an open online discussion forum was made available on the Topcoder platform, along with a downloadable competitor pack which included detailed information and instructions, data dictionaries, sample data, and scripts for registered participants.

3.6.1.1 Challenge Rules

In addition to NIST official rules posted on Challenge.gov, NIST also reserved the right to adjust provisional and final scoring methodologies during the contest, in a manner fair for all competitors. [17] This statement in the rules enabled adjustment for flaws in the original protocol or methodologies.

3.6.1.2 Milestones

- Pre-registration: October 1, 2018 – October 30, 2018
- Match 1
  - Development Period: October 31, 2018 – November 29, 2018
  - Progressive Prize Award: November 15, 2018
  - NIST Evaluation and Judging: November 30, 2018 – December 31, 2018
  - Awards: January 2, 2019
- Match 2
  - Development Period: January 11, 2019 – February 9, 2019
  - Progressive Prize Award: January 26, 2019
  - NIST Evaluation and Judging: February 10, 2019 – March 6, 2019
  - Awards: March 7, 2019
- Match 3
  - Development Period: March 10, 2019 April 23, 2019
  - Progressive Prize Award: April 8, 2019
  - NIST Evaluation and Judging: April 22, 2019 – May 20, 2019
  - Awards: May 23, 2019

3.6.2 Differential Privacy Definition Relaxed

For the empirical phase, a common relaxation of the DP definition was chosen, ϵ, δ -differential privacy. When δ is bounded to a small amount ($> 1/n^2$) a strong practical privacy guarantee can be retained, allowing the use of several techniques and improving accuracy by reducing the need for long-tailed noise distributions, which may sporadically result in large added-noise values. Given two neighboring datasets $D1, D2$ that differ in the data of a single individual, a data publication scheme satisfies ϵ, δ -differential privacy if its published result R satisfies the following constraint.

$$\Pr\left[R(D0) \in S\right] \leq e^{\varepsilon} \cdot \Pr\left[R(D1) \in S\right] + \delta$$

Adapted Definition of Epsilon Delta (ε/δ) Differential Privacy [18]

3.6.3 Data and Data Preparation

The challenge was designed with the objective of spurring fast-paced research and development of practical solutions to release a privatized synthetic dataset to the public. The challenge's data

collection/publication paradigm assumed a central data owner's collection of a complete dataset of tabular (event or survey) data and an essentially arbitrary amount of time and computing power to process the data for release. Data considerations identified in Section 2.1 and challenge preparation timelines drove data selection. Data was broken into two parts: A testing dataset that was considered to be publicly released data for the purposes of algorithm development, and a sequestered dataset that was considered sensitive**,** private data to be used for final scoring. Often**,** two datasets were used in final scoring**;** one whose distribution resembled the public testing data and one with a significantly different distribution, to check an algorithms**'** generalizability.

The first and second marathon matches leveraged open source data downloaded from the City and County of San Francisco's Open Data Portal and the third match utilized United States Census Bureau Public Use Microdata Sample (PUMS) of the 1940 U.S. Census Data, fetched from the IPUMS USA website[3]. PUMS data was selected to increase complexity while still providing public safety planning relevance for the third challenge. Additional details on data size, number of columns, and variables are included in the Fig.s in Section 3.6.6.1.

The zip file 'Competitor Pack' for each match included the provisional training datasets, provided in .csv format, along with their corresponding data dictionaries, as JSON files, which described all of the columns to be privatized, and additional details in a readme file. Provisional training datasets were accompanied by a training ground truth dataset. The U.S. Census PUMS data codebook was also provided to competitors for the third match.

- *Field types* for each column for all matches included:
    - `enum` - categorical data. The count field provided the number of possible values in each of such columns; and the possible values were from 0 (inclusive) to N (exclusive).
    - `integer` and `float` types denote columns with integer and float values. In both cases, the dictionary provided their minimum and maximum values (both inclusive); along with optional Boolean field, which told whether the value was optional (may be empty). For the columns with optional equal `false`, each record in the dataset required a numeric value; while for the columns with an optional field equal to `true` the record may have had either numeric value, or be empty.

- *Data Preparation for Matches 1 & 2:* For the sake of simplicity all original data values were converted to numeric formats as follows:
    - Categorical values (string literals) were replaced by consecutive integer numbers from 0 (inclusive) to N (exclusive), where N was the total number of possible values.
    - Date/time values were parsed and converted into integer Unix timestamps (number of seconds from 00:00:00 UTC, January 1, 1970).
    - Geographical coordinates were split into two separate columns containing real numbers for latitude and longitude.
    - Data columns were sorted so that sizes of value domains for each column increases from the first to the last column; i.e. the first and second columns contain categorical data with two possible values; the 3rd column contains

[3] https://usa.ipums.org/usa/

categorical data with five possible values; etc. Numerical (both integer and float) columns were placed along with the categorical data columns containing 100 possible values. Count values were provided for each column.

- *Data Preparation for Match 3:* The original dataset was converted from its fixed column width format to .csv for ease of use and to correspond to the data formats used in the first two matches. Not all of the columns mentioned in the US Census codebook were included in the dataset.
    - Leading zeros were removed from all codes in the dataset.
    - For numerical columns the values like '99998' corresponded to the 'N/A' value. The 'N/A' value in a certain column has as many '9' digits as necessary for the value to fill the full width of the original column.
    - For scoring purposes, the third match considered all columns categoric and unlike in the previous matches, the values of categorical columns were not restricted to continuous ranges from '0' to 'count-1', where 'count' values were given. For this data set, competitors were permitted to use the input data to determine the set of possible values for each columns.
    - The provided count values specified the total number of distinct values found in each column of the dataset; and 'maxval' specified the maximum value found in each column.

Final scoring occurred on a sequestered dataset in the same schema as the original data. Both the provisional and sequestered datasets were small partitions (selected by state or year) of the same large, publicly available dataset. The particular choice of subset that would be used for final evaluation was not disclosed. Sequestered datasets were tightly retained, and password protected throughout the challenge and restricted access was provided only to selected NIST, Knexus, and Topcoder staff directly involved in data development or execution of the final scoring stage of the matches.

3.6.4 Python Scripts

Python scripts and instructions for each match were also contained in the Competitor Pack to enhance understanding of DP, enable iterative development, and facilitate submission and scoring. These included:

- a sample naive implementation of a simple ε-differential data privacy algorithm.
- an auxiliary script for preparing the Topcoder submission for the competitor's DP algorithm. This script ran the algorithm on the specified number of columns for the match and generated a .csv file for each of the three levels of ε identified in the challenge rules, checked that the .csv files satisfied limit requirements, then packed output files with the specified name.
- stochastic ground truth generators for each scoring method used by the match. These produced randomized sets of scoring tests used for provisional scoring.
- a test scoring script which detected the number of columns and returned a score in the range 0 to 1,000,000 for each dataset at the generated level of ε, then averaged the three score results.

3.6.5 Measurement and Validation Techniques

In this section, we delve further into the scoring measurement and DP validation techniques applied during the empirical phase of the challenge.

3.6.5.1 Synthetic Data Quality Metrics

NIST utilized three bespoke metrics to measure accuracy of synthetic data generated by the contestant algorithms. A new scoring metric was introduced in each match and applied in addition to the previous metrics to progressively increase difficulty for each match. The first two were developed by Topcoder, in conjunction with the NIST technical lead, and the third was derived from suggestions and interviews with data users and subject matter experts. The Competitor Packs for each match included the code for generation of tests, and subsequent scoring of the synthetic datasets, along with instructions on how to use them for local scoring.

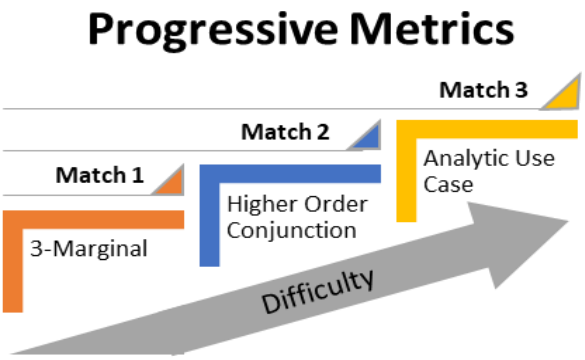


Figure 12 – New scoring metrics were introduced in each match

The scoring metrics used for this challenge were designed to provide both good coverage and good discriminative power, while being efficiently computable and generally applicable to any tabular data schema. Only the original dataset could achieve a perfect score against any metric. Each metric benchmarked competitor solutions against the original dataset.

3.6.5.2 *k*-Marginal

The *k*-Marginal evaluation metric is a randomized heuristic that measures similarity between two high dimensional datasets, by considering all correlations of k or fewer variables. The Synthetic Data challenge's first metric captured correlations that existed between three or fewer features.

How k-Marginals work:

1. Numerical features are grouped into range bins.
2. A set set of k-marginals, e.g. variables from the available columns in the dataset, are selected in accordance with the specified strategy (for example – choose marginals uniformly random at a sampling rate of 0.1).
3. Count and density for each bin is determined for the selected k-marginals from both the real and synthetic datasets.
4. For each selected k-marginal, the difference between the real and synthetic data densities are calucated and converted to an absolute value.
5. The sum of the absolute values of each bin provides the score. The total score is derived by averaging all test scores and then converted to a human readable score.

REAL

Bins	B	E	G	Counts	Density
1	b1	e1	g1	10	0.10
2	b1	e1	g2	11	0.11
3	b1	e2	g1	16	0.16
4	b1	e2	g2	15	0.15
5	b2	e1	g1	9	0.09
6	b2	e1	g2	8	0.08
7	b2	e2	g1	14	0.14
8	b2	e1	g2	17	0.17

VS.

SYNTHETIC

Bins	B	E	G	Counts	Density
1	b1	e1	g1	8	0.08
2	b1	e1	g2	10	0.10
3	b1	e2	g1	17	0.17
4	b1	e2	g2	18	0.18
5	b2	e1	g1	8	0.08
6	b2	e1	g2	10	0.10
7	b2	e2	g1	13	0.13
8	b2	e1	g2	16	0.16

SCORE

Bin Number	Real Density	Synthetic Density	Difference	Absolute Value
1	0.10	0.08	0.02	0.02
2	0.11	0.10	0.01	0.01
3	0.16	0.17	-0.01	0.01
4	0.15	0.18	-0.03	0.03
5	0.09	0.08	0.01	0.01
6	0.08	0.10	-0.02	0.02
7	0.14	0.13	0.01	0.01
8	0.17	0.16	0.01	0.01
SUM				0.12

Figure 13 - Overview of k-marginal scoring

3.6.5.2.1 3-Marginal Scoring Methodology

A single test worked on three marginals, picked randomly. The domain of possible values in the random columns was split into bins. For example, the algorithm selected two categorical columns with 50 and 200 possible values, and a third numeric column with integer values between *a* and *b*, and empty values. The scoring algorithm then divided the domain of the numerical column into 100 equal ranges, of size *(a – b)/100* each. As the column allowed empty values, it added "virtual range" for its empty values. After, it created 50 × 101 × 200 buckets, plus one special bin for records outside of the valid value range. Then for both the original and submitted datasets it counted the number of records falling into each bin. Resulting counts were divided by the total number of records in each dataset to get the density distribution of records. Then, the scoring algorithm calculated the absolute difference of density distributions for the original and submitted datasets by taking the sum of absolute differences of density values in corresponding pairs of buckets. Due to normalization of density distributions to 1.0, the resulting difference was a number *s*, belonging to the range between 0.0 (perfect match of density distributions) and 2.0 (density distributions for the original and synthetic dataset do not overlap at all). The resulting single test score was defined as:

Three Marginals Output from Step 1: *Actual and Synthetic Person Data Sources*

Gender (M/F)	Income (Number)	Attended University (T/F)	Actual Count	Synthetic Count
M	$0-33K	F		
F	$0-33K	F		
M	$0-33K	T		
F	$0-33K	T		
M	$34-66K	F		

Figure 14 - Illustration of a 3-marginal distribution

$$S = x\ 10^6\ \left(1 - \frac{s}{2}\right)$$

To calculate the score shown in the provisional leaderboard, we created a set of 100 tests, described above, with randomly picked columns for each test. The scores from separate tests were averaged; and if the submitter had been approved in the pre-screening procedure, the score was multiplied by 1000.

