Network Connection Anomaly Detection and Analysis for Potential Attacks.

Huan Zhao
hzhao2@memphis.edu
U00879231

***Abstract*—In a world where digital interactions are essential, analyzing log tables, which track system activities, is key to understanding user behavior and spotting potential risks or opportunities. This project examines network connection logs to find anomalies—unusual patterns that might signal malicious actions or new opportunities. Using a Gaussian model, we detect unusual data entries, focusing on behaviors like frequent server connections. We analyze specific examples and display the results through graphs and tables, providing clear insights into the detected unusual activities. This method helps improve system security and better understand user behavior.**

I. Introduction

In today's world, where digital interactions and transactions dominate, the study of log tables has become essential for identifying potential clients and understanding how users behave. These tables are organized data sets that capture interactions within a system, like attempts to connect to a server or email communications. The goal of this project is to analyze network connection logs, gathered from an anomaly detection platform found at http://www.secrepo.com/.

For beginners in this field, you can think of logs as records that computers keep, documenting every event, such as server connection attempts or email exchanges. When reviewing these records, we are particularly focused on entries that stand out from typical patterns, known as anomalies, as they may highlight potential risks or untapped opportunities.

Anomalies, in this context, could involve a user connecting to a server much more frequently than usual during a specific time frame. Such irregular behavior may indicate suspicious activities or other important events, making it essential to detect these patterns as quickly as possible.

The challenge we are tackling is the detection of these abnormal patterns, such as excessive connection attempts, within a large volume of regular data entries. Identifying such outliers is key to maintaining system security, uncovering unusual user behaviors, and pinpointing areas where improvements or actions may be needed.

II. Related Work

In today's highly connected world, where digital services and platforms are deeply embedded in daily life, the significance of studying log tables has greatly increased. These tables serve as organized datasets, carefully documenting every interaction or transaction occurring within a system, from server connection attempts to email exchanges, embodying the core of network communication. [1]

A. Understanding Network Connections

Network connections refer to the links created between multiple devices or computers to facilitate the sharing of information and resources. These connections are essential for the smooth operation of various digital services and platforms. Every time such a connection is attempted, it is logged in data tables, which provide a comprehensive record of network interactions.

B. Significance of Detecting Anomalies

Research has extensively focused on detecting unusual patterns in these log tables, as such anomalies often indicate deviations from normal behaviors. Identifying these anomalies is vital for promptly uncovering potential threats or opportunities, thus helping to protect the integrity of the systems in question. [2]

C. Exploring Log Tables for Anomalies

Academic studies have delved into the detailed examination of log tables to find irregularities within the data. These logs, much like detailed journals kept by computers, can reveal abnormal or unexpected behaviors when thoroughly analyzed, potentially highlighting security gaps or untapped opportunities within the digital space. [3]

D. Gaussian Model in Anomaly Detection

The Gaussian model is widely used in anomaly detection for its ability to differentiate between typical and atypical data points in logs. By calculating the average values and measuring data variance, this model helps identify significant outliers. This approach plays a critical role in maintaining system security by detecting potential threats or areas for improvement through abnormal patterns.

E. Enhancing Security and Gaining Insights

Detecting anomalies not only protects system security but also provides valuable insights into user behavior, highlighting areas for improvement or intervention. By applying advanced models and analyzing deviations in log data, important information can be revealed, which aids in both optimizing user experience and enhancing system security.

This project builds on established methodologies, utilizing the Gaussian model to examine network connection logs from an anomaly detection platform. The goal is to identify abnormal behaviors, such as frequent server connections, thereby

contributing to the ongoing efforts to improve system security and better understand user behavior.

III. Proposed Approach

The primary objective of this project is to develop a robust approach for identifying abnormal entries within the network connection log data. Utilizing the Gaussian model, the aim is to distinguish between typical and unusual patterns, which could indicate possible risks or new opportunities.

- Data Collection: Gather network connection logs from the source at http://www.secrepo.com/.
- Feature Identification: Analyze the structure of each table and its fields to choose features that reflect the behavior of the source IP. The goal is to isolate features that may point to unusual or suspicious activities.
- Data Processing: Focus on features such as byte_in and byte_out, and compute metrics like daily maximum, average, and total traffic. These metrics can help identify anomalies and provide clues regarding potential attack targets.
- Data Transformation: Adjust the data to ensure each feature and column aligns with a Gaussian distribution.
- Model Training: For each log entry, calculate the likelihood based on the Gaussian distribution, and assign an anomaly score. Entries with lower scores are more likely to be anomalies. The threshold is set at the lowest 10% of all scores, and key parameters used in these computations are saved.
- Model Prediction: Process and transform new data similarly to the training phase. Compute anomaly scores for each record, and flag any that fall below the predefined threshold as potential anomalies.

A. Data Collection

The log data, retrieved from http://www.secrepo.com/, act as digital records tracking various activities within a system, such as client-server connection attempts or email communications. For network connections, we focus on a specific user and analyze half a month's worth of data for that source (src), examining daily connections along with the byte_in and byte_out activity.

B. Feature Identification

A clear understanding of the network connection log structure and field meanings is essential. The logs detail how the src IP interacts with various destination IPs. With a wealth of data on byte_in and byte_out available, these metrics are selected to represent the behavior of the src IP. Sudden increases in data transmission on any given day may help identify potential security threats.

C. Data Analysis

Once the key features are chosen, we calculate the daily maximum, average, and total data flow. A significant rise in total traffic could be an indication of anomalous behavior, while the maximum and average flows might help in identifying the target of suspicious activity.

D. Data Transformation

Assuming that the data features follow a normal distribution, we apply transformations to ensure that each feature or column conforms to a Gaussian distribution.

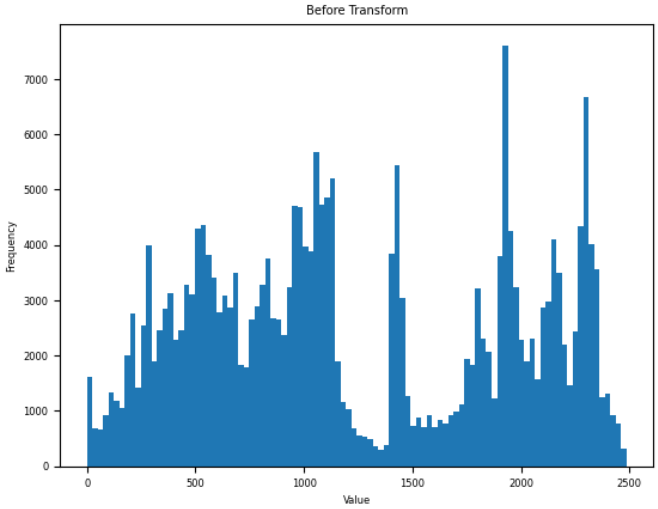


Fig. 1. The Original Data.

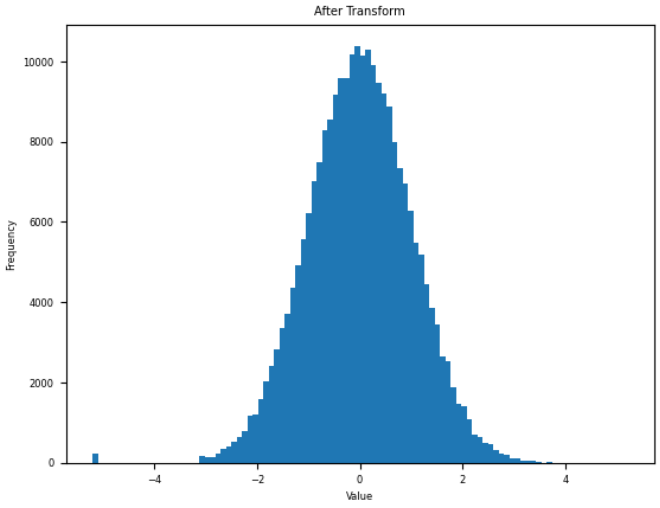


Fig. 2. The Transformed Data.

E. Gaussian Model Training

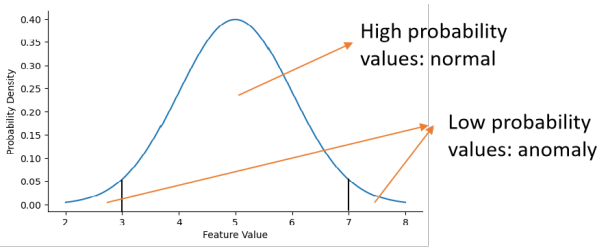
- Gaussian Model Implementation

  Compute the Mean: Establish the central point of the Gaussian distribution.

  Measure the Dispersion: Determine the standard deviation to understand how data varies around the mean.

  Detect Anomalies: Identify data points that deviate significantly from the mean, classifying them as anomalies.

- The Gaussian model is trained using at least two weeks of data to calculate the probability for each feature. Each day receives an anomaly score, and a threshold is defined. Scores significantly higher or lower than the threshold signal anomalies. Days with lower scores are treated as normal (negative samples), while higher scores indicate potential anomalous activities (positive samples), potentially pointing to suspicious behavior or attacks.



F. Gaussian Model Prediction

In this step, the data first undergoes pre-processing, where it is transformed similarly to the training phase. Anomaly scores are then computed for the newly processed data, and any scores below the set threshold are flagged as anomalies, marking significant deviations from expected behavior.

G. Anomaly Analysis

The analysis looks at large differences from normal behavior, like a user connecting to a server much more often than usual, which could suggest harmful actions. After finding these anomalies, they are studied to understand if they show possible risks, new opportunities, or areas that need improvement.

IV. Experiments and Results

In the evaluation section, we thoroughly test how well our solution can identify unusual entries in the network connection logs.

Our evaluation method is unique because we don't follow the usual metrics. This is because we don't have prior knowledge of which IP addresses might be attackers. Instead, we designed a process to find and confirm suspicious behaviors that could indicate potential attackers.

The goal is to identify possible attackers based on the actions of source IP addresses. We first choose a group of IP addresses for case studies. These IPs are run through the network model, and the results are shown as score graphs to help spot unusual activity. Then, we show an alert table with specific details about the detected anomalies, providing proof of why these points are considered unusual.

In our experiment, we focused on detecting anomalies from a source IP '192.168.74.10'. We trained our model with data from July 22, 2011, to June 30, 2013, and predicted results for July 1, 2013, to March 24, 2014. After loading the data, we analyzed the average, maximum, and total values for bytes, packets, and conversation lengths per day. This comprehensive approach helped us detect anomalies from different angles for this specific source.

A. Data Processing and Analysis

We first showed the raw data, then moved on to the processed data, which included the daily averages, highest values, and totals for bytes, packets, and conversation durations. This was an important step to establish the normal network activity and prepare for detecting any unusual behavior.

We assumed the data followed a normal (Gaussian) distribution, so we adjusted the processed data to match this assumption. Then, we showed the Gaussian distribution before and after the transformation to demonstrate how the data was adjusted.

B. Gaussian Model Training and Anomaly Detection

With the transformed data, we trained a Gaussian model and assigned a score to each day. We set a threshold at the 10th percentile, meaning we flagged the top 10% of the highest-scoring days as anomalies. This threshold was chosen to focus on the most extreme outliers and reduce the number of false positives.

C. Anomaly Logging and Feature Analysis

Next, we created logs for the days that were identified as anomalies. When we examined these logs, we noticed that many features on those days had significantly higher values compared to their normal averages. This was important because it not only confirmed the anomalies but also provided details about the nature and size of these unusual activities.

D. Baseline: Isolation Forest for Anomaly Detection

In addition to our primary method based on Gaussian distribution, we implemented a machine learning baseline using the Isolation Forest algorithm. [4] This method is well-suited for anomaly detection in high-dimensional data and operates without assuming a specific distribution for the data.

We trained the Isolation Forest model using the same dataset from July 22, 2011, to June 30, 2013, and used it to predict anomalies in the subsequent period. The model assigns an anomaly score based on how "isolated" a particular day's activity is from the rest of the dataset. Days that are more isolated (outliers) are flagged as anomalies.

E. Comparison of Results

The Isolation Forest baseline provided a useful point of comparison to our Gaussian-based method. While the Gaussian method was focused on detecting anomalies that deviate from a normal distribution, Isolation Forest captured anomalies based on broader patterns in the data, such as unusual combinations of multiple features (bytes, packets, duration, etc.).

Both methods produced anomaly logs, but the anomalies detected by the two approaches did not entirely overlap. This suggested that each method captured different types of anomalies: the Gaussian method focused on extreme outliers within a single feature distribution, whereas Isolation Forest flagged more complex outliers across multiple features.

	id.resp_h	duration_tot	duration_ave	duration_max	orig_bytes_tot	orig_bytes_ave	orig_bytes_max	resp_bytes_tot	resp_bytes_ave	resp_bytes_max	orig_pkts_tot	orig_pkts_ave	orig_pkts_max	resp_pkts_tot	resp_pkts_ave	resp_pkts_max	S0	S1	REJ
8/17/2011	24	861	5.937931034	35	14550	100.3448276	1919	14550	100.3448276	1919	660	4.551724138	195	525	3.620689655	289	29	0	0
9/10/2011	25	251	0.731778426	60	168507	491.2740525	1520	168507	491.2740525	1520	3555	10.36443149	277	4350	12.68221574	392	1	17	0
10/3/2011	18	334	7.767440186	64	21853	508.2093023	3904	21853	508.2093023	3904	930	21.62790698	265	1294	30.09302326	401	1	2	0
10/9/2011	14	458	13.878788788	63	43862	1329.151515	9039	43862	1329.151515	9039	1407	42.63636364	423	1995	60.45454545	627	1	13	0
10/18/2011	8	148	10.571428857	62	8857	632.6428571	2433	8857	632.6428571	2433	649	46.35714286	208	942	67.28571429	313	1	1	0

Fig. 3. Network Connection Log Processed Table.

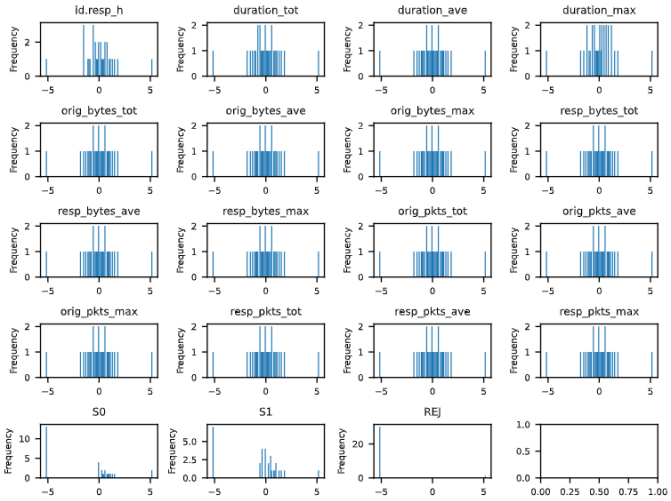


Fig. 4. Transformed Processed Table.

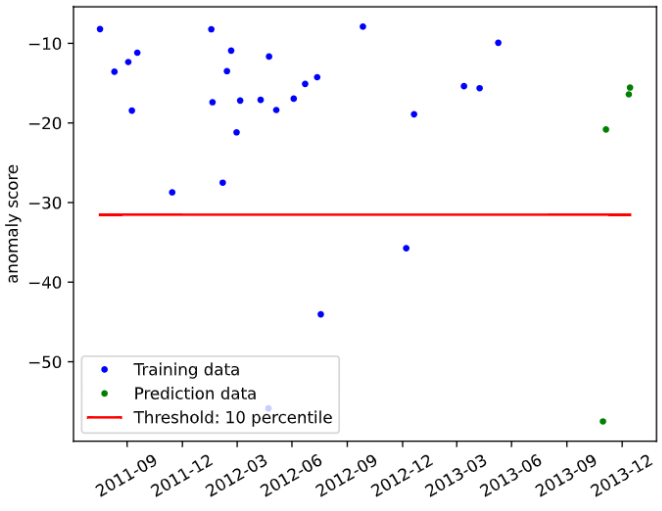


Fig. 5. Anomaly Score For Each Date.

F. Anomaly Logging and Feature Analysis

Next, we created logs for the days that were identified as anomalies by both the Gaussian method and the Isolation Forest baseline. When we examined these logs, we noticed that many features on those days had significantly higher values compared to their normal averages. This confirmed the anomalies and provided detailed insight into the nature and size of the unusual activities, further supporting the accuracy of both models in identifying potential attackers.

V. Discussion

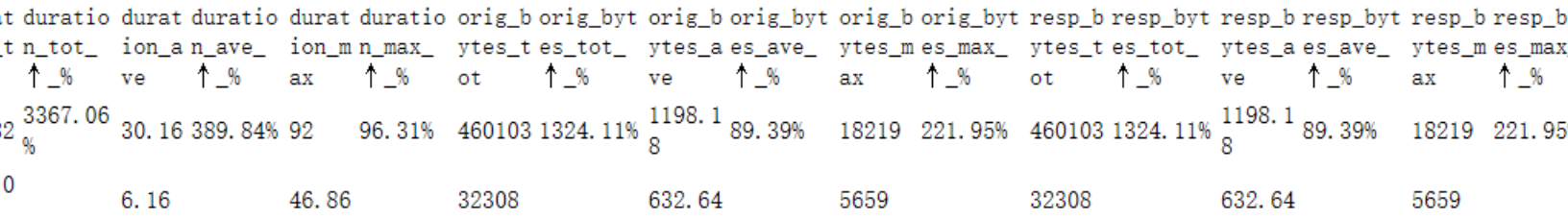
In this approach, while both the Gaussian distribution and Isolation Forest methods effectively detected anomalies in the network traffic, there are some limitations. First, assuming the data follows a Gaussian distribution may not always be accurate, especially in complex network environments. Additionally, setting the threshold can lead to a trade-off between detection accuracy and false positives. Although Isolation Forest captures more complex patterns, it lacks interpretability, making it difficult to pinpoint the exact cause of an anomaly. In the future, we could explore using time-series models or more advanced unsupervised methods like autoencoders to capture more intricate patterns and improve detection performance.

VI. Conclusions

In this project, we focused on analyzing network connection logs to detect unusual patterns, or anomalies, that might signal security threats or provide insights into user behavior. We applied a Gaussian model to distinguish between normal and abnormal activities in the logs. By defining what typical behavior looks like, we were able to identify and investigate any significant deviations. This approach not only helps to protect our systems from potential cyber risks but also gives us a better understanding of user behavior, which can be used to improve our services.

References

[1] Dumais S, Jeffries R, Russell DM, Tang D, Teevan J. Understanding user behavior through log data and analysis. Ways of Knowing in HCI. 2014:349-72.



t	duratio n_tot_ ↑_%	durat ion_a ve	duratio n_ave_ ↑_%	durat ion_m ax	duratio n_max_ ↑_%	orig_b ytes_t ot	orig_byt es_tot_ ↑_%	orig_b ytes_a ve	orig_byt es_ave_ ↑_%	orig_b ytes_m ax	orig_byt es_max_ ↑_%	resp_b ytes_t ot	resp_byt es_tot_ ↑_%	resp_b ytes_a ve	resp_byt es_ave_ ↑_%	resp_b ytes_m ax	resp_b es_max ↑_%
32	3367.06 %	30.16	389.84%	92	96.31%	460103	1324.11%	1198.18	89.39%	18219	221.95%	460103	1324.11%	1198.18	89.39%	18219	221.95
0		6.16		46.86		32308		632.64		5659		32308		632.64		5659	

Fig. 6. Alert table.

[2] Krivchenkov A, Misnevs B, Pavlyuk D. Intelligent methods in digital forensics: state of the art. InReliability and Statistics in Transportation and Communication: Selected Papers from the 18th International Conference on Reliability and Statistics in Transportation and Communication, RelStat'18, 17-20 October 2018, Riga, Latvia 18 2019 (pp. 274-284). Springer International Publishing.

[3] Davis JJ, Clark AJ. Data preprocessing for anomaly based network intrusion detection: A review. computers & security. 2011 Sep 1;30(6-7):353-75.

[4] Liu, F.T., Ting, K.M. and Zhou, Z.H., 2008, December. Isolation forest. In 2008 eighth ieee international conference on data mining (pp. 413-422). IEEE.