NISTIR 8325

Media Forensics Challenge Image Provenance Evaluation and State-of-the-Art Analysis on Large-Scale Benchmark Datasets

Xiongnan Jin
Yooyoung Lee
Jonathan Fiscus
Haiying Guan
Amy N. Yates
Andrew Delgado
Daniel F. Zhou

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8325

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

NISTIR 8325

Media Forensics Challenge Image Provenance Evaluation and State-of-the-Art Analysis on Large-Scale Benchmark Datasets

Xiongnan Jin
Yooyoung Lee
Jonathan Fiscus
Haiying Guan
Amy N. Yates
Andrew Delgado
Daniel F. Zhou
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8325

October 2020

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

National Institute of Standards and Technology Interagency or Internal Report 8325
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8325, 30 pages (October 2020)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8325

Abstract

With the development of storage, transmission, editing, and sharing tools, digital forgery images are propagating rapidly. The need for image provenance analysis has never been more timely. Typical applications are content tracking, copyright enforcement, and forensics reasoning. However, large-scale image provenance datasets, which contain diverse manipulation history graphs with various manipulation operations and rich metadata, are still needed to facilitate the research. It is one of the major factors that hinders the development of techniques for image provenance analysis. To address this issue, we introduce large-scale benchmark datasets for provenance analysis, namely Media Forensics Challenge-Provenance (MFC-Prov) datasets. Two provenance tasks are designed along with evaluation metrics. Furthermore, extensive analysis is conducted for system performance in terms of accuracy on our datasets.

Key words

Benchmark Dataset; Image Provenance Analysis; Media Forensics Challenge; MFC-Prov; Provenance Evaluation.

Disclaimer

Certain commercial equipment, instruments, software, or materials are identified in this article in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor is it intended to imply that the equipment, instruments, software or materials are necessarily the best available for the purpose.

The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Acknowledgement

The authors gratefully acknowledge the members of the DARPA Media Forensics (MediFor) Program and members of the Air Force Research Lab for managing the program and weekly data team meetings; special thanks go to Matthew Turek, Neil Johnson, David Doermann, and Rajiv Jain for their instructions and strong support. PAR Government conducted this work under DARPA sponsorship via Air Force Research Laboratory (AFRL) contract FA8750-16-C-0168. NIST conducted this work under NIST Interagency Agreement Number 1505-774-08-000.

Table of Contents

List of Tables

List of Figures

1. Introduction

With the proliferation of image data and the development of storage, transmission, editing, and sharing tools, the volume and diversity of forged images are increasing rapidly. Even non-experts can create, modify, and distribute manipulated images with minimal efforts using these tools. Also, new techniques for manipulating images are being developed continuously, e.g., Generative Adversarial Networks (GANs, "deepfakes") [1] and Computer Generated Imagery (CGI) [2]. Our trust in digital images is being enormously challenged.

Detection of manipulated content in digital images has been actively studied in the literature. The related approaches have reached a stage of maturity, which means they can sufficiently detect whether a manipulation operation is applied to an image in general [3, 4]. A logical next step is to tackle a more complicated and practical problem, which is called image provenance analysis (IPA) [5]. In this paper, IPA is defined to retrieve a set of related images for a query image (i.e., target image), as well as to construct their relationships (i.e., the sequence of manipulation operations) among retrieved images. Manipulation history graphs are used to represent such relationships. A manipulation history graph is a directed graph whose nodes (i.e., vertices) are original or manipulated images and links (i.e., edges) are manipulation operations, e.g., additive noise.

IPA is very important to image processing and computer vision [5]. With the increasingly frequent occurrence of image compositions on the Internet and social media, the need for IPA has never been more timely. IPA is extremely useful in applications such as content tracking, copyright enforcement, and forensics reasoning about the possible intent of the manipulation [6, 7]. The manipulation intent can be inferred based on the history of manipulations and paths of dissemination among sites and players involved in the manipulation graph [8].

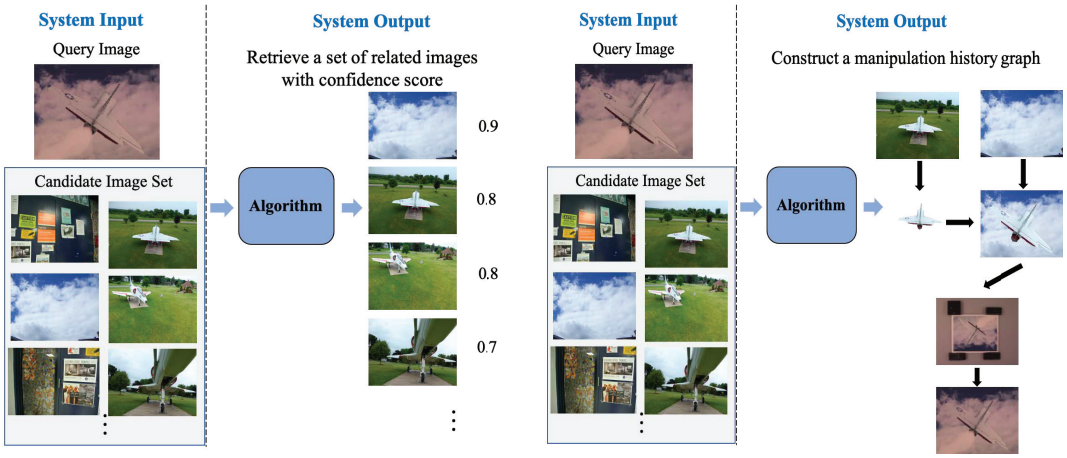
Recently, researchers began to actively study IPA, and a few approaches are introduced in [5, 8–11]. But the IPA systems still have much room to improve their performance. One of the main bottlenecks is the lack of proper benchmark datasets. The Reddit dataset [5] is an IPA dataset that is collected from an online Reddit community. The professional splicing dataset [12] can be used to test IPA systems since the manipulation of images are recorded. However, large-scale image provenance datasets, which contain diverse manipulation history graphs with various manipulation operations and rich metadata, are still needed to further facilitate the image provenance analysis research.

To address above mentioned issues, we introduce large-scale benchmark datasets namely Media Forensics Challenge-Provenance (MFC-Prov) datasets. MFC-Prov is one of the MFC evaluations [13, 14] that are supported by the DARPA MediFor program[1]. MFC contains a series of tasks, which are manipulation/splice detection and localization, camera verification, event verification, provenance filtering (PF) and provenance graph building (PGB). In this paper, we focus on IPA related tasks PF and PGB and present corresponding dataset, evaluation design, and state-of-the-art evaluation results and data analysis. Since the IPA task is very challenging, instead of designing a single task directly, the IPA task is

[1] https://www.darpa.mil/program/media-forensics

decoupled into two sub-tasks: PF and PGB. In this way, the performance of each step can be measured and reported separately.

The objective of the PF task is to retrieve related images for a query image from the millions of candidate images. PF results can be thought of nodes for the PGB task. The recall-based metrics are used to measure the accuracy of PF. PGB is our final objective. The target of the PGB task is to construct the relationships among the retrieved images, i.e., build a manipulation history graph for a query image. For the PGB task, similarity-based metrics [15] are adopted. Figure 1[2] shows the examples of PF and PGB. As shown in Figure 1(a), an IPA system outputs top-*k* related candidate images for the PF task and a manipulation history graph for the PGB task.



(a) PF (b) PGB

Fig. 1. Example of Provenance Filtering (PF) and Provenance Graph Building (PGB) tasks. For the PF task, IPA systems are given a query image and a set of candidate images as input and required to produce a set of related images with confidence score. For the PGB task, the input is the same as the PF task but require a manipulation history graph as output. A manipulation history graph consists of related images as nodes and manipulation operations as links.

The aim of our MFC-Prov evaluation is to answer not only a basic question "How well does a state-of-the-art IPA system perform?", but also advanced questions such as "What major factors affect the system performance?" and"Which system performs better in a certain situation?". Therefore, we collected, annotated, and composed a large corpus of images to build large-scale benchmark datasets for IPA together with PAR Government[3] (hereafter, PAR for short). 4 datasets were produced, which are NC17-Prov, MFC18-Prov, MFC19-Prov, and MFC20-Prov. Nimble Challenge (NC) is the former name of MFC. The latest MFC20-Prov dataset includes 5,926 query images, 1,571 manipulation history graphs, and 2 million candidate images (including distractors, which are images not related

[2]All images, graphs, and charts are original works created for DARPA MediFor Program.
[3]https://www.pargovernment.com

to any query image) with rich metadata such as create date, GPS latitude, camera model, editing processing software, exposure time, lens type, color filter, light source, thumbnail offset, etc. Manipulation history graphs in our dataset incorporate more than 20 types of manipulation operations. Each manipulation history graph contains up to 75 nodes and 121 links.