3.6.5.2.2 Higher Order Conjunctions (HOC)

For the higher order conjunction metric, target rows were randomly selected from the real data and synthetic data to create a pool of "similar" rows, and the relative size of the two pools were compared. This process used a randomly generated similarity function for each feature, and the solution was scored across the many, randomly selected target rows. The results were averaged to create the final score.

Figure 15 - Illustration of a Higher Order Conjunction

3.6.5.2.2.1 Higher Order Conjunction Scoring Methodology

A single test consisted of a set of rules for different columns. Each column had a 33% chance to be included into a set, thus, on average, a single test rule was ~11 columns. For categorical columns, the rule was a randomly selected subset of its possible values (from 1 to a maximum number of values); for numeric columns, the rule was a randomly selected range of values. A dataset record satisfied the set of test rules if all categorical columns included in the test, included values corresponding to the rule's subsets; and if all numeric columns included in the test had values within the selected ranges. Tests were generated to guarantee that in the original dataset there was at least a single record matching the test rules.

The *i*-th test calculated the fraction of records satisfying the test rules in the original $(f_{o,i})$ and in the synthetically privatized ($f_{p,i}$) datasets. Their mismatch was then quantified using the following formulas:

$$d_i = \ln(\max(f_{p,i};\ 10^{-6})) - \ln(f_{o,i})$$

$$\Delta = \sqrt{\frac{1}{N}\sum_{i=l}^{N}(d_i)^2},\ \text{where N = 300 is the total number of tests}$$

$$SCORE = \max(0,\ 10^3\left(1 + \frac{\Delta}{\ln(10^{-3})}\right)$$

Submissions that had been submitted for pre-screening and approved were multiplied by 1000.

3.6.5.2.3 Applied Analytic Use Case

The third measurement approach pushed beyond enabling comparison of synthetic data in columns or rows to provide a heuristic for validating data extracted from a dataset.

3.6.5.2.3.1 Applied Analytic Use Case Scoring Methodology

A single test for each level of ε of the applied analytic use case heuristic utilized two component scores derived from the SEX, INCWAGE, and CITY columns of the synthetic and original datasets.

- *Score 1* - The measure of income distribution across each city, or Gini Index[4], was calculated for the synthetically privatized dataset and original dataset. The mean-square deviation between the two datasets was calculated, then averaged by the number of cities present in the CITY column to provide the Score 1 result.

- *Score 2* - The gender pay gaps of the synthetic generated dataset and original dataset were ranked by city, and the calculation of the mean-square deviation of the two datasets was utilized for the second score.

The two score components were averaged to produce an overall score for each level of ε. The resulting three scores for the various levels of ε were then averaged together to provide the third measurement technique input for the Provisional Leaderboard Scores.

Sequestered scores for the Analytic Use Case heuristic were done with repeated trials and additional values of ε as needed, and the final score was computed as a privacy/accuracy AUC (Area Under Curve).

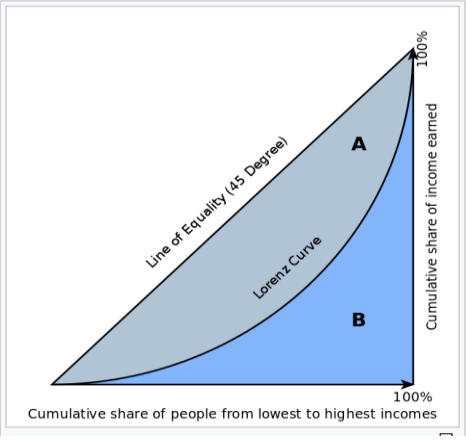


Figure 16 - Illustration of a Lorenz Curve used for determining Area Under Curve

3.6.5.3 Privacy Budgets

As indicated in Section 1.1, a very small privacy budget can require the probability distribution to be so wide (and the added random noise to be sufficiently large) that the published results no longer bear any resemblance to the true data and provide no utility for analysis. Balancing privacy and utility with reasonable error depends on the size and distribution of the data and number of properties to be estimated. For the first match, NIST set the three values of epsilon to (10.0, 1.0, 0.1). This included a generously large privacy budget of 10 which along with the large data size and smaller number of features, reduced the problem complexity in order to encourage contestants. However, the second match sought to distinguish between solutions and dramatically reduced the privacy budgets to better enable prize ranking (1.0, 0.1, 0.01). The third match aimed to move contestants and solutions towards currently applied practical levels of ε similar to those used in applications such as On the Map or being tested by the U.S. Census.[5] After consulting with additional advisors with federal and commercial experience, provisional testing ε levels for Match 3 were held at 8, 1.0, and 0.3, the same as for the Match 2 sequestered testing.

3.6.5.4 Test Harness

The Topcoder python test script powered the live leaderboard once contestant synthetic data was uploaded to Topcoder in a Docker container environment and ran against a small pre-generated set of tests, not known to the competitors.

[4] https://www.census.gov/topics/income-poverty/income-inequality/about/metrics/gini-index.html
[5] https://onthemap.ces.census.gov/

Topcoder chose to run final scoring on a standalone laptop. Most competitor solutions were reasonably fast even generating a single privatized dataset in approximately 30 minutes, using 1 processor core. GAN submissions, which required more resources and more runs, were also run in the Topcoder AWS cloud. Scoring of a test case set at each level of ε, with selected parameters, took between a few to ten minutes to run. Results were generated running four jobs in parallel overnight for approximately seven hours. The second match reduced the number of ε levels for the sequestered stage, decreasing the testing load. The Topcoder staff ran as many test cases as necessary to be sure that the score accuracy variance is smaller than the score difference between placements. Average scores and standard deviations were checked until the results were accurate enough for fair ranking of competitor solutions.

The batches of the resulting calculations and scores were accumulated into a table and passed to the SME panel for human review. Overall, the scoring process for each match took between seven and fourteen days.

3.6.5.5 Differential Privacy Verification

Two checkpoints in the matches were introduced to prevent intentional or accidental violations of DP. Violations, which could be variations of anonymization that were not DP or solutions that were hardcoded for the targeted datasets, could result in high scores and change leaderboard results during the provisional stage, thus incorrectly awarding prizes in the sequestered stage. A manual SME review approach was selected for validation of both its DP affordability and ease of implementation. In addition to averting violations, the verification process provided feedback to the teams to advance learning and development. General feedback about common problems was posted on the forum, enabling newcomers of DP to avoid obvious pitfalls. Specific detailed feedback, and requests for more information, were directly emailed to contestants to protect intellectual property.

3.6.5.5.1 Provisional Stage Pre-screening

To earn a 1000× score boost, contestants were required to submit clear, complete algorithm specifications and privacy proofs to pass a DP Pre-screen. The pre-screen was conducted by a minimum of three members of the SME panel during a weekly one-half hour-long SME review teleconference. The pre-screen served as a quick check to ensure that the contestant made a good faith effort to satisfy DP and to identify obvious errors.

3.6.5.5.2 Sequestered Stage DP Validation

SMEs confirmed that algorithms satisfied DP and that the code was an earnest best-effort implementation of that algorithm, without significant errors. Contestants invited to the sequestered stage were required to submit source code, code guide/documentation, updated algorithm specification, and privacy proof for a thorough final pass/fail DP Validation by the SME review panel. Two reviewers were assigned to each submission. Submission concerns were introduced by the assigned reviewers and discussed during ad hoc and weekly teleconferences amongst all panel members until a pass/fail decision was derived. Solutions failing validation were eliminated from prize eligibility for that match. Detailed feedback was provided by email. Contestants could fix or change strategies for the next match.

3.6.5.6 Judging Process

Similar to the conceptual phase, a manual technical review by experts proved critical to the final judging of each match. At the end of each match, the Topcoder and SME review panel verified the scores, source code, and documentation from the sequestered stage and provided their findings for review by the judge panel. The NIST team consolidated SME comments and finalist recommendations for the judge determination meeting. In this meeting, the judge panel was presented with the challenge data, contestant solutions and SME comments for evaluation. The NIST-appointed judge reviewed solutions for adherence to the challenge rules and made the final decision on final rankings and awards. This process was followed for each of the three matches.

3.6.6 Marathon Match Results

In this section, we describe the outcomes of the empirical phase in this challenge. We summarize the overall response; detail the focus, participation, scoring approach, and highlight issues faced by solutions during each marathon match and end with a brief description of each solution.

3.6.6.1 Response and Participation

The Differential Privacy Synthetic Data contest drew many participants and teams from the previous conceptual phase, and like the conceptual phase, participants' initial interest far exceeded the number of actual competitors.

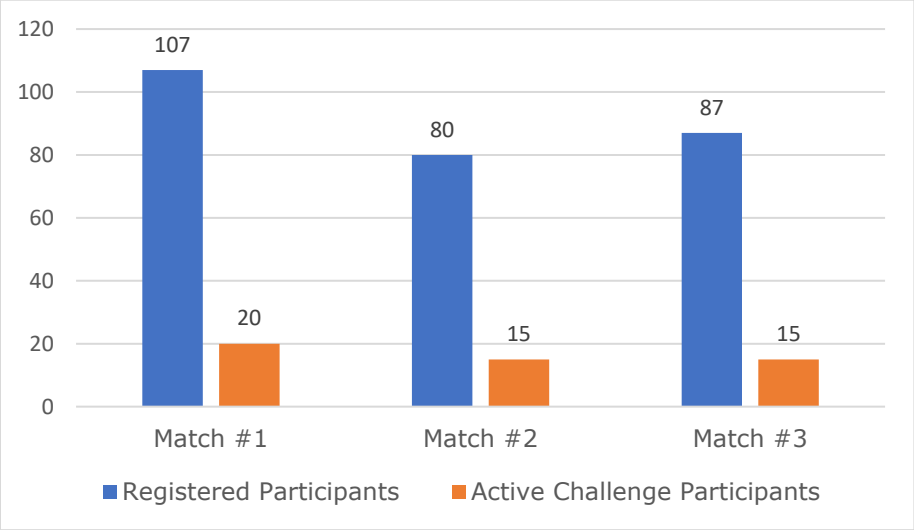


Figure 17 - Empirical Phase Participation

Tracking of contest submission numbers proved difficult in the empirical phase, skewing well over the number of active participants, since the challenge rules allowed contestants to upload datasets every four hours to test and spur further development. Sixty-one submissions were recorded for the first match, 93 for Match 2, and 61 for Match 3. HeroX and Topcoder's final report cited only the total number of data uploads. These numbers were attributed only to the number of data uploads in the HeroX and Topcoder's final report, preventing further analysis on the contestant development process.

