Poster: Data Quality for Security Challenges: Case Studies of Phishing, Malware and Intrusion Detection Datasets

Rakesh M. Verma, Victor Zeng and Houtan Faridi
University of Houston, Texas
{rverma,vzeng,hfaridi}@uh.edu

ABSTRACT

Techniques from data science are increasingly being applied by researchers to security challenges. However, challenges unique to the security domain necessitate painstaking care for the models to be valid and robust. In this paper, we explain key dimensions of data quality relevant for security, illustrate them with several popular datasets for phishing, intrusion detection and malware, indicate operational methods for assuring data quality and seek to inspire the audience to generate high quality datasets for security challenges.

KEYWORDS

Semiotics; data quality; data difficulty; data poisoning

ACM Reference Format:
Rakesh M. Verma, Victor Zeng and Houtan Faridi. 2019. Poster: Data Quality for Security Challenges: Case Studies of Phishing, Malware and Intrusion Detection Datasets. In *2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19), November 11–15, 2019, London, United Kingdom.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3319535.3363267

1 INTRODUCTION

Researchers have been applying data mining and machine learning techniques to security challenges for the past 20 years or so. A search of the bibliographic database DBLP[1] with the query *intrusion detection neural network* led to 216 matches on August 21, 2019. Note that this result is for just one machine learning technique and just one security challenge. The mind boggles when one considers the plethora of data mining and machine learning techniques and the variety of security challenges, which have no good separations between legitimate instances and attack instances. These include phishing, malware, stepping-stone detection, intrusion detection, and denial of service attacks, just to name a few.

However, some researchers had observed that the overwhelming majority of this research is simply not being deployed [17]. So, the question arises: "Why is there such a gap between theory and practice in security?" The reasons seem to be mainly two:

- Researchers, in their work, have missed at least some of the unique needs of the security domain. For a discussion on these aspects, see [3, 20].
- The general lack of trustworthy datasets for security challenges. Although companies are collecting mountains of data, they are reluctant to share for fear of the consequences.[2] Datasets collected by academic researchers could be plagued with problems of data quality: recency, diversity, class ratio, scale, errors and consistency. Dataset poisoning attacks must also be considered.

1.1 Poster Outline

We present the unique needs of the security domain, data quality assessment, and issues with existing datasets. We examine ways of discovering data quality problems and suggest some solutions for the issues that arise. Illustrative examples of these topics include:

(1) Unique needs of security domain relevant to datasets: nonstationarity arising from an active attacker or streaming data; lack of large, diverse and representative datasets; potential poisoning of datasets, etc.
(2) Existing dimensions of data quality relevant to security domain from a semiotics standpoint [18], and new dimensions, e.g., data difficulty and poisoning, inspired from our work.
(3) Illustration of data quality problems using specific datasets for phishing, malware and intrusion detection.
(4) Methods for finding data quality problems.
(5) Suggestions on how to handle the issues that are found.

2 CASE STUDY: PHISHING DATASETS

A systematic review of phishing datasets [7] (used in almost 300 papers from CORE B or better venues published during 2010-2018), including URL, website and email datasets, shows that they are: (i) generally not recent, and, when recent, usually not available publicly, (ii) tend to be small-scale (typically less than 10^6 instances), (iii) almost balanced with respect to class labels, and (iv) sometimes suffer from other issues. For example, several phishing URL detection papers have used domains from Alexa.com and URLs from Phishtank with features involving URL length. In the phishing email detection literature, we find the Nazario dataset of approximately 5000 phishing emails collected during early 2000s and a smaller one from 2015-17. The Enron dataset has sanitized headers and the SpamAssassin dataset has "lightly" sanitized headers. In previous work [19], we had designed the "nonsense" filter for two kinds of potential poisoning attacks, based on analysis of the Nazario dataset. Due to these issues, we created four datasets for phishing email detection, with and without email headers, that are public.

2.1 Creation of IWSPA-AP Email Dataset

Our objective for the IWSPA-AP dataset was to ensure diversity, i.e., different types of attacks and legitimate emails as well as a mix

[1] https://dblp.uni-trier.de/

[2] At least one company seems to be going against the grain [3].

CCS '19, November 11–15, 2019, London, United Kingdom
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-6747-9/19/11…$15.00
https://doi.org/10.1145/3319535.3363267

of new and classical attacks, so we gathered recent and historical emails from as many sources as possible [2]. Legitimate emails were relatively easy to find compared to phishing ones, thanks to Wikileaks. Phishing emails were collected from the IT departments of different universities. We also included some emails from the popular Nazario phishing corpora. Note that the emails collected from universities' IT departments usually do not have a full header, so we only used these sources for the no-header subtask.

