Digital Investigation 22 (2017) S94–S105

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

DFRWS 2017 USA — Proceedings of the Seventeenth Annual DFRWS USA

Availability of datasets for digital forensics — And what is missing

Cinthya Grajeda, Frank Breitinger*, Ibrahim Baggili

Cyber Forensics Research and Education Group (UNHcFREG), Tagliatela College of Engineering, ECECS, University of New Haven, 300 Boston Post Rd., West Haven, CT 06516, USA

Keywords:
Availability
Data collection
Dataset
Origin
Experiment generated
User generated
Repository

A B S T R A C T

This paper targets two main goals. First, we want to provide an overview of available datasets that can be used by researchers and where to find them. Second, we want to stress the importance of sharing datasets to allow researchers to replicate results and improve the state of the art. To answer the first goal, we analyzed 715 peer-reviewed research articles from 2010 to 2015 with focus and relevance to digital forensics to see what datasets are available and focused on three major aspects: (1) the origin of the dataset (e.g., real world vs. synthetic), (2) if datasets were released by researchers and (3) the types of datasets that exist. Additionally, we broadened our results to include the outcome of online search results. We also discuss what we think is missing. Overall, our results show that the majority of datasets are experiment generated (56.4%) followed by real world data (36.7%). On the other hand, 54.4% of the articles use existing datasets while the rest created their own. In the latter case, only 3.8% actually released their datasets. Finally, we conclude that there are many datasets for use out there but finding them can be challenging.

Introduction

Research may or may not require datasets. For instance, if one wants to construct an e-mail parser, perform Android malware analysis or improve facial recognition algorithms, one would need access to e-mails, malware samples or facial images, respectively. On the other hand, creating an encryption scheme, post-quantum key exchange or side-channel attacks may not necessarily require a particular dataset. This article focuses on the former type of research. In order to produce high-quality research results, we argue that three critical features must be examined:

1. *Quality* of the datasets. This helps guarantee that results are accurate and generalizable. Researchers need data that is correctly labeled and similar to the real world or originates from the real world.
2. *Quantity* of the datasets. This ensures that there is sufficient data to train and validate approaches/tools which is especially important when utilizing machine learning techniques.
3. *Availability* of data. This is critical as it allows the research to commence and ensures reproducible results helping in improving the state of the art.

For instance, a comparison/improvement of results is only possible if the identical input data sources are used. Therefore, researchers either need access to the tool/algorithm or the data source. As test-runs can be time consuming and require familiarity with someone else's approach, one usually favors access to datasets. We therefore contend that is important to have easily accessible datasets. This was also pointed out by Penrose et al. (2013) who stated "in the scientific method it is important that results be reproducible. An independent researcher should be able to repeat the experiment and achieve the same results. […] Most research has been done with private or irreproducible corpora generated by random searches on the WWW."

The importance of available datasets is now also addressed by granting agencies, government and other three letter agencies. Precisely, "The Obama Administration is committed to the proposition that citizens deserve easy access to the results of research their tax dollars have paid for" (Stebbins, 2013). Consequently, some federal granting agencies now require a data management plan, e.g., NIST (2014). On the other hand, agencies sponsored online repositories such as the Computer Forensic Reference Data Sets (CFReDS, cfreds.nist.gov.[1]) from NIST or the Information Marketplace for Policy and Analysis of Cyber-risk & Trust (IMPACT, impactcybertrust.org) program from the Department of Homeland Security that "supports global cyber risk research & development by coordinating, enhancing and developing real world

* Corresponding author.
E-mail addresses: Cgraj1@unh.newhaven.edu (C. Grajeda), FBreitinger@newhaven.edu (F. Breitinger), IBaggili@newhaven.edu (I. Baggili).
URL: http://www.unhcfreg.com/, http://www.FBreitinger.de/, http://www.Baggili.com/.

[1] All links provided in this article were last accessed 2017-01-20.

http://dx.doi.org/10.1016/j.diin.2017.06.004

data, analytics and information sharing capabilities, tools, models, and methodologies."

In this work we analyzed a total of 715 cybersecurity and cyber forensics research articles from the years 2010–2015 from five different conferences/journals with respect to the utilization of datasets. We first categorized the dataset's origin (i.e., computer generated, experiment generated or real world), then analyzed its availability (i.e., if a dataset was released). Lastly, we examined the different kinds of datasets (e.g., malware, disk images, etc.).

Our findings illustrate that the majority of available datasets were experiment generated (over 1/2) and only around 1/3 originated from real world data. Furthermore, we show that researchers (re-)use available datasets frequently but when they have to create their own dataset, it is rarely shared with the community (less than 4%). Besides these findings, a major contribution of this work is a comprehensive list of available repositories/datasets which may be employed in research and are summarized on http://datasets.fbreitinger.de[2] (a less comprehensive version of our findings is provided in Appendix A). Secondly, we provide an overview of the top 7 used in Table B.6 (in Appendix B).

Limitations

All of our data analysis was performed by manual inspection. We note that human error might have been introduced, but we attempted to alleviate the errors by conducting multiple runs. Due to time constraints, our dataset of research articles included only papers from 2010 up to 2015 from selected venues and does not include every single paper published worldwide in the cyber forensics domain. We do however believe that our research paper dataset is representative in both breadth and depth. We argue that our results are still applicable and our findings paint the picture of the state of the domain with regards to datasets.

Related work

Our study was inspired by Abt and Baier (2014) who published an article named availability of ground-truth in network security research. In their article, the authors analyzed 106 network security papers over four years (2009–2013) and concluded with three main findings: (1) many researchers manually produced their datasets, (2) datasets are often not released after the work is completed and (3) there is a lack of standardized datasets that are labeled that can be used in research. These weaknesses combined, produced one of the major disadvantages facing the cybersecurity/forensics community to this day, which is low reproducibility, comparability and peer validated research.

Penrose et al. (2013) (as mentioned in the introduction) and Fitzgerald et al. (2012) also argued that it is poor common practice to perform research and not publish the underlying dataset. Another example comes from Axelsson (2010) who stated that it is "difficult to compare the results we obtain with previous results, since the data was not available for comparison". To encourage comparative research in the field, he performed his experiment on the open Digital Corpora (see next paragraph). Hence, researchers that want to validate the study can access the dataset. Additional datasets from their work were also made available upon request. A proactive approach was taken by Garfinkel et al. (2009) who outlined the restrictions put on forensic research due to the lack of freely available, standardized datasets. Consequently, Garfinkel lead the creation of the Digital Corpora (digitalcorpora.org) – one of the first free online dataset repositories for digital forensics. Despite its popularity, it seems like the platform is no longer updated – at the time writing, the last post was from September 2014.

Methodology

While this work was influenced by Abt and Baier (2014), the difference between both studies is that we do not exclusively focus on network traffic but on all kinds of datasets that may be useful for cybersecurity/forensics research, e.g., malware, disk images or memory dumps. Moreover, our study expands to a broader number of articles, results from Google searches and provides an overview of existing datasets. To analyze the availability of datasets which we define in Sec. Definition of a dataset, we first investigated peer-reviewed articles from several conferences/journals and then performed online searches. The details of both steps are discussed in Sec. Analyzing peer-reviewed articles and Sec. Online searches, respectively.

Definition of a dataset

For this work we define a dataset as a collection of related, discrete items that has different meanings depending on the scenario and was utilized for some kind of experiment or analysis. For instance, valid datasets would be but are not limited to files, memory dumps, raw images, pcap files, log files, outputs from `/dev/urandom` that were analyzed/processed. In contrast, here are some examples that we did not consider as datasets: an input that was only used to measure runtime efficiency, results written to log files, or a tool that outputs data which is never used.

Analyzing peer-reviewed articles

The first phase entailed the collection and analysis of publications from digital forensics and security conference proceedings as well as journal publications[3] spanning six years (from 2010 to 2015). The decision for these conferences/journals was based on our familiarity, experience, access to articles and quality of the venue (which may be considered subjective). For each article utilizing a dataset, we asked the following questions:

1. **Origin of datasets**: Is the dataset *computer generated* (e.g., an algorithm, bot, `/dev/urandom`), experiment generated (e.g., a user creates specific scenarios) or user generated (e.g., real world data). Results are discussed in Sec. Origin of datasets.
2. **Availability of datasets**: Are datasets available to the community?
   - Was the utilized dataset available prior to the research? (re-usage)
   - If the dataset was created, was it released? (availability)
   - If the dataset was available prior to the research, is the origin disclosed/is it freely available? (proprietary to one 'group')

   Findings are presented in Sec. Availability of datasets.
3. **Kinds of datasets**: What datasets exist and can be used by researchers?
   - Were any third party databases, services or online tools used in the creation of datasets?