The data from the challenge implementer report regarding global reach and interest in the 2018 NIST Differential Privacy Synthetic Data Challenge was limited to summary information on the nationality of contestants. The first match drew competitors from six countries and reduced to four in the second and third match.

Additional information on the effectiveness of the marketing strategy, specifically the number of hits and top sources of views of the webpages associated with the Differential Privacy Challenge may be found in Appendix A.

3.6.6.2 Match 1

Match 1		
Prize	**Team Name**	**Points**
1st Place $10,000 + Progressive Prize $1,000	jonathanps (Marginal)	781,953
2nd Place $7,000	ninghui (Marginal)	736,780
3rd Place $5,000 + Progressive Prize $1,000	rmckenna (Marginal)	664,623
4th Place $2,000	manisrivatava (GAN)	93,955
5th Place $1,000	privbayes (PGM)	82,414
Progressive Prize $1,000	brettbj	
Progressive Prize $1,000	eceva	

Figure 18 - Match 1 Results

Focus: The first match was designed to support research prototypes and then increase software requirements over the course of the final two matches. This match focused on conserving the clustering characteristics in the synthetic data.

Participation: Out of all the submissions, there were nine teams who passed the DP pre-screenings during the match. These nine teams achieved the highest provisional scores and were invited to participate in the sequestered stage. Only seven chose to submit code and documentation for the final round and, of those five, all passed the DP validation review and moved on for sequestered scoring.

Scoring: The sequestered scoring measured the synthetic dataset performance against columns in the 2017 San Francisco Fire dataset for 273 test cases, generated with a uniform chance of each column being selected. Sequestered scoring was repeated 100 times for the five different values of ε identified in Fig. 13 below. The average score of five separate runs produced the final sequestered score.

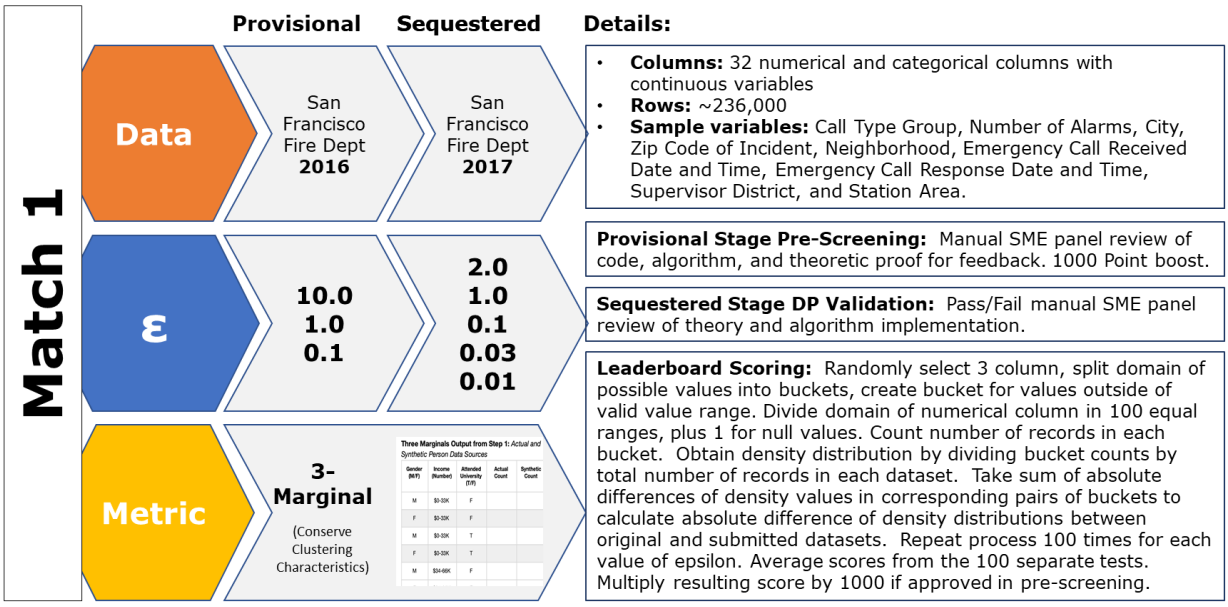


Figure 19 - Summary of Match 1 Methodology

The challenge implementer underestimated the learning curve for competitors and complexity of the challenge, initially setting the minimum prize eligibility score at 250,000 points, a target level which only three of five teams met. Topcoder's scoring official, NIST SME panel, and judges chose to throw out this requirement in the first match to maintain participation and maximize awards.

Solution Progression: In the first match, three solutions required multiple passes to generate each of the five required runs. The number of attempts varied for each solution, eliminating at least one team in the sequestered stage after 100 attempts, and failing to have their algorithm load. Two solutions in the sequestered stage required multiple passes ranging in number between three and 15 attempts for each level of ε. The first match also identified bugs in two solutions resulting in null or insufficient output for scoring.

3.6.6.3 Match 2

Match 2		
Prize	**Team Name**	**Points**
1st Place $15,000 + Progressive Prize $1,000	jonathanps (Marginal)	748,427
2nd Place $10,000 + Progressive Prize $1,000	ninghui (Marginal)	705,843
3rd Place $5,000	privbayes (PGM)	641,671
4th Place $3,000 + Progressive Prize $1,000	rmckenna (Marginal)	639,887
5th Place $2,000 + Progressive Prize	gardn999 (Marginal)	604,066

Figure 20 - Match 2 Results

Focus: The second match sought to capture a data distribution across multiple features of the dataset. The approach measured the synthetic dataset performance across rows of the original dataset.

Participation: In the second match, seven out of nine teams that submitted documentation for pre-screening received favorable feedback from the reviewers. All nine teams advanced to the sequestered stage but only six passed the DP validation review. The second match was the first time the validation process required further input from SMEs to reach consensus. Four of the six submissions passed after a full panel discussion. These six teams advanced for sequestered scoring and competed for prizes.

Scoring: Sequestered scoring was conducted against the two datasets at the three levels of ε identified in Fig. 14. The test set size was set at 300 for the 3-Marginal method, and at 1000 for the Higher Order Conjunction method.

Topcoder conducted three runs of each level of ε for each dataset and calculated their average scores, standard deviation, and percent standard deviation. A weighted mean average was applied with a double weight assigned to the $\varepsilon = 0.1$ condition which created an overall score for each dataset. Aggregated final scores were the average of the two overall scores.

Teams were allowed to process and submit subsets of the dataset consisting of the first $N < 34$ consecutive columns. Teams selecting this strategy received a zero score for any 3-Marginal test that relied upon columns outside of the submitted subset. Columns not present were assumed to

satisfy rule criteria for Higher Order Conjunctions (HOC). On average, contestants could not expect a favorable bias by eliminating columns because, if all columns beyond the ones in the submitted dataset were included in the test, then the test would include all of the records in the submitted data. This would skew the count and result in a poor score.

Minimum score requirements for prize eligibility were significantly reduced from the 250,000 requirement in the first match to a nominal 10,000 points for Match 2.

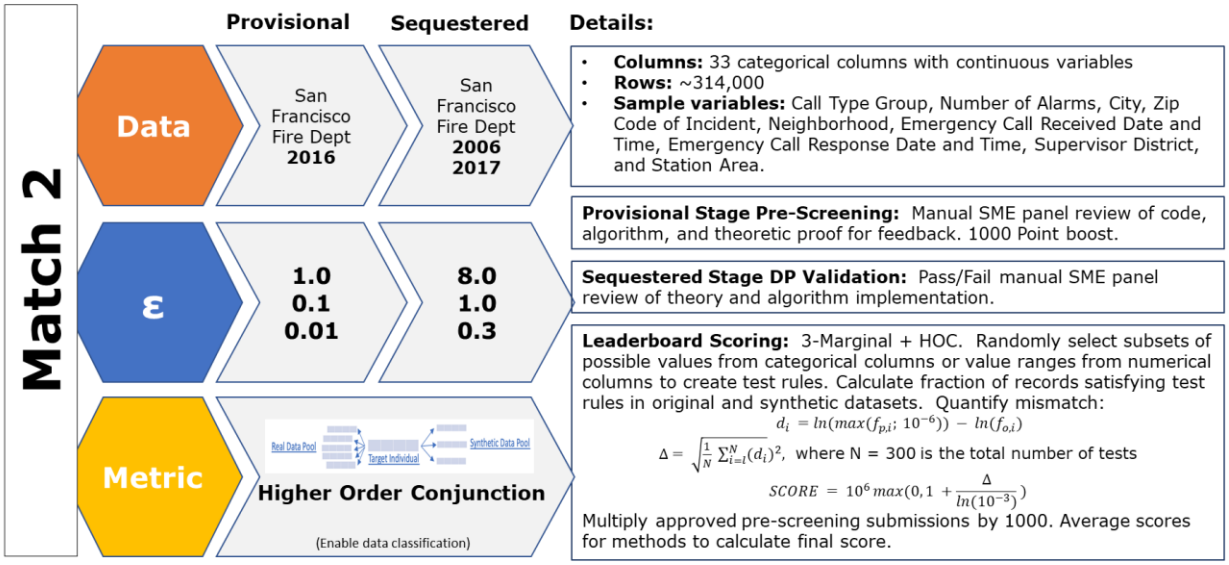


Figure 21 - Summary of Match 2 Methodology

Solution Progression: Solution improvements made during the second match reduced the scoring complications experienced in Match 1, eliminating bugs resulting in null or incomplete data output. Teams that experienced difficulties with messy code in the first match did not submit in the second match. The Match 2 methodology utilized two datasets, however, it highlighted new bugs that resulted in very slight variances in scores between years, values of ε, and runs; as well as; one solution's failure on all runs and all levels of the 2006 dataset. Those solutions that appeared to produce deterministic scoring results also failed DP validation upon review by the SME panel.

3.6.6.4 Match 3

Match 3		
Prize	**Team Name**	**Points**
1st Place $25,000 + Progressive Prize $1,000	rmckenna (Marginal)	902,307
2nd Place $15,000 + Progressive Prize $1,000	ninghui (Marginal)	870,097
3rd Place $10,000 + Progressive Prize $1,000	privbayes (PGM)	823,513
4th Place $5,000 + Progressive Prize $1,000	gardn999 (Marginal)	768,802
5th Place $3,000	manisrivatava (GAN)	541,494

Figure 22 - Match 3 Results

Focus: The methodology for the third match mimicked a sample economic use case, exploring pay and gender differences, to evaluate the ability of submitted solutions to conserve some amount of derived dataset characteristics. The sequestered 'area under curve' accuracy scoring was done using the same year in the U.S. Census dataset, but for two different states. This allowed for data from a different set of individuals and ensured that the algorithm refinement process, from the data made available during the provisional stage, did not violate DP guarantees.

