Network Threat Analysis for detection of Emerging Threats

Submitted in partial fulfillment of the requirements of

BDA Mini Project (CSC702)

for

Fourth Year of Computer Engineering

By

Bhavesh Dhake 19102A0027
Harshada Chavan 19102A0025
Devish Gawas 19102A0024

Under the Guidance of

Prof. Pankaj Vanwari
Department of Computer Engineering

Vidyalankar Institute of Technology
Wadala(E), Mumbai-400437

University of Mumbai

2021-22

CERTIFICATE OF APPROVAL

This is to certify that the project entitled

“Network Threat Analysis for detection of Emerging Threats”

is a bonafide work of

Bhavesh Dhake 19102A0027
Harshada Chavan 19102A0025
Devish Gawas 19102A0024

submitted to the University of Mumbai in partial fulfillment of

BDA Mini Project (CSC702)

for

Fourth Year of Computer Engineering

Guide (Name) | Head of Department | Principal

Acknowledgements

This Project wouldn’t have been possible without the support, assistance, and guidance of a number of people whom we would like to express our gratitude to. First, we would like to convey our gratitude and regards to our mentor Prof. Pankaj Vanwari for guiding us with his constructive and valuable feedback and for his time and efforts. It was a great privilege to work and study under his guidance.

We would like to extend our heartfelt thanks to our Head of Department, Dr. Sachin Bojewar for overseeing this initiative which will in turn provide every Vidyalankar student a distinctive competitive edge over others.

We appreciate everyone who spared time from their busy schedules and participated in the survey. Lastly, we are extremely grateful to all those who have contributed and shared their useful insights throughout the entire process and helped us acquire the right direction during this research project

Abstract

Analysis and prediction of network traffic has several applications and has recently garnered a substantial number of research. Experiments of various types are carried out and analyzed in order to discover various flaws in existing computer network applications. The study and forecast of network traffic is a proactive strategy to ensuring safe, reliable, and quality network communication. Various strategies for evaluating network traffic are developed and tested, ranging from neural network-based techniques to data mining techniques. Similarly, many linear and non-linear models for network traffic prediction are offered. Several intriguing combinations of network analysis and prediction approaches are used to get efficient and effective outcomes.

Table of Contents

Sr No	Description	Page No
1	Introduction	**6**
2	Problem Definition	**8**
3	Literature Survey	**9**
4	Dataset	**10**
5	Implementation	**11**
6	Conclusion	**16**
7	Future Work	**19**
8	References	**20**

Introduction

Threat analysis is a cybersecurity strategy that aims to assess an organization's security protocols, processes and procedures to identify threats, vulnerabilities, and even gather knowledge of a potential attack before they happen.

By studying various threats staged against one's organization in detail, security teams can gain a better understanding of the level of sophistication of threats staged against the organization, the exploitation strategies, and identify areas in the organization's security posture that may be vulnerable to these threats.

Threat analysis is categorized as a reactive strategy in IT cybersecurity since the organization is assessing threats in real-time as they are staged against their security perimeter. Even though this strategy relies on attacks being staged against the organization, when done properly, this strategy can greatly reduce the scope of damages sustained in an unforeseen cyber-attack.

Types of Threats Found in a Threat Analysis

A successful threat analysis strategy can uncover various types of threats within an organization. Some of the categorization of threats is as follows:

- Accidental Threats
  Whether it's a misconfiguration of a security process, or an accident that leaves an organization exposed, one of today's leading causes of cyberattacks, is unfortunately an exploitation facilitated through human error. By performing a threat analysis, organizations can identify and remediate accidental errors before bad actors have the chance to exploit them.

- Intentional Threats
  The threat that every organization is worried about is the intentional threat. Intentional threats are those conducted by malicious entities to gain access and exploit sensitive data within an organization for profit.

- Internal Threats
  One of the most worrisome threats is not actually what you'd expect. Often, organizations worry about external threats and build sophisticated security architectures to keep bad actors out, however, the real concern resides inside the security perimeter of the organization. Unfortunately, when an employee

decides to act in a malicious way it can be catastrophic as they may have easier access to sensitive information.

Enterprise networks are monitored using network traffic analysis systems. Companies use network traffic analysis to capture and evaluate network traffic and inter-asset communications trends. This information is then utilized to detect and mitigate security threats. Network traffic analysis monitors network availability and activity and detects operational and security abnormalities.

Traditional network traffic analysis establishes a baseline for 'normal' network functioning. This data is then utilized to notify users of abnormalities, such as connections from a foreign country or an unknown device type. However, such an approach may generate too much noise, particularly in the post-pandemic period, when company IT is continuously developing to keep up with a more dynamic, remote work environment. As a result, cutting-edge network traffic analysis systems are designed to operate intelligently by taking into consideration historical trends and entity behavior.

Problem Definition

A network threat analysis is a process for studying various threats to your network in great detail - which is why the process can also be called a cyber threat analysis. When protecting your business, you need to monitor your network and cloud-based devices. But is that enough? Monitoring is just the first step. An organization needs to know what the data is telling you. What signals are they seeing about a potential threat? Where exactly are the weak spots in their network? And thus, we need Network Traffic/Threat Analytics that takes an organization from simply monitoring cyber security threats to active threat analytics, management, and prevention.

Literature Survey

From the inception of networking, network traffic measurement and analysis were seen as essential. Network traffic monitoring assists administrators in a variety of ways, including spotting bottlenecks and malicious network activity. With fast network growth and increasingly diversified applications, network bandwidth is expanding to thousands of gigabytes. When there is a low packet loss rate, it becomes difficult to monitor and analyze traffic in a timely and effective manner. Network traffic monitoring mostly entails collecting packets using network monitoring tools such as WIRESHARK, NTOP, Microsoft Message Analyzer, and PRTG and doing comprehensive packet analysis. Before analyzing network traffic, one needs understand the tool's structure, how to utilize it, and its limits in order to get the required findings.

There are Commercial Network Traffic/Threat Analysis Tools as well which allow you to monitor incoming and outgoing network data packets to uncover key insights such as network performance, security, and bandwidth utilization. These tools are available for every enterprise use case and budget. What these tools have in common is that they cover key metrics, have an intuitive GUI, support smart data and trend visualization, generating actionable insights along with integration with other apps and hosts.

Currently, the best tools in the market based on publicly available information are Cisco Network Analysis Module Traffic Analyzer, ExtraHop Network Traffic Analysis, Packetbeat Network Traffic Analysis and much more. All these tools are used to continuously observe, track, and analyze real-time traffic on a network. These tools include features to visualize traffic flows through network maps so users can swiftly address bottlenecking and other IT environment issues. These tools typically provide more robust traffic conflict solutions than you’d find in other categories.

This project is an attempt to only visualize and analyze an existing captured dataset using the capabilities of a certain visualization tool.

Dataset

LUFlow is a flow-based network intrusion detection data set which contains a robust ground truth through correlation of malicious behaviour. LUFlow contains telemetry containing emerging attack vectors through the composition of honeypots within Lancaster University's address space. The labelling mechanism is autonomous and is supported by a robust ground truth through correlation with third part Cyber Threat Intelligence (CTI) sources, enabling the constant capture, labelling and publishing of telemetry to this repository. Flows which were unable to be determined as malicious, but are not part of the normal telemetry profile are labelled as outliers. These are included to encourage further analysis to discover the true intent behind their actions. Normal traffic is also captured from production services.

This data set was created with the intention to promote research into detection mechanisms suitable for emerging threats. Critically, this data set is *continuously* updated due to the automatic labelling mechanism, therefore, it is able to reflect novel and emerging attack patterns.

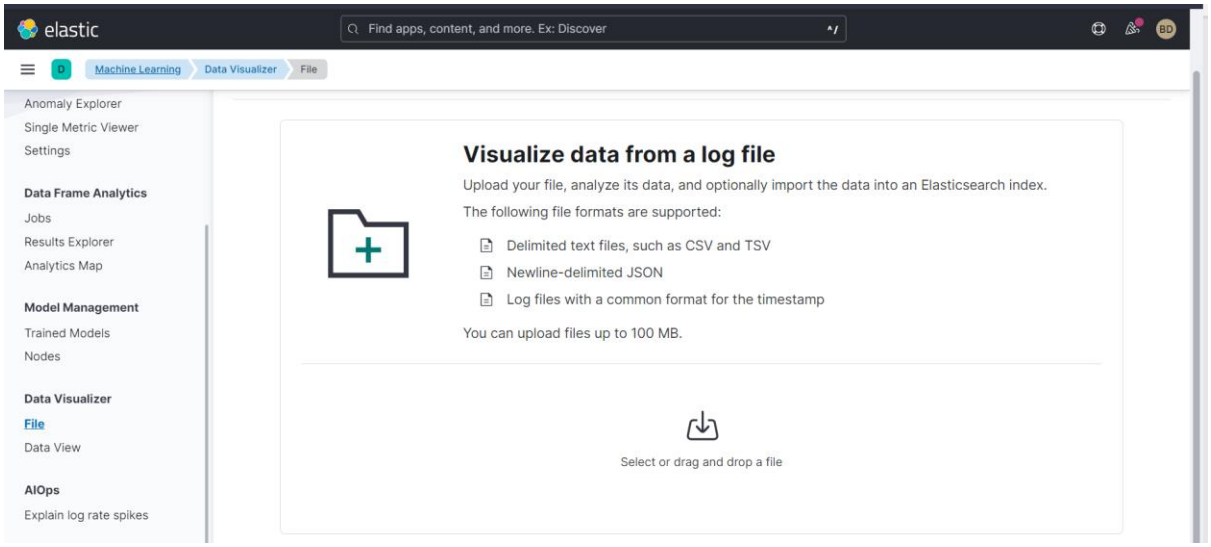
Name	Description
src_ip	The source IP address associated with the flow. This feature is anonymized to the corresponding Autonