Collecting, Labeling, and Using Networking Data: the Intersection of AI and Networking

John Heidemann* Jelena Mirkovic* Wes Hardaker* Michalis Kallitsis**
*USC/ISI **Merit Network, Inc.

1 INTRODUCTION

Networks face many threats, ranging from DDoS attacks that overwhelm services, to compromised IoT devices, to vulnerability scanning and intrusion attempts. Many of these problems can be framed as anomalies that deviate from regular traffic, when attackers strive to blend in. Further, network managers face the perpetual need for accurate traffic forecasting to assist them in capacity planning and other traffic engineering tasks.

The AI revolution provides Machine Learning the opportunity to address long-standing problems like these that could not be resolved through deterministic algorithms. In fields such as text recognition, machine translation, image labeling, and computer games, machine learning has solved a number of long-term problems. But machine learning needs rich, diverse and huge datasets for training. Networking historically lacks such datasets in the public domain.

Can we use Machine Learning to address longstanding problems in networking and cybersecurity? What are the biggest challenges to do so? We suggest that there are three barriers to overcome: (1) broadening collection and distribution of data, (2) improving and sharing of labels on that data, and (3) evaluating and developing network-specific features. Our new project, CLASSNET [10], hopes to contribute to lowering these barriers.

2 DATA COLLECTION AND AVAILABLE DATA

To apply data-centric ML techniques to networking, one must start with interesting data! To be interesting, first the data must be "real", accurately characterizing some aspect of real-world traffic (the "background"), often including typical natural variation (diurnal or seasonal patterns). Second, it must contain events relevant to research: DDoS attacks, phishing attempts, lateral movement, or other detection targets.

Challenges: Both of these requirements are challenging. Getting real traffic requires access to a live network, but with the widespread use of the Internet for all kinds of purposes (personal, private, medical, educational, social), general network traffic is usually considered privacy sensitive. Ethical research standards require researchers to obtain informed consent from users whose traffic they leverage. In a small network (say, a research laboratory), consent of each user may be possible, but the observed data will necessarily be specialized, narrow and likely not representative. Data from a large network (say, a university) will be diverse, but it will certainly contain sensitive traffic, and it is impossible to obtain consent from all campus users. (In fact, identifying network users to obtain consent itself may increase the privacy risk.) While some networks include "data available for research" in their standard terms-of-use, but operational networks often remain hesitant to share data without additional safeguards. Finally, we know that Internet traffic and network conditions vary greatly [4], so researchers need data from multiple, diverse, large networks.

Second, we must identify research-interesting events and highlight them. While small events occur often, identifying them in time to collect and curate data, on top of other operational requirements, remains a burden.

Potential Approaches: We suggest that *careful data collection* and *controlled data sharing* can make more data available to the community.

Thanks to pervasive use of encryption, privacy of most users is already well-protected in network traffic. While in prior years the "the Internet" was regarded as unobserved, many general networks are no longer considered secure, and pervasive monitoring is recognized as a risk that should be addressed through protocol design, as recommended by RFC7258 [3]. We see evidence of this assumption through widespread use of TLS for nearly all web traffic, e-mail exchange, and recently even for DNS [11]. We suggest that most users understand that wireless networks (e.g. public wifi) require link-level security, and that encrypted user traffic can be used for research, if the research is done in a controlled manner.

We also point to non-traditional data sources to complement live network traffic. Data from "capture-the-flag" hacking competitions is attractive because consent is available from all parties, and because the contest operator can provide some ground truth. Partially or completely synthetic datasets can even be of interest for research use. Mixing known attacks with anonymized real-world traces can protect user privacy and provide ground truth. While fully synthetic datasets risk missing important traffic features, they are still valuable for some research tasks.

Finally, *controlled sharing* plays an important role in responsible dataset use. While in prior years data was posted on the public Internet for use, the richness of modern data raises potential risks. We recognize that even anonymized data risks leaking some information, even if it cannot be traced to an individual [7], suggests that multiple steps should be taken to protect data sharing. We recommend basic anonymization in all cases, scrambling at least part of the IP address and removing any unencrypted payloads. If a particular research question requires only specific packet fields or features, dropping other fields and payloads is clearly safer than preserving them.

An additional layer of defense atop anonymization is to provide data only to researchers with a clear research need and with a contractual agreement. While it is important to share data broadly with the research community, having a written contract between the data provider and researchers helps identify precisely how the

Workshop on Data for AI in Network Systems, Oct. 2021, Virtual

data will be used, and by whom. Such an agreement should include clear rules forbidding de-anonymization of the data. Formal legal agreements can be difficult to carry out, but provide safety through legal ramifications for researchers that fail to comply. Written agreements between researchers without formal legal endorsement do not go as far, but at least clearly state expectations.

We see some hope for providing real-world network data with controlled sharing. The CRAWDAD project provided wireless data for a number of years with NSF support [9]. The DHS IMPACT program operated for nearly a decade, providing networking traces, DDoS, darknet, and topology datasets for network and security research [8]. In the CLASSNET project, we hope to build on these prior efforts to continue to collect and provide these types of data to researchers—we plan to work with multiple network operators to provide diverse network data to the community.

3 LABELING DATA

An enabler to recent success in machine learning has been the use of huge datasets labeled with ground truth—ImageNet [2] for computer vision and Argoverse [1] for autonomous driving. Unfortunately, networking and security datasets do not have such large datasets, nor do they contain high quality labels that identify the malicious actors in the traffic. Building a large, labeled dataset is very labor-intensive.