Participation: Nine teams provided documentation for pre-screening, however, only six of them provided it by the deadline for the progressive prize awards. All were invited to the sequestered stage, but only seven of the nine teams submitted documentation and source code for DP validation, and underwent sequestered scoring.

Scoring: The match used a combination of three scoring methods, identified in Section 3.6.4.1. Each scoring method was run five times for each state at each level of ε noted in Fig. 16 below. Average scores of the three methods were calculated for each run, and an overall average score was calculated at each level of ε. Then, the final score was generated by averaging the three levels of ε.

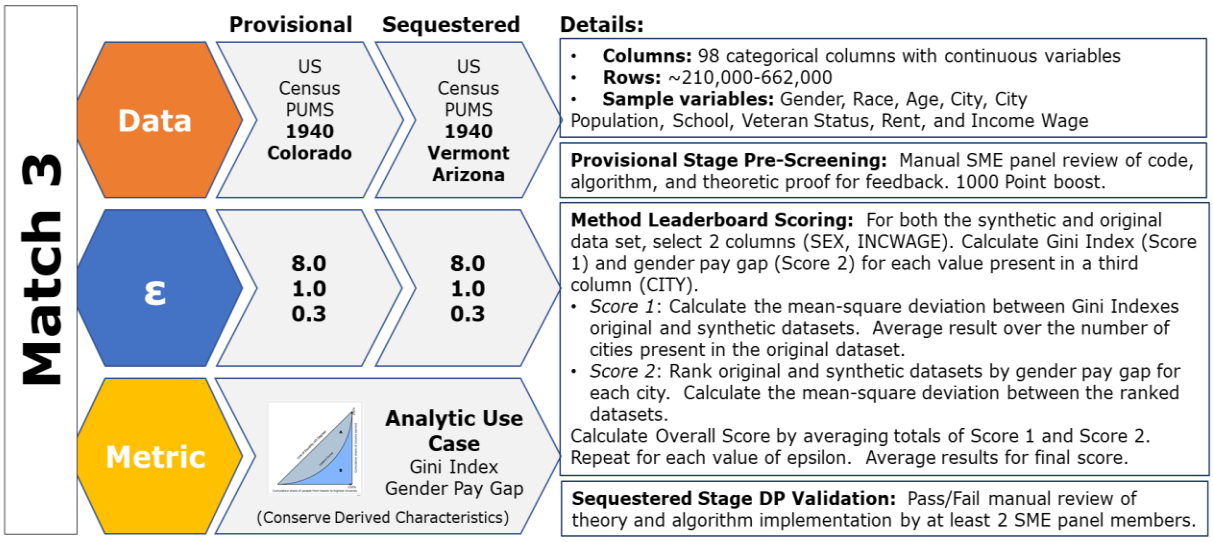


Figure 19 – Summary of Match 3 Methodology

Solution Progression: Software improvements naturally progressed throughout the matches as teams identified new requirements that could enhance performance. By the third match, the test harness revealed only one bug which had hardcoded the dataset size to small values. The associated team was informed of the error and was able to provide an updated solution in time for sequestered scoring.

3.6.7 Solutions

This section categorizes and very briefly describes promising solutions that performed well during the challenge. Insight on the approaches is drawn from both the contest scoring and the documentation submitted by the contestant teams.

The five final winning solutions fell into three basic categories of approaches: Marginals, Probabilistic Graphical Models, and Generalized Adversarial Networks. This category list is not exhaustive for the problem of differentially private synthetic data, and over the course of the competition, there was additional participation both within and outside of these categories; however, during this challenge, these were the approach categories that produced the best solutions. Alternative categorization approaches and more detailed descriptions on methodology are offered by Bowen and Snoke. [6]

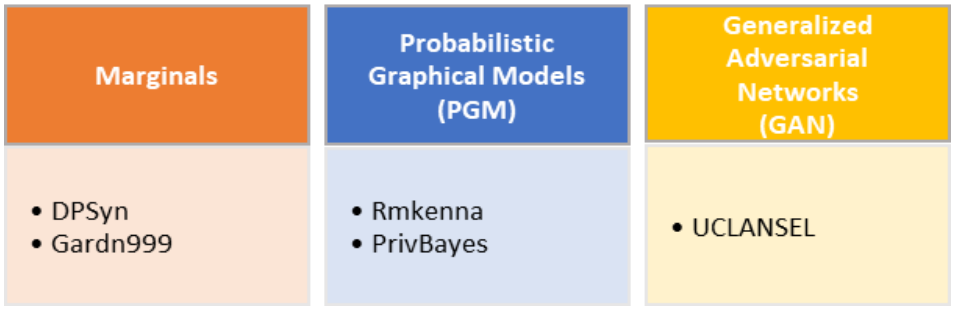


Figure 23 - Categorization of Winning Solutions

3.6.7.1 Marginal-Based Approaches

For the purposes of this analysis the term marginal-based approach is used to describe those solutions which determined the probability distribution of the variables contained in the ground truth data by using the marginal distribution of a subset of a collection of the random variables. These solutions give the probabilities of various values of the variables in the subset without reference to the values of the other variables.

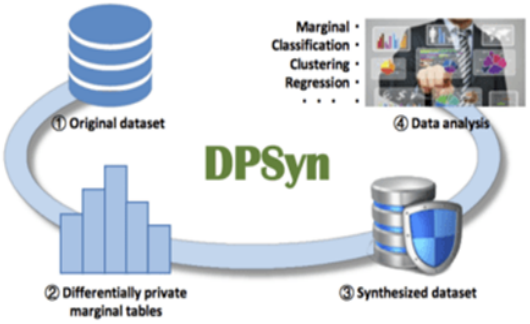
3.6.7.1.1 Ninghui - DPSyn

DPSyn, submitted by Ninghui Li (Purdue University), Zhikun Zhang (Zhejiang University), Tianhao Wang (Purdue University) built upon the team's previous work, outside of the NIST challenge, on PriView.

Ninghui - DPSyn

Solution Summary: Generate a synthetic dataset that approximates many randomly chosen marginal distributions of the input dataset.

Overview: The DPSyn synthetic dataset has similar low-degree marginals to the original dataset. Each marginal is specified by a subset of attributes and can be viewed as a projection from the full contingency table to those attributes.



PURDUE UNIVERSITY

DPSyn Team members: Ninghui Li, Zhikun Zhang, Tianhao Wang

Match Performance:

736,780 | 705,843 | 870,097

Details: Given a dataset as input, the algorithm has four steps:

1. Select the set of marginals to be published depending on factors such as, the size of the dataset, the total privacy budget, the relationships between the attributes, and any specific utility objectives. The description of the data schema and other available datasets of similar schema are used to support marginals selection.
2. Generate noisy marginals on the dataset. The algorithm computes these marginals and adds noises to them so that differential privacy is satisfied. Noises drawn from the Gaussian distribution and recent results on the composition of private algorithms are used to prove the privacy of the algorithm.
3. Utilize techniques developed in PriView to ensure that all noisy marginals are consistent with each other.
4. Generate a synthetic dataset, given these consistent marginals. This step requires a new algorithm, which starts with a randomly generated dataset and iteratively changes it to be consistent with the noisy marginals.

Figure 24 - Summary of DPSyn solution

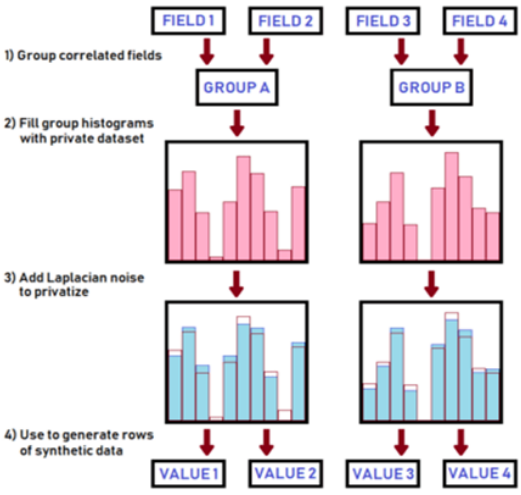
3.6.7.1.2 Gardn999 - DPField Groups

The Gardn999 team DPField Groups algorithm submission from John Gardner, an unaffiliated recruit from the Topcoder community, produced exact repeated scores for each run in its Match 2 debut. While this raised eyebrows amongst scoring officials, the algorithm passed DP Validation after full SME panel review and consensus.

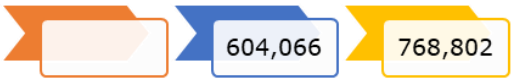
Gardn999 – DPField Groups

Solution Summary: Differentially Private Disjoint Marginals

Overview: a straightforward, marginal approach with minimal additional processing and is able to run very efficiently on datasets with a large number of fields with a wide range of possible values.



Match Performance:



Details: Team Gardn999 took the following steps to create their solution:

1. Group Correlated Fields: Public data similar to that being privatized is studied to identify highly correlated fields and to understand the field-value distributions. These fields are then identified in the data to be privatized, and highly correlated fields are grouped together.
2. Fill Group Histograms with Private Dataset: For each group, a marginal histogram is created for the purpose of counting the number of occurrences of every possible combination of values of all fields in the group. These bins are then filled with the data to be privatized.
3. Add Laplacian noise to privatize: For privatization, Laplacian noise is added to every bin with scale proportional to the number of groups. With highly correlated fields, it is expected that most of the bins will be empty before noise is added. To prevent a massive increase in the percentage of non-zero bins after noise is added, a threshold cut is made. The bin count after adding noise must be at least as large as the threshold or be set to zero. Although this adds a risk of removing some rare true occurrences, it greatly reduces the negative effect of the added noise.
4. Use to Generate Rows of Synthetic Data: Synthetic data is generated by selecting a random bin for each group with probability weighted by the noisy bin counts. Field values corresponding to each group's selected bin are written as a single row of synthetic data.

Figure 25 - Summary of Gardn999 approach

3.6.7.2 Probabilistic Graphical Models

Probabilistic Graphical Models (PGMs) rely on both statistical probability and computer science decision theory techniques and the dependencies between the different variables in a domain are considered using a graphical representation. The approach allows construction of interpretable models that can be learned automatically from data which are then manipulated by reasoning algorithms. [19] Bayesian networks are a common implementation of PGMs which depict the joint probability distribution from a directed acyclic graph in which random variables represent the nodes and probabilistic relationship between variables as edges, and a set of conditional probability densities for each variable. [20]

3.6.7.2.1 RMcKenna

The first-place winning solution submitted by Ryan McKenna from University of Massachusetts at Amherst, Employed an elegant mixed approach of marginal technique with PGM that was refined through the course of the three matches.

