Report of Project 2

Segurança em Redes de Comunicações
Universidade de Aveiro
DETI

Adalberto Júnior, (105589)

Conteúdo

1 Introduction

The main objective of this project is to define and implement SIEM (Security Information and Event Management) rules that can effectively identify unusual network activities that may indicate potential threats. The project uses historical traffic flow data from a “corporate” network, knowing that one of the data sets provides typical network behavior and the other contains anomalous network behaviors. Based on the analysis of these data sets, the project’s output will be a set of SIEM rules, designed from the analysis, that can identify potential threats, either flagging them as an alarm or blocking actions based on the severity of the danger.

2 Methodology

I began the project by analyzing the “data9.parquet” dataset, which contains a full day’s worth of data on typical network behavior. This data, already certified as free of illicit behavior, forms the basis for our understanding of normal network operations. I then delved deeper into the “test9.parquet” dataset. This dataset, mirroring the structure of the previous dataset but potentially containing anomalous behavior, allows us to contrast normal and abnormal network activity and identify unique patterns that signal potential threats. And finally, I analyzed "servers9.parquet". This dataset contains a full day’s worth of external access to the corporation’s servers (on the 200.0.0.0/24 network) from a small group of clients on the same network, which may contain external users interacting with the corporation’s servers in an anomalous manner. The data analysis will utilize Python, leveraging the pandas library for data analysis.

2.1 Data Analysis

In the data analysis part of the project, I started by collecting simple information from the datasets using pandas from both the normal dataset and the test dataset and servers such as the ports and protocols used, evaluating that the ones used were: UDP on port 53, TCP on port 443 which represent DNS and HTTPS respectively. You can check the average usage of the protocols from the Figure 1 and the ports in the Figure 2.

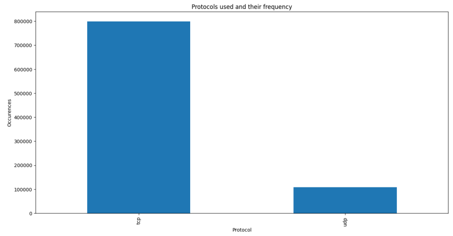


Figura 1: Protocols average use.

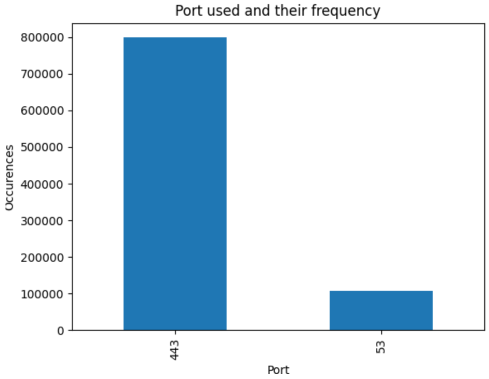


Figura 2: Port average use.

And analyzing the servers. Parquet, only TCP on Port 443 is used. You can check the average usage of the protocols and the ports in the Figure 3 and 4 respectively.

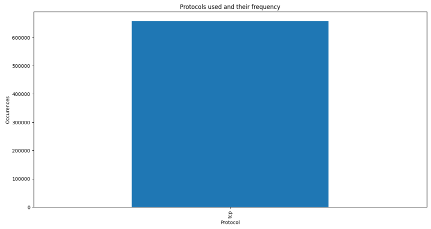


Figura 3: Protocols average use in servers.