One challenge for data labeling is the difficulty in finding interesting events in datasets. Although security events occur frequently, *identifying* them in traffic may be difficult. Some may be stealthy, and some may leverage new attack techniques. Further, if security events were easy to identify, they would already be removed and no longer of interest. In addition to the challenge of labeling security events, the data labeler must also consider benign network incidents (like flash crowds) that are atypical but not malicious.

We see three paths towards obtaining large and labeled network datasets. First is to construct synthetic datasets that mix real traffic with known anomalies. By adding in the anomalies, we can label them, and by mixing them with real traffic we can provide a rich background to avoid overfitting during training. Other types of artificial datasets, such as capture-the-flag data, can also provide traffic where attacks and background traffic are externally known. Fully synthetic datasets such as the DARPA IDS dataset are still in use after more than a decade [6]; while dated and imperfect, they are relatively large and labeled.

Second, one can leverage commercial tools already running in a network to label some interesting events, such as cloud-based DDoS defenses, intrusion detection systems, etc.

Finally, in CLASSNET, we plan to explore collaborative, community-driven labeling. We accept that there is no perfect ground truth possible in real data, since one can never guarantee that all events of interest have been discovered. From this viewpoint, we envision allowing researchers to apply different algorithms to label the data and submit their labels into a shared repository. Ultimately each record will end up having multiple labels, and the users can decide which labels to use for ground truth. We expect that shared use of a few specific datasets and their examination by multiple researchers will help build confidence in label quality, as well as spread the effort at evaluating these labels.

4 FEATURE ENGINEERING

A final interaction in networking and ML is *feature engineering*—selecting which features to extract from the data before used as input for classification or other machine learning tasks. While classification algorithms are quite powerful, ultimately they perform efficient mathematical clustering in some high-dimensional space. Those dimensions are derived from features extracted from the data, so careful selection of those features is critical their success.

Networking experts are best prepared to identify potential features that exist in networking data and know to extract and possibly normalize them. As a simple example, one can easily apply n-gram clustering algorithms by treating fields in packet headers as words. However, domain expertise in networking may suggest that some fields are uninteresting. For example, the IP version number and packet checksums have little value, either because they take on too few values (IP version is 4 or 6), or they are effectively random (the checksum). Other fields, like MAC address, are nominally opaque 48-bit numbers, but the upper three bytes indicate a vendor code, while the rest serve to identify a specific device, – thus, its parts have very different values when classifying data on a LAN. Similarly, domain-specific knowledge can help distinguish between the random and payload parts of spam or malware.

On the other hand, ML experts can use their tools to reveal information about networking. Given labeled data and a classification algorithm, one can reverse the process to determine which features provide the greatest discriminative power, perhaps revealing patterns in network protocols about which we were unaware. And of course, we look to AI to contribute new ML-based classification techniques that are improve efficiency or scalability to large datasets. These will leverage domain-specific features.

As an example, in ML-based classification of reverse DNS data, we found a mix of rate-based properties (such as queries per time) and protocol-specific properties (such as IP address entropy) helped to classify traffic [5].

5 CONCLUSIONS AND FUTURE DIRECTIONS

AI- and ML-based approaches are already being applied to networking Our goal, and the goal of CLASSNET, is to enable deeper collaboration with relevant data, backed up by labels, to develop new techniques, features, and methods. To accomplish these goals, we will careful record or create new rich network datasets, labeled with known ground-truth or discovered through our manual and algorithmic label-sharing research platform.

Acknowledgments: This work is partially supported by the project "Community Labeling and Sharing of Security and Networking Test datasets (CLASSNET)" " (NSF CRI-8115780).

REFERENCES

[1] Argo AI. Argoverse public datasets. https://www.argoverse.org, 2021.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, June 2009. IEEE.

[3] S. Farrell and H. Tschofenig. Pervasive monitoring is an attack. RFC 7758, Internet Request For Comments, May 2014. (also Internet BCP-188).

[4] Sally Floyd and Vern Paxson. Difficulties in simulating the Internet. *ACM/IEEE Transactions on Networking*, 9(4):392–403, August 2001.

[5] Kensuke Fukuda, John Heidemann, and Abdul Qadeer. Detecting malicious activity with DNS backscatter over time. *ACM/IEEE Transactions on Networking*, 25(5):3203–3218, August 2017.

[6] Manaf Gharaibeh and Christos Papadopoulos. Darpa 2009 intrusion detection dataset. Technical report, Colorado State University, August 2014.

[7] Basileal Imana, Aleksandra Korolova, and John Heidemann. Institutional privacy risks in sharing DNS data. In *Proceedings of the Applied Networking Research Workshop*, Virtual, July 2021. ACM.

[8] DHS IMPACT Program. The IMPACT portal. Website https://impactcybertrust.org/, 2016.

[9] David Kotz, Tristan Henderson, and Ilya Abyzov. CRAWDAD: A community resource for archiving wireless data at Dartmouth. web site http://crawdad.cs.dartmouth.edu/, December 2004.

[10] Jelena Mirkovic, Michalis Kallitsis, Wes Hardaker, and John Heidemann. Community LAbeling and Sharing of Security and NETworking test datasets (CLASSNET). website https://ant.isi.edu/classnet/, October 2021.

[11] Liang Zhu, Zi Hu, John Heidemann, Duane Wessels, Allison Mankin, and Nikita Somaiya. Connection-oriented DNS to improve privacy and security. In *Proceedings of the 36th IEEE Symposium on Security and Privacy*, pages 171–186, San Jose, Californa, USA, May 2015. IEEE.