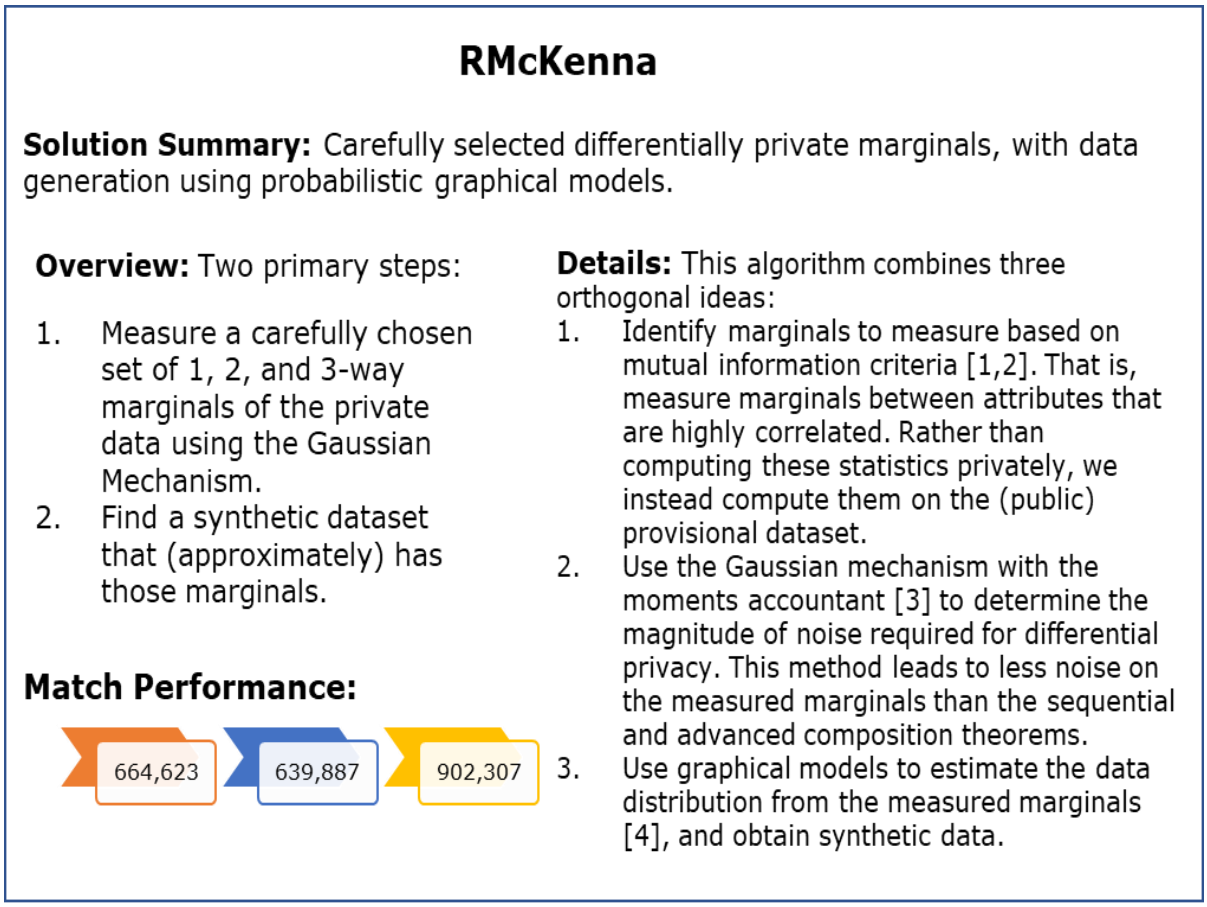


Figure 26 - Summary of Rmckenna Solution

3.6.7.2.2 PrivBayes

The PrivBayes solution created by Boling Ding, Xiaokui Xiao, Jun Zhao, Ergute Bao and Xuejun Zhao had the most dramatic score improvement over the challenge.

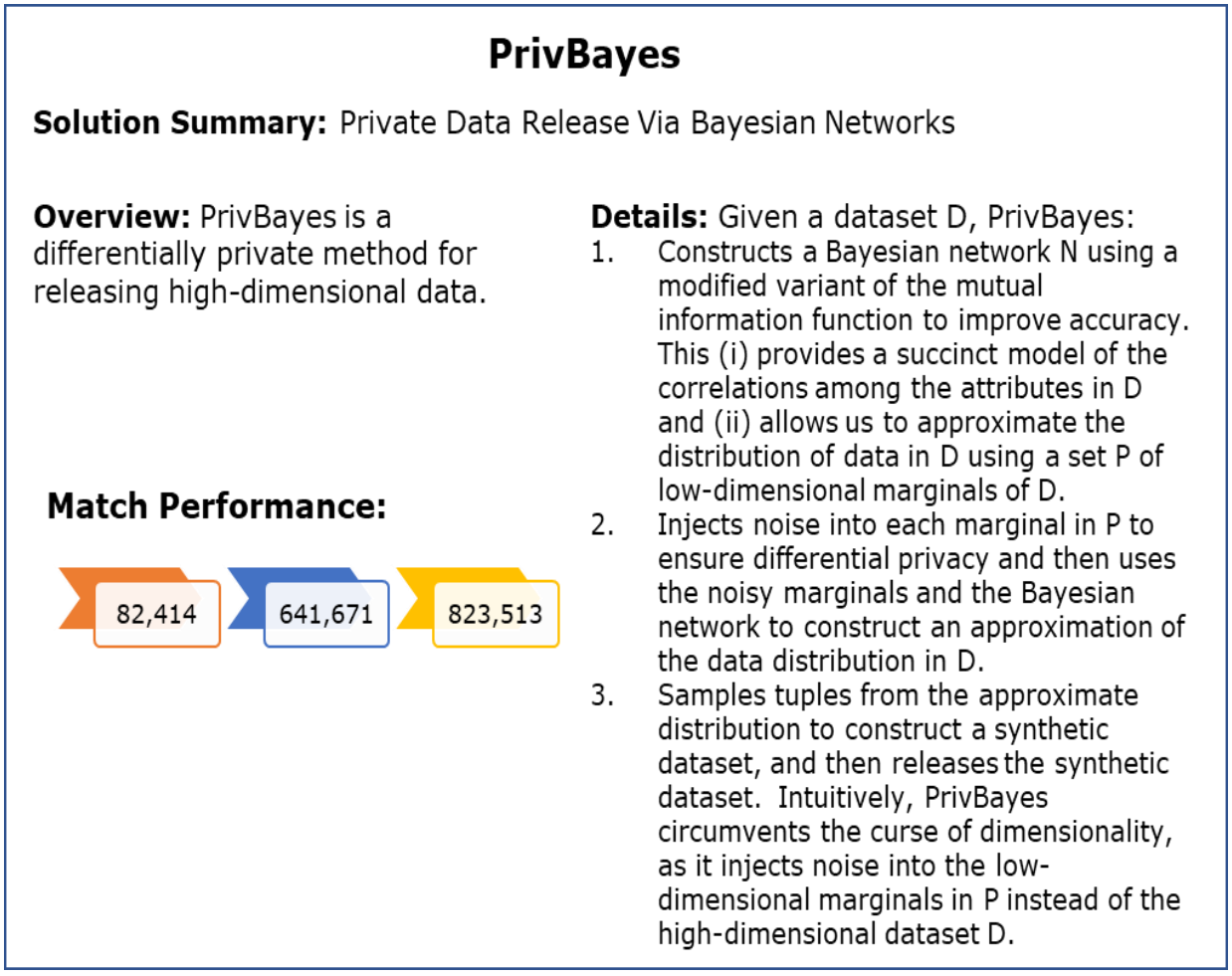


Figure 27 - Summary of PrivBayes approach

3.6.7.3 Generalized Adversarial Networks (GANs)

GANs utilize a generator and a discriminator which are trained under adversarial learning approach to estimate the potential distribution of original data samples and generate new synthetic data samples from that distribution. [21] GANs hold great potential for a variety of public safety use cases and have been applied with increasing success computer vision and image analysis problems. [22] However, applied to DP in this challenge, this novel approach suffered from the significant computational burden which resulted in clipping and algorithm termination as the privacy budget expended. [6]

3.6.7.3.1 Srivastava - UCLANESL

The UCLANESL team of Mani Srivastava (UCLA), Moustafa Alzantot (UCLA), Supriyo Chakraborty (IBM Research) and Nathaniel Snyder (UCLA) did not participate in the second match but made marked improvement in the third match after significant model training.

Manisrivastava - UCLANESL

Solution Summary: Differentially Private Synthetic Data Generation using Wasserstein GANs

Overview: Combines two state-of-art methods for synthetic data generation and accounting of differential privacy loss. Namely, the WGAN method for training GAN (generative adversarial networks) models and the moments accountant approach for calculating the bound of the privacy loss.

Match Performance:

Details: The approach is:

1. The provided real dataset is used to train a generative model (GAN) using the improved Wasserstein GAN training algorithm. After training, the generator model can be used to produce synthetic dataset by feeding noise samples from a prior distribution (e.g. normal distribution) into it.
2. In order to satisfy the differential privacy constrains, the training of the WAN models is made private by sanitizing the gradient before applying gradient descent steps. Sanitation of gradient occurs in two steps:
    1. The l^2 norm of the computed gradient is clipped to ensure that the gradient function has bounded l^2 sensitivity.
    2. Gaussian noise is added to the sum of clipped gradient.
3. Total privacy loss is tracked using the moments accountant method.

Figure 28 - Summary of UCLANESL approach

4. Technical Observations

Continually improving solutions requires participants to improve their understanding of the problem space, and detailed exploration of a problem space often produces new insights. In this section, we provide a short, informal list of unexpected technical observations drawn from the results of the challenge, which we believe may be worthy of more formal investigation in downstream research.

4.1 High Performance Despite Unspecified Workload

The synthetic data quality metrics used for scoring in the challenge were intentionally designed to have good coverage, as described in Section 2.1, rather than giving preference to any specific workload of data queries. The *k*-marginal and HOC scoring metrics are randomized heuristics, making it impossible for competitors to improve performance by biasing their algorithms to maximize accuracy on specific features or combinations of features. Although the data space was relatively large for both datasets (98 features in Match 3), and the scoring metrics forced

competitors to address the full space equally, they were nonetheless able to develop solutions that produced very high-quality results. It is possible that the properties of the underlying data may explain this; high scoring solutions took advantage of the strong correlations between features that often occur in human data. [23]

4.2 High Performance Despite Small Epsilon (ε)

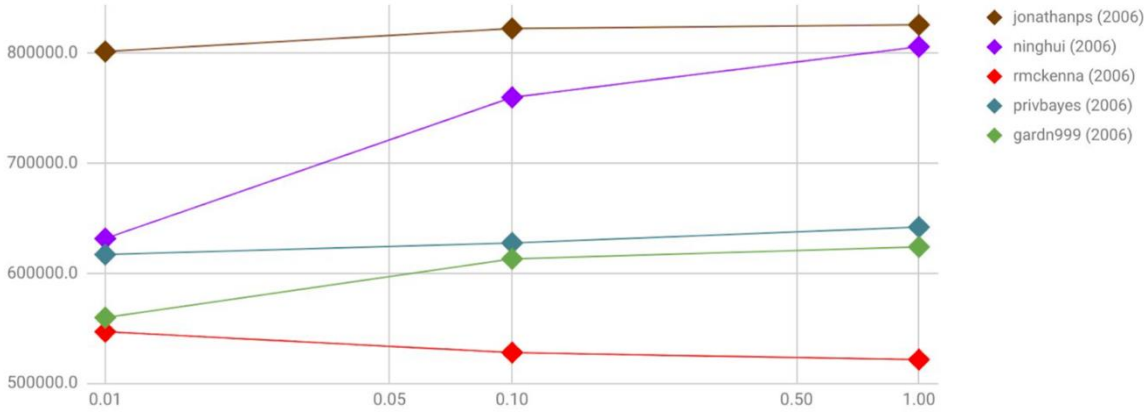


Figure 21 - Algorithm performance with respect to ε during Match 2.

