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

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. 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.