Trace-Share: Towards Provable Network Traffic Measurement and Analysis

Milan Cermak
Masaryk University, Institute of Computer Science
Brno, Czech Republic
cermak@ics.muni.cz

Keywords. Network traffic analysis, Semi-labeled dataset, Research provability.

Abstract

Research in network traffic measurement and analysis is a long-lasting field with growing interest from both scientists and the industry. However, even after so many years, results replication, criticism, and review are still rare. We face not only a lack of research standards, but also inaccessibility of appropriate datasets that can be used for methods development and evaluation. Therefore, a lot of potentially high-quality research cannot be verified and is not adopted by the industry or the community.

The aim of our research is to overcome the mentioned controversy with focus on the whole issue covering all areas of data anonymization, authenticity, recency, publicity, and their usage for research provability. We believe that these challenges can be solved by utilization of semi-labeled datasets composed of real-world network traffic and annotated units with interest-related packet traces only. While the real-world traffic capture needs to be kept private, the annotated units can be freely shared since they only contain the interest-based trace of traffic with a minimum of private information. Our approach enables to insert such annotated events to an unlabeled real-world network traffic dataset and create semi-labeled dataset providing a ground truth used for the development of analytical methods as well as their validation. We do not claim that semi-labeled datasets provide a universal solution to all problems related to dataset usage. However, we aim to show, that it offers more benefits than other current approaches.

The most crucial part in creating a semi-labeled dataset is the adjustment of inserted annotated units so that their features, such as TTL values or packets delay, are indistinguishable from features of the real-world network traffic dataset. For this purpose, we are developing Trace-Mix[1] tool based on ID2T toolkit [2]. The tool analyzes real-world network traffic capture and calculates all necessary features of each connection. Based on these features, it allows to insert selected annotated unit at the specified time and adjust it to match the original dataset (e.g., according to common characteristics of all connections with the target IP address). The IP addresses of the annotated unit can be fully adjusted to match addresses in the original capture. Alternatively, the original IP address distribution of annotated units can be preserved. In this case, original IP addresses from annotated units can serve as a natural label of the inserted traffic. Our approach makes possible to insert any annotated unit into a real-world network traffic and create datasets for development and verification of various measurement and analysis methods.

Semi-labeled datasets can be used for the development of new analysis methods, or adaptation of deployed methods to specifics of a given network, as well as for the verification of their correctness. Figure 1 demonstrates the use of the semi-labeled dataset on the example of a development of network threat detection inspired by PDSA methodology. A similar approach can be used for tweaking of measurement

[1] Trace-Mix tool is publicly available at https://github.com/Trace-Share.

and analysis methods so that they work properly within a given network. In this case, the annotated units serve as ground truth, which must always be recognized while false positives are reduced. In the case of method validation, it is possible to use different annotated units containing a similar event, for example various types of DDoS attack, and test whether they are correctly recognized by the tested method.

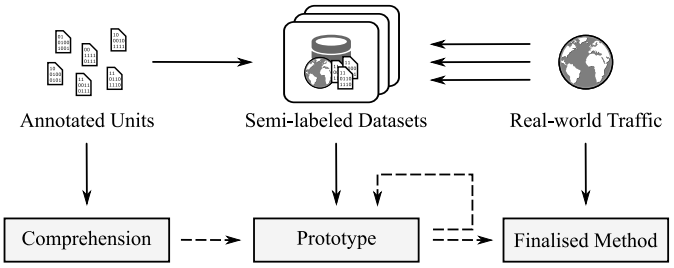

Figure 1: Usage of annotated units and semi-labeled datasets for analysis method development [1].

Sharing of annotated units and cooperation represents a vital component of the whole approach. We are currently developing a sharing platform called Trace-Share intended as a community hub. The platform is built upon essential functions of uploading, searching, downloading, and mixing of annotated units. The unified procedures for normalization with anonymization included will narrow the heterogeneity of the shared data and, at the same time, build trust in the sharing platform. Furthermore, the community-based approach, such as commenting or tagging of units, will help to alleviate some workload from the hub managers and ensure project sustainability with regular updates.

This article is a brief introduction of the concept of semi-labeled datasets, and we are aware that many challenges need to be addressed in further research. Our goal is not to deal with all identified problems at this point, but to present a general solution to start a discussion of its usability. We hope that the follow-up discussions will help us to move forward to a solution that will be accepted by the research community, help us to establish better research conditions, and make research more accessible to other researchers and the industry as well.

Paper origin

The original paper has been accepted and presented at the Network Traffic Measurement and Analysis Conference (TMA 2018) [1].

Acknowledgment

This research was supported by ERDF "CyberSecurity, CyberCrime and Critical Information Infrastructures Center of Excellence" (No. CZ.02.1.01/0.0/0.0/16_019/0000822).

References

[1] M. Cermak, T. Jirsik, P. Velan, J. Komarkova, S. Spacek, M. Drasar and T. Plesnik. "Towards Provable Network Traffic Measurement and Analysis via Semi-Labeled Trace Datasets," in *2018 Network Traffic Measurement and Analysis Conference (TMA)*. Vienna, Austria: IEEE, 2018. doi:10.23919/TMA.2018.8506498.

[2] E. Vasilomanolakis, C. G. Cordero, N. Milanov, and M. Muhlhauser, "Towards the creation of synthetic, yet realistic, intrusion detection datasets," in *NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium*. IEEE, apr 2016, pp. 1209–1214.