The second match required competitors to use only very small values of ε: [1.0, 0.1, 0.01]. Ensuring a very strong privacy guarantee requires significant amounts of added noise, and small values of ε can cause differentially private algorithms to generate output that no longer resembles the original data. Smaller values of ε like those applied in Match 2 revealed a degradation in solution performance. However, as shown in Fig. 21, the degradation effect was smooth rather than catastrophic, and relative performance between solutions was generally maintained. Most competitors identified strategies to improve their solutions on the small ε values during the provisional part of the second match, and these solutions performed very well on the larger and more complex data space in the third match. Understanding and improving algorithm performance on small values of ε may have value for improving algorithm design overall.

Understanding and improving algorithm performance on small values of epsilon may have value for improving algorithm design overall.

4.3 High Performance of Marginal-Based Approaches

The challenge saw a variety of approaches applied to the problem of differentially private synthetic data. However, marginal-based approaches, which rely on histograms of counts across small sets of features, performed with notable success throughout the challenge. It is possible that these techniques are well-suited to the patterns of feature correlations that occur in tabular data, while neural network approaches are better suited to the patterns of feature correlations that occur in image or sound

Evaluation of differential privacy approaches against non-tabular types and representations of human-generated data may have value for improving performance and extending privacy protections.

data. A better understanding of the properties of specific types and representations of human-generated data might produce more significant insights for improving the performance of differentially private algorithms.

5. Lessons Learned and Considerations for Future DP Challenges

DP verification is an unavoidably complex part of any prize challenge in DP. The two-prong approach with pre-screening and final validation offered some benefits. It was not overly burdensome on SME schedules, it helped prevent clearly non-private solutions from interfering with the leaderboard rankings during the provisional stage of each match, and it was effective in catching a variety of subtle mistakes during final validation. However, it posed difficulties as well. In one case, a misleading function name led to a source code being misidentified as containing a DP violation. In other cases, violations were easily and correctly identified, but complex codebases made it difficult to confidently find and clearly explain *all* issues, resulting in delayed or incomplete feedback to contestants. Misconceptions about the definition of DP also led to a few tense interactions on the challenge forum; these required a combination of careful, clear explanations (provided by the NIST technical lead) and good forum moderation skills (provided by the Topcoder technical lead) to resolve.

However, the overall 'bootcamp' strategy used for this challenge, which led contestants to evaluate, improve, and document their solutions from a conceptual phase through an increasingly difficult sequence of empirical matches, was well-suited to the problem and was a generally successful approach for moving solutions from theory to practice. As shown in **Section 4**, scores improved over the course of the challenge, and the top-ranked, final winning solutions produced high-quality results in a difficult, real-world use case. The challenge design was also helpful for addressing skepticism, both within the privacy research community and the data user community, about the feasibility of the problem. By publicly putting the solutions to the test on real-world data and applications, and enabling sharing of the solutions through an open-source privacy forum after the challenge, progress was made towards answering many unanswered questions and points of concern about the problem of differentially private synthetic data. The very nature of the challenge problem, showing that synthetic datasets are practically solvable, provides valuable information to the DP research community that this problem is practically solvable.

In this section we revisit our initial considerations and assumptions, as well as, detail some observations that may serve future prize challenges or efforts in the DP field.

- *Concept maturity* – Although the concepts of DP date back to 2006, the practical implementations of them were few and these remained experimental at the time of the contest. Rather than developing and refining specific public safety customer-driven solutions to a problem, this challenge addressed a gap between basic and applied research to prove whether the problem could be solved with current technology. The amount of limited practical expertise in the subject required heavy reliance on DP SME expertise. For even a highly experienced, organized challenge implementation team the challenge was atypical.

  The challenge implementers underestimated the difficulty of the problem set and number of potential problem solvers in the DP field. Though interest was high, the contest registration and number of actual competitors fell below implementer goals. Typical data science

marathon matches average 400 solution submissions from 50 data scientists. [23] The conceptual phase proved beneficial in raising awareness in the field, but it also revealed a high-level of confusion between anonymization and DP techniques. We sought to overcome these through the addition of public webinars and outreach efforts to better define the problem set. Additional measures that could ease challenge recruitment and implementation include:

- In order to attract a good diversity and quality of candidate solution strategies, effective outreach is vital. Traditional social media outreach strategies can help generate public interest in the challenge and recruit participants from outside the field but may not effectively reach the DP research community.
    - Direct SME involvement in outreach for problems that are covered by a relatively small research community. SME assistance can be invaluable for assembling targeted email lists and drafting outreach emails that will be meaningful for other researchers in the field.
    - Attending or hosting workshops that allow challenge organizers to directly interact with members of the privacy research and data-user communities is helpful. Being available at an in-person event allows the organizers to ask and answer questions and provide more detailed explanation of the challenge.

- *Data* – Datasets for the second and third matches were selected and prepared during the course of preceding matches. This allowed limited time for identifying datasets with more robust features and continuous Internal Review Board updates.
    - To fully understand the contents of the data, even publicly available data from online repositories, it is necessary to download it and use appropriate software to review it. The data must be fully vetted, a schema (such as .json format) must be created such that contestants' solutions can be taken as input, and a human-readable data dictionary would need to be produced to explain in detail the variables in the dataset. If only a particular partition of the data is used (such as a particular state or a particular time period), the schema should reflect that constraint so that algorithms are not forced to privatize a much larger data space than the data actually occupies. These tasks are best performed with both technical and SME expertise.
    - It was useful to have at least two sequestered datasets during the final scoring stage and create the final score by averaging the results. We chose one dataset that resembled the provisional data and one that differed, in order to evaluate the ability of solutions to generalize across realistic differences in data distribution under the given schema. In the third match, we provided the Colorado Census data for the provisional stage and then completed final scoring using a state with a similar data distribution, Arizona, and one with a very different distribution, Vermont. Generalizability is an important solution quality, and it can be a good distinguisher between algorithms with otherwise similar performance.
    - Focusing on real-world applications over real-world data has a variety of advantages. It is generally valuable for benchmarking progress towards the development of practically usable solutions and, as such, the value of the challenge results is easy for the public to understand and evaluate. Working with data users to understand their needs, both in data use cases and quality metrics, can help significantly with this. In addition, specifically in the context of DP, real data may be a significantly more valuable test case than randomly

generated data because the patterns of correlations presented in human-generated data can result in lower entropy datasets that, once understood and leveraged, may produce significant performance benefits.

- *Benchmarking tools and metrics* – One of the most difficult issues in the challenge was the driving need for iterative development to advance the application of DP theory within the scheduled timeframe and resources. The need to verify formal differential privacy guarantees were satisfied required an approach which balanced a typical scoring evaluation with one to two manual DP validations.
    - On a high level the scoring metrics compared distribution of values between original and privatized datasets. Our metrics constructed random divisions of histograms in two or three dimensions, for data from a few randomly selected columns or rows from the test and synthetic datasets. The multiple dimensions took care of verification that cross-correlations between different columns and rows are conserved during the privatization. Having just two or three dimensions kept it computationally feasible, and well-defined for the size of datasets used in the contest. Since resulting average score are better defined as the number of generated tests and scores increase, we chose three phases for scoring for each match – contestant self-scoring during early development, provisional leaderboard scoring to refine solutions, and final scoring.
    - During the final validation code review, we still encountered points of uncertainty (such as misleading function names, or variable definitions that differed from the algorithm write-ups) that would have been much simpler to resolve with an explicitly interactive process that permitted direct requests for clarification from the contestants. In many academic science conferences, once editorial acceptance is approved the peer review process consists of three steps: initial reviews are given to the authors, the authors are given a short period to respond to reviewers' questions and comments, and then a final decision is made. Adopting a peer review process during validation would be helpful. Additionally, during the response period, contestants could be given the opportunity to fix small implementation bugs that could affect their privacy guarantee and then have their solution's accuracy score recomputed.
    - In addition to the SME review process, an automated black-box validation of DP guarantees would help as a second source of verification, to catch the mistakes that the SMEs might have either missed or wrongly identified. However, there are a few hurdles that need to be overcome for automated validation to be effective in this context. Research prototype code is often not robustly engineered; if there exists any input which causes the program to crash, that is, a technical violation of DP that can be identified by an automated validation process, it is generally not the type of flaw that a challenge is concerned with. If used, an automated validation would be necessary during final scoring, and the total computational time and resources would need to be viable within the scheduling constraints of the challenge.

- *Motivation* – Progressive prizes were an effective way to push the contestants to develop their first iteration of code quickly. With a deadline built into the schedule roughly two weeks after the start of the match, teams were forced to start early with the best version of their solution using their preferred DP technique. Progressive prizes were only awarded to

teams if the solution satisfied DP. It also gave the SME panel an early view of the solutions and prepared them for the final reviews.

- *Scheduling* – The pace of the empirical phase proved arduous for challenge staff, challenge implementers, and competitors. This was compounded by a 60+ day government shutdown in 2019. Uncertainty leading up the shutdown led to loss of a key staff member and also slowed award processing. Contracted staff were relied upon to maintain momentum throughout. A general recommendation would be to account for budget uncertainties as a schedule risk in the project plan and balance the staff mix as a mitigation strategy.
  - The approach of five weeks per match did not allow enough time for data development and algorithm development; and three weeks created a lot of pressure on the SME panel to comfortably evaluate both the inevitable complexities of final scoring and the design tasks necessary to launch the next match. For future challenges, extending the schedule for each match to three months may produce better results for everyone involved.
  - Due to the number of challenge participants from academia, the contest schedule should be aligned with the academic school year, including holidays, midterms and final exams. Launching the challenge in early September, when students and faculty are deciding on their obligations for the year, can make it easier for participants to integrate the challenge into their schedules. Collecting the final submissions for the last match in late spring, either before or after 'finals', ensures that students can complete the work on their challenge solutions before beginning their summer internships.

