Threat Detection via Network Flows and Logs Correlation

Stanislav Spacek[1,2] and Pavel Celeda[2]

[1] Faculty of Informatics, Masaryk University, Brno, Czech Republic
[2] Institute of Computer Science, Masaryk University, Brno, Czech Republic
{spaceks|celeda}@ics.muni.cz

Abstract. A rising amount of mutually interconnected and communicating devices puts increasing demands on cybersecurity operators and their tools. With the rise of end-to-end encryption, it is becoming increasingly difficult to detect threats in network traffic. With such motivation, this Ph.D. proposal aims to find new methods for automatic detection of threats hiding in encrypted channels. The focus of the proposal is on correlating the data still available in the encrypted network flows with the data contained in the logs of network applications. Our research is in the initial phase and will contribute to a Ph.D. thesis in four years.

Keywords: intrusion detection, network flows, network logs, encrypted traffic

1 Introduction

The network traffic constantly evolves as new user applications and services are being introduced. On the one hand, they provide new possibilities for users, on the other hand, they pose new challenges for cybersecurity experts and the Network Intrusion Detection Systems (NIDS) designed for network threat detection. One of the latest challenges the NIDS must cope with is encrypted communication. We detected that approximately 60 % of the traffic generated by nearly 43 000 devices on our campus network used encryption in September 2017. This share will surely rise even higher in the future [2].

Current NIDS mostly use threat detection based on the network flow analysis [12]. It is a fine strategy suitable for detection of previously unknown attacks, but as attacks evolve and patterns in network traffic change, it must be continuously adapted to remain reliable. Initially, the network flow analysis was only able to analyze network traffic based on its source, destination, timestamp and the amount of transferred data. Recently, it allows for more fine-grained analysis based on the data from the application layer [3]. This way, it can identify applications and protocols in transfer and thus better distinguish between malicious and benign traffic.

Unfortunately, content analysis is not straightforward for encrypted traffic. The lack of information lost in encryption is recently approached from two directions. The first approach lies in the analysis of information that remain

unencrypted, for example, initial handshake to establish encryption parameters. The second approach encompasses the analysis of statistical data that we can measure even for encrypted traffic. The measurements include parameters like length and volume of the transmission, packet inter-arrival times and frequency of sent and received packets. Kovanen et al. provide a whole set of statistical features, which might indicate a threat [7]. The experiments with a combination of the above approaches were conducted by Anderson et al. [1]. We plan to achieve better results by adding another data source.

A different approach to security analysis of network traffic is event log monitoring. Logs are generated either on the key infrastructure elements such as the servers and routers or on the endpoints such as user devices. Some of these devices like IoT and mobile devices might be unable or unwilling to provide their logs. However, they still leave traces in the logs of the network infrastructure. Security Information and Event Management (SIEM) systems currently in use are designed specifically for threat detection by log analysis. They monitor and correlate events that are generated inside their constituency. This way, they are able to not only detect anomalous behavior, but also reconstruct a sequence of actions, for example, all actions of a user in a specific time window.

To our knowledge, intrusion detection is currently done separately in network flows and logs. Gu et al. [5] combined both approaches for the purpose of botnet detection. However, logs and flows were analyzed separately to achieve better detection accuracy. In this Ph.D. proposal, we intend to use the logs of network applications as an additional information input along the network flow data. When appropriately correlated, this new base of data will provide new insight into network traffic and extend the range of detectable attacks. Specifically, we expect that including logs will improve detection rate for attacks in encrypted traffic.

2 Research Questions

The main objective of the proposed research is to detect advanced threats in evolving network traffic. We formulated the following research questions that need to be answered in order to achieve our goal.

1. **How can we ameliorate the network flow analysis with logs to detect threats in a constantly evolving environment?**
   The monitoring of network flows provides insight into network events from a different perspective than the monitoring of network logs. We plan to investigate possibilities of correlating information contained in both network flows and logs so that a common base of data can be established. The parameters we identified for the correlation so far are the timestamp and source and destination addresses that can help to identify specific service and its corresponding log. We expect that when starting from this extended base, detection methods can achieve higher accuracy and detect threats that currently remain hidden.

2. **How can the network flow and log correlation improve the threat detection rates in encrypted traffic?**
   The content analysis and the deep packet inspection is not possible for encrypted communication, but information can still be gained from the unencrypted handshakes and statistical analysis. For example, the SSL/TLS handshake may be extracted from a flow of a client-server communication and correlated with the corresponding server log based on the client IP address and timestamp. The handshake then provides list of available cyphers and the log provides the service name and the actions executed by the client, thus extending the base of data available for analysis.

3 Proposed Approach

3.1 Stream Representation of Logs and Flows

There are many different ways to store logs, and log entries themselves differ from application to application as they usually combine runtime variables with human-readable text. Therefore, the logs must be first transformed into a unified form and then directed to central storage where further analysis takes place. The unified form must contain all the information from the original log entry. It must be processable automatically and, should be similar to the form in which network flows are processed.

A log processing model is naturally represented by the event-driven architecture [4]. When the event-driven architecture is applied, a log entry corresponds to an event entity. The application that generates network logs fits the role of an event producer, and the central log storage acts as an event consumer.

The event entity unifies log entries generated by diverse log producers through the use of parameters. The parameters are either generic, for example the timestamp and source, or producer-specific. The producer-specific parameters differ by a producer and also by log entry type. For example, an SSH log producer may generate authentication success, failure, disconnection and other messages that each have a different set of specific parameters. Nevertheless, the result is a unified data structure that represents various log entries. It can be automatically processed and does not omit any information contained in the original log entry.