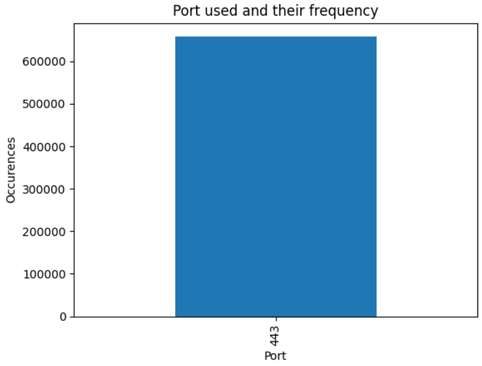
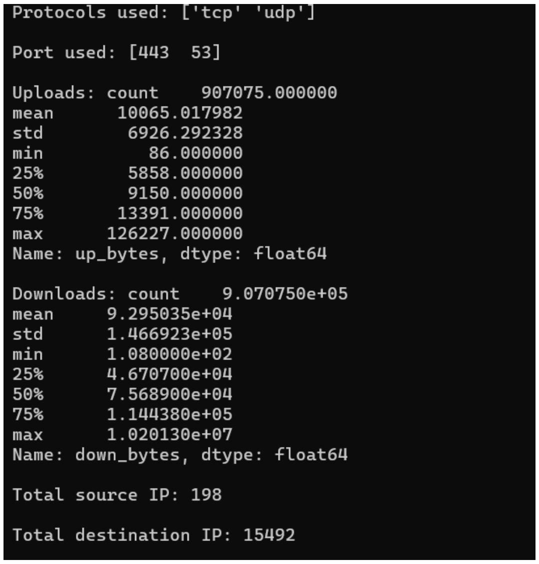


Figura 4: Port average use in servers.

I also analyzed the number of flows per source IP, bytes uploaded and bytes downloaded per flow and total, both for internal communications (private IP to private IP) and external communications (private IP to public IP). I also collected metrics on mean value, standard deviation and variance. Additionally, some country statistics were also collected, taking into account all the metrics mentioned above.



```
Protocols used: ['tcp' 'udp']

Port used: [443  53]

Uploads: count    907075.000000
mean      10065.017982
std        6926.292328
min          86.000000
25%        5858.000000
50%        9150.000000
75%       13391.000000
max      126227.000000
Name: up_bytes, dtype: float64

Downloads: count    9.070750e+05
mean     9.295035e+04
std      1.466923e+05
min      1.080000e+02
25%      4.670700e+04
50%      7.568900e+04
75%      1.144380e+05
max      1.020130e+07
Name: down_bytes, dtype: float64

Total source IP: 198

Total destination IP: 15492
```

Figura 5: Bytes uploaded and bytes downloaded per flow.

3 Non-Anomalous Behavior Analysis

Discuss the typical behavior of network devices based on the analyzed data. Taking this normal data set as a starting point for what is typical network behavior when there is no malicious activity, I gathered some information from it to later compare it with the test data set that has anomalous communication patterns.

3.1 Internal communications

By exploring the internal communications, we found several normal behaviors, such as the list of IP addresses that normally use the network and their communications within it. Regarding the protocols and ports used, we observed that there are no TCP connections on port 53. This port is typically used for DNS traffic. This finding suggests that DNS traffic on this network is using UDP instead of TCP, which is typical for DNS queries and responses. Also in this part, I analyzed the number of UDP and TCP flows for each source IP, where I discovered which IPs belonged to the servers on the network. As can be seen in Figure 6, the IPs with the most flows are the potential addresses of the servers, so we are left with the server IPs: 192.168.109.227, 192.168.109.224, 192.168.109.230 and 192.168.101.225. In the following figure we can see the top 4 private IPs with the most flows.

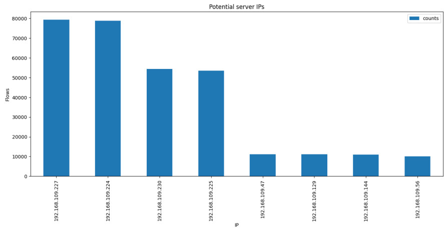


Figura 6: Potential Server IPs.

3.2 External Communications

In this section, I took the same approach as before, looking at what protocols, ports, IP addresses, and flow density existed. However, I was also able to collect which countries and organizations were being contacted from the internal network and how much of it was happening, so that I had a baseline of what the internal to external network communications were. I also got values like flows per country, flows per organization, etc.

For reference, I looked at the top 5 countries with the most streams, bytes sent, and bytes downloaded. The countries with the most streams, bytes sent, and bytes downloaded were the United States of America, Portugal, the Netherlands, Namibia, and Germany, respectively.

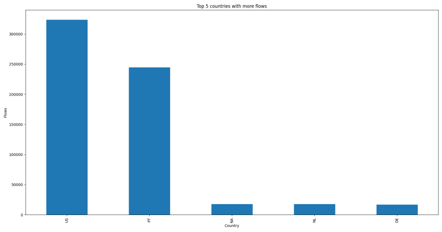


Figura 7: Top 5 countries with the most flows.