- *Test Harness Considerations* - Although the bespoke test algorithms and platform provided the insight necessary to accomplish the challenge, the following lessons learned, and alternative courses of action could improve future challenges or evaluations.
  - During the provisional stage, contestants only submitted their output datasets for scoring on the leaderboard. Final scoring required the various contestant source codes to be run from Topcoder's Docker containers and the laptop to confirm its performance. Running prototype code on any new system involves confronting a variety of hurdles, from crashes due to hardcoded file paths to solutions that assume more or differently configured computational resources than what is available. These issues can cause significant difficulties for scoring in an effective and efficient manner.
  - Consider developing the test harness in a containerized cloud environment and allow contestants the opportunity to run and test their code in the final judging environment before the final evaluation stage begins.
  - Specify the available configurations and computing power of the instance that will be used for final judging.
  - Set reasonable restrictions for a solution run time on the specified environment (scoring should complete within a given *k* hours).
  - Finally, DP, or other newly maturing technology challenges, can require careful adaptations to the traditional coding contest process, and a combined expertise is required to make decisions on everything from data vetting and cleaning to automated scoring code. Blended challenge teams, such as the one comprised of DP SME advisors and multiple contractors and government staff for this contest, need open, close, and continuing engagement to readily exchange ideas and develop evaluation techniques and systems while still ensuring effective oversight of contractors. Future efforts may

benefit from Agile development practices such as daily standup meetings, sprint cycles, and joint scrum meetings. Partnership in development and external SME support roles should be highlighted in all contracts and agreements.

- *Challenge Monitoring & Analysis* -
  - Variance between the HeroX and Topcoder's metrics led to dissimilar data, making it difficult to track global interest and participation across both phases of the challenge. Also, their final report cited only the total number of data uploads during the matches rather than the number of uploads per team, preventing further analysis on the contestant development process. We recommend additional forethought and planning on post challenge analysis measures including evaluation of participant feedback and impacts.

6. Open Questions

In this section, we describe a few of the remaining open questions and problems in privacy-preserving synthetic data that were not addressed within the scope of this challenge.

6.1. Future Research Directions

A challenge provides structure, benchmarking tools, and motivation to make rapid progress on addressing a defined problem, and periods of rapid advancement often leave many loose threads worthy of more in-depth investigation at a more deliberate pace after the sprint towards the final stage is completed. A few potential research directions directly related to this challenge that we hope will continue into the future:

- Continue to improve the synthetic data solutions that were used in the challenge.
- Evaluate other approaches to privatizing tabular (synthetic and event) data that were not represented in this challenge. These can be benchmarked against challenge solutions using code, data, and tools available on the NIST Privacy Engineering Collaboration Space.
- Analyze other approaches regarding data quality metrics for synthetic data, and further formal research on the metrics used during the challenge.
- Create a better formal understanding of the tuning process, including work on automated tuning over publicly available data and further exploration of the technical observations described in **Section 4**.

6.2. Unexplored Data Paradigms and Modalities

The scenario used for this challenge is common in public safety and government, but other use cases, paradigms, and data modalities pose privacy issues worthy of exploration. These are a few examples of data privatization, and use cases relevant to public safety communications applications, that have yet to be addressed:

- Time Series Sequence Data - a sequence of data points belonging to the same individual, rather than each individual contributing only to a single event or record.
- Data that is privatized and shared in real time, to support immediate analysis and response, rather than data that is published only after it is centrally collected and processed.

- Data that is privatized at the point of collection on an individual device or sensor, so that no central trusted data owner is needed (local DP).
- Measure DP solutions against additional richer data types such as image, video, sound, and unstructured data. Solutions that performed well in this challenge addressed tabular (event and survey) data, however solutions performance could vary on datasets with significantly different properties.

7. Conclusions

The NIST PSCR Differential Privacy Synthetic Data challenge results convincingly established that the DP theory can be successfully implemented with current technology, and that the high ranked final solutions provide very meaningful insight as to how it can be done. Both, the feasibility of the problem and the set of successful techniques, were far from expected results anticipated by the DP academic community at the outset of the challenge. The challenge effectively garnered global participation and support from the small circle of researchers accelerating advancement in the field, as well as, expanding the knowledge of DP techniques to public safety data owners and technologists.

The end results of any early tier TRL research and development effort like this are early prototype solutions rather than production-ready applications. Despite their early readiness levels, the challenge's winning solutions not only sparked interest in the privacy and statistics circles which resulted in invited talks at conferences, but they also caught the attention of Fortune 500 companies looking to leverage demographic rich data in an era of growing privacy restrictions. While further work and investment will required in the areas of automated algorithm tuning, software and computer engineering, and user-interface development to create a commercial application that can be used to produce synthetic data, the PSCR challenge served as a necessary bridge over the wide development gap.

Future research by the competing teams, as well as new researchers, and collaborations with the public safety and commercial sector can continue to build and improve further on these solutions. New techniques may also use these challenge results as a benchmarking tool. The challenge brings with it a new dawn where privacy and data access are not bianary. DP solutions can offer a measurable level of confidence for the privacy of individuals and vastly expand access to public safety, government, and other industrial sector information.

8. References

[1] Felts R, Leh M, McElvany, T (2017) *Public Safety Analytics (PSA) Research and Development (R&D) Roadmap*. (U.S. Department of Commerce, Washington, D.C.), NIST Technical Note TN 1917 Available at http://dx.doi.org/10.6028/NIST.TN.1917

[2] Dwork C, (2011) A Firm Foundation for Private Data Analysis. *Communications of the ACM*, 54(1): 86-95.

[3] Dallas Police Public Data (2020) *RMS Incidents* Available at https://www.dallasopendata.com/Public-Safety/Police-Incidents/qv6i-rri7.

[4] Ochoa S, Rasmussen J, Robson C, Salib M (2001) *Reidentification of Individuals in Chicago's Homicide Database: A Technical and Legal Study.* Available at https://www.researchgate.net/publication/2838440_Reidentification_of_Individuals_in_Chicago's_Homicide_Database_A_Technical_and_Legal_Study.

[5] Near J, Darius D, Boeckl K (2020) NIST, Cyber Security Insights, Differential Privacy for Privacy-Preserving Data Analysis: An Introduction to our Blog Series," Available:

[6] Bowen C, Snoke J (2020) *Comparative Study of Differentially Private Synthetic Data Algorithms From the NIST PSCR Differential Privacy Synthetic Data Challenge*. Available at https://arxiv.org/abs/1911.12704.

[7] United States Congress (2011) *PUBLIC LAW 111–358—JAN. 4, 2011*. Available at https://www.congress.gov/111/plaws/publ358/PLAW-111publ358.pdf.

[8] National Institute of Standards and Technology, (2019) *Global City Teams Challenge.* (Department of Commerce, Washington, D.C.) Available at https://www.nist.gov/el/cyber-physical-systems/smart-americaglobal-cities/global-city-teams-challenge.

[9] National Institute of Standards and Technology (2020) NIST Open Cross-Lingual Information Retrieval Prize Challenge. (Department of Commerce, Washington, D.C.) Available at https://openclir.nist.gov/.

[10] National Institute of Standards and Technology (2018) *NIST Differential Privacy Synthetic Data Challenge - 11/13/2018 Webinar.* Available at https://www.youtube.com/watch?v=6sU-NFTsR-I&feature=youtu.be.

[11] National Institute of Standards and Technology (2019) *NIST Differential Privacy Synthetic Data Challenge - 1/15/2019 Webinar.* Available at https://www.youtube.com/watch?v=f5ig0qBByFo&feature=youtu.be.

[12] National Institute of Standards and Technology (2019) *NIST Differential Privacy Synthetic Data Challenge - 3/14/2019 Webinar.* Available at https://www.youtube.com/watch?v=dxvyaZwYJeQ&feature=youtu.be.

[13] Department of Commerce - National Institute of Standards and Technology, (2018) *The Unlinkable Data Challenge - Advancing Methods in Differential Privacy.* Available at https://www.challenge.gov/challenge/the-unlinkable-data-challenge-advancing-methods-in-differential-privacy/.

[14] NIST PSCR (2018) *The Unlinkable Data Challenge: Advancing Methods in Differential Privacy*. Available at https://www.herox.com/UnlinkableDataChallenge/updates.

[15] Information Technology Laboratory/Applied Cybersecurity Division (2020) "Privacy Engineering Program De-Identification Tools," (Department of Commerce, Washington, D.C.) Available at https://github.com/usnistgov/PrivacyEngCollabSpace/tree/master/tools/de-identification.

[16] usnistgov, "PrivacyEngCollabSpace," GitHub, 6 June 2019. [Online]. Available: https://github.com/usnistgov/PrivacyEngCollabSpace/tree/master/tools/de-identification/Differential-Privacy-Synthetic-Data-Challenge-Algorithms.

[17] Department of Commerce - National Institute of Standards and Technolgy (2019) *Differential Privacy Synthetic Data Challenge.* Available at https://www.challenge.gov/challenge/differential-privacy-synthetic-data-challenge/.

[18] Meiser S (2018) *Approximate and Probabilistic Differential Privacy Definitions.* Available at https://eprint.iacr.org/2018/277.pdf.

[19] Koller D, Friedman N (2009) *Probablistic Graphical Models* (MIT Press, Cambridge, MA).

[20] Pearl, J (1988) Probablistic reasoning in intelligent systems. *Networks of Plausible Inference*, Palo Alto, CA.

[21] Wang K, Gou C, Duan Y, Lin Y, Zheng X, Wang F (2017) Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica,* 4(4) 588-598.

[22] Shah S, Mantini P, Stroup J, Weldon T (2019) Video Analytic based Alerting in Public Safety. *PSCR 2019*, Chicago, 2019.

[23] Zhang J, Cormode G, Procpiuc C, Srivastava D, Xiao X (2017) PrivBayes: Private data release via Bayesian networks. *ACM Transactions on Database Systems,* 42(4):41.

[24] Bays J, Goland T, Newsum J (2009) *Using prizes to spur innovation*. Available at https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/using-prizes-to-spur-innovation#.

Appendix A – Webpage Views

Appendix A provides a high-level overview of views of the webpages associated with the Differential Privacy Challenge. These are derived from the final HeroX report. Metrics are broken down into four sections:

1. Topcoder Minisite Page Views
2. NIST DP #1 Page Views
3. NIST DP #2 Page Views
4. NIST DP #3 Page Views

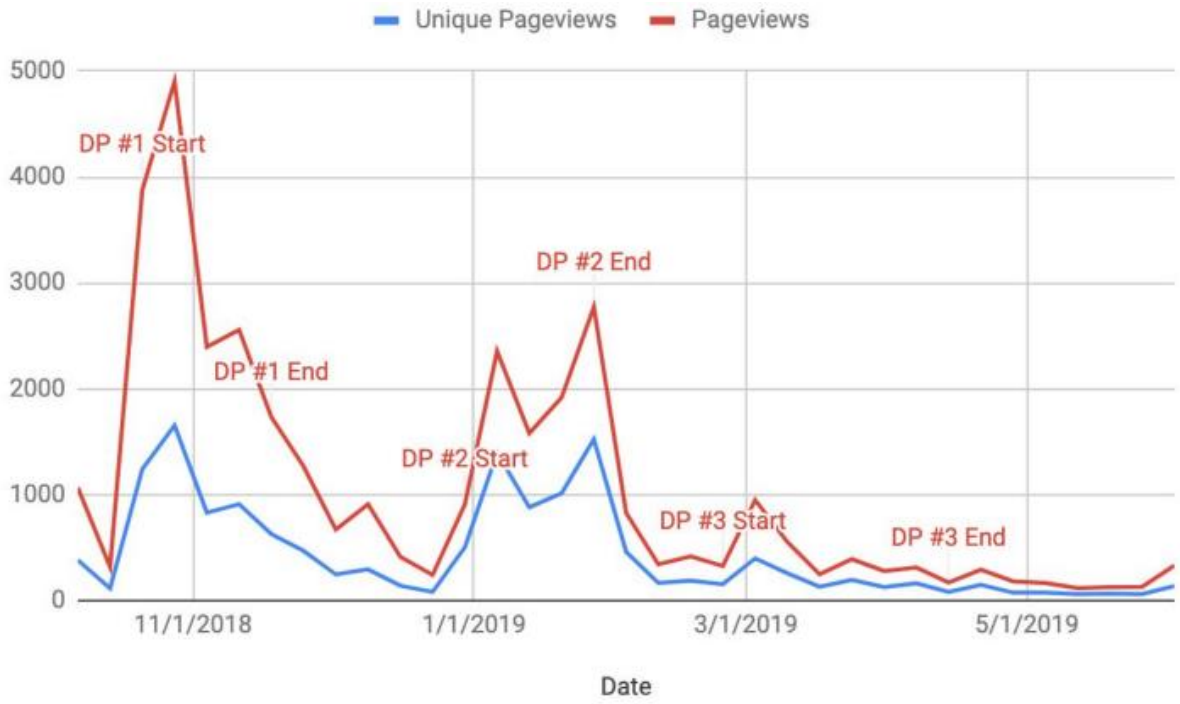


Figure 29 - Topcoder Minisite Page Views - Date Range: October 7, 2018 (pre-registration page launch) through June 1, 2019 (one week post-winner-announcement)

Table 1 - Top web traffic sources for NIST Challenge

Source	Pageviews	Users	New Users	Bounce Rate
Herox	18,429	4,136	3,908	16%
(direct)	4,105	704	621	8%
Google	2,635	554	326	8%
NIST	2,221	269	217	2%
Challenge.gov	1,698	259	160	20%
nist.gov	2,190	248	171	5%
community-app-main	1,051	186	-	0%
Topcoder Members	1,553	104	31	4%
mail.google.com	192	36	16	0%

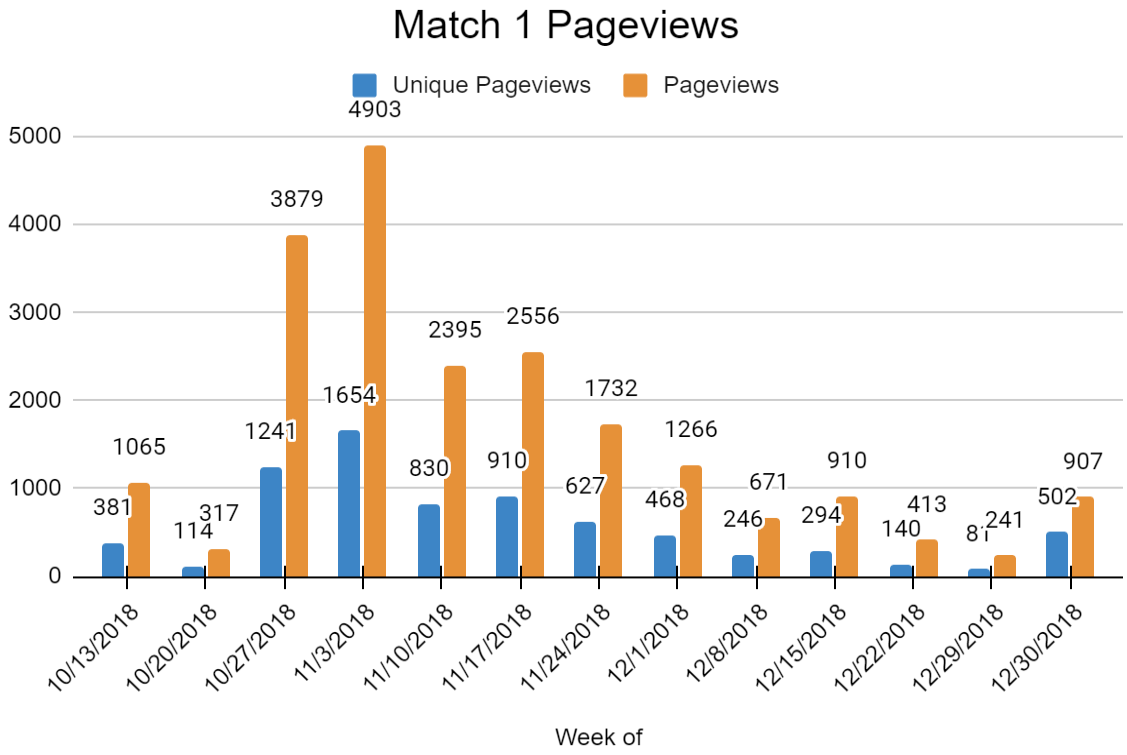


Figure 30 - Match 1 Page Views - Date Range: October 31, 2018 (Match 1 Start) through November 30, 2018 (Match 1 End)

Table 2 - Top sources of traffic for Match 1

Source	**Pageviews**	**Users**	**New Users**
(direct)	595	228	36
Google	466	202	-
Herox	544	135	-
community-app-main	305	114	-
Challenge.gov	311	62	-
wipro	31	26	-
nist.gov	62	26	-
m.facebook.com	47	21	10
NIST	269	21	-
t.co	21	21	5

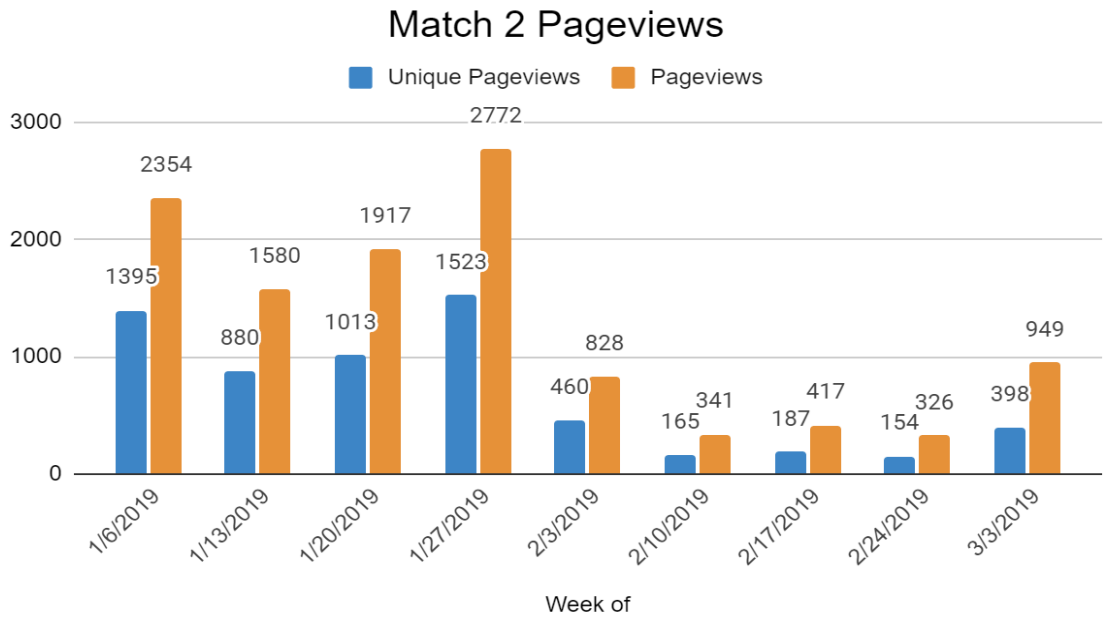


Figure 31 - Match 2 Page Views - Date Range: January 6, 2019 (Match 2 Start) through March 6, 2019 (Match 2 End)

Table 3 - Top sources of traffic for Match 2

Source	Pageviews	Users	New Users
(direct)	595	228	36
Google	466	202	-
Herox	544	135	-
community-app-main	305	114	-
Challenge.gov	311	62	-
wipro	31	26	-
nist.gov	62	26	-
m.facebook.com	47	21	10
nist.gov	269	21	-
t.co	21	21	5

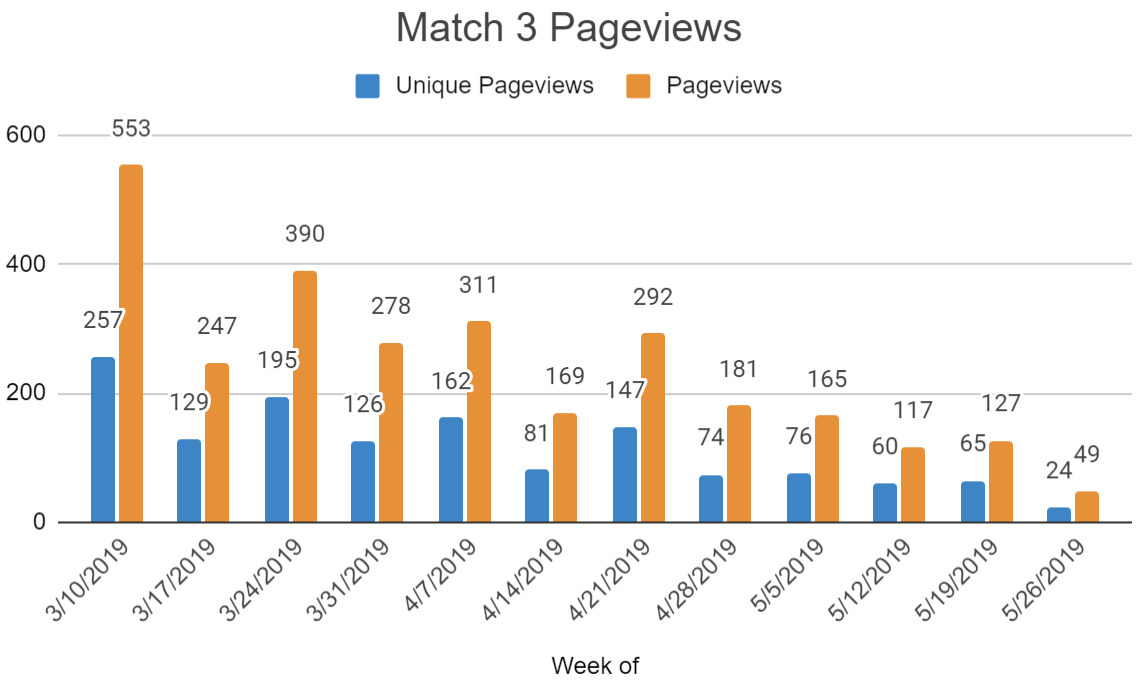


Figure 32 - Match 3 Page Views - Date Range: March 10, 2019 (Match 3 Start) through May 20, 2019 (Match 2 End)

Table 4 - Top sources of traffic for Match 3

Source	Pageviews	Users	New Users
Google	21	21	0
nist.gov	16	10	0
community-app-main	16	5	0
Topcoder Members	57	5	0
Challenge.gov	5	5	0
(direct)	5	5	0