Typically, network applications generate a large number of log events. These events may be grouped into streams by the originating network application or by the actual device that was the source. This allows specifying the scope of monitoring either locally or on the whole network. The possibilities of event grouping into streams were researched by Tovarňák [11].

The stream-based processing of network flows is possible, as shown by Jirsik et al. [6]. The log event stream can be correlated to the flow stream based on similar properties that will be the subject of our research. A possible approach is through timing-based correlation since timestamp is a parameter present both in flows and log events. However, time synchronization is rarely accurate over the network so a tolerance scope must be defined. While a large tolerance scope will lead to an inaccurate association of logs to flows, a too small tolerance scope will

struggle to associate any logs to flows. This is also an issue that we plan to focus on in our research, along with exploring other parameters for correlation.

3.2 Threat Detection in Encrypted Traffic

The challenge of the encryption poses for threat detection might be demonstrated on botnets. Botnets often use domain names instead of IP addresses to communicate [9]. In network flows, the domain name might be learned either from the Server Name Indication (SNI) parameter of application layer or the DNS communication. Currently, the SNI is unencrypted in SSL/TLS communication. However, recent drafts of TLS 1.3 propose encrypting the SNI to improve communication privacy [10], and DNS communication might also be encrypted by DNSSEC. Adding logs to flow analysis helps to resolve this issue. The domain name can be obtained by correlating the communication flows with the DNS log.

Most of the current threat detection methods for encrypted traffic use machine learning techniques, particularly clustering algorithms, to identify anomalies in the network flow [13]. The k-means and k-nearest neighbor algorithms are the most and second most used one respectively. The popularity of k-means is due to its variability, which allows it to be fine-tuned for various purposes. The algorithm variability is an important property since the detection method must meet several criteria. First, it must be fast enough to process a huge amount of data supplied by the monitoring infrastructure. Second, it must be accurate enough to detect as many attacks as possible. It will pose a challenge to balance speed and accuracy, as both may prove mutually exclusive.

3.3 Testing Methodology

The most prevalent public datasets currently used for detection method testing are DARPA98 and KDD99, released in 1998 and 1999 respectively [12], despite the fact that the limitations of the DARPA dataset were extensively covered by McHugh et al. [8]. Since network traffic changed drastically during last 20 years, a newer dataset will be used to better represent the current environment. The testing will be done by injecting manually labeled malware samples into background traffic and logs captured on a real network. This approach allows the creation of different datasets by changing the background traffic and belittles the issue of overfitting. We will share the samples to ensure results reproducibility in the future.

4 Conclusion

In this research, we aim to expand the base of data used by current detection methods to discover threats in network traffic. We consider logs as an information-rich source that, when properly correlated with network flow data, will provide new insight into the network traffic. We expect that our research will bring following contributions. Firstly, we will define new methods for log and flow correlation. Secondly, we will use the correlated data to improve NIDS threat

detection rates. Lastly, our testing datasets will be available for future detection methods verification.

Acknowledgement

This research was supported by the Security Research Programme of the Czech Republic 2015 - 2020 (BV III / 1 VS) granted by the Ministry of the Interior of the Czech Republic under No. VI20172020070 Research of Tools for Cyber Situation Awareness and Decision Support of CSIRT Teams in the Protection of Critical Infrastructure.

References

1. Anderson, B., McGrew, D.: Identifying Encrypted Malware Traffic with Contextual Flow Data. In: Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security. pp. 35–46. AISec '16, ACM, New York, NY, USA (2016)
2. Cisco Systems: Encrypted Traffic Analysis. Whitepaper (January 2018), https://www.cisco.com/c/dam/en/us/solutions/collateral/enterprise-networks/enterprise-network-security/nb-09-encrytd-traf-anlytcs-wp-cte-en.pdf, accessed in April 2018.
3. Claise, B., Trammell, B., Aitken, P.: Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information. RFC (September 2013), https://tools.ietf.org/html/rfc7011, accessed in April 2018.
4. Etzion, O.: Event Processing in Action. Manning, Greenwich Conn (2011)
5. Gu, G., Perdisci, R., Zhang, J., Lee, W., et al.: BotMiner: Clustering Analysis of Network Traffic for Protocol-and Structure-Independent Botnet Detection. In: USENIX security symposium. vol. 5, pp. 139–154 (2008)
6. Jirsik, T., Cermak, M., Tovarnak, D., Celeda, P.: Toward Stream-Based IP Flow Analysis. IEEE Communications Magazine 55(7), 70–76 (2017)
7. Kovanen, T., David, G., Hämäläinen, T.: Survey: Intrusion Detection Systems in Encrypted Traffic. In: Internet of Things, Smart Spaces, and Next Generation Networks and Systems, pp. 281–293. Springer (2016)
8. McHugh, J.: Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory. ACM Transactions on Information and System Security 3(4), 262–294 (nov 2000)
9. Nazario, J., Holz, T.: As the net churns: Fast-flux botnet observations. In: Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on. pp. 24–31. IEEE (2008)
10. Rescorla, E., Huitema, C.: SNI Encryption in TLS Through Tunneling. RFC Draft (March 2018), https://tools.ietf.org/html/draft-ietf-tls-sni-encryption-02, accessed in April 2018.
11. Tovarňák, D.: Normalization of Unstructured Log Data into Streams of Structured Event Objects. Dissertation thesis, Masaryk University, Faculty of Informatics, Brno (2018), https://is.muni.cz/th/rjfzq/
12. Umer, M.F., Sher, M., Bi, Y.: Flow-based Intrusion Detection: Techniques and Challenges. Computers & Security 70, 238–254 (2017)
13. Velan, P., Čermák, M., Čeleda, P., Drašar, M.: A Survey of Methods for Encrypted Traffic Classification and Analysis. International Journal of Network Management 25(5), 355–374 (2015)