Cybersecurity Datasets: A Mirage

Jelena Mirkovic, Stephen Hayne, Michalis Kallitsis, Wes Hardaker, John Heidemann, Christos Papadopoulos, Devkishen Sisodia

1 Introduction

Cybersecurity research has a great, unmet need for datasets that meet several specific requirements: (1) datasets must contain *real security and "peace-time" events*, (2) researchers should be able to *adequately access both recent and curated* datasets, (3) some reasonable number of datasets must be *accurately labeled* with regard to security events, (4) some datasets should have *varying levels of event sophistication*, and (5) it must be possible to *cross-correlate datasets* with other datasets from public or private domain. This paper examines these requirements, discusses on why they are difficult to meet and why they are crucial for advancements in cybersecurity research, and discusses some forward directions.

2 Dataset Requirements

In various sub-fields of cybersecurity research there is often endeavor to develop new ways of detecting security events, such as spam and phishing campaigns, denial-of-service attacks, network scanning or other reconnaissance activities, DNS poisoning, BGP route hijacking, botnet coordination, data exfiltration, etc. We can abstract these research goals as *security event detection*. To perform security event detection research, cybersecurity researchers need access to *real-world datasets* that meet the requirements we outlined in the Introduction. We elaborate more on these requirements here.

Both peace-time and security events are needed, and in sufficient quantities so that researchers can train and test their approaches and evaluate their efficacy (e.g., detection accuracy, misclassification rate, detection timeliness, etc). Having only security events (e.g., only phishing emails) cannot calibrate detection systems and make them robust against changes in normal data patterns (e.g., regular emails). Having only a handful of security events biases detection systems towards these specific events, especially since missing even one of them will greatly inflate a false negative rate.

Both curated and recent datasets are needed. Curated, snapshot datasets facilitate research community's growth since researchers can build on each other's work and compare approaches against the same benchmark datasets. However, networks and their usage evolve with time and recent datasets facilitate ongoing evaluation of detection approaches. They also offer a wealth of corner cases in both normal and abnormal event patterns, which help mature detection approaches. Further, researchers must have **adequate access** to both datasets and the computational resources required to process it.

Labeled datasets are sorely needed, since labels both establish ground truth for both training and evaluation of detection accuracy. However, in cybersecurity it is often impossible to establish an absolute ground truth. A frequently employed approach labels data records (e.g., traffic flows, email messages, file system accesses, binaries) using a commercial tool (e.g., VirusTotal for malware research). Yet, commercial tools are not 100% accurate and have poor recall rates. They may miss security events, mis-identify their attributes (e.g, malware family), erroneously label peace-time events as security events (e.g., benign binaries as malware) and their detection may lag behind the actual security event (e.g., attack detection tools frequently raise alerts some time after the onset of an attack).

Finally, researchers may seek to **correlate events** across datasets, for example by combining scanning traffic datasets with IP addresss blocklists of repeat offenders or Mirai-infected hosts. This correlation is only possible if datasets are aligned over the same time period and are not anonymized, or at least anonymized in the same way. In some cases, cross-correlation across datasets also requires data providers to collect those datasets at the same vantage points. However, due to various reasons, meeting this requirement may not always be possible for data providers. The need for cross-correlation and the difficulties in providing datasets that allow for cross-correlation raises another significant tussle between the research community and data providers.

There are a small number of datasets today being used for cybersecurity research, but these datasets fail to meet one or several criteria we outlined above. Data providers face challenges when sharing their data, because they need to **protect privacy** of data sources, which often necessitates anonymization and/or access via a "code-to-data" approach requiring the data provider's infrastructure. These arrangements do not support adequate access or correlation requirements, and may limit dataset sizes. Data providers are also rarely compensated for data access, and thus have limited resources to dedicate to data collection, processing and ensuring flexible data access is available. Hence, many datasets available today are only opportunistic and poorly labeled.

3 Paths Forward

We outline some paths forward, given the tussle between researcher needs and providers abilities to meet them. First, we need a way to vet researchers and to assign an objective reputation measurement. This will enable a tiered approach with non-sensitive data sharing with novices, and sensitive data access for experienced researchers. For equitable access, the reputation measure should depend only on the dynamics of dataset usage and not external research visibility. Datasets are most useful to students and junior researchers.

Second, we need a way to share data in a way that is user-specific and auditable, so that leaked datasets can be traced to the user that leaked them. Third, we need a way to express a flexible ground truth, perhaps containing multiple labels with version control and attribution. Then we can engage the research community to contribute these labels so that we can keep enriching datasets, and thus producing better cybersecurity research.

Finally, we should design methods for researchers to cost-share in data collection, processing and storage. This would enable a collaborative dialog between data providers and researchers, leading to co-design of an infrastructure that is easy to maintain and that meets researcher needs. It is our hope that this approach to data sharing will lead to interesting discoveries and security solutions that will be useful to both data providers and researchers. In such an ecosystem, data providers would be incentivized to share since sharing helps them benefit from research that ultimately improves their own security posture. Publication outlets in other fields often require that datasets be included with manuscript submission and curated with acceptance. For example, the Journal of Informatics (Elsevier) highly recommends sharing datasets with reviewers, and suggests that "sharing your data or code publicly supports research reproducibility, supports you with receiving credit for your work through citations and others with reuse"