[2] If you want to contribute, please submit your dataset information to the authors by using the contact form on the website.

[3] The following conferences were examined: IEEE Security and Privacy, Digital Forensic Research Workshop (DFRWS – USA, EU), International Conference on Digital Forensics & Cyber Crime (ICDF2C), and Association of Digital Forensics, Security and Law (ADFSL). The following journal was looked at: Digital Investigation.

The results of this question are shown in Sec. Kinds of datasets.

4. **What is missing**: What datasets or other things are currently missing? This will be addressed in Sec. What is missing.

Additionally, we collected the following information (when possible): publication name, author(s), conference/journal, published year, dataset description, dataset size, method of gaining access to the dataset, and the dataset's location (URL).

Online searches

For the second phase, we worked in reverse order and queried Google for available datasets/repositories that may have not been used to their full potential in our field or appeared in any articles that we had analyzed. We specifically used four queries related to the following: 'available digital forensics dataset repositories', 'available cybersecurity and forensics dataset repositories', 'available malware dataset repositories', and 'available computer dataset repositories'. In our analysis, we focused on the first 100 results for each query. Once a repository/dataset was identified, we gathered data similar to ones found referenced in academic articles. Additionally, we attempted to identify where possible, articles that had already used such datasets/repository or that had analyzed such data in some manner. The results are shown in Sec. Datasets found through Google research.

Results overview and origin

A total of 715 articles were analyzed in this study from conferences and journals listed in Sec. Analyzing peer-reviewed articles where approximately 49% employed datasets. Our analysis started with the conference proceedings of IEEE Security & Privacy (S & P) where 76 out of 240 ($\approx$32%) articles utilized datasets. Thus, the majority of the articles did not involve datasets as they focused on studies informing the community about standards, techniques, policies and laws but also about topics on programming, algorithms, cryptography, hardware and system flaws, etc. Given the fairly small number of articles utilizing datasets in S & P, we surveyed the digital forensics domain where we hypothesized more datasets would be employed. Our starting point was the Digital Forensic Research Workshop (US & EU) which yielded 78 out of 91 (~86%) articles that included datasets. Due to the significantly higher adoption of datasets in the digital forensics domain, the remaining analysis focused on conferences/journals that embodied digital forensics as a main thematic topic. In summary, we found the following ratios: (i) International Conference on Digital Forensics & Cyber Crime (ICDF2C) had 60 out of 107 that used datasets; (ii) Association of Digital Forensics, Security & Law (ADFSL, Conference) contained 29 out of 87 articles that utilized datasets; and (iii) Digital Investigation (Journal) contained 108 out of 190 articles that employed datasets.

Origin of datasets

The first aspect we analyzed was the origin of the datasets and how they were created. A summary of our findings is shown in Table 1 which will be discussed throughout the upcoming subsections. Note, the 'mixed sets' row holds articles we could not mark with a single category. For instance, Mohamed and Yampolskiy (2012) developed a new facial recognition approach and their tests were executed on the ORL dataset (now known as the database of faces) as well as two sets of avatar images. Given that 'mixed sets' represented only a small part of all the utilized datasets, we focused our analysis on datasets marked with a single category.

Table 1
Overview of the origin of the 351 identified datasets out of the 715 analyzed articles.

Articles	Total	
Experiment generated	56.4%	198
User generated	36.7%	129
Computer generated	4.6%	16
Mixed sets (user, experiment & computer)	2.3%	8

Experiment generated datasets

Over half of the datasets found in this study were experiment generated, where researchers created specific scenarios to conduct their experiments. There are several reasons for having such a heavy shift towards this kind of data. First, in many cases, there is a lack of real world datasets available to the digital forensics community (Baggili and Breitinger, 2015). Another reason is that using experiment generated data allows researchers to test and verify such data, especially when conducting experiments on new technologies as that is common within the area of cybersecurity and digital forensics (Garfinkel et al., 2009). For instance, Lee et al. (2014) investigated the possibility of stealing webpages from the browser by exploiting vulnerabilities of Graphical Processing Units (GPUs) where memory dumps from both attackers and victims were created and collected.

User generated datasets

With over 36%, user generated datasets (a.k.a. real world datasets) were the second most used type of data. According to Baggili and Breitinger (2015), experimenting on real world data is crucial for developing reliable algorithms and tools – "how can we learn from our past when we do not have real, accessible data to learn from?" One of the major reasons is clearly copyright and privacy laws which prohibit sharing with the community (Abt and Baier, 2014). If real world data was used, we found the following different origins:

Dataset was released: A prominent example of a real-world dataset is the Enron e-mail dataset (A.4.6.1)[4] which was posted online by the Federal Energy Regulatory Commission after its investigation and later on purchased by the Massachusetts Institute of Technology (MIT). Eventually, the dataset was stripped of private user information and e-mail attachments to avoid violating user privacy rights.[5] Note, this was one of the most frequently used sets.

User data was collected before research: Some institutions (especially Universities) collected real world data upfront, e.g., from students where the individual signs an agreement and then researchers capture the desired information. One example is Spam Data Mine, a research project under The Center for Information Assurance and Joint Forensics Research (CIS-JFR), which generates information about currently on-going campaigns by spammers. It archives spam e-mails received from numerous sources and honeypots, and collects approximately 1 million spam e-mails each day (Khan et al., 2014). Another example is work by Guido et al. (2016),[6] that investigated user behavior on mobile devices over a three month period. Mobile devices were handed out to college students after following

[4] These references will be found throughout the paper and refer to the overview tables in Appendix A. Precisely, there is more information at point (A.3.6.1) in the ref-column.

[5] https://www.technologyreview.com/s/515801/the-immortal-life-of-the-enron-e-mails/.

[6] Note, this article was not part of our study but we believe it provides a good example.

human review board practices and the mobile phone images were then collected to be used by researchers at MITRE.

Collaboration with law enforcement: According to our results, there were at least eight collaborations between law enforcement agencies and academia. One example is Rogers and Seigfried-Spellar (2014) where law enforcement asked researchers to investigate Internet artifacts from a suspect's Internet Browsing History/Bookmarks to identity any trends in pornography use. Liberatore et al. (2010) provided a second example where they developed a tool named RoundUp for law enforcement to analyze peer-to-peer (P2P) file sharing networks. "Using leads and evidence from RoundUp, a total of 558 search warrants have been issued and executed during that time."

Source of data is online: A significant amount of data is also publicly available online although the intent was never research. For instance, the Apache Mail Archives (A.4.6.3) is a mail archive used by Armknecht and Dewald (2015). Other examples are applications through Google Play, Twitter, YouTube or Google images where one can access real world data. Within digital forensics, the Digital Corpora online repository is popular as it offers disk images purchased from various countries in the world, files crawled from the Internet and so on.

One may argue that there are more than the four aforementioned categories or that a set falls into several classes. For instance, Drebin (A.4.7.3) is a collection of over 5500 Android malware applications collected from disparate sources. Thus far, this source was only used once based our research article analysis, but overall according to their website, it has been utilized by at least 157 universities and organizations around the world. Other examples are from the National Institute of Standards & Technology. They provide massive collections of data across these categories, e.g., the National Software Reference Library (NSRL, nsrl.nist.gov) which was leveraged by Rowe (2013) and is a list of over 100 million hashes of applications; or the National Vulnerability Database nvd.nist.gov which was utilized by Liu et al. (2014).

Computer generated datasets

The final category is computer generated datasets or synthetic data which may have several origins, e.g., an algorithm, bots, `/dev/urandom` or simulators. Our analysis revealed that almost 5% of the analyzed articles employ those datasets which is not necessarily a surprise – often researchers in digital forensics want to solve real world problems and therefore cannot use simulated or generated data. One argument for generated data is the exact knowledge of the ground truth. For instance, Breitinger et al. (2014c) utilized pseudo-random data from `SecureRandom.random_bytes` to analyze the precision & recall rates of approximate matching algorithms. Their challenge was that no labeled real world data existed and therefore the only possibility was generated data. Other forms of computer generated datasets could be acquired from websites such as Creative Crash[7] and CG Society.[8] For instance, Farid and Bravo (2012) used these types of datasets to measure the difference between real and computer generated images.

Usage of third party databases, services or online tools

In our research, we realized that about 20.4% (39/191) articles used third party databases, services or online tools to retrieve information. For instance, Conti et al. (2010) created fragment data using random data from the website random.org. The motivation for the researchers was the high quality of the random data; it is based on atmospheric noise. Others that employed databases and services are Al-Shaheri et al. (2013) who queried openMalware.org to acquire malware for their research and Nappa et al. (2015) who utilized virusTotal.com to scan for malicious files.

