BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2024
http://excel.fit.vutbr.cz

Rating Log Events Using Reputation and Anomaly Scores

Bc. Jan Zbořil*

Abstract
The aim of this paper is to develop a method for scoring logs containing enhanced NetFlow data using anomaly and reputation scores (AS, RS).
Principle Component Analysis statistical method is used for detection of records which are anomalous regarding to the normal traffic. RS, computed from features related to security and bad behaviour, is combined with the AS, for each log record to filter it out or to create an alert.
Both resulting scores are evaluated, their relation is described, analysis of the results in comparison with the training data is performed. RS is compared to existing solutions. The combination of AS and RS successfully reduced more than 727 000 events to 743 alerts, a reduction to 0.1 % of the original number of log records.

*xzbori20@stud.fit.vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Introduction

Traffic logs and flows hide many patterns, which might be useful for analysis by network and security engineers or system administrators. Reducing the number of events shown to them enables easier, faster, and more pleasant analysis and problem-solving.

By combining anomaly score (AS) and reputation score (RS) computed for each IP address from historical data, one can rate present events recorded in the logs. Rated logs are easier to filter, or they can be used for alerting.

Both anomaly detection and reputation scoring are well studied fields. Many methods for anomaly detection exist, utilising statistical methods [1, 2], machine learning [3], or other approaches. The methods mainly differ by the outcome and the suitable input for each method. RS is successfully utilised in commercial projects (Cisco Talos [4]) or academic research (CESNET NERD [5]).

This project shows the entire process from data analysis, selecting the correct method for the dataset, choosing the right features, data preprocessing, anomaly detection, reputation scoring, log enhancement, to an evaluation of results using experiments.

Using the combination of anomaly detection and reputation scoring for rating logs is, however, a novelty.

2. Anomaly Detection

Anomaly score is the first metric used for scoring log events. AS represents the amount of deviation between the current network traffic and learned baseline. It is computed separately for each network node detected in the data, which is represented by an IP address. Computing the score for all IP addresses would skew the score towards the most prominent IPs (by traffic volume) to be dominant.

Features are selected according to an analysis conducted upon a dataset consisting of Suricata IDS logs of flow records. The features are: date, type, day of week, hour, source & destination ports, L4 protocol, packets & bytes to server, packets & bytes to client, HTTP hostname, HTTP user agent, HTTP content type, HTTP method, HTTP status, HTTP length, DNS flags, DNS type, MQTT host, MQTT type, anomaly protocol, alert category.

Dataset for each IP was preprocessed using categorical transformation (text features to numeric), dropping singular columns and feature normalisation to L2 norm [6]. These steps are shown in Fig 1.

Mahalanobis distance [7] is calculated for every data

point to determine its position according to the rest of the dataset. 0.5 % of values with the highest distance are marked as outliers.

The dataset is sorted by timestamp. The first two thirds of the data are used for training, the last third is regarded as testing data. Data points marked as outliers are removed from the training part, while they are kept in the testing dataset.

The Principal Component Analysis model is fitted with the training data. Each data point in the testing dataset is transformed, and then inverse transformed by the previously fitted PCA model. PCA, as a linear transformation, is reversible, although some amount of information is inherently lost. This loss of information, measured by mean square error, represents the AS of a given data point. An iterative algorithm working with confusion matrices is used to determine the best threshold. Entire process is shown in Alg 1.

3. Reputation Scoring

These features were used for calculation of RS: IP, date, type; HTTP hostname, url, user agent; DNS rrname, rrtype; anomaly protocol, type, event; alert signature, category, severity. RS baseline is 0 and RS is increased each time an IP address behaves badly. Subscores are computed either with data available from the EVE logs only (alerts), or external information is needed (DNS, HTTP) Daily score of an IP address is computed according to Eq 1. The curve defined in Eq 1 is demonstrated in Fig 2. Overall score is computed as a weighted average of daily scores across the last 14 days, simulating ageing. The formula for the overall RS is shown in Eq 2.

4. Dataset

The unlabelled dataset used for this work is composed of EVE JSON logs with extended flow records generated by Suricata IDS deployed at FIT. Analysis of the data was carried out to determine the most suitable features and methods. Features were extracted from most prominent protocols as seen in Fig 3. Fig 4 shows the distribution of flows by length. Information provided in the figure is important in relevance to the AS results. Fig 5 shows the original data format and its transformation suitable for further processing.

5. Results

The goal of this work was to score log records in the Suricata EVE logs. When the AS was applied, 0.62 % of events were marked as anomalous. Combining the AS with RS with threshold 0.3, the number of alerts drops to 0.1 %, as seen in Tab 1.

The behaviour of AS is shown in Fig 6. The shape created by the data is caused by the distribution of events in time - most events happened close to noon each day. Anomalies are mainly extremely long MQTT records, spanning up to 10 days. For the constant use of the same ports, the communication is reported as one long flow, instead of being split. TLS, HTTP and SSH flows longer than average flow length are also considered anomalous.

RS for worst behaving nodes is demonstrated in Fig 7. It shows the slow start and ageing behaviour of the score. RS cannot rise by more than 0.1 per day, for it is composed of multiple daily scores. If the node stops behaving maliciously, RS slowly fades to zero for up to 14 days. RS can stagnate or oscillate, when an IP address continuously misbehaves.

Table Tab 2 compares results of my RS with existing solutions. Although the available dataset is composed of year-old records, the bad behaviour of nodes on the Internet is stable for long periods of time. More than half of the worst behaving nodes detected by my method were, or are, also presented as IPs with bad reputation, either by Talos or NERD. Addresses with terrible reputation according to NERD, which were also reported by me, but with low score were caused by small amount of traffic in the dataset. RS of such nodes would most likely be higher, if they communicated more.

No significant correlation between AS and RS was observed. The scoring with enhancement of the original logs took 130 seconds for dataset with 1 000 000 data points. The preprocessing and score computation itself took 55 seconds.

6. Conclusions

A method for rating log events using AS and RS was proposed. The implemented method showed how log events from Suricata can be used to train machine learning models for anomaly detection and how RS can be computed.

A proposed method of matching RS to the original logs, together with AS, can serve as base for an entirely automatic way of deciding, whether a given IP address is malicious, or if it behaves extraordinarily.

A reduction of these log files by including only records with scores above a specified threshold was proposed, implemented and demonstrated. Experiments with the results were conducted and observed behaviour of scores was explained.

References

[1] Mei-Ling Shyu, Shu-Ching Chen, Kanoksri Sarinnapakorn, and LiWu Chang. A novel anomaly detection scheme based on principal component classifier. *Proceedings of International Conference on Data Mining*, Jan 2003.

[2] Zhaoli Liu, Tao Qin, Xiaohong Guan, Hezhi Jiang, and Chenxu Wang. An integrated method for anomaly detection from massive system logs. *IEEE Access*, 6:30602–30611, 2018.

[3] Marta Catillo, Antonio Pecchia, and Umberto Villano. Autolog: Anomaly detection by deep autoencoding of system logs. *Expert Systems with Applications*, 191:116263, 2022.

[4] Cisco Systems. Reputation center. online, 2024. https://www.talosintelligence.com/reputation.

[5] Václav Bartoš. Network entity reputation database (nerd). online, 2020. https://nerd.cesnet.cz.

[6] Eric W. Weisstein. L^2-norm. online, 2024. https://mathworld.wolfram.com/L2-Norm.html.

[7] StatisticsHowTo.com. Mahalanobis distance. online, 2023. https://www.statisticshowto.com/mahalanobis-distance/.