The dataset was cleaned by replacing all the URLs in the emails with $\langle link \rangle$, since the URLs in legitimate emails tended to be quite revealing. Another concern was the recognizability of the sources. So we tried to remove from the emails, as much as feasible, any signs that could hint at the origin of the datasets. For this purpose, we included in the preprocessing steps the normalization of organizations' or universities' names, recipients' names, domain names, signatures, threading, and removed non-English emails.

We also removed emails that are too big (more than 1 MB) or too small and all base64 encoded text. To remove as much noise as possible, we attempted to remove leftover HTML tags and empty spacing that resulted from parsing the body of the email using an HTML parser. As a final check before release, a logistic regression model was trained and run on the test subsets.

2.2 Rigorous Analysis of IWSPA-AP Datasets

We ran a state of the art part-of-speech tagger on the email bodies of the IWSPA-AP dataset and collected the 50 most frequent nouns. The software revealed that, despite the extensive cleaning operation, there were still significant number of nouns tied to the sources of the emails in the phishing subset.

In 2019, researchers managed to achieve an accuracy of 99.848% on the IWSPA-AP dataset using their THEMIS model [9]. We hypothesized that one of the factors contributing to this high accuracy was the "difficulty" of the dataset. To test this hypothesis, we ran the PhishBench benchmark [1] with 69 header features and 48 body features on the *full-header* subset of the dataset using only the first 50 legitimate and 10 phishing emails for training. We found that five out of nine classifiers managed to achieve accuracy greater than 99%. This prompted a more thorough experiment where we performed 20 round Monte-Carlo cross validation on the full-header dataset using the same 69 header and 48 body features with a training set size of 50 legit and 10 phish emails. In this second experiment, three out of nine classifiers managed to achieve mean accuracy greater than 99%, thus confirming the results of our initial experiment.

These findings led to another round of cleaning of the dataset emails, resulting in Version 2.0. On this version, we normalized variations of recipient organization names and domains missed in the first round of cleaning. We then ran the 50-10 experiment on Version 2.0 of the dataset and found that there was no significant change in classifier performance. Consequently, we performed another round of cleaning that removed various normalization artifacts left by the previous round and normalized any IP addresses to 254.254.254.254. This gave us Version 2.1.

Despite the additional rounds of data cleaning, running the 50-10 experiment on Version 2.1 of the dataset showed no significant change in classifier performance over Version 1.0. This suggests that recipient information may be derivable from more subtle aspects of the dataset such as the structure of the headers. We intend on testing this hypothesis in future research. This leads to a new data quality dimension, *data difficulty*, which can be defined as the minimum distance between two data instances from different classes for binary classes, and the minimum or average of the pairwise minimums for multiple classes.

3 CASE STUDY: MALWARE DATASETS

There are many factors in producing a robust malware dataset. These include: (i) sufficient coverage of malware types and attack vectors, (ii) recency and relevance of samples, (iii) multiple methods of feature analysis, (iv) sufficiently large families, (v) strong coverage of multiple obfuscation and permutation techniques, and (vi) proper labels and groupings. While there exist a few publicly available malware collections, e.g., VXHeavens, VirusShare, and theZoo, most of these are not proper datasets and thus lack adequate labeling, balanced and well distributed types of malware, or recent samples. For these reasons we decided to analyze an industrial dataset that contained ground-truth labels.

3.1 Analysis: Industrial Dataset

Engineers at a security company provided us with a labeled dataset of malware features. The dataset consisted of JSON files representing the behavior of a malicious sample during execution. To generate these JSON files, a team collected malicious Windows Portable Executables that attempted to propagate themselves over a network. Each malware was then executed within a Cuckoo Sandbox to extract sets of features. To group the malware into families, labels were constructed using Suricata network intrusion detection (NID) signatures. All malicious samples that triggered the same signature were grouped into the same family. Upon receiving the dataset, a few steps had to be taken to clean the data. First, several malware JSON files contained no usable data within them and were removed. Next, we discovered that many samples were classified in two or more groups, so these samples needed to be removed also. After filtering, we were left with 5,673 usable samples. Discarding problematic samples caused group sizes to become highly unbalanced. The largest family contained 772 samples while the smallest groups only contained 1, and the mode of all groups was merely 3 samples. Once the dataset was cleaned, we grouped the samples using multiple clustering techniques to observe if we could reproduce the provided family groupings [10, 11]. In forming clusters, we tested several different groupings of malware features, including system and network interactions, to observe which combinations produce the best results.