There were three teams: Columbia University from the Kitware team (Col+Kit) [8], the Notre Dame component of the Purdue team (ND+Pur) [5], and University of Southern California Information Sciences Institute (USCISI) [16], who participated in both PF and PGB tasks of our MFC20-Prov. We firstly present their overall performance in terms of recall and graph similarity. Then in-depth study of MFC20-Prov evaluation results is conducted based on sensitivity analysis by changing one-factor-at-a-time (OAT). We carefully choose five factors for IPA that are most relevant to our interest. Furthermore, correlation between IPA systems is studied based on two aspects.

The contributions of this paper are summarized as follows:

- MFC-Prov datasets are designed and constructed for image provenance analysis (IPA) with the help of PAR, which are large-scale benchmark datasets including diverse manipulation history graphs with various manipulation operations and rich metadata. MFC-Prov datasets are available upon request via email: mfc_poc@nist.gov.
- Two tasks are introduced for IPA, which are provenance filtering (PF) and provenance graph building (PGB), together with PAR. Corresponding evaluation metrics are presented as well.
- Extensive data analysis about the state-of-the-art IPA system performance on MFC-Prov is conducted to gain insight of system behavior. The overall system performance in terms of accuracy, one-factor-at-a-time (OAT) sensitivity analysis [17] and correlation between systems are presented to answer the questions such as "How well does an IPA system perform?", "What major factors affect the system performance?" and "Which IPA system performs better in a certain situation?"

The rest of the paper is organized as follows. Section 2 reviews the related work[4] in the literature, and Section 3 introduces the benchmark dataset and evaluation design. Evaluation results and data analysis are presented in Section 4. Section 5 concludes the paper.

2. Related Work

Provenance analysis is a well-known topic in data-centric domains such as the data annotation, management, and warehousing [18–21]. However, the problem of IPA is not

[4] Any mention of commercial products or reference to commercial organizations in this paper is for information only; it does not imply recommendation or endorsement by NIST nor does it imply that the products mentioned are necessarily the best available for the purpose.

intensively studied so far in the literature. The research areas that are related to IPA are near duplicate detection [22, 23], and image splicing detection [24–29]. Most of these works are designed to classify whether a candidate image is a near duplicate to a given query image. However, these approaches are not designed to determine which images are original. Detecting original images is important to IPA systems.

The most similar problem to IPA is image phylogeny. The target of image phylogeny is to construct the kinship relationships between different versions of an image [6]. Dias et al. [6, 30] propose to construct an image phylogeny tree to describe relationships among near duplicate images based on a dissimilarity matrix and phylogeny tree reconstruction algorithm. In [7], Dias et al. introduce an image phylogeny forest to represent relationships among a set of semantically similar images. However, the base structure of an image phylogeny tree and forest is a single-root tree whose root is an original image. Thus an image phylogeny tree and forest cannot be applied to forgeries with multiple original images. The difference between IPA and image phylogeny is that image phylogeny strictly follows the form of single-root tree, whose root is the only original image that contributing to the query image, while in IPA a query image can have multiple source original images.

Recently, several approaches have been introduced for IPA by Col+Kit and ND+Pur teams of the DARPA MediFor program. Moreira et al. [5] introduce an image indexing scheme and a clustering algorithm for provenance filtering and graph building. Bharati et al. [9] propose an undirected graph to show the relationships between images based on spatial information given by representative key points and match consistency. Pinto et al. [10] present a provenance filtering method to improve retrieval of candidate images by incorporating the context of top results. More recently, Bharati et al. [11] design an IPA approach that utilizes commonly present file metadata tags, e.g., date, location, camera, editing, and thumbnail related metadata. Zhang et al. [8] demonstrate an approach to learn a pairwise ancestor-offspring classifier for detecting related images, as well as a graph building algorithm that combines local feature matching and pixel similarity scores.

Benchmark datasets are essential to facilitate an area of research. Several datasets are created for general media forensics evaluation. The EU REWIND digital forensics project[5] collects three datasets: a Realistic dataset that includes 69 manually manipulated images and corresponding 69 original images; a Synthetic dataset that is composed of 4,800 automatically manipulated images; and a dataset that contains 200 images that are taken from a Nikon D60 camera. The first Image Forensics Challenge [31] fetches thousands of images of diverse indoor and outdoor scenes using 25 cameras.

There are also datasets constructed for particular manipulation operations. The Columbia database of automatically spliced images [32] is divided into two parts: a grayscale image dataset that contains 933 authentic and 912 spliced images; and a color image dataset that includes 183 authentic and 180 spliced images. CASIA's database for image tampering detection evaluation [33] is composed of 7,491 authentic and 5,123 tempered images. For copy-move detection, there also exists datasets such as MICC F220, MICC F2000 [34], FAU Erlangen image manipulation datasets [35]. UMDfaces [36] contains 367,888 face

[5] http://www.rewindproject.eu/

Table 1. Statistics of datasets for image provenance analysis

dataset	$\lvert\mathcal{MG}\rvert$	$\lvert\mathcal{N}\rvert$	$\lvert\mathcal{L}$ types$\rvert$	$\lvert\mathcal{C}\rvert$	**PF**	**PGB**
Professional splicing [12]	80	75	8	164	$\times$	oracle
Reddit [5]	184	89	-	10,421	$\times$	oracle
Our MFC20-Prov	1,571	75	57	2,000,000	$\surd$	full

images with annotation for 8,277 subjects. FaceForensics dataset [37] has about half a million manipulated images that is generated using state-of-the-art face editing approach from over 1,000 videos. The images are annotated with classification and segmentation references.

There are limited datasets for IPA existing in the literature. The Reddit dataset [5] is an IPA dataset that was collected from an online Reddit community known as Photoshop battles[6]. The Reddit dataset consists of 184 manipulation history graphs and 10,421 candidate images. The professional splicing dataset [12] can be used to test IPA systems since it is a work of professional artists that tried to make the images as credible as possible. The professional splicing dataset has 80 available manipulation history graphs generated using 164 candidate images. Each manipulation history graph contains 75 nodes, and each node has always two parent nodes.

However, large-scale IPA datasets, which contain diverse manipulation history graphs with various manipulation operations and rich metadata, are still needed to further facilitate the IPA research. Table 1 shows the comparison of statistics between existing datasets and our MFC20-Prov dataset. $\mathcal{MG}$ denotes the set of manipulation history graphs, $\mathcal{N}$ and $\mathcal{L}$ indicate the sets of nodes and links in a manipulation history graph, and $\mathcal{C}$ means a set of candidate images. $\mathcal{L}$ types of Reddit dataset are unknown since reddit users simply post modified images without explanation of their modifications. As shown in Table 1, existing IPA datasets do not support the PF task since there are no distractors for candidate images, i.e., related images are fixed and given for a query image. Therefore, existing IPA datasets merely support the oracle condition of PGB, where PF results are given.

3. Image Provenance Benchmark Dataset and Evaluation Design

The target of our Media Forensics Challenge-Provenance (MFC-Prov) evaluation is to answer not only a primary question "How well does a state-of-the-art IPA system perform?", but also advanced questions such as "What major factors affect the system performance?" and "Which IPA system performs better in a certain situation?". Table 2 lists the notations used in this paper.

To answer these advanced questions, the first important part is the manipulation history graphs that record where and how the manipulated images come from. The formal definition of a manipulation history graph $\mathcal{MG}$ is described as follows:

[6]https://www.reddit.com/r/photoshopbattles

Table 2. Notation for MFC-Prov

Notation	Description
IPA	image provenance analysis
PF	provenance filtering
PGB	provenance graph building
$\mathcal{MG}$	manipulation history graph
$\mathcal{N}$	a set of nodes in a manipulation history graph
$\mathcal{L}$	a set of links in a manipulation history graph
$\mathcal{C}$	a set of candidate images
$\mathcal{Q}$	a set of query images
q	a query image for the PF or PGB task

Table 3. Example metadata of the MFC-Prov datasets

Metadata category	Metadata
Date	datetime original, modify date, create date
Location	GPS latitude, GPS latitude ref, GPS longitude, GPS longitude ref
Camera	make, model, software, orientation
Editing	processing software, artist, host computer, image resources
Exposure	time, program, compensation, mode
Lens	type, spec, zoom position, mount, firmware version
Color	temperature, compensation filter, mode, space
Light	source, value
Thumbnail	offset, length, image

Definition 1 (Manipulation history graph $\mathcal{MG}$) A manipulation history graph is a directed graph $\mathcal{MG} = (\mathcal{N}, \mathcal{L})$, where $\mathcal{N}$ is a set of nodes (i.e., vertices) and $\mathcal{L}$ is a set of links (i.e., edges). Each node *n* indicates an image from a set of candidate images $\mathcal{C}$, and each link *l* denotes a manipulation operation, e.g., color balance.