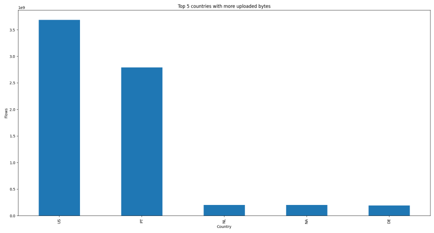


Figura 8: Top 5 countries with the most uploaded bytes.

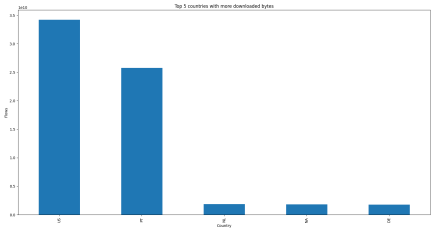


Figura 9: Top 5 countries with the most downloaded bytes.

4 Anomalous Behavior Detection

4.1 Botnet activity

A botnet is a collection of compromised devices that are under the control of a central command. These compromised devices, often referred to as “bots,” are typically infected with malicious software without their owners’ knowledge. Based on this, I took the IP addresses involved in new internal communications and considered them suspicious. I then analyzed whether their communications are normal or malicious by looking at factors such as unusually high data volume, presence of multiple streams, etc. The IPs that we detected as suspicious were 192.168.109.20 and 192.168.109.72. To determine whether these IP addresses are part of a botnet, further analysis is required. As an example, we can refer to Figure 10, which shows a new IP making connections to some of the servers. However, these connections have a small time window for a human to make these many requests. The other types of anomalies are easy to detect with a rule. While some anomalies can be easily detected with automated rules, identifying unusual connection patterns may require manual analysis. Isolation provides time for this analysis, allowing security teams to gather more information.

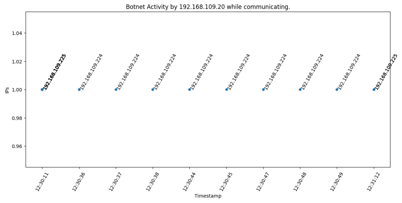


Figura 10: Botnet Behaviour.

4.2 CC attacks

In the section dedicated to the Command and Control (CC) assessment of the project, it is crucial to highlight that there were no relevant events that occurred during the monitoring phase. Through a meticulous analysis of the data, it was realized that the servers were not under attack. Therefore, there was no unusual change in the number of flows compared to the normal data, as can be seen in Figure 11. Typically, network traffic, or "flows", remains within a certain expected range, as is the case in my case.

In a CC attack, compromised systems frequently communicate with a command server to receive instructions or send data. Increased flows of specific IP addresses to servers may imply that these systems are compromised and part of a botnet used for the CC attack. This type of behavior is symptomatic of systems that have been commanded to perform tasks under the control of a remote attacker. Suspicious IPs should be isolated and thoroughly analyzed for any malicious content or behavior.

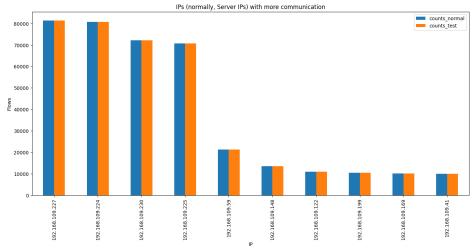


Figura 11: Increase of flows of the IPs with the most flows on the test dataset.

4.3 Data Exfiltration

In this section of our analysis, we focus on identifying potential data exfiltration and anomalous external communications. Data exfiltration can be a serious threat because it involves the unauthorized transfer of data from within the organization to an external location, potentially compromising sensitive information. The analysis first involves identifying new IP addresses that have started communicating externally. The code separates these IP addresses by comparing the list of source IPs in the current data (external flows test) with the source IPs in the baseline data (external flows). This process is also replicated for target IPs. This helps identify any new or unusual communication patterns with external IP addresses that were not present before and may be indicative of a security breach or unauthorized connection. Once the new externally communicating IP addresses are identified, the analysis delves deeper by investigating how much data these IPs are uploading and downloading. This is particularly important for identifying data exfiltration activities.