3.2 Issues with Malware Labeling

An important issue is the difficulties in producing proper ground-truth labels for malware. Often it may be necessary to identify or group malware based on families of similar samples. Unfortunately, labeling malware into distinguishable groups can be a difficult task. Most anti-virus vendors strongly disagree with how strands of malware should be labeled and often produce groupings with vastly different levels of specificity [4]. Due to this difficulty, most malware datasets do not come with classification labels or a ground-truth with which to compare results. To work around this, many researchers have implemented anti-virus majority voting systems to construct labels. This technique calls for only using malware samples that have the same label across a majority of anti-virus vendors. However, this method can have multiple drawbacks. First, it can significantly reduce the number of samples within a dataset as most malware do not have universal labels. Out of 14,212 malicious samples in the ANUBIS dataset, only 2,658 were deemed usable through anti-virus voting [6]. Furthermore, only using malware whose labels are universally agreed on, may bias towards positive results as a model could potentially only be classifying malware that are easiest to group or distinguish [13]. Efforts to produce standardized malware classification labels are ongoing, e.g., MAEC [12]. Until standardized methods are produced, manual labeling may be necessary to accurately test the performance of a model.

4 CASE STUDY: INTRUSION DATASETS

For intrusion detection, we analyzed the CICIDS2017 dataset since the KDD Cup/DARPA dataset is quite old and has been analyzed already [8]. The CICIDS2017 dataset is a synthetic dataset consisting of a complete capture of all send and receive traffic from the main switch of the Victim Network [15]. It contains both raw packet capture files and 3119345 network flows analyzed by their CICFlowMeter and labeled by attack type. In our exploration of the network flow information, we noticed several issues:

Missing Information. The dataset contains 288602 completely empty records and 1358 instances that are missing the number of bytes sent. We removed these empty/incomplete instances.

Duplicates. We found 202 duplicate instances in the dataset. Of these duplicates, 201 are benign and 1 is a DoS Hulk attack. If we ignore the timestamp, then the number of duplicates increases to 12981. Of these 12981 duplicates, 5393 instances are benign, 7561 instances are DoS Hulk, and 27 are DoS slowloris.

Attack Diversity. After removing instances with missing information and duplicates not ignoring timestamps, the dataset contained 2829183 instances. Of these instances, 80.32% were labeled benign, and the rest were spread out over the various attacks. Moreover, even amongst the attacks, we see a skewed distribution. The three most common attacks constitute 92.87% of the attack instances, and the least common attacks have less than 50 instances each.

Dataset Difficulty. As an indirect measure of dataset difficulty, we perform a small data experiment, where we used a Decision Tree classifier to identify malicious traffic with only a randomly selected training set of 0.1% (2828 instances) of the dataset. Our test set consisted of 100000 randomly selected instances, which were not used in the training set. Over 10 iterations, we managed to achieve a mean accuracy of 92.88%.

5 RELATED WORK

The DBLP query, security data quality, produced 35 results over the period 1994-2015. However, the relevant papers numbered fewer than 10. We repeated the query on ACM Digital Library (DL), Google Scholar (allintitle query), and IEEE Xplore. Many of the retrieved results from DBLP, DL, and Xplore had security in the title of the journal or conference (e.g., Journal of Computer Security, etc.) A summary of the most relevant work is below.

A nice taxonomy of data quality terms ("dimensions") is presented in [18]. In [14], the goal was to integrate security and accuracy into data quality evaluation. Data quality challenges in sharing threat intelligence were discussed in [16]. A general survey is [5].

6 CONCLUSIONS

Through three different case studies we have illustrated the issues with datasets for security challenges. We have shown how to find them and how to address them. We defined dataset difficulty as a measure of dataset quality. Much remains to be done, for example, when to stop cleaning. One possibility is to see if dataset sources or other key metadata can be identified accurately based on the data.

ACKNOWLEDGMENTS

Research supported in part by NSF grants CNS 1319212, DGE 1433817, DUE 1356705 and IIS 1659755.

REFERENCES

[1] Ayman El Aassal, Shahryar Baki, Avisha Das, and Rakesh M. Verma. 2019. An In-Depth Benchmarking Evaluation of Phishing Detection Research for Security Needs. (2019). To be submitted.

[2] Ayman El Aassal, Luis Moraes, Shahryar Baki, Avisha Das, and Rakesh Verma. 2018. Anti-Phishing Pilot at ACM IWSPA 2018: Evaluating Performance with New Metrics for Unbalanced Datasets. In *Proc. of IWSPA-AP Pilot*. CEUR, 2–10.

[3] Idan Amit, John Matherly, William Hewlett, Zhi Xu, Yinnon Meshi, and Yigal Weinberger. 2018. Machine Learning in Cyber-Security - Problems, Challenges and Data Sets. (Dec. 2018). Online.

[4] Michael Bailey, Jon Oberheide, Jon Andersen, Z Morley Mao, Farnam Jahanian, and Jose Nazario. 2007. Automated classification and analysis of internet malware. In *RAID*. Springer, 178–197.

[5] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. Methodologies for Data Quality Assessment and Improvement. *ACM Comput. Surv.* 41, 3 (July 2009), 16:1–16:52.

[6] Ulrich Bayer, Paolo Milani Comparetti, Clemens Hlauschek, Christopher Kruegel, and Engin Kirda. 2009. Scalable, behavior-based malware clustering.. In *NDSS*, Vol. 9. USENIX, 8–11.

[7] Avisha Das, Shahryar Baki, Ayman El Aassal, Rakesh Verma, and Arthur Dunbar. [n. d.]. SoK: A Comprehensive Reexamination of Phishing Research from the Security Perspective. ([n. d.]). Under review.

[8] Abhishek Divekar, Meet Parekh, Vaibhav Savla, Rudra Mishra, and Mahesh Shirole. 2018. Benchmarking datasets for Anomaly-based Network Intrusion Detection: KDD CUP 99 alternatives. (2018). Online.

[9] Y. Fang, C. Zhang, C. Huang, L. Liu, and Y. Yang. 2019. Phishing Email Detection Using Improved RCNN Model With Multilevel Vectors and Attention Mechanism. *IEEE Access* 7 (2019), 56329–56340. https://doi.org/10.1109/ACCESS.2019.2913705

[10] H. Faridi, S. Srinivasagopalan, and R. Verma. 2018. Performance Evaluation of Features and Clustering Algorithms for Malware. *IEEE ICDMW (ADMiS)* (2018).

[11] H. Faridi, S. Srinivasagopalan, and R. Verma. 2019. Parameter Tuning and Confidence Limits of Malware Clustering. In *Proc. 9th CODASPY*. ACM, 169–171.

[12] Ivan Kirillov, Desiree Beck, Penny Chase, and Robert Martin. 2011. Malware attribute enumeration and characterization. (2011).

[13] Peng Li, Limin Liu, Debin Gao, and Michael K Reiter. 2010. On challenges in evaluating malware clustering. In *RAID*. Springer, 238–255.

[14] Leon Reznik and Elisa Bertino. 2013. POSTER: Data Quality Evaluation: Integrating Security and Accuracy. In *Proc. of CCS (CCS '13)*. ACM, 1367–1370.

[15] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A Ghorbani. 2018. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *ICISSP*. SciTePress, 108–116.

[16] Christian Sillaber, Clemens Sauerwein, Andrea Mussmann, and Ruth Breu. 2016. Data Quality Challenges and Future Research Directions in Threat Intelligence Sharing Practice. In *Proc. Workshop on Information Sharing and Collaborative Security (WISCS '16)*. ACM, 65–70.

[17] Robin Sommer and Vern Paxson. 2010. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In *31st IEEE Symp. on Security and Privacy, S&P 2010, 16-19 May 2010*. IEEE, 305–316.

[18] Gurvirender Tejay, Gurpreet Dhillon, and Amita Goyal Chin. 2005. Data Quality Dimensions for Information Systems Security: A Theoretical Exposition (Invited Paper). In *Security Management, Integrity, and Internal Control in Information Systems*. Springer US, Boston, MA, 21–39.

[19] Rakesh Verma and Nabil Hossain. 2014. Semantic Feature Selection for Text with Application to Phishing Email Detection. In *Information Security and Cryptology – ICISC 2013*, Hyang-Sook Lee and Dong-Guk Han (Eds.). Springer International Publishing, Cham, 455–468.

[20] R.M. Verma, M. Kantarcioglu, D.J. Marchette, E.L. Leiss, and T. Solorio. 2015. Security Analytics: Essential Data Analytics Knowledge for Cybersecurity Professionals and Students. *IEEE Security & Privacy* 13, 6 (2015), 60–65.