$\mathcal{C}$ is a collection of publicly available images acquired off the Internet and the images taken by cameras that are physically accessible to us. A large amount of $\mathcal{MG}$s are generated by professional human manipulators. PAR and NIST developed a manipulation journaling tool [38] to assist human experts in generating $\mathcal{MG}$ with annotation, metadata, and reference data. Since $\mathcal{MG}$ annotation cost is very high, to maximize the usage we share the manipulation $\mathcal{MG}$s between the PF and PGB tasks.

In addition, tremendous number of candidate images $\mathcal{C}$ with metadata are collected and produced. Examples of metadata are listed in Table 3.

3.1 Dataset

MFC-Prov datasets were firstly developed in 2017 and since evolved annually. Table 4 lists the statistics of the MFC-Prov datasets. $\mathcal{Q}$ indicates a set of query images. Note that the

Table 4. Statistics of the MFC-Prov datasets

dataset	$\lvert\mathcal{Q}\rvert$	$\lvert\mathcal{MG}\rvert$	$\lvert\mathcal{C}\rvert$
NC17-Prov	1,000	406	1,000,000
MFC18-Prov	10,000	641	1,000,000
MFC19-Prov	9,420	1,025	2,000,000
MFC20-Prov	5,926	1,571	2,000,000

datasets are not based on the orthogonal design.

The number of $\mathcal{Q}$ increased and then dropped recently. We did this by reducing the non-target query images $\mathcal{Q}$, i.e., $\mathcal{Q}$ that are non-manipulated images, which are not evaluated by the provenance evaluation metrics.

The number of $\mathcal{MG}$ increased continuously. $\mathcal{MG}$ can be references of multiple manipulated images. Abundant manipulation operations and techniques are adopted in $\mathcal{MG}$, such as splice, clone, crop, resize, global blur/smooth, GAN [1], face manipulation, etc.

The size of $\mathcal{C}$ reached 2 million images. Over 500 distinct camera models are included to ensure the diversity of data. $\mathcal{C}$ includes all ancestors and descendants of the manipulated images w.r.t. $\mathcal{MG}$. In addition, $\mathcal{C}$ contains 176,000 self-produced images. Self-produced images are collected using physically accessible devices by PAR. Thus all device-relative metadata are precisely recorded along with self-produced images. Self-produced images are original images that guarantee there is no previous manipulation with no copyright conflict. Our self-produced images are released under Creative Commons 0 (CC0) license to make them completely public.

3.2 Task Definition and Evaluation Metrics

Our image provenance evaluation consists of two tasks: provenance filtering (PF) and provenance graph building (PGB). In this section, we describe the definitions and evaluation metrics for the two tasks.

3.2.1 Provenance Filtering

The target of the PF task is to search for a potential pool of related images from $\mathcal{C}$. The formal definition is given as follows:

Definition 2 (Provenance Filtering: PF) Given a query image q and a set of candidate images $\mathcal{C}$, PF aims to find top-k related nodes (i.e., node images) from $\mathcal{C}$ with confidence scores.

The high confidence score means a node n is considered by a system to be highly related to q. The confidence score is used to sort the output nodes of a system.

The evaluation metric of PF is $recall@k$, which is described as follows:

$$recall(k,q,s) = \frac{|\mathcal{N}_{\mathcal{MG}_{(q,s,k)}} \cap \mathcal{N}_{\mathcal{MG}_q}|}{|\mathcal{N}_{\mathcal{MG}_q}|} \tag{1}$$

$\mathcal{N}_{\mathcal{MG}_{(q,s,k)}}$ denotes the top-k nodes returned by a system s for a query image q. And $\mathcal{N}_{\mathcal{MG}_q}$ indicates the nodes of the reference $\mathcal{MG}$ of q. We set k as 50, 100, 200, and 300 to test system performance. Among them, $recall@300$ is the primary metric.

3.2.2 Provenance Graph Building

The target of PGB task is to construct the relationships among retrieved related images along with finding the ancestor and descendants sequences, as well as applied manipulation operations. The PGB task is formally defined as follows:

Definition 3 (Provenance Graph Building: PGB) Given a query image q and a set of candidate images $\mathcal{C}$, PGB aims to create a manipulation history graph for q, $\mathcal{MG}_q$ that describes the relationships among the related images with the manipulation sequences. The nodes of $\mathcal{MG}_q$ are retrieved from $\mathcal{C}$ and the links represent manipulation operations.

To measure the similarity between a manipulation history graph for q produced by a system $\mathcal{MG}_{(q,s)}$ and a reference $\mathcal{MG}_q$, we employ three evaluation metrics [15] namely node overlap sim_{NO}, link overlap sim_{LO}, and node and link overlap sim_{NLO}. Among them, sim_{NLO} is the primary metric. The metrics measure the similarity between $\mathcal{MG}_{(q,s)}$ and $\mathcal{MG}_q$. The following equations describe the evaluation metrics:

$$sim_{NO}(q,s) = 2\frac{|\mathcal{N}_{\mathcal{MG}_{(q,s)}} \cap \mathcal{N}_{\mathcal{MG}_q}|}{|\mathcal{N}_{\mathcal{MG}_{(q,s)}}| + |\mathcal{N}_{\mathcal{MG}_q}|} \tag{2}$$

$$sim_{LO}(q,s) = 2\frac{|\mathcal{L}_{\mathcal{MG}_{(q,s)}} \cap \mathcal{L}_{\mathcal{MG}_q}|}{|\mathcal{L}_{\mathcal{MG}_{(q,s)}}| + |\mathcal{L}_{\mathcal{MG}_p}|} \tag{3}$$

$$sim_{NLO}(q,s) = 2\frac{|\mathcal{N}_{\mathcal{MG}_{(q,s)}} \cap \mathcal{N}_{\mathcal{MG}_q}| + |\mathcal{L}_{\mathcal{MG}_{(q,s)}} \cap \mathcal{L}_{\mathcal{MG}_q}|}{|\mathcal{N}_{\mathcal{MG}_{(q,s)}}| + |\mathcal{N}_{\mathcal{MG}_q}| + |\mathcal{L}_{\mathcal{MG}_{(q,s)}}| + |\mathcal{L}_{\mathcal{MG}_q}|} \tag{4}$$

where $\mathcal{N}_{\mathcal{MG}_{(q,s)}}$ and $\mathcal{N}_{\mathcal{MG}_q}$ indicate the nodes of system output and reference $\mathcal{MG}$ for q. Besides, $\mathcal{L}_{\mathcal{MG}_{(q,s)}}$ and $\mathcal{L}_{\mathcal{MG}_q}$ denote the links of system output and reference $\mathcal{MG}$ for q. Three metrics range in [0,1] and the larger value means the more similar to a reference $\mathcal{MG}$ are, i.e., the more accurate $\mathcal{MG}_{(q,s)}$ is.

Here, we give an example of measuring the accuracy of (i.e., scoring) $\mathcal{MG}_{(q,s)}$.

Example 1 (Scoring $\mathcal{MG}_{(q,s)}$) Figure 2 shows a manipulation history graph generated by a system s for a query image q $\mathcal{MG}_{(q,s)}$. A rectangle indicates a node and an arrow denotes a link. The green, red, and grey nodes (links) represent correctly detected, falsely detected, and missed nodes (links), respectively. And the bold green rectangle stands for q. Therefore, the sim_{NLO} comes to $(5+2)/(10+11) = 0.333$.