Specifically, the analysis calculates the total volume of data in megabytes uploaded and downloaded by these suspicious IPs. A sudden increase in uploaded data could be indicative of network data exfiltration, as seen in Figure 12

Additionally, the analysis involves checking the percentage increase in external communication by merging the baseline and test data sets and calculating the increase in the number of flows. IPs with an increase of more than 150% and with more than 50 flows are targeted for closer examination. In the Data Exfiltration subsection, the emphasis is on the volume of data sent to external sources.

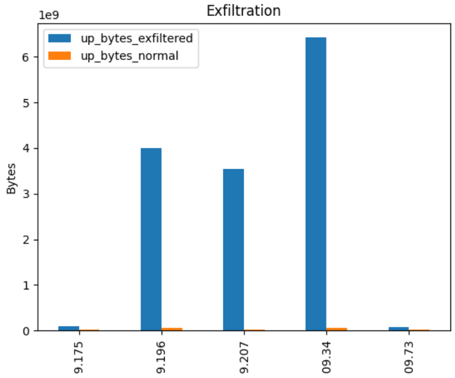


Figura 12: Average uploaded bytes.

4.4 Suspicious Country Communications

In this section of the report, I examined network communication data to analyze interactions with external countries. This involves understanding the distribution of network flows, bytes uploaded, and bytes downloaded for different countries and identifying any unusual or anomalous communications.

First, I aggregated the data by destination country code (dst cc) and calculated the total counts of network flows, as well as the sum of bytes sent (top bytes) and bytes downloaded (bottom bytes) for each country. I created 3 bar charts: one showing the top countries with the most communication flows. This chart shows which countries have the most frequent communication. Another bar chart is created to show the top countries where the most data is sent to, and finally a bar chart showing the top countries from which the most data is received. To detect any significant changes or anomalies, I compared the test dataset with the normal dataset. A country is considered to have unusual communications if it meets the following criteria: The number of flows is greater than 200 and there is at least a 50% increase in the number of flows or there is at least a 50% increase in the uploaded/downloaded bytes. Based on the observations in Figure 13, I have found that only the communications with the United States, despite showing a slight increase in update bytes, can still be considered "normal"and should undergo further analysis. However, it is recommended to block the remaining communications, since there is a significant amount of traffic to these countries in the anomalous dataset, compared to almost no communications to this country in the normal dataset.

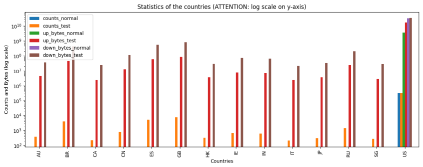


Figura 13: Countries Statistics.

5 SIEM Rules Definition

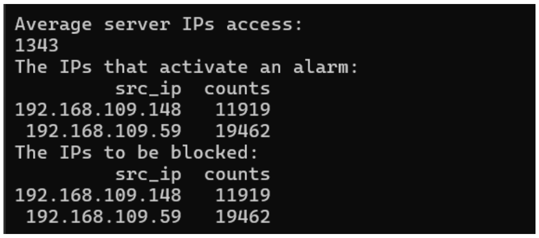
5.1 Server Access Increase

The following rule is intended to address the issue of server access, which will also detect distributed denial of service (DDoS) attacks.

The first step of the rule involves calculating the average number of hits to specific server IPs, based on the provided dataset. This average serves as a baseline or reference point to evaluate subsequent access patterns. To identify potential alarm trigger events, the rule considers any access count that exceeds three times the calculated average. This threshold indicates a significant increase in server access, which may require further attention. Additionally, the rule recommends a more stringent response when an IP address has an access count five times higher than the average. This level of activity is considered a more severe threat, indicating potentially malicious intent. In such cases, the rule advises blocking the IP address to prevent further unauthorized access and mitigate the risk of a potential DDoS attack.