Availability of datasets

The second part of our study analyzed the availability and re-use of datasets. A summary of our findings is depicted in Table 2 and will be discussed in the following subsections.

Creating vs. re-using datasets

The *first row* in Table 2 provides an overall summary and indicates that 45.6% of the articles analyzed produced their own datasets in their experiments while 54.4% of the articles utilized datasets that existed (re-use of an existing set). This almost equal-share seems reasonable as researchers often train algorithms based on simulated/experiment data while on the other hand for evaluating performance/comparing two algorithms often real world datasets are favored, e.g., Laskov et al. (2014) used the Contagio (A.3.2.1) malware sets to measure and compare algorithm accuracy.

Coming to the high usage of self-made datasets, some researches clearly stated they were required to create their own dataset since nothing was available. For instance, Roussev and Quates (2013) created the msx-13 corpus (A.4.8.3) because they could not find an appropriate sample for their experiment. This corpus contains 22,000 MS Office 2007 random files (e.g., docx, xlsx, pptx) crawled from the Internet. This indicates that researchers re-use datasets if they are available and do not necessarily favor building their own. Similar to the introduction and Penrose et al. (2013)'s statement, Fitzgerald et al. (2012) mentioned that "most of the previous work on this problem [file fragment classification] exclusively uses private datasets, making it more difficult for other researchers to reproduce experimental results." Both recommended that all studies should use the freely available sets from Digital Corpora (A.4.8.5) in order to encourage friendly competition and aid other researchers interested in reproducing results.

Currently available datasets

The current availability is discussed in the *second row* of Table 2 – only 29.0% (102) of all sets are available for research and thus allow reproducible results. The vast majority (96) of the sets already existed where on the other hand only 3.8% of the newly created ones were released. Examining the origin of currently available sets revealed, that 59.8% (61/102) employed real world datasets. Subsequently, 38.2% of available datasets were recognized as experiment generated and 2.0% as computer generated datasets.

Table 2
Results of 715 analyzed articles with 351 containing datasets.

Articles	Total	
Created through research	45.6%	160/351
– Existed prior to research (re-use)	54.4%	191/351
Currently available sets	29.0%	102/351
– Existed and available (re-use)	50.3%	96/191
– Created and released	3.8%	6/160
Exist and not available	29.3%	56/191
Available as services[a]	20.4%	39/191

[a] This was discussed in the Sec. Usage of third party databases, services or online tools.

[7] https://www.creativecrash.com/.
[8] http://www.forums.cgsociety.org/.

Most of the *user generated datasets* originated from four different major online repositories, ranked in descending order (Digital Corpora, Enron E-mail Dataset, the t5-corpus (A.4.8.4) and Android Malware Genome Project (no longer available)). Subsequently, the majority of *experiment generated datasets* originated from four main repositories, including the already mentioned Digital Corpora (this repository contains both kinds).

One example is the M57-patents scenario by Woods et al. (2011) which offers a variety of experiment generated dataset samples (e.g., RAM data, e-mails, disk drive images, etc.). This is followed by The CFReDS Project, BOSS (A.3.3.1), and Digital Forensics Tool Testing Images (DFTT).[9] On the other hand, articles that used *computer generated* datasets were the most scarce and generally relied on random.org or `/dev/urandom` to create the data. That being said, we were only able to identify the particular source but not the dataset itself, e.g., random.org will produce a different output for every query. Furthermore, only 3.8% of newly created sets were released for future research where almost all identified sets were classified as experiment generated datasets with the exception of one, classified as user generated.

Non available datasets

This section focuses on datasets that exist but were not available. Specifically, we discovered 29.3% (56/191) articles with datasets that we were unable to verify and classify as currently available. We organized this set of articles into three groups:

Source is unknown: With a total of about 39.3% (22/56), this is the most common reason for dataset unavailability. This is a major problem because not knowing the source of the datasets may raise questions about the quality and integrity of such data. Moreover, it completely hinders researchers from reproducing experimental results. Two examples that did not clearly state where the data originated from are Boukhtouta et al. (2015) and Brown (2011). They used specific services online to acquire datasets. However, the articles failed to specify if the acquired data was freely given to them or if a fee was required for such service. We rated this as unavailable as funds in research (especially in academia) are limited.
Source has privacy restrictions: The second most common reason with over 46.4% (26/56) for not releasing the datasets were privacy concerns as discussed in Sec. User generated datasets. In other words, these were mostly real world datasets generated by Universities, Government agencies and law enforcement and could not be released.
Source not accessible: About 1/7 of the articles had accessibility problems, such as temporarily unavailable, download link broken or not maintained anymore. For instance, Zhou and Jiang (2012) created and maintained a highly utilized free Android Malware Genome Project[10] (according to their website it was shared with 460 entities). However, the repository is not available anymore as the students graduated. Note, we expect the number of 'source not accessible' to grow for older articles (recall in this study we focused on 2010–2015).

Kinds of datasets

This section clusters the identified datasets together. Therefore, we created sections for each of the major dataset types we found and briefly summarize what was found. Since some sources (e.g., Digital Corpora) is a collection of datasets, the reference to Digital Corpora will show up in multiple sections. The subsections contain datasets that were directly used in research or found to be utilized within the given sources, e.g., if a source provides sets *A*, *B* and *C*, but only *A* was used in an article, we will still name *B* and *C* in the appropriate subsection.

In summary, we found over 70 different datasets though our article analysis and organized them in 21 categories with major ones discussed in the following subsections. Each subsection will provide references/links to the available datasets, and provide a brief overview, e.g., origin, amount of samples, total size, etc. (when obtainable). Additionally, we provide our detailed results in Appendix A; the latest version of the datasets' table can be found on the project website.

Malware datasets (computer and mobile)

In total, seven real world data online repositories were found throughout this study that offer computer and mobile malware samples (note, there are additional 'services' as mentioned in Sec. Usage of third party databases, services or online tools).

Android. In total, three repositories were frequently used. (1) *Drebin* (Arp et al., 2014) is a collection of 5560 Android samples from 179 different malware families collected between 2010 and 2012 and was used by Talha et al. (2015) to test permission based malware detection. (2) *Contagio Mobile Mini-Dump* (A.4.7.1) is part of the larger computer malware repository Contagio Malware Dump. In contrast to other repositories, this website is more like a traditional blog with an upload/download functionality. Thus, users can download the repository but also extend it. According to the website, there are over 200 malware posts and each post might contain more than one malware sample, collected from 2011 to 2016. Lastly, (3) Jang et al. (2015) possess a dataset (A.4.7.2) of 9990 malware samples which can be requested for research purposes. Part of this dataset included samples from the repository Contagio Mobile Mini-Dump and Virus Share (A.3.2.2) (exact amount not mentioned in article).
Computer malware. In total, four repositories were utilized in the analyzed articles: (1) *Contagio Malware Dump* is similar to its counterparts and has around 400 posts. (2) *VX Heaven* (A.3.2.3) which is a virus information website that contains over 271,000 computer malware samples. However, it is unknown how often the website is updated and as the website states, the last time the malware collection was scanned was by Kaspersky Anti-Virus in 2006. (3) *Virus Share* which was the most comprehensive malware collection that was referenced with over 27 million samples. Although not stated, it seems that this repository is a mix of mobile and computer malware. Additionally, it is one of the most updated sites with new entries every month. Consequently, this malware site is one of the most secure in relation to the acquisition of malware since access to the site is by invitation only. If access is needed an e-mail is required to be sent to the admin stating reasons to be added. Lastly (4), the forum *KernelMode.info* (A.3.2.4) was mentioned by Al-Shaheri et al. (2013). According to the post dates which range from 2010 to 2016, this forum seems still active but registration is required. Unfortunately, the amount of malware samples in this forum is unverifiable but it seems to have a mix of mobile and computer malware as well.

E-mail datasets

In total, three e-mail datasets were found. The Enron E-mail Dataset version 2015 introduced by Schmid et al. (2015) which

[9] http://dftt.sourceforge.net/.
[10] http://www.malgenomeproject.org.

consists of over 619,000 real world messages belonging to 158 users. Besides that, Armknecht and Dewald (2015) used about 75,724 real world e-mails from the Apache online e-mail repository which was never intended to be a dataset but provides real world examples. Lastly, we found about 12 e-mails in Digital Corpora's experiment generated scenarios which however were never utilized.

File sets/collections

File sets are collections of files with various types like text, html, pdf, doc, ppt, jpg, xls, gif, zip or csv. They are frequently used for different purposes (e.g., to test/improve forensic file formats like AFF4 (Schatz, 2015)). The most prominent and comprehensive dataset may be the GovDocs1 corpus from Digital Corpora which consists of ~1 million documents gathered by crawling the .gov domain. Given that massive size, a common subset is the t5-corpus which was created by Roussev (2011) and contains 4457 files of various types and is commonly used for testing approximate matching, e.g., by Breitinger and Roussev (2014). Lastly, Roussev and Quates (2013) also created the msx-13 corpus which contains 22,000 MS Office 2007 user generated random files (e.g., docx, xlsx, pptx) crawled from the Internet.

