Availability of datasets for digital forensics e And what is missing

서울과학기술대학교
디지털 포렌식
남기훈
17510137

abstract

- 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.

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

- In order to produce high-quality research results, we argue that three critical features must be examined:

Introduction

- 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 reprod

Introduction

- 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)

Limitations

- 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 (2009e2013) 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.

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.

Methodology

- To analyze the availability of datasets which we define in Sec. Definition of a dataset, we first investigated peerreviewed 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.

Methodology

- Analyzing peer-reviewed articles
- 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?

Methodology

- Analyzing peer-reviewed articles

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

Online searches

- 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 manne

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 (z32%) 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

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

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 dataset

- First, in many cases, there is a lack of real world datasets available to the digital forensics community

- 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

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 e "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:

User generated datasets

- 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).

User generated datasets

- Computer generated da

- 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 e 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.

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

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.

Availability of 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 equalshare 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

Availability of datasets

- 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

- 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.

Non available datasets

- his 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:

Non available datasets

- 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

Kinds of datasets

- 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

- 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)

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. S98 C. Grajeda et al. / Digital Investigation 22 (2017) S94eS105 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

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)

RAM dumps

- 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

Images of other devices

- 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”

Images of other devices

- 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.

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)
- Coming to the high usage og self-made datasets, dome researches clearly stated and mxc-13 corpus.

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)

Datasets found through Google research

- 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.

Datasets found through Google research

- 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

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)

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

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

Centralized repository

- 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

- 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 deidentification as pointed out by Garfinkel et al. (2009). If we find ways to un-personalize data by removing, changing or manipulating names, phone numbers

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

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

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.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

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.

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