5.1.1 Rule Testing and Device Identification

When testing the rule, we obtained the following results: The effectiveness of the rule is evident, as it successfully detected and alerted certain IPs, including those associated with malicious activity. The rule proved valuable in blocking IPs that were clearly abusing access to the server, ensuring that only high-risk entities were blocked, and providing alerts for IPs that required further analysis.



```
Average server IPs access:
1343
The IPs that activate an alarm:
         src_ip  counts
192.168.109.148   11919
 192.168.109.59   19462
The IPs to be blocked:
         src_ip  counts
192.168.109.148   11919
 192.168.109.59   19462
```

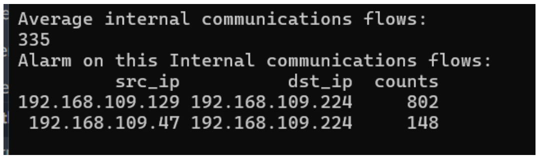
Figura 14: Server Access Increase Rule Result.

5.2 Internal Communications

The rule provided focuses on monitoring and responding to internal communications within a network, primarily targeting new internal communications that may require blocking. The rule follows a series of steps: first, it identifies internal communications between private IP addresses in the normal dataset and stores the information in the normal_internal dataframe. Next, it calculates the average count of internal communications by taking the average of the communication counts in the normal_internal dataframe. This average serves as a baseline for evaluating subsequent internal communications. The rule then scans the test dataset to identify new internal communications between private IP addresses. By filtering the test data and grouping it based on source and destination IP addresses, the rule calculates the size of each communication group, storing the information in the test_internal dataframe. To determine new internal communications in the test dataset, the rule compares the internal communications in the normal dataset (normal_internal) with those in the test dataset (test_internal). It selects and stores communications that exist only in the test dataset, indicating new internal communications, in the internal_diff dataframe. The rule outputs the internal communications that triggered an alarm. The information includes the source IP address, destination IP address, and communication count for each internal communication in the internal_diff dataframe.

5.2.1 Rule Testing and Device Identification

When testing the rule, the following results were obtained:



```
Average internal communications flows:
335
Alarm on this Internal communications flows:
         src_ip          dst_ip  counts
192.168.109.129 192.168.109.224     802
 192.168.109.47 192.168.109.224     148
```

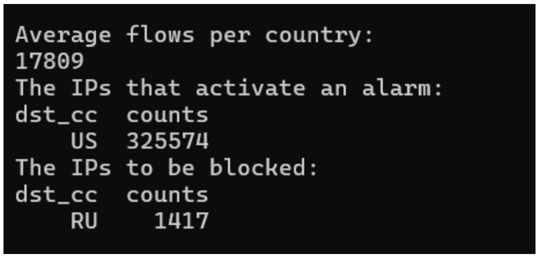
Figura 15: Internal Communication RuleResult.

5.3 Naughty Countries

The following rule analyzes the average number of flows per country by grouping the data based on the destination country code (dst_cc). It calculates the average flow count across countries, providing an understanding of the normal flow level for each country. It then checks the test dataset for countries with flow counts greater than 18 times the average or with an infinite number of flows. These countries are flagged as potential sources of concern or anomalies and an alarm is raised. The rule further examines the test dataset to identify countries that exist in the test dataset but not in the normal dataset. Among the newly identified countries, the rule selects those with flow counts greater than 500.

5.3.1 Rule Testing and Device Identification

When testing the rule, the following results were obtained:



```
Average flows per country:
17809
The IPs that activate an alarm:
dst_cc  counts
    US  325574
The IPs to be blocked:
dst_cc  counts
    RU    1417
```

Figura 16: Naughty Countries Rule Result.

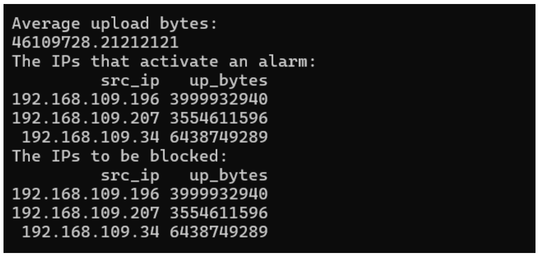
The rule has proven effective in raising alarms about the increase in flows observed in the US. This increase is likely normal given the country's role as a hub for numerous data centers and widely used platforms. However, the rule also identified and blocked flows from Russia, a country that had not previously been involved in any communications.

5.4 Data Exfiltration

This data exfiltration rule aims to analyze the average upload bytes in the normal dataset. It calculates the total upload bytes for each source IP address (src_ip), calculates the average upload bytes across all specified IPs, and determines the average upload bytes. The rule then examines the test dataset to identify flows that exceed three times the average upload bytes. It groups the data by source IP address, calculates the total upload bytes for each IP address, and selects the flows that exceed the threshold. Additionally, the rule checks for flows in the test dataset that exceed five times the average upload byte

5.4.1 Rule Testing and Device Identification

When testing the rule, the following results were obtained:



```
Average upload bytes:
46109728.21212121
The IPs that activate an alarm:
         src_ip   up_bytes
192.168.109.196 3999932940
192.168.109.207 3554611596
 192.168.109.34 6438749289
The IPs to be blocked:
         src_ip   up_bytes
192.168.109.196 3999932940
192.168.109.207 3554611596
 192.168.109.34 6438749289
```

Figura 17: Data Exfiltration Rule Result.

The rule proved effective in triggering an alarm for three IPs and blocking them due to discrepancies. Among them, one IP was uploading almost 4 GB of data, which is remarkable considering the average upload byte is around 50 MB.

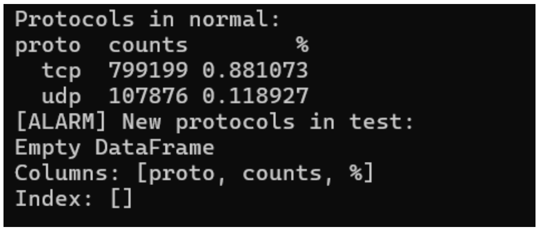
5.5 New Protocols and New Ports

To enhance monitoring capabilities, we implemented a rule specifically designed to identify the use of new protocols and new ports that have not been previously observed. The presence of new protocols may indicate new types of communication or modifications to the system. This rule starts by examining the protocols used in the normal dataset. It organizes the data based on protocol type (proto), calculates the count for each protocol, and determines the percentage representation of each protocol in the dataset. The rule then analyzes the test dataset to detect new protocols. Following a similar procedure as before, it groups the data by protocol type, calculates the count and percentage for each protocol, and merges this information with the protocol data from the normal dataset. The rule retains only those records where the protocol count is zero in the normal dataset but has a positive count in the test dataset.

For the port, the same methodology that was used in the rule for protocols is followed.

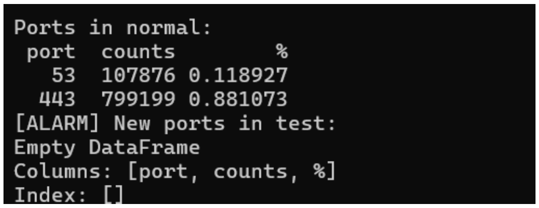
5.5.1 Rule Testing and Device Identification

When testing the rules, the following results were obtained:



```
Protocols in normal:
proto  counts        %
  tcp  799199 0.881073
  udp  107876 0.118927
[ALARM] New protocols in test:
Empty DataFrame
Columns: [proto, counts, %]
Index: []
```

Figura 18: New Protocols Rule Result.



```
Ports in normal:
 port  counts        %
   53  107876 0.118927
  443  799199 0.881073
[ALARM] New ports in test:
Empty DataFrame
Columns: [port, counts, %]
Index: []
```

Figura 19: New Ports Rule Result.

As we can see, no results were found because no protocols or new ports were used in the given datasets.

6 Conclusion

The project demonstrated the critical importance of robust SIEM rules in monitoring a network and identifying potential threats. By analyzing a typical day's network traffic, I gained valuable insights into normal network usage and used this to identify deviations that could indicate potential threats. When testing these SIEM rules against the "test9.parquet" dataset, I found that the rules were effective in highlighting anomalous network behavior, confirming the value of a data-driven approach. Despite analyzing the "servers9.parquet" dataset, I was unable to think of any rules that could be applied to this data, in part due to the lack of insight into this data and its correlation with other datasets.