RAM dumps

Our study found six repositories having over 90 dumps where all of them were experiment generated (obviously RAM cannot be fully controlled and therefore it can be considered as a mixture of user and experiment data). The first set was published by Minnaard (2014) where the authors acquired their own RAM data from different operating systems and devices. The authors state the complete RAM archive is available on request, but a sample with over 1 GB of data can be downloaded (A.4.9.1). A second set consisting of five 1 GB RAM dumps (Windows, 2000, 2003, Vista Beta 2, and XP) is provided by the CFReDS Project (A.4.9.3). According to the website, the "systems were not engaged in any malicious or even network based activity at the time of imaging." Two more dumps of WinXP 32-bit machines were released by the DFRWS' forensic challenge (A.4.9.2). Another experiment generated dataset which was used by Case and Richard (2015) originates from The Art of Memory Forensics book (Ligh et al., 2014) and can be downloaded from the corresponding website (A.4.9.4). This single dump has a size of 3.8 GB. Lastly and the most comprehensive collection of memory dumps with 88 samples and a total size of over 44 GB can be downloaded from Digital Corpora (A.4.9.6).

Images of computer drives

Especially in digital forensics, complete disk images are valuable to create and test tools as well as procedures. Leading the way is the Real Data Corpus (RDC) from Digital Corpora which according to their website[11] "is a collection of raw data extracted from data-carrying devices that were purchased on the secondary market around the world." As of 2011, the non-U.S corpus contained 1289 hard drive images ranging in size from 500 MB to 80 GB. According to Garfinkel et al. (2009) there is also a U.S RDC which contains 1228 hard disk images, however, we could not locate it on the website nor does it say anything about it at the time of writing. A second but way smaller set is provided by the CFReDS Project (A.5.15.3) which contains three images extracted with different imaging tools (Encase, iLook, & Compressed dd). The original image was made with 5 partitions (OS Extended Journaling, OS Extended, another OS Extended, OS Standard & UNIX File System) created on a MAC OS X. According to the website, the purpose of having images extracted from 3 different tools was to test if those tools would recognize the file systems created on the Mac OS X.

Images of other devices

Besides hard-drive images, we found a series of other images which will be briefly explained in the following:

Cell Phones: In total, we found 26 images within the two repositories CFReDS (A.4.11.1) and Digital Corpora (A.4.11.2). The former one contains 14 images; 7 from a Nexus One and 7 from a Nexus S-1 while the latter one has 12 images from Black Berry Torch 9800, HTC One V, iPhone 3GS and the Nokia 6102i.
Gaming systems: Although there are a variety of consoles out there which get analyzed, we only identified 2 sets with Xbox images. The first one 3.1.1 was released by Moore et al. (2014) and according to them it was released so the "forensic community may expand upon our work". The second one 3.1.2 came through the nps-2014 XBox-1 scenario comprising of 4 disks; 2 originals and 2 modified by experiments. No other game console image was found.
SIM card: SIM card images were not utilized in any article, nonetheless, we discovered at least 3 images in the CFReDS (A.4.14).
Apple iPod & Tablet: Although not utilized in any of the articles, Digital Corpora offers a total of 10 iPod disk images (A.5.18) and 25 disk images of various tablets (A.5.19) (brands not disclosed).
Flash Drives: As far as real world flash drive images go, Digital Corpora offers a total of 643 flash images (e.g., USB, Memory Stick, SD and other), with sizes from 128 MB to 4 GB with real world data. Furthermore, it offers the nps-2009-canon2 (A.5.16) and nps-2013-canon1 sets which is a collection of 7 images of 32 MB SD cards which were used by Lambertz et al. (2013) & Garfinkel et al. (2010) for testing image/picture carving tools.

Network traffic

This section summarizes a variety of different network traffic sources which include PCAP files acquired through tools such as Wireshark or logs (i.e., port and protocol data, IP and operating systems source information and so on). The following datasets were found through our study: The first set was generated for the DFRWS 2009 forensic challenge (A.4.12.2) and thus contains experiment generated PCAP files where most of the traffic is HTTP traffic on port 80. A second shared PCAP dump (A.4.12.3) was created by Karpisek et al. (2015). The dataset was compiled by the researchers for the purpose of acquiring WhatsApp traces that they were able to decrypt. The dataset is comprised of 3 PCAP files containing WhatsApp register and call traffic. A wireless network repository named CRAWDAD was discovered in our study (A.4.13) from which datasets of mobility traces of taxi cabs in San Francisco were acquired. This website also contains hundreds of other types of wireless network traffic (e.g., TCP traces, Bluetooth, accelerometer, 802.11p packets, etc.) released since 2002.

Scenarios/cases for analysis

We identified three scenarios or cases for analysis. The first one is the nps-2009-domexusers on Digital Corpora which is a disk image of two users (domexuser1 and domexuser2) who communicate with a third user (domexuser3) via IM and e-mail. The disk image is of a Windows XP SP3 system (NTFS format and used twice in our study). The second comprehensive scenario is the 2009-m57-patents created by Woods et al. (2011) for digital forensics

[11] http://digitalcorpora.org/corpora/disk-images/real-data-corpus.

and security educational purposes. According to the website, the "scenario tracks the first four weeks of corporate history of the M57 Patents company". It consists of redacted drive images, USB drive images, RAM Images, network traffic and documentation. While this scenario was originally designed for education purposes, it was also utilized by Garfinkel and McCarrin (2015)'s experiment where it served as sample input to test hash carving techniques. The last scenario consists of three network log traces plus a USB device image from the CFReDS Rhino Hunt scenario. Additionally, this source comes with a answers.pdf which allows to fully understand the scenario.

Mixed and others

Some sets that could not be classified are summarized in the following:

Pictures: Besides finding a great amount of real pictures, we also found computer generated graphics and forged images tainted with steganography. Some of these datasets come from websites such as 'Break our Steganography System' (BOSS, A.3.3.1), which hosts a challenge that contains a testing database of 1000 512×512 pgm greyscale images and a training database of 9074 cover images.
Language corpus (*text*): Language corpora are often used for Statistical Machine Translation. A common collection is the European Parliament Proceedings Parallel Corpus 1996–2011 (A.3.5.6) which contains about 21 European language versions and 60 million words per language.
Chat logs: The dataset (A.5.20) is comprised of 1100 chat logs from 11,143 chat sessions from a single computer and recorded between 2010 and 2012 using Messenger Plus!.
Password lists: These sets are commonly used for probabilistic password research such as work by Ma et al. (2014). Some comprehensive dictionaries are listed on a security wiki page (A.5.21) and have millions of leaked passwords from websites such as RockYou, Myspace, and Hotmail. According to this website, these datasets are useful "to generate or test password lists". Note, any type of private information such as name or e-mail is redacted.

Datasets found through Google research

While the previous sections focused on articles only, this section summarizes the results from our Google searches (see Sec. Online searches). Overall, we identified ten sources providing datasets: Four of the sources are websites provided links to other online repositories and six sources pertained to network traffic, text files, and machine learning data. Note: only a few of the sources found were chosen to be discussed in this section, however, the rest of them can be found in our website.[12]

Security Repo: secrepo.com is a comprehensive list of samples of security related data. As stated on the website, "this is my attempt to keep a somewhat curated list of Security related data I've found, created, or was pointed to". This source contains about 100 links to datasets or third party references. This includes samples of networking scanning/recon, shell traffic, security incidents, system logs, ssl certs, malware, and more.
Note, the following three repositories were only found through this website. Our Google search did not lead us to either of them which shows how cumbersome finding repositories can be.

Mid-Atlantic Collegiate Cyber Defense Competition (*MACCDC*): netresec.com has PCAP files of three MACCDC competitions from 2010 to 2012 which comes to a total of 59 PCAP files where the 2010 competition was analyzed and summarized by Carlin et al. (2010). Additionally, this website includes links to other websites hosting cyber challenges, malware datasets, networking traffic, etc.
The Cyber Systems and Technology Group of MIT Lincoln Laboratory[13]: According to the website, this is "the first standard corpora for evaluation of computer network intrusion detection systems" which was collected by MIT Lincoln Laboratory. The three datasets (from 1998 to 2000) are composed of file system dumps, pcap files, NT event log audit data, outside TCP dump Data, as well as "the first formal, repeatable, and statistically significant evaluations of intrusion detection systems". The 1999 evaluation dataset was also analyzed by Mahoney and Chan (2003).
The Black Market Archives[14]: As its name implies, this data was acquired from Dark Net Markets (DNM) usually hosted in Tor hidden networks. The DNMs operate on selling and buying drugs, guns, and any other type of illegal or government regulated goods. The author of the site claims he collected 1.6 TB of data comprising 89 DNMs from 2013 to 2015; we found 15 papers that have cited the website/dataset.
Malware samples[15]: This personal website lists about 12 links directed at other malware repositories/services like malshare.com or thezoo.morirt.com. The former one is an open source malware repository that permits users to download 1000 samples per day with a requested public API Key (if more samples are necessary, it requires to contact the admin). The second website is a malware repository which aims at collecting all versions of malware available for download directly from the site with no restrictions.
PeekaTorrent: peekatorrent.org contains about 3.2 billion hash values from 2.65 million torrent files totaling 66 GB of compressed data (84 GB raw) and was collected by Neuner et al. (2016).
Impact Cyber Trust: Sponsored by the U.S. Department of Homeland Security (DHS) and other technology and cybersecurity organizations, this website hosts a central database of ground truth and synthetic data available for research. The data provided was donated by at least 10 organizations and ranges from 2009 to 2016, some of them include, Georgia Tech, Packet Clearing House, etc. Note, most of the datasets relate to network traffic (e.g., IDS/Firewall, DNS, IP, BGP routing data, etc.).

What is missing?

Our study shows that many researchers prefer not to *share* their datasets which could be for several reasons. Note, the following are our assumptions and feedback that we received from two authors that we asked for the reason(s) why the datasets were not released when the article was published and if they were willing to share those datasets with the community if asked (A comprehensive survey study is necessary to verify the feedback we received).

First, researchers may not have the capability of sharing the set (e.g., the dataset is too comprehensive and one does not have the online resources available) which could be solved by a centralized, community based repository (see Sec. Centralized repository). For instance, some authors said that 'at the time of publishing, we did

[12] http://datasets.fbreitinger.de/.

[13] http://www.ll.mit.edu/ideval/data/index.html.
[14] http://www.gwern.net/Black-market%20archives.
[15] https://zeltser.com/malware-sample-sources/.

not have a stable platform through which we could provide access to our data'. Furthermore, they also faced the problem of collecting data (images in this case), so they agree about the worth of dataset sharing in research communities. They also would be willing to share upon request.

A second factor may be related to privacy concerns as discussed in Sec. Data de-identification research. Thirdly, researchers might simply not have thought of the importance of sharing their data. This was noted from feedback we received from a researcher we queried that said 'initially I did not exactly have in mind how important it was to curate and share such data'. As far as sharing this specific paper's datasets the answer was, 'I probably wouldn't want to share them (at least not in a publicly accessible manner) because when I picked the content off the Internet, I didn't take into consideration that there might be some privacy or copyright issues that may come up'. This author also agrees with our thoughts – making datasets publicly available is definitely important.

Lastly, we believe that many researchers do not want to share their datasets for intellectual property reasons. They view the ownership of the dataset as a way of having something that other researchers do not have. Besides sharing, we identified some additional shortcomings discussed in the following subsections.

Variety

While we found a good amount of sets online, this study also revealed on what is missing in regards to actual datasets. For instance, despite published work, we could not find samples of PlayStation Vita and the PlayStation 4 although they have been used in crimes, e.g., a PlayStation might have been used during the ISIS Paris attack (Tassi, 2015). A second group of devices we could not find data for were Smart-TVs. Coming to a world where everything is connected (IoT), there are many more devices we should try to acquire data from, e.g., Unmanned Aerial Vehicle (UAV), streaming devices, such as Roku or Apple TV.

Updates and upgrades

Having a closer look revealed that there are massive differences in the number of items per dataset, e.g., while there are 27 million malware samples, we only found 26 smartphone images. However, smartphones are frequently used and require extensive research (e.g., recall the San Bernardino iPhone case.[16]). A second aspect is the age of the datasets. While some sets like files are timeless (to a certain extend), other require frequent updates and need to be maintained, e.g., malware or smartphone images. For example, the 2009-domexusers scenario used by Garfinkel et al. (2010) includes disk images of a Windows XP SP3 operating system. On the other hand, we did not find any Windows 10 images. It looks like whenever a first dataset is released, researchers stop releasing new sets/samples to expand existing corpora. In fact, besides malware and network traffic which we have found to have the most up to date datasets out there, no other dataset found was being completely and continuously updated.

Centralized repository

We believe that the community is missing a *single* centralized, maintained and well organized repository. Our study showed that whenever a repository is created (e.g., Digital Corpora, CFReDS, Virus Share or Impact Cyber Trust) it is appreciated and frequently used by researchers. However, often these repositories are not maintained and become outdated. For instance, the Digital Corpora was updated the last time in 2014; the Android Malware Genome Project (Zhou and Jiang, 2012) announced after 3.5 years "due to limited resources and the situation that students involving in this project have graduated, we decide to stop the efforts of malware dataset sharing." We see a possible solution in either a government funded endeavor (as started by the DHS with their impact project) or managed jointly by the complete community (e.g., a 'github' of datasets).

Data de-identification research

One of the main problems impeding datasets from being released is privacy and proprietary concerns. We believe that this could be addressed by expanding research in the domain of de-identification as pointed out by Garfinkel et al. (2009). If we find ways to un-personalize data by removing, changing or manipulating names, phone numbers, addresses, and other personalized data, datasets could be shared and utilized for research. There are already guidance methods provided by HIPPA (Office of Ethics and Compliance, 2016) for de-identification of data.

Strategies to share complex data

As we are moving more and more into the cloud (Platform as a Service, Software as a Service), we need strategies on how to share this kind of data among researchers. In order words, how can we ensure that results are reproducible by other researchers if it takes place in a cloud environment. Our study discovered at least 25 articles that focused on cloud research. Some articles targeted areas on how to acquire and analyze data from Apple's iCloud, targeted ways on how to build trustworthy cloud systems for storing criminal evidence, and methods on how to discover illegal sharing of copyright materials over the cloud, e.g., Google Drive or DropBox. Others, for instance, Dykstra and Sherman (2012) or Pichan et al. (2015) specifically focused on the forensics aspect and offered options on how to acquire and share datasets. Nonetheless, none of the articles mentioned offered any datasets acquired through their investigations.

Publisher support

Lastly, sharing secondary information (i.e., datasets) is mostly not well supported by publishers. A step into the right direction would be to enable sharing data or even force researchers to submit secondary information. For example, in journals in Elsevier or IEEE, a dataset may be attached to a paper similar to what third party like researchgate.net do.

Discussion

Research that requires datasets currently faces several challenges as data is barley shared among the community. Our results show that less than 4% shared their dataset while on the other hand almost 50% make use of existing datasets. In other words, whenever a repository or a sophisticated dataset is available, researchers appreciate and utilize it. Beside the lack of sharing datasets, maintenance and availability are major issues. Many repositories/datasets are outdated and not maintained. Given that they are spread throughout the Internet, single individuals might be responsible for maintaining which is simply not feasible. As pointed out in Sec. Centralized repository, we believe that this could be solved through a centralized and community based repository, e.g., a github for datasets where everyone can share datasets. Another challenge is the availability of real world data which is of

[16] Details about the case can be found at http://www.cnbc.com/2016/03/29/apple-vs-fbi-all-you-need-to-know.html.

importance for researchers to produce high quality results—only about 1/3 of the datasets originated from real users. In order to allow reproducibility, improvements and faster research progress, we believe the mindset of researchers need to change and data should be released. Besides the aforementioned points, this will also enable competition and then ultimately lead to better results.

Conclusion & future work

For this article we analyzed 715 research articles and performed Google searches to summarize the availability of datasets for the community. While this study comes with a comprehensive list of available datasets and repositories which can be leveraged by researchers, we also show that there is a lack of sharing data which we believe is key to improve the quality and pace of research especially in domains like digital forensics. In the What Is Missing? section we highlight six points that we believe are needed in order to solve those current challenges: variety of datasets, updates & upgrades of repositories/datasets, a centralized repository, more research in de-identification, strategies to share complex data such as 'cloud services' and publisher support. On the other hand, we see first steps towards solutions, e.g., by DHS and their Impact Cyber Trust project. Our hope is that this article raises the awareness and importance of sharing information/dataset. For our next steps we plan on contacting some of the repositories to understand why they stopped maintaining the sites. Additionally, we will try to raise the awareness of our webportal with the hope that researchers contribute and keep our list up to date.

Acknowledgements

We like to thank graduate researcher Mateusz Topor for his help in creating the dataset website.

Appendix A. Overview of the datasets

As discussed throughout the article, there are three major findings: First, we identified several datasets by reviewing articles, second we identified several sets by running Google searches and a third we identified third party services that we found in our articles' analysis. All of the findings are presented on our website http://datasets.fbreitinger.de/ which allows to contribute to the collection. In addition, we attached Tables A.3–A.5 which contain the available dataset repositories.

Table A.3
Available datasets.

Dataset type	Ref.	Source	Available datasets	Total size	Origin	Date created/last modified
Video Game Console Disk Images	1.1	University of New Haven cFREG	5 Xbox One partitions	476 GB	Experiment Generated	2014
	1.2	Digital Corpora	4 disk images	11.9 GB		2013–2014
	1.3	DFRWS 2009 Challenge	1 PS3 Linux partition	N/A		2009
Computer Malware	2.1	Contagio Malware Dump	11,960 malware samples	N/A	User Generated	2008–2016
	2.2	Virus Share	27,518,833 malware samples			2016
	2.3	VX Heaven	271,092 malware samples			2006–2016
	2.4	KernelMode.info	N/A			2016
Media (Pictures)	3.1	BOSS – Break Our Steganographic System	10,074 images	N/A	Experiment Generated	2010
	3.2	BOWS2 – Break Our Watermarking System	10,000 images	1.6 GB		2007–2008
	3.3	Columbia University DVMM Laboratory	3600 images	N/A	User & Computer Generated	2005
	3.4	Image Communication Laboratory	2988 images		Experiment Generated	2012
	3.5	King Saud University – Image Forensics	>10 images			2010
	3.6	NRCS Photo Gallery – USDA Natural Resources Conservation Service	13,483 images		User Generated	2016
	3.7	The Berkeley Segmentation Dataset and Benchmark	>300 images	50 MB	User & Computer Generated	2003–2013
	3.8	AT&T Laboratories Cambridge – The Database of Faces	400 images	4.5 MB	Experiment Generated	1992–1994
	3.9	Columbia University – TrustFoto	2218 images	N/A	Experiment Generated	2004–2006
	3.10	The Dresden Image Database	>25,137 images			2010
Media (Videos)	4.1	Region-Level Video Forgery	18 video sequences	48 MB	Experiment Generated	2013
	4.2	YUV Video Sequences	26 video test sequences	N/A		N/A
	4.3	NRCS Photo Gallery – USDA Natural Resources Conservation Service	11 videos		Computer Generated	2014–2016
	4.4	Columbia University – Consumer Video (CCV) Database	9317 YouTube videos		User Generated	2011
World Languages/Text	5.1	Drexel University – Privacy, Security and Automation Lab	Text files with 352,500 words	N/A	User Generated	2009–2012
	5.2	Sentiment Word Net	1298 English & Arabic words			2015
	5.3	Openwall Wordlists Collection	4 million words with wordlists for 20+ languages			2012–2015
	5.4	Reuters Corpora (RCV1, RCV2, TRC2) – Reuters Ltd NIST	3,097,370 Reuters news stories			2004–2015
	5.5	SCOWL (Spell Checker Oriented Word Lists)	250,000 English words	2.4 MB		2016
	5.6	European Parliament Proceedings Parallel Corpus	60 million words per language of 21 European languages	>2 GB		1996–2011

Table A.4
Available datasets.

Dataset type	Ref.	Source	Available datasets	Total size	Origin	Date created/last modified
Email Datasets	6.1	Enron Email Dataset	619,446 messages from 158 users	>423 MB	User Generated	2015
	6.2	Digital Corpora	12 Emails	34.8 KB	Experiment Generated	2012
	6.3	Apache Mail Archives	N/A	N/A	User Generated	2006–2016
	6.4	DFRWS 2009 Rodeo	Outlook PST file		Experiment Generated	2009
Mobile Malware for Android	7.1	Contagio Mobile	>237 malware samples	N/A	User Generated	2011–2016
	7.2	University of Korea Hacking and Countermeasure Research Lab – Andro-AutoPsy	9990 malware samples			2013–2014
	7.3	University of Göttingen, Germany – The Drebin Dataset	5560 malware samples			2010–2012
Different Types of Computer Files	8.1	DFRWS 2006 Challenge	Various types of files	48 MB	Experiment Generated	2006
	8.2	DFRWS 2007 Challenge	Various types of files	330 MB		2007
	8.3	The MSX-13 Corpus	22,000 MS Office 2007 files	24 GB	User Generated	2013
	8.4	The t5 Corpus	4457 different types of files	1.9 GB		2011
	8.5	Govdocs1 – Digital Corpora	1 million files	N/A		2009
Ram Dumps	9.1	Article – Wicher Minnaard	Note: memory sample is directly linked to a tar file	>1 GB	User Generated	2014
	9.2	DFRWS 2008 Rodeo	Laptop memory image	N/A	Experiment Generated	2008
	9.3	The CFReDS Project – NIST	5 memory images	>2 GB		2005–2007
	9.4	The Art of Memory Forensics	N/A	4 GB		2014
	9.5	Digital Corpora	88	44.1 GB		2014
	9.6	DFRWS 2009 Challenge	1 PS3 Linux physical memory dump	N/A		2009
Apk Files	10.1	Secure-Software-Engineering/DroidBench	119 APK files	N/A	User Generated	2015
	10.2	Digital Corpora	2128 APK files			2012
Smartphone Disk Images	11.1	The CFReDS Project – NIST	12 mobile images	N/A	Experiment Generated	2016
	11.2	Digital Corpora	14 mobile images			2011
	11.3	DFRWS 2009 Rodeo	1 mobile image	59 MB		2009
Network Traffic (Logs/pcap)	12.1	Digital Corpora	50 pcap files	N/A	Experiment Generated	2008–2016
	12.2	DFRWS 2009 Challenge	3 pcap files			2009
	12.3	University of New Haven cFREG	1 pcap file	876 KB		2015
	12.4	The CFReDS Project – NIST	3 trace logs	3.8 MB		2016
Wireless Network Traces	13	Crawdad – Resource for Archiving Wireless Data At Dartmouth	133 datasets	N/A	User Generated	2012–2016
Subscriber Identity Module – SIM Card Images	14	The CFReDS Project – NIST	3 SIM images	130 KB	Experiment Generated	2016

Table A.5
Available datasets.

Dataset type	Ref.	Source	Available datasets	Total size	Origin	Date created/last modified
Hard Disk Images	15.1	Digital Corpora	169 disk images	1.106 TB	User/Experiment Generated	2008–2015
	15.2	Computer Forensic Tool Testing (CFTT) – NIST	11 dism images	150 MB	Experiment Generated	2003
	15.3	The CFReDS Project – NIST	53 disk images	12.2 GB		2016
Secure Digital Card – SD Images	16	Digital Corpora	7 SD images	174 MB	Experiment Generated	2015
Universal Serial Bus – USB Flash Drive Images	17.1	Digital Corpora	20 USB images	10.9 GB	Experiment Generated	2009–2015
	17.2	Computer Forensic Tool Testing (CFTT) – NIST	1 USB image	124 MB		2005
	17.3	The CFReDS Project – NIST	3 USB images	462 MB		2016
	17.4	DFRWS 2008 Rodeo	1 USB image	N/A		2008
	17.5	DFRWS 2009 Challenge	1 USB image			2009
Apple iPod Disk Images	18	Digital Corpora	10 iPod images	55 GB	Experiment Generated	2010–2015
Tablet Images	19	Digital Corpora	25 images	16.7 GB	Experiment Generated	2012–2014
Chat Logs	20	Article – Tarique Anwar & Muhammad Abulaish	1100 chat logs	715 MB	User Generated	2010–2012
Leaked Passwords	21	Skull Security Wiki	30 sets	N/A	User Generated	2009–2010

Appendix B. Top 7 most frequently used datasets

Table B.6 presents the top 7 most used datasets from our study. The first row shows the rank, followed by the name of the actual dataset. The 'articles-column' shows the references organized by conference.

The first eye-catching fact is that sometimes researchers might need multiple sets, as in the following three cases

- (Garfinkel and McCarrin, 2015) and (Roussev et al., 2013) utilized the govdocs as well as the M57-patents scenario in their studies.
- (Breitinger et al., 2014b) utilized the govdocs as well as the t5 file corpus (note, t5 is a subset of govdocs).

This example clearly demonstrates how convenient it is to have a centralized dataset repository, which at the same time could benefit research in more than one form.

Another interesting observation from the table is the fact that some of these datasets were reused more than once by the same authors in different occasions. For instance, the t5 File Corpus was referenced from seven articles, however, there are only four different names: Breitinger, Roussev, Gupta and Baggili which had several collaborations. Other examples are Beebe & Liu and the M57-patents scenario or Lu et at./Quach and the pictures/BOSS set.

Table B.6
Top 7 datasets used in 45 papers.[a]

Rank	Dataset/Repository	Articles
1st	Govdocs File Corpus/Digital Corpora	DFRWS: (Schatz, 2015), (Garfinkel & McCarrin, 2015), (Fitzgerald et al., 2012), (Axelsson, 2010); ICDF2C: (Karabiyik & Aggarwal, 2014), (Breitinger et al., 2014b); DI: (Breitinger et al., 2014c), (Penrose et al., 2013), (Roussev et al., 2013), (Savoldi et al., 2012)
1st	Emails/Enron	DFRWS: (Schmid et al., 2015), (Shields et al., 2011); ICDF2C: (Crabb, 2014); DI: (Magalingam et al., 2015), (Quick & Choo, 2013b), (Quick & Choo, 2013a) (Al-Zaidy et al., 2012), (Cheng et al., 2011), (Iqbal et al., 2010); IEEE S & P: (Naveed et al., 2014)
3rd	t5 File Corpus/Roussev	DFRWS: (Breitinger & Roussev, 2014), (Breitinger et al., 2014a), (Breitinger et al., 2013), (Roussev, 2011); ICDF2C: (Gupta & Breitinger, 2015), (Breitinger & Baggili, 2014), (Breitinger et al., 2014b)
4th	M57-patents Scenario/Digital Corpora	DFRWS: (Garfinkel & McCarrin, 2015), (Beebe & Liu, 2014b); ADFSL: (Woods et al., 2011); DI: (Beebe & Liu, 2014a), (Marturana & Tacconi, 2013), (Roussev et al., 2013)
4th	Real Drive Corpus/Digital Corpora	DFRWS: (Brown, 2011), (Beverly et al., 2011); ICDF2C: (Schwamm & Rowe, 2014), (Rowe, 2013), (Rowe & Garfinkel, 2011); DI: (Noel & Peterson, 2014)
6th	Android Malware Genome Project[b]	DFRWS: (Guido et al., 2013); DI: (Talha et al., 2015); IEEE S & P: (Xia et al., 2015), (Bianchi et al., 2015), (Zhou & Jiang, 2012)
7th	Pictures/BOSS – Break Our Steganographic System	DFRWS: (Quach, 2014); DI: (Lu et al., 2015), (Lu et al., 2014), (Quach, 2012)

[a] Note: Three papers used more than one dataset.
[b] Site is no longer available. See Sec. Non available datasets for details.

References

Abt, S., Baier, H., 2014. Are we missing labels? A study of the availability of ground-truth in network security research. In: Proceedings of the 3rd Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS'14) (September 2014). IEEE.

Al-Shaheri, S., Lindskog, D., Zavarsky, P., Ruhl, R., 2013. A forensic study of the effectiveness of selected anti-virus products against ssdt hooking rootkits. In: Proceedings of the Conference on Digital Forensics, Security and Law, pp. 137–160.

Al-Zaidy, R., Fung, B.C., Youssef, A.M., Fortin, F., 2012. Mining criminal networks from unstructured text documents. Digit. Investig. 8, 147–160.

Armknecht, F., Dewald, A., 2015. Privacy-preserving email forensics. Digit. Investig. 14, S127–S136.

Arp, D., Spreitzenbarth, M., Hubner, M., Gascon, H., Rieck, K., 2014. Drebin: effective and explainable detection of android malware in your pocket. In: NDSS.

Axelsson, S., 2010. The normalised compression distance as a file fragment classifier. Digit. Investig. 7, S24–S31.

Baggili, I., Breitinger, F., 2015. Data sources for advancing cyber forensics: what the social world has to offer. In: 2015 AAAI Spring Symposium Series.

Beebe, N.L., Liu, L., 2014a. Clustering digital forensic string search output. Digit. Investig. 11, 314–322.

Beebe, N.L., Liu, L., 2014b. Ranking algorithms for digital forensic string search hits. Digit. Investig. 11, S124–S132.

Beverly, R., Garfinkel, S., Cardwell, G., 2011. Forensic carving of network packets and associated data structures. Digit. Investig. 8, S78–S89.

Bianchi, A., Corbetta, J., Invernizzi, L., Fratantonio, Y., Kruegel, C., Vigna, G., 2015. What the app is that? deception and countermeasures in the android user interface. In: 2015 IEEE Symposium on Security and Privacy. IEEE, pp. 931–948.

Boukhtouta, A., Mouheb, D., Debbabi, M., Alfandi, O., Iqbal, F., El Barachi, M., 2015. Graph-theoretic characterization of cyber-threat infrastructures. Digit. Investig. 14, S3–S15.

Breitinger, F., Baggili, I., 2014. File detection on network traffic using approximate matching. J. Digital Forensics Secur. Law JDFSL 9, 23.

Breitinger, F., Roussev, V., 2014. Automated evaluation of approximate matching algorithms on real data. Digit. Investig. 11, S10–S17.

Breitinger, F., Stivaktakis, G., Baier, H., 2013. Frash: a framework to test algorithms of similarity hashing. Digit. Investig. 10, S50–S58.

Breitinger, F., Baier, H., White, D., 2014a. On the database lookup problem of approximate matching. Digit. Investig. 11, S1–S9.

Breitinger, F., Rathgeb, C., Baier, H., 2014b. An efficient similarity digests database lookup-a logarithmic divide & conquer approach. J. Digital Forensics Secur. Law JDFSL 9, 155.

Breitinger, F., Stivaktakis, G., Roussev, V., 2014c. Evaluating detection error trade-offs for bytewise approximate matching algorithms. Digit. Investig. 11, 81–89.

Brown, R.D., 2011. Reconstructing corrupt deflated files. Digit. Investig. 8, S125–S131.

Carlin, A., Manson, D., Zhu, J., 2010. Developing the cyber defenders of tomorrow with regional collegiate cyber defense competitions (ccdc). Inf. Syst. Educ. J. 8.

Case, A., Richard, G.G., 2015. Advancing mac os x rootkit detection. Digit. Investig. 14, S25–S33.

Cheng, N., Chandramouli, R., Subbalakshmi, K., 2011. Author gender identification from text. Digit. Investig. 8, 78–88.

Conti, G., Bratus, S., Shubina, A., Sangster, B., Ragsdale, R., Supan, M., Lichtenberg, A., Perez-Alemany, R., 2010. Automated mapping of large binary objects using primitive fragment type classification. Digit. Investig. 7, S3–S12.

Crabb, E.S., 2014. "Time for some traffic problems": enhancing e-discovery and big data processing tools with linguistic methods for deception detection. J. Digital Forensics Secur. Law JDFSL 9, 167.

Dykstra, J., Sherman, A.T., 2012. Acquiring forensic evidence from infrastructure-as-a-service cloud computing: exploring and evaluating tools, trust, and techniques. Digit. Investig. 9, S90–S98.

Farid, H., Bravo, M.J., 2012. Perceptual discrimination of computer generated and photographic faces. Digit. Investig. 8, 226–235.

Fitzgerald, S., Mathews, G., Morris, C., Zhulyn, O., 2012. Using nlp techniques for file fragment classification. Digit. Investig. 9, S44–S49.

Garfinkel, S.L., McCarrin, M., 2015. Hash-based carving: searching media for complete files and file fragments with sector hashing and hashdb. Digit. Investig. 14, S95–S105.

Garfinkel, S., Farrell, P., Roussev, V., Dinolt, G., 2009. Bringing science to digital forensics with standardized forensic corpora. Digit. Investig. 6, S2–S11.

Garfinkel, S., Nelson, A., White, D., Roussev, V., 2010. Using purpose-built functions and block hashes to enable small block and sub-file forensics. Digit. Investig. 7, S13–S23.

Guido, M., Ondricek, J., Grover, J., Wilburn, D., Nguyen, T., Hunt, A., 2013. Automated identification of installed malicious android applications. Digit. Investig. 10, S96–S104.

Guido, M., Brooks, M., Grover, J., Katz, E., Ondricek, J., Rogers, M., Sharpe, L., 2016. Generating a corpus of mobile forensic images for masquerading user experimentation. J. Forensic Sci. 1467–1472.

Gupta, V., Breitinger, F., 2015. How cuckoo filter can improve existing approximate matching techniques. In: International Conference on Digital Forensics and Cyber Crime. Springer, pp. 39–52.

Iqbal, F., Binsalleeh, H., Fung, B.C., Debbabi, M., 2010. Mining writeprints from anonymous e-mails for forensic investigation. Digit. Investig. 7, 56–64.

Jang, J.-w., Kang, H., Woo, J., Mohaisen, A., Kim, H.K., 2015. Andro-autopsy: anti-malware system based on similarity matching of malware and malware creator-centric information. Digit. Investig. 14, 17–35.

Karabiyik, U., Aggarwal, S., 2014. Audit: automated disk investigation toolkit. J. Digital Forensics Secur. Law JDFSL 9, 129.

Karpisek, F., Baggili, I., Breitinger, F., 2015. Whatsapp network forensics: decrypting and understanding the whatsapp call signaling messages. Digit. Investig. 15, 110–118.

Khan, R., Mizan, M., Hasan, R., Sprague, A., 2014. Hot zone identification: analyzing effects of data sampling on spam clustering. In: Proceedings of the Conference on Digital Forensics, Security and Law. Association of Digital Forensics, Security and Law, p. 243.

Lambertz, M., Uetz, R., Gerhards-Padilla, E., 2013. Resurrection: a carver for fragmented files. In: International Conference on Digital Forensics and Cyber Crime. Springer, pp. 51–66.

Laskov, P., et al., 2014. Practical evasion of a learning-based classifier: a case study. In: 2014 IEEE Symposium on Security and Privacy. IEEE, pp. 197–211.

Lee, S., Kim, Y., Kim, J., Kim, J., 2014. Stealing webpages rendered on your browser by exploiting GPU vulnerabilities. In: 2014 IEEE Symposium on Security and Privacy. IEEE, pp. 19–33.

Liberatore, M., Erdely, R., Kerle, T., Levine, B.N., Shields, C., 2010. Forensic investigation of peer-to-peer file sharing networks. Digit. Investig. 7, S95–S103.

Ligh, M.H., Case, A., Levy, J., Walters, A., 2014. The Art of Memory Forensics: Detecting Malware and Threats in Windows, Linux, and Mac Memory. John Wiley & Sons.

Liu, C., Singhal, A., Wijesekera, D., 2014. Relating admissibility standards for digital evidence to attack scenario reconstruction. J. Digit. Forensics Secur. Law JDFSL 9, 181.

Lu, J., Liu, F., Luo, X., 2015. A study on jpeg steganalytic features: co-occurrence matrix vs. Markov transition probability matrix. Digit. Investig. 12, 1–14.

Lu, J.-c., Liu, F.-l., Luo, X.-y., 2014. Selection of image features for steganalysis based on the fisher criterion. Digit. Investig. 11, 57–66.

Ma, J., Yang, W., Luo, M., Li, N., 2014. A study of probabilistic password models. In: 2014 IEEE Symposium on Security and Privacy. IEEE, pp. 689–704.

Magalingam, P., Davis, S., Rao, A., 2015. Using shortest path to discover criminal community. Digit. Investig. 15, 1–17.
Mahoney, M.V., Chan, P.K., 2003. An analysis of the 1999 darpa/lincoln laboratory evaluation data for network anomaly detection. In: International Workshop on Recent Advances in Intrusion Detection. Springer, pp. 220–237.
Marturana, F., Tacconi, S., 2013. A machine learning-based triage methodology for automated categorization of digital media. Digit. Investig. 10, 193–204.
Minnaard, W., 2014. Out of sight, but not out of mind: traces of nearby devices' wireless transmissions in volatile memory. Digit. Investig. 11, S104–S111.
Mohamed, A., Yampolskiy, R.V., 2012. Face recognition based on wavelet transform and adaptive local binary pattern. In: International Conference on Digital Forensics and Cyber Crime. Springer, pp. 158–166.
Moore, J., Baggili, I., Marrington, A., Rodrigues, A., 2014. Preliminary forensic analysis of the xbox one. Digit. Investig. 11, S57–S65.
Nappa, A., Johnson, R., Bilge, L., Caballero, J., Dumitras, T., 2015. The attack of the clones: a study of the impact of shared code on vulnerability patching. In: 2015 IEEE Symposium on Security and Privacy. IEEE, pp. 692–708.
Naveed, M., Prabhakaran, M., Gunter, C.A., 2014. Dynamic searchable encryption via blind storage. In: 2014 IEEE Symposium on Security and Privacy. IEEE, pp. 639–654.
Neuner, S., Schmiedecker, M., Weippl, E.R., 2016. Peekatorrent: leveraging p2p hash values for digital forensics. Digit. Investig. 18, S149–S156.
NIST, 2014. NIST Open Data Strategy Implementation and Processes. In: https://www.nist.gov/pba/nist-open-data-strategy-implementation-and-processes. last accessed 2017-01-20.
Noel, G.E., Peterson, G.L., 2014. Applicability of latent dirichlet allocation to multi-disk search. Digit. Investig. 11, 43–56.
Office of Ethics and Compliance, 2016. Hipaa Requirements and Forms for Research. http://irb.ucsf.edu/hipaa. last accessed 2017-01-20.
Penrose, P., Macfarlane, R., Buchanan, W.J., 2013. Approaches to the classification of high entropy file fragments. Digit. Investig. 10, 372–384.
Pichan, A., Lazarescu, M., Soh, S.T., 2015. Cloud forensics: technical challenges, solutions and comparative analysis. Digit. Investig. 13, 38–57.
Quach, T.-T., 2012. Locating payload embedded by group-parity steganography. Digit. Investig. 9, 160–166.
Quach, T.-T., 2014. Extracting hidden messages in steganographic images. Digit. Investig. 11, S40–S45.
Quick, D., Choo, K.-K.R., 2013a. Dropbox analysis: data remnants on user machines. Digit. Investig. 10, 3–18.
Quick, D., Choo, K.-K.R., 2013b. Forensic collection of cloud storage data: does the act of collection result in changes to the data or its metadata? Digit. Investig. 10, 266–277.
Rogers, M.K., Seigfried-Spellar, K.C., 2014. Using internet artifacts to profile a child pornography suspect. J. Digit. Forensics Secur. Law 9, 57–66.
Roussev, V., 2011. An evaluation of forensic similarity hashes. Digit. Investig. 8, S34–S41.
Roussev, V., Quates, C., 2013. File fragment encoding classification an empirical approach. Digit. Investig. 10, S69–S77.
Roussev, V., Quates, C., Martell, R., 2013. Real-time digital forensics and triage. Digit. Investig. 10, 158–167.
Rowe, N.C., 2013. Identifying forensically uninteresting files using a large corpus. In: International Conference on Digital Forensics and Cyber Crime. Springer, pp. 86–101.
Rowe, N.C., Garfinkel, S.L., 2011. Finding anomalous and suspicious files from directory metadata on a large corpus. In: International Conference on Digital Forensics and Cyber Crime. Springer, pp. 115–130.
Savoldi, A., Piccinelli, M., Gubian, P., 2012. A statistical method for detecting on-disk wiped areas. Digit. Investig. 8, 194–214.
Schatz, B.L., 2015. Wirespeed: extending the aff4 forensic container format for scalable acquisition and live analysis. Digit. Investig. 14, S45–S54.
Schmid, M.R., Iqbal, F., Fung, B.C., 2015. E-mail authorship attribution using customized associative classification. Digit. Investig. 14, S116–S126.
Schwamm, R., Rowe, N.C., 2014. Effects of the factory reset on mobile devices. J. Digit. Forensics Secur. Law 9, 205–220.
Shields, C., Frieder, O., Maloof, M., 2011. A system for the proactive, continuous, and efficient collection of digital forensic evidence. Digit. Investig. 8, S3–S13.
Stebbins, M., 2013. Expanding Public Access to the Results of Federally Funded Research. https://obamawhitehouse.archives.gov/blog/2013/02/22/expanding-public-access-results-federally-funded-research. last accessed 2017-01-20.
Talha, K.A., Alper, D.I., Aydin, C., 2015. Apk auditor: permission-based android malware detection system. Digit. Investig. 13, 1–14.
Tassi, P., 2015. How ISIS Terrorists May Have Used PlayStation 4 to Discuss and Plan Attacks. http://www.forbes.com/sites/insertcoin/2015/11/14/why-the-paris-isis-terrorists-used-ps4-to-plan-attacks/#33b2a3ec731a. last accessed 2017-01-20.
Woods, K., Lee, C.A., Garfinkel, S., Dittrich, D., Russell, A., Kearton, K., 2011. Creating realistic corpora for security and forensic education. In: Proceedings of the Conference on Digital Forensics, Security and Law. Association of Digital Forensics, Security and Law, p. 123.
Xia, M., Gong, L., Lyu, Y., Qi, Z., Liu, X., 2015. Effective real-time android application auditing. In: Security and Privacy (SP), 2015 IEEE Symposium on. IEEE, pp. 899–914.
Zhou, Y., Jiang, X., 2012. Dissecting android malware: characterization and evolution. In: 2012 IEEE Symposium on Security and Privacy. IEEE, pp. 95–109.