Fig. 2. Scoring example of $\mathcal{MG}_{(q,s)}$

4. Evaluation Results and Analysis

We conducted evaluations on MFC-Prov datasets to answer the primary and advanced questions for IPA systems. As mentioned in Section 3, the questions include "How well does a state-of-the-art IPA system perform?", "What major factors affect the system performance?", "Which IPA system performs better in a certain situation?", etc. Three performer teams participated in our MFC-Prov: Columbia University from the Kitware team (Col+Kit) [8], the Notre Dame component of the Purdue team (ND+Pur) [5], and University of Southern California Information Sciences Institute (USCISI) [16]. All 3 teams have taken part in the MFC-Prov all through the years from NC17-Prov to MFC20-Prov.

The remainder of this section presents the results of overall performance, one-factor-at-a-time (OAT) sensitivity analysis, and correlation analysis of PF and PGB tasks.

4.1 Overall Performance

To address the primary question "How well do a state-of-the-art provenance analysis system perform?" , we present the PF and PGB overall results across the 4 years of the evaluations. Note that MFC-Prov datasets and participant systems were different from year to year. It may be meaningless to compare inter-year performance.

Figure 3 shows the PF results over years. The evaluation metric is *recall*@200 since there was no *recall*@300 for NC17-Prov. The performance of systems did not grow with time. It may be due to that the evolving volume and complexity of evaluation datasets. And no single team achieved the highest score across the all years. Col+Kit achieved the highest score on NC17-Prov, MFC19-Prov, and MFC20-Prov, but performed slightly worse than ND+Pur on MFC18-Prov.

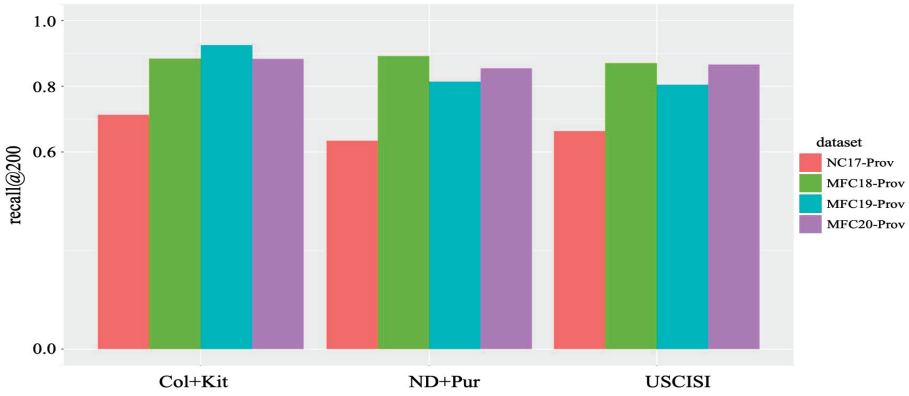


Fig. 3. PF results over years in terms of *recall*@200

Figure 4 shows the PGB result over years using the metric sim_{NLO}. Col+Kit did not participated in the PGB task for NC17-Prov. After then, Col+Kit scored the highest sim_{NLO} for MFC18-Prov, MFC19-Prov, and the recent MFC20-Prov. The scores of PGB were

Table 5. Factors of OAT sensitivity analysis for MFC20-Prov

Factor	Description
Manipulation count	how many manipulations are conducted to a query image q. Manipulation count ranges in [1,8]
Face manipulation	whether q contains face manipulation
GAN	whether q includes manipulation operation GAN
Anti-forensics	whether the manipulations to q are made after applying anti-forensics applications. Anti-forensics applications modify, conceal or destroy information to inhibit or prevent the effectiveness of forensic science examinations [39]
Target Operations	target operations indicate manipulation operation types and 20 target operations are selected. If q contains a target operation, then q will be added to $\mathbb{Q}$ with the target operation

much lower than that of PF, and tended to decrease over the years when excluding NC17-Prov.

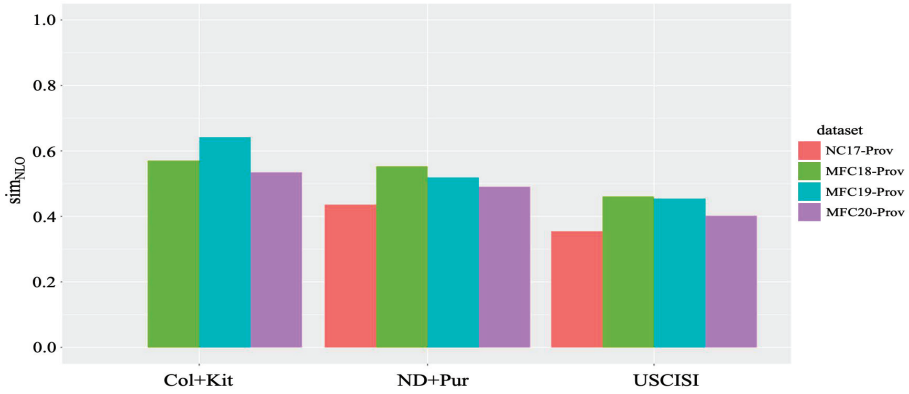


Fig. 4. PGB results over years in terms of sim_{NLO}

4.2 OAT Sensitivity Analysis for MFC20-Prov

We aim to answer the advanced questions "What major factors affect the system performance?" and "How well does a given system perform under a specific condition?". OAT sensitivity analysis [17] is a logical approach as any change observed in the output will unambiguously be due to the single variable changed. OAT sensitivity analysis is adopted since it fits to find out which factor affects the performance of IPA systems. Five factors are carefully selected according our interests for IPA. The factors are listed in Table 5.

In OAT sensitivity analysis, we kept the sample size of $\mathbb{Q}$ larger than 70 to minimize uncertainty of the data analysis. OAT sensitivity analysis was conducted using the latest MFC20-Prov dataset, which is not orthogonally designed to collect data. The primary metrics are $recall@300$ for the PF task and sim_{NLO} for the PGB task.

Table 6. Average rank of manipulation count for PF from Figure 5

Manipulation count	8	7	2	6	4	5	3	1
Average rank	2.00	3.00	3.33	3.67	4.33	5.33	6.33	8.00

4.2.1 Manipulation Count

Figure 5 shows the PF results of MFC20-Prov with varying manipulation counts, where the x axis is the team name and the y axis is *recall*@300 value. The numbers in the boxes indicate the manipulation count. Table 6 lists the average ranks of the manipulation count shown in Figure 5. The lower average rank means the higher *recall*@300 score and better performance.

As shown in Figure 5, single manipulation is harder to retrieve than multiple manipulations for all the systems, and the systems achieved the highest score with 7 or 8 manipulations. In addition, the *recall*@300 scores of ND+Pur did not vary much with different manipulation counts. The manipulation count had the largest impact on USCISI. However, if we exclude the single manipulation, then USCISI becomes the most robust system against manipulation count.

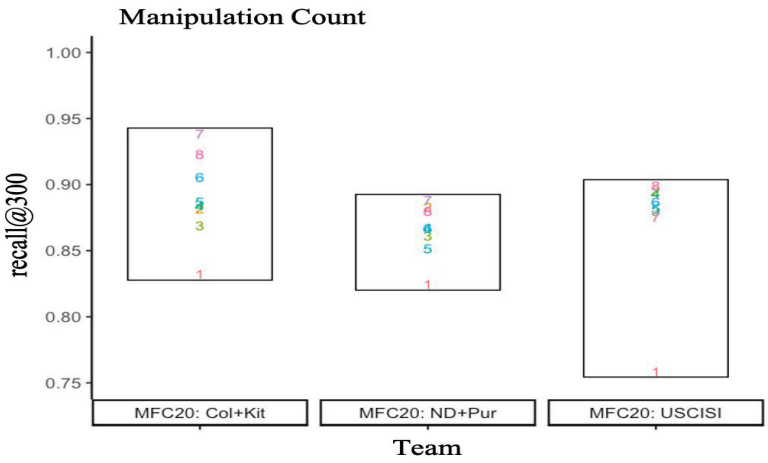


Fig. 5. PF results of MFC20-Prov with varying manipulation counts. The digits in the boxes denote the numbers of manipulations applied

Figure 6 shows the PGB results of MFC20-Prov with varying manipulation counts and Table 7 lists the corresponding average ranks of the manipulation count. Similar to the PF task, systems performed better on PGB with a large manipulation count, e.g., 7 and 8. In contrast, systems struggled on PGB with median manipulation count, e.g. 3 and 4. Manipulation count had greatest impact on Col+Kit and smallest impact on USCISI, but the difference of impact is not as broad as that of PF.

Table 7. Average rank of manipulation count for PGB from Figure 6

Manipulation count	8	7	6	5	1	2	3	4
Average rank	1.00	2.00	3.33	4.33	4.67	6.00	7.33	7.33

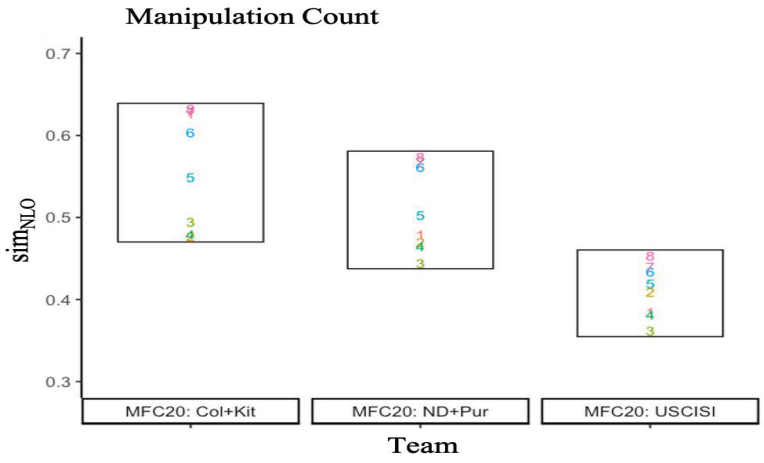


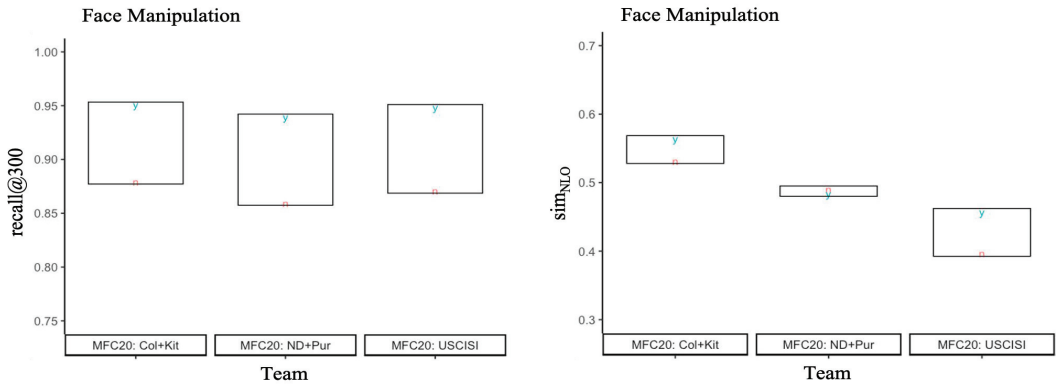
Fig. 6. PGB results of MFC20-Prov with varying manipulation counts. The digits in the boxes denote the numbers of manipulations applied

4.2.2 Face Manipulation

Figure 7 shows the PF and PGB results of MFC20-Prov with/without face manipulation. y indicates the score for query images $\mathcal{Q}$ with face manipulation and n denotes that without face manipulation. The face manipulation affected the PF performance greatly. As shown in Figure 7(a), $\mathcal{Q}$ with face manipulations were easier to retrieve compared to the non-face manipulations across the three systems. For the PGB task, the face manipulation had a great influence on USCISI. However, the sim_{NLO} of ND+Pur was barely affected by the face manipulation. And ND+Pur even performed better with non-face manipulations for PGB.

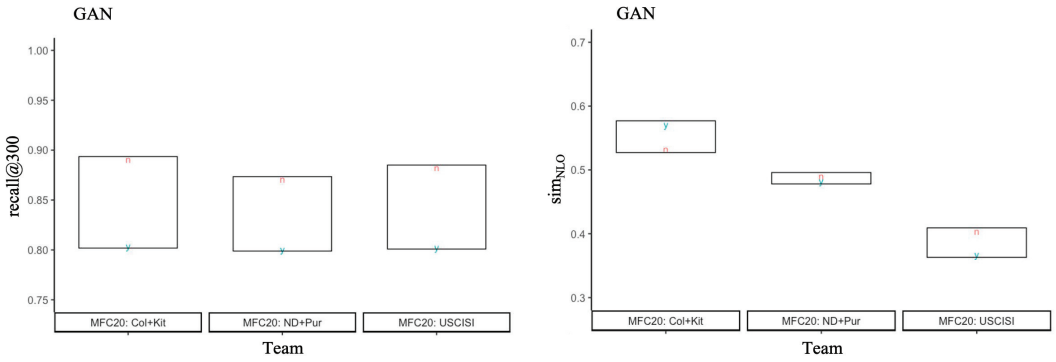
4.2.3 GAN

Figure 8 shows the PF and PGB results of MFC20-Prov with/without GAN. Here, GAN denotes the method proposed in [1] and its related techniques. y indicates the score for $\mathcal{Q}$ with GAN manipulation and n denotes that without GAN manipulation. As shown in Figure 8(a), $\mathcal{Q}$ with GAN were much harder to detect for all three systems. For the PGB task, the performance of ND+Pur seems robust to the GAN. Note that Col+Kit even scored higher for $\mathcal{Q}$ with GAN manipulations.



(a) PF (b) PGB

Fig. 7. PF and PGB results of MFC20-Prov with/without face manipulation. The green y indicates that the face manipulation is applied and the red n denotes that no face manipulation is employed

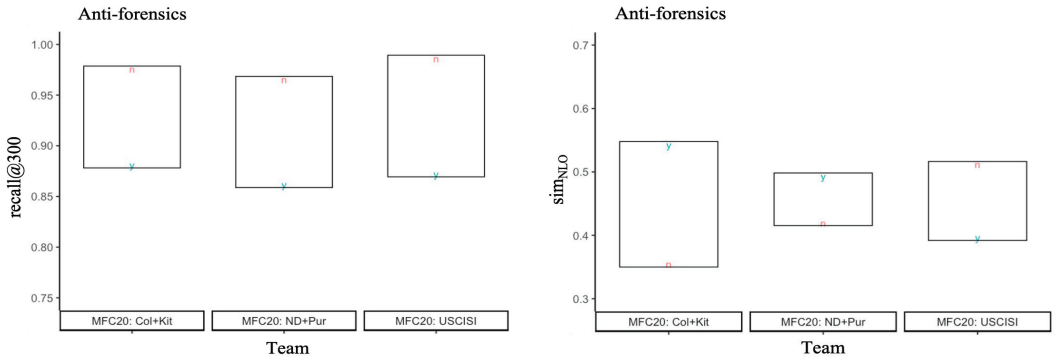


(a) PF (b) PGB

Fig. 8. PF and PGB results of MFC20-Prov with/without GAN. The green y indicates that the GAN is applied and the red n denotes that no GAN is employed

4.2.4 Anti-forensics

Figure 9 shows the PF and PGB results of MFC20-Prov with/without anti-forensics. *y* indicates the score for Ω with anti-forensics and *n* denotes that without anti-forensics. Ω without anti-forensics were easier to detect for the PF task and the impact was large on all the three systems, as shown in Figure 9(a). In contrast, Col+Kit and ND+Pur performed better with anti-forensics for PGB. Furthermore, the gap of sim_{NLO} between with and without anti-forensics was big for Col+Kit but relatively small for ND+Pur.



(a) PF (b) PGB

Fig. 9. PF and PGB results of MFC20-Prov with/without anti-forensics. The green y indicates that the anti-forensics is applied and the red n denotes that no anti-forensics is employed

Table 8. PF (PGB) results of MFC20-Prov in terms of *recall*@300 (sim_{NLO}) with various target operations

to	TW	TD	CGI	DPD	CrF	ConAF	AN
ND+Pur	.95 (.46)	.92 (.47)	.91 (.46)	.91 (.46)	.92 (.44)	.90 (.42)	.88 (.45)
Col+Kit	.96 (.55)	.94 (.64)	.94 (.65)	.94 (.61)	.93 (.61)	.92 (.57)	.93 (.59)
USCISI	.95 (.48)	.94 (.57)	.91 (.59)	.91 (.55)	.90 (.53)	.91 (.53)	.89 (.54)
to	PSa	Cu	ColBal	TRes	E	ArtS	Blu
ND+Pur	.89 (.42)	.87 (.40)	.87 (.40)	.86 (.43)	.86 (.44)	.85 (.43)	.85 (.42)
Col+Kit	.88 (.53)	.90 (.49)	.92 (.52)	.89 (.54)	.90 (.60)	.90 (.62)	.89 (.57)
USCISI	.92 (.50)	.91 (.48)	.87 (.48)	.90 (.51)	.87 (.55)	.87 (.55)	.88 (.49)
to	Lev	Sat	SelRem	LayOp	PSp	Hue	
ND+Pur	.84 (.36)	.83 (.41)	.81 (.41)	.83 (.37)	.83 (.38)	.81 (.43)	
Col+Kit	.90 (.47)	.90 (.61)	.87 (.55)	.89 (.47)	.87 (.50)	.86 (.60)	
USCISI	.87 (.45)	.83 (.56)	.88 (.52)	.83 (.42)	.87 (.47)	.84 (.54)	

4.2.5 Target Operation

We tested 20 target operations, which are add noise (AN), artificial shadow (ArtS), blur (Blu), CGI fill (CGI), color balance (ColBal), content aware fill (ConAF), creative filter (CrF), curves (Cu), digital pen draw (DPD), exposure (E), hue (Hue), layer opacity (LayOp), levels (Lev), paste sampled (PSa), paste splice (PSp), saturation (Sat), select remove (SelRem), transform distort (TD), transform resize (TRes), transform warp (TW). Note that the target operations are not strictly disjoint.

Table 8 shows the PF and PGB results of MFC20-Prov in terms of *recall*@300 and sim_{NLO} with various target operations *to*s. Table 9 lists average rank *ar* of the target operation *to* in the PF task. The lower rank indicates the higher *recall*@300. TW was the easiest operation to filter for the PF task, then followed by TD and CGI. LayOp, PSp, and

Table 9. Average rank of target operation for PF from Table 8

to	TW	TD	CGI	DPD	CrF	ConAF	AN	PSa	Cu	ColBal
ar	1.00	2.33	4.33	4.33	5.33	6.00	7.67	9.00	10.00	11.00
to	TRes	E	ArtS	Blu	Lev	Sat	SelRem	LayOp	PSp	Hue
ar	11.33	11.67	13.00	13.33	13.33	15.67	16.00	17.67	17.67	19.33

Table 10. Average rank of target operation for PGB from Table 8

to	CGI	TD	DPD	ArtS	CrF	E	AN	Sat	Hue	ConAF
ar	1.67	1.67	3.67	5.33	6.33	6.67	7.33	8.00	8.33	10.33
to	TW	TRes	Blu	SelRem	PSa	Cu	ColBal	PSp	Lev	LayOp
ar	10.67	11.67	12.33	12.67	13.67	16.33	16.67	17.67	19.33	19.67

Hue appeared to be the hardest operations. The impact of the target operation *to* was large for all the three PF systems.

Table 10 lists average rank *ar* of the target operation *to* in the PGB task of MFC20-Prov. The lower rank denotes the higher sim_{NLO}. CGI was the easiest operation for the PGB task, then followed by TD and DPD. On the other hand, PSp, Lev, and LayOp were the hardest operations for PGB.

Combining the two results, we could conclude that CGI, TD, DPD, and CrF are the easy operations to process and LayOp and PSp are common hard operations for provenance analysis.

4.2.6 OAT Sensitivity Analysis Summary

We first summarize the OAT sensitivity analysis for the PF task of MFC20-Prov. Manipulation count, anti-forensics, and target operation are the main factors for the Col+Kit system, which caused difference in *recall*@300 more than 0.1. Manipulation count is the most noteworthy factor for the USCISI system. For the ND+Pur system, target operation comes out to be the most influential factor. The Col+Kit system gained the highest score in most cases. However, the USCISI system performed the best when there is no anti-forensics being executed or the target operation is PSa, TRes, or SelRem.

OAT sensitivity analysis for the PGB task of MFC20-Prov is summarized as follows. Manipulation count, anti-forensics, and target operation are the most influential factors for the Col+Kit system, the same result as the PF task. For the ND+Pur system, manipulation count and target operation are the main factors. Manipulation count, anti-forensics, and target operation are the key factors for the USCISI system. The Col+Kit system resulted in the highest sim_{NLO} in most cases. However, USCISI scored higher when no anti-forensics had been performed.

4.3 Correlation Analysis for MFC20-Prov

In this section we analyze the correlation between provenance analysis systems. The following analysis is conducted to demonstrate how to utilize systems in a certain application with minimal computational cost. Correlation can be used to predict the performance and select a proper system. For instance, if a system s_1 scored high on a provenance dataset, a system s_2 is positively correlated with s_1 and another system s_3 is negatively correlated with s_1, then s_2 would score high and s_3 would have low performance on the same dataset with a high probability. In this case, s_2 may be considered as a better option without actually running all the systems. It could save time and computational cost.

q score is chosen as the factor for correlation analysis. q score indicates the scores of a system for each q. We have 2926 q scores for each PF system. Note that not all $\mathcal{Q}$ were assigned with scores. Only the manipulated $\mathcal{Q}$ (i.e., $\mathcal{Q}$ with reference $\mathcal{MG}$s) are evaluated. If two systems s_1 and s_2 have strong positive correlation w.r.t. q scores and s_1 performed great on a certain dataset, then s_2 would perform well on the same dataset with high probability. q score factor works for both PF and PGB tasks.

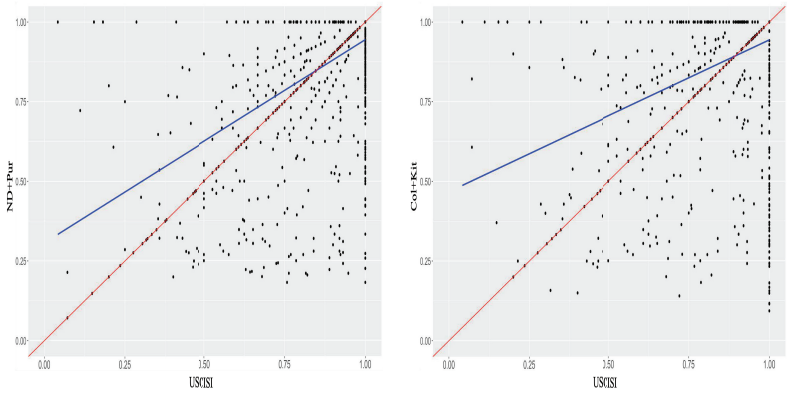
We first analyzed the correlation between system pairs $\langle s_1, s_2 \rangle$. Figure 10 shows the results for the PF task of MFC20-Prov. The x axis is the $recall@300$ score of s_1, the y axis is the $recall@300$ score of s_2, each point represents q, a red line denotes the perfect positive correlation, and a blue line indicates the linear regression. The linear regression lines was computed using the *lm* function provided by geom_smooth R package[7].

There were 2,926 points for each system pair. The closer a red line and a blue line were located, the more positively two systems s_1 and s_2 are correlated. As shown in Figure 10, the correlation between ⟨ND+Pur, Col+Kit⟩ was the highest, followed up by ⟨USCISI, ND+Pur⟩ and ⟨USCISI, Col+Kit⟩, which were loosely correlated.

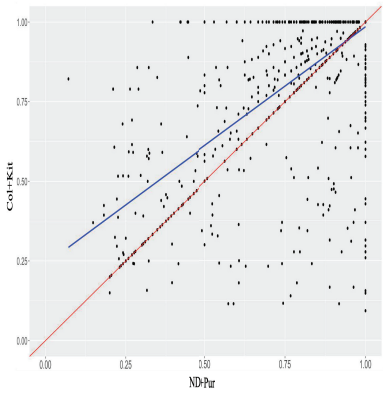
Figure 11 shows the correlation heatmap for PF systems of MFC20-Prov based on q score. The value denotes the Pearson correlation coefficient that ranges in [-1,1]. Pearson correlation coefficient is widely used in correlation analysis, 1 means two systems are perfectly positively correlated, -1 stands for perfect negative correlation, and 0 implies that there is no linear correlation. The result shown in Figure 11 is consistent with that of Figure 10. Specifically, $Cor(ND+Pur, Col+Kit) > Cor(USCISI, ND+Pur) > Cor(USCISI, Col+Kit)$, where $Cor(s_1, s_2)$ indicates the Pearson correlation coefficient between s_1 and s_2 in terms of q score.

For the PGB task, Figure 12 shows the system pair correlations of MFC20-Prov based on q score. The x axis is the sim_{NLO} score of s_1 and the y axis is the sim_{NLO} score of s_2. There were only few $\mathcal{Q}$ with score larger than 0.75 of USCISI. In contrast, ND+Pur and Col+Kit had many $\mathcal{Q}$ with high score including those sim_{NLO} =1. ⟨USCISI, Col+Kit⟩ was less correlated than the other two system pairs since its linear regression (blue) line and the perfect positive correlation (red) line was relatively far away. In addition, it is hard to tell the difference between ⟨USCISI, ND+Pur⟩ and ⟨ND+Pur, Col+Kit⟩ in terms of the linear regression line. Note that the points in Figure 12 tended to lie in two clusters. One cluster is

[7] https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf



(a) ⟨USCISI, ND+Pur⟩ (b) ⟨USCISI, Col+Kit⟩



(c) ⟨ND+Pur, Col+Kit⟩

Fig. 10. System pair correlations based on q scores for PF of MFC20-Prov

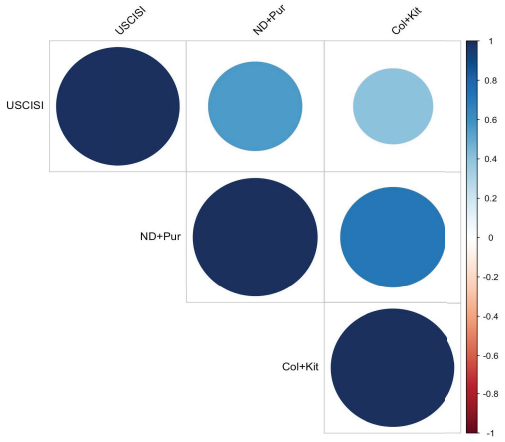
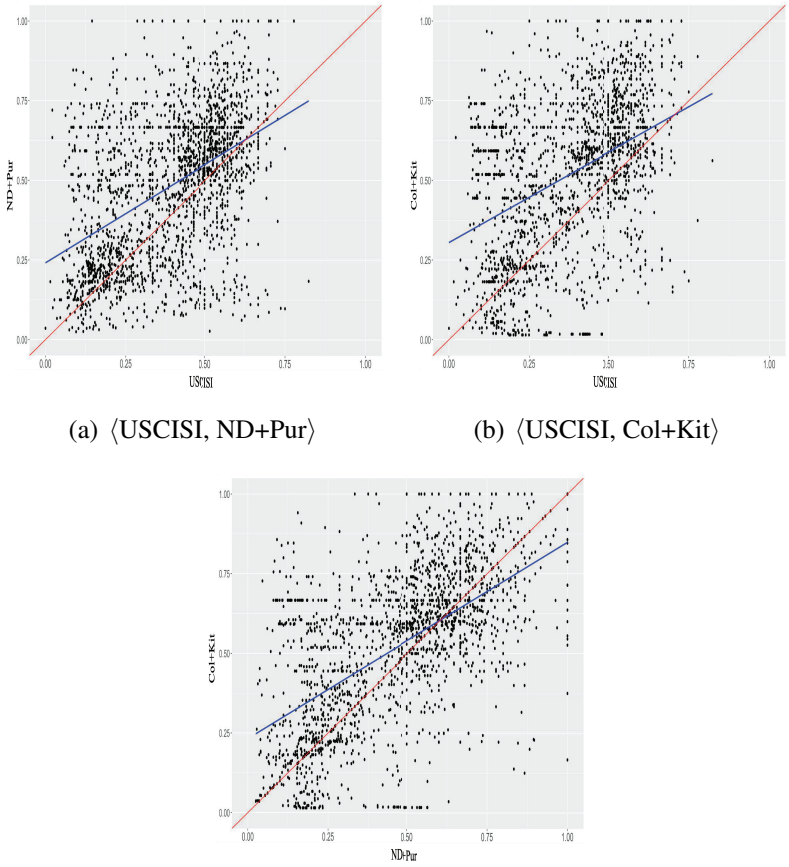


Fig. 11. Pearson correlation heatmap for PF systems of MFC20-Prov based on q score

low-low where two systems both scored low, and the other one is medium-medium where two systems both scored in the medium range. It may due to that the PGB q scores were mainly located in two ranges [0.2,0.3] and [0.5,0.8], as shown in Figure 13. In Figure 13, count denotes the number of query images that lie in a score range.



(a) ⟨USCISI, ND+Pur⟩ (b) ⟨USCISI, Col+Kit⟩

(c) ⟨ND+Pur, Col+Kit⟩

Fig. 12. System pair correlations based on q score for PGB of MFC20-Prov

Figure 14 shows the correlation heatmap for PGB systems of MFC20-Prov based on q score. The value is Pearson correlation coefficient. ⟨USCISI, Col+Kit⟩ appeared as the least correlated system pair, which is the same result as shown in Figure 12. However, $Cor(ND+Pur, Col+Kit)$ was obviously larger than $Cor(USCISI, Col+Kit)$. Such a difference might be caused by that the linear regression function lm, which is provided by the geom_smooth R package, could not express the relation of q scores clearly.

In summary, ⟨ND+Pur, Col+Kit⟩ had the highest positive q correlation for both PF and PGB tasks of MFC20-Prov. It means that if ND+Pur (Col+Kit) system performed well on a certain dataset for the PF or PGB task, then Col+Kit (ND+Pur) system would also perform well with a high probability.

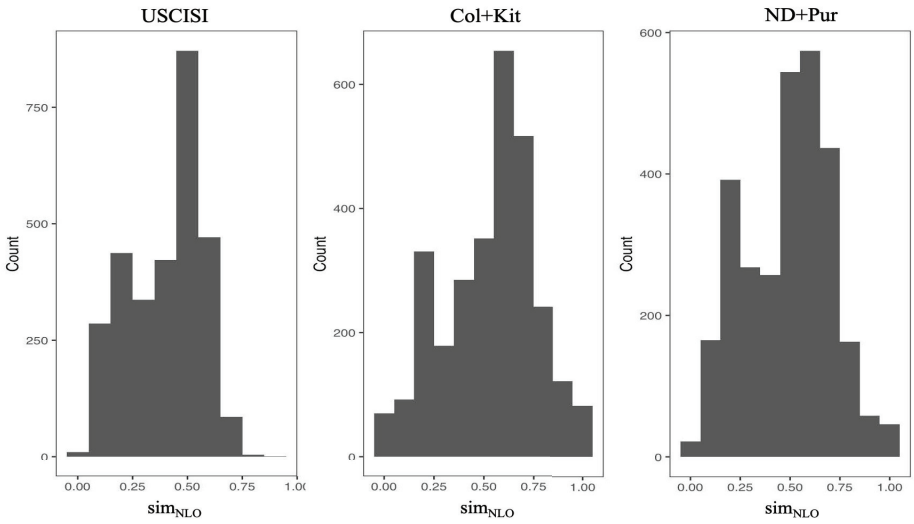


Fig. 13. PGB results on MFC20-Prov w.r.t. q score count

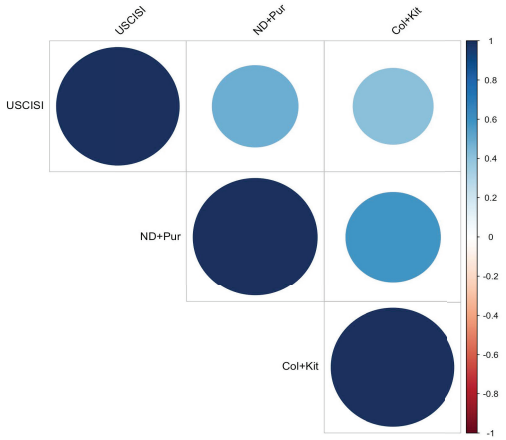


Fig. 14. Pearson correlation heatmap for PGB systems of MFC20-Prov based on q score

The ND+Pur and Col+Kit share commonalities such as local feature-based image representation/retrieval, dataset indexing, query refinement, (dis)similarity calculation, and clustering. Among the commonalities, the USCISI system only adopts dataset indexing and similarity calculation. This may be the reason of the correlation difference between system pairs.

5. Conclusion

The aim of this paper is to answer not only the primary question for overall performance of image provenance analysis (IPA) systems but also advanced questions. These advanced

questions include "What major factors affect the system performance?", "Which IPA would perform better in a certain situation?", "How are the systems correlated?".

Specifically, we introduced large-scale benchmark datasets, namely MFC-Prov, to facilitate the IPA research. The latest MFC-Prov dataset includes 5,926 query images, 1,571 manipulation history graphs, and 2 million candidate images with rich metadata. The benchmark datasets are available upon request via email: mfc_poc@nist.gov. Two tasks, which are provenance filtering and provenance graph building, are designed along with corresponding evaluation metrics. Furthermore, in-depth study was conducted on the evaluation results w.r.t. overall performance, OAT sensitivity analysis, and correlation analysis.

References

[1] Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:171010196* .

[2] Aubry M, Russell BC (2015) Understanding deep features with computer-generated imagery. *Proceedings of International Conference on Computer Vision (ICCV)*, , pp 2875–2883.

[3] Farid H (2017) How to detect faked photos. *American Scientist* 6.

[4] Rocha A, Scheirer W, Boult T, Goldenstein S (2011) Vision of the unseen: Current trends and challenges in digital image and video forensics. *ACM Computing Surveys (CSUR)* 43(4):1–42.

[5] Moreira D, Bharati A, Brogan J, Pinto A, Parowski M, Bowyer KW, Flynn PJ, Rocha A, Scheirer WJ (2018) Image provenance analysis at scale. *IEEE Transactions on Image Processing* 27(12):6109–6123.

[6] Dias Z, Rocha A, Goldenstein S (2011) Image phylogeny by minimal spanning trees. *IEEE Transactions on Information Forensics and Security* 7(2):774–788.

[7] Dias Z, Goldenstein S, Rocha A (2013) Toward image phylogeny forests: Automatically recovering semantically similar image relationships. *Forensic science international* 231(1-3):178–189.

[8] Zhang X, Sun ZH, Karaman S, Chang SF (2020) Discovering image manipulation history by pairwise relation and forensics tools. *IEEE Journal of Selected Topics in Signal Processing* .

[9] Bharati A, Moreira D, Pinto A, Brogan J, Bowyer K, Flynn P, Scheirer W, Rocha A (2017) U-phylogeny: Undirected provenance graph construction in the wild. *Proceedings of IEEE International Conference on Image Processing (ICIP)*, , pp 1517–1521.

[10] Pinto A, Moreira D, Bharati A, Brogan J, Bowyer K, Flynn P, Scheirer W, Rocha A (2017) Provenance filtering for multimedia phylogeny. *Proceedings of IEEE International Conference on Image Processing (ICIP)*, , pp 1502–1506.

[11] Bharati A, Moreira D, Brogan J, Hale P, Bowyer K, Flynn P, Rocha A, Scheirer W (2019) Beyond pixels: Image provenance analysis leveraging metadata. *Proceedings of Winter Conference on Applications of Computer Vision (WACV)*, , pp 1692–1702.

[12] De Oliveira AA, Ferrara P, De Rosa A, Piva A, Barni M, Goldenstein S, Dias Z,

Rocha A (2015) Multiple parenting phylogeny relationships in digital images. *IEEE Transactions on Information Forensics and Security* 11(2):328–343.

[13] Guan H, Kozak M, Robertson E, Lee Y, Yates AN, Delgado A, Zhou D, Kheyrkhah T, Smith J, Fiscus J (2019) Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. *Proceedings of Winter Conference on Applications of Computer Vision Workshop (WACVW)*, , pp 63–72.

[14] Yates AN, Guan H, Lee Y, Zhou D, Delgado A, Kheyrkhah T, Fiscus J (2019) Media forensics challenge 2019 evaluation plan. Available at https://www.nist.gov/system/files/documents/2019/03/12/mfc2019evaluationplan.pdf.

[15] Papadimitriou P, Dasdan A, Garcia-Molina H (2010) Web graph similarity for anomaly detection. *Journal of Internet Services and Applications* 1(1):19–30.

[16] Cao H, Abd-Almageed W (2020) Improving near-duplicate image cluster detection for provenance filtering. Available at https://mediforprogram.com/wiki/download/attachments/11508299/D2_1520_provenance_filtering_slides_for_2020_pi_meeting_USC-ISI.pptx?version=1&modificationDate=1587584942990&api=v2.

[17] Lee Y, Filliben JJ, Micheals RJ, Phillips PJ (2013) Sensitivity analysis for biometric systems: A methodology based on orthogonal experiment designs. *Computer Vision and Image Understanding* 117(5):532–550.

[18] Niu X, Kapoor R, Glavic B, Gawlick D, Liu ZH, Krishnaswamy V, Radhakrishnan V (2017) Provenance-aware query optimization. *Proceedings of IEEE International Conference on Data Engineering (ICDE)*, , pp 473–484.

[19] Buneman P, Tan WC (2019) Data provenance: What next? *ACM SIGMOD Record* 47(3):5–16.

[20] Stork L, Weber A, Miracle EG, Verbeek F, Plaat A, van den Herik J, Wolstencroft K (2019) Semantic annotation of natural history collections. *Journal of Web Semantics* 59:100462.

[21] Chen A, Wu Y, Haeberlen A, Loo BT, Zhou W (2017) Data provenance at internet scale: Architecture, experiences, and the road ahead. *Proceedings of Conference on Innovative Data Systems Research (CIDR)*, , .

[22] Morra L, Lamberti F (2019) Benchmarking unsupervised near-duplicate image detection. *Expert Systems with Applications* 135:313–326.

[23] Zhang C, Lin Y, Zhu L, Yuan X, Long J, Huang F (2019) Hierarchical one permutation hashing: efficient multimedia near duplicate detection. *Multimedia Tools and Applications* 78(21):30537–30560.

[24] Iuliani M, Fabbri G, Piva A (2015) Image splicing detection based on general perspective constraints. *Proceedings of IEEE International Workshop on Information Forensics and Security (WIFS)*, , pp 1–6.

[25] Huh M, Liu A, Owens A, Efros AA (2018) Fighting fake news: Image splice detection via learned self-consistency. *Proceedings of European Conference on Computer Vision (ECCV)*, , pp 101–117.

[26] Cozzolino D, Poggi G, Verdoliva L (2015) Splicebuster: A new blind image splicing detector. *Proceedings of IEEE International Workshop on Information Forensics and*

Security (WIFS), , pp 1–6.

[27] Chen C, McCloskey S, Yu J (2017) Image splicing detection via camera response function analysis. *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, , pp 5087–5096.

[28] Brogan J, Bestagini P, Bharati A, Pinto A, Moreira D, Bowyer K, Flynn P, Rocha A, Scheirer W (2017) Spotting the difference: Context retrieval and analysis for improved forgery detection and localization. *Proceedings of IEEE International Conference on Image Processing (ICIP)*, , pp 4078–4082.

[29] Bahrami K, Kot AC, Li L, Li H (2015) Blurred image splicing localization by exposing blur type inconsistency. *IEEE Transactions on Information Forensics and Security* 10(5):999–1009.

[30] Dias Z, Rocha A, Goldenstein S (2010) First steps toward image phylogeny. *Proceedings of IEEE International Workshop on Information Forensics and Security (WIFS)*, , pp 1–6.

[31] Rocha A, Piva A, Huang J (2012) The first ifs-tc image forensics challenge. *IEEE Inf Forensics Security Tech Committee* .

[32] Ng TT, Chang SF, Sun Q (2004) A data set of authentic and spliced image blocks. *Columbia University, ADVENT Technical Report* :203–2004.

[33] Dong J, Wang W, Tan T (2013) Casia image tampering detection evaluation database. *Proceedings of IEEE International Conference on Signal and Image Processing (ICSIP)*, , pp 422–426.

[34] Amerini I, Ballan L, Caldelli R, Del Bimbo A, Serra G (2011) A sift-based forensic method for copy–move attack detection and transformation recovery. *IEEE transactions on information forensics and security* 6(3):1099–1110.

[35] Christlein V, Riess C, Jordan J, Riess C, Angelopoulou E (2012) An evaluation of popular copy-move forgery detection approaches. *IEEE Transactions on information forensics and security* 7(6):1841–1854.

[36] Bansal A, Nanduri A, Castillo CD, Ranjan R, Chellappa R (2017) Umdfaces: An annotated face dataset for training deep networks. *Proceedings of International Joint Conference on Biometrics (IJCB)*, , pp 464–473.

[37] Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2018) Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:180309179* .

[38] Robertson E, Guan H, Kozak M, Lee Y, Yates AN, Delgado A, Zhou D, Kheyrkhah T, Smith J, Fiscus J (2019) Manipulation data collection and annotation tool for media forensics. *Proceedings of Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, , pp 29–37.

[39] (2013) Standard terminology for digital and multimedia evidence examination. Available at http://materialstandard.com/wp-content/uploads/2019/11/E2916-13.pdf.