Challenges in Network Security, Privacy, and the Law

Erick Galinkin*

October 4, 2021

1 Introduction

No small amount has been written to date on the topic of network security – a subject that is taught to countless undergraduate and graduate students each year and powers an industry expected to exceed $150 billion this year [4]. Despite this, the adoption of artificial intelligence in network security systems, has lagged behind other security disciplines such as malware detection and user behavior analytics. The availability of freely available datasets has been cited as a rationale for this difficulty since, in contrast with the likes of EMBER [2] and SoReL-20m [5], datasets for network security tend to be from single-sources and due to the rapidly changing nature of the web, are quickly outdated. In this paper, we discuss the data challenges in detail and offer solutions for those who may want to make this sort of data freely available.

2 Features of Network Data

The overwhelming amount of modern networking protocols leverage the TCP/IP stack, meaning that the "natural" state of network data is in the form of packets [3] and their corresponding packet capture files or "pcaps". Pcaps are the basis for much of network intrusion detection and forensics, and are distributed[1] by threat researchers to help identify threats. These packet captures have structured data that is defined per-protocol, and parsers like those built into Wireshark[2] can help researchers identify files, streams, and so on.

Like natural language, the "meaning" of a packet is highly context dependent, which is why for many protocols, the full stream must be reconstructed as packets can and will be sent out of order, fragmented, or lost. If the full stream is not able to be reconstructed, this is akin to losing not a single word in a sentence, but rather a bunch of letters from scattered

*erick_galinkin@rapid7.com

[1] https://www.malware-traffic-analysis.net/

[2] https://www.wireshark.org/

words throughout the sentence, making it difficult or impossible to actually construct the sentence. This also presumes that if encryption is present, the traffic is able to be decrypted. Adoption of encryption across the internet is increasing, so the need for inline observation and decryption is crucial to the task. Assuming the full stream can be reconstructed, we can begin to analyze the data – a daunting task.

The body of the packets which contain the data sent back and forth between the client and server may itself require separate parsing, as application-specific instructions may be sent over a protocol that supports serving that application *e.g.* HTTP, SSH. This means that in order to extract features about network activity, we need knowledge of not only the protocol, but *the application on the protocol.* Using HTTP as an example, many hundreds of web applications and languages are transmitted over HTTP through normal web transactions. However, we also see entire protocols like DNS [6] are actually able to be implemented on top of the HTTP standard. Thus, in order to derive meaningful features from HTTP requests, we must not only parse the HTTP standard, but be able to parse out application-specific data and identify protocols implemented on top of it. Consequently, any dataset that we would wish to collect and provide must instead be preceded by a definition of what the sort of traffic we seek to collect and share should look like, allowing parsers to be set up so the data can be appropriately featurized.

3 Privacy and the Law

Another tremendous complication in the derivation of network datasets is the implementation of data privacy laws such as the European Union's General Data Protection Regulation [1] (GDPR), which establishes the rights of individuals over their data. Collecting traffic without the express consent of individuals would not be permissible, so any organization wishing to collect this traffic would need express consent. The Hawthorne effect [8] – the phenomenon of individuals behaving differently when they know they are being watched – comes into play here, and will bias the behavior of individuals subject to collection. Furthermore, in order to distribute the data responsibly, personal information like usernames, passwords, and other sensitive data would need to be scrubbed.

In addition, GDPR explicitly specifies IP address as a form of personally identifiable information, even if that IP address is dynamically assigned. This means that any collected data which could possibly contain traffic associated with European Union citizens is subject to the GDPR and must be treated accordingly. Whether or not pseudonymization is sufficient in this context remains unclear [7] and so the collection and distribution of this sort of data is legally fraught. In order to collect and distribute these sorts of data, an organization with sufficient reach, funding, and motivation would need to do so and also work with lawyers to ensure that there is no personal information conatined.

References

[1] General data protection regulation.

[2] Anderson, H. S., and Roth, P. Ember: an open dataset for training static pe malware machine learning models. *arXiv preprint arXiv:1804.04637* (2018).

[3] Fall, K. R., and Stevens, W. R. *TCP/IP illustrated, volume 1: The protocols.* addison-Wesley, 2011.

[4] Gartner. Gartner forecasts worldwide security and risk management spending to exceed $150 billion in 2021. https://www.gartner.com/en/newsroom/press-releases/2021-05-17-gartner-forecasts-worldwide-security-and-risk-managem, 2021. Online; accessed 3 September 2021.

[5] Harang, R., and Rudd, E. M. Sorel-20m: A large scale benchmark dataset for malicious pe detection. *arXiv preprint arXiv:2012.07634* (2020).

[6] Hoffman, P., and McManus, P. Dns queries over https (doh). RFC 8484, RFC Editor, October 2018.

[7] Mourby, M., Mackey, E., Elliot, M., Gowans, H., Wallace, S. E., Bell, J., Smith, H., Aidinlis, S., and Kaye, J. Are pseudonymiseddata always personal data? implications of the gdpr for administrative data research in the uk. *Computer Law & Security Review 34*, 2 (2018), 222–233.

[8] of Bias Collaboration, S. C., Spencer, E., Mahtani, K., et al. Hawthorne bias. *Catalogue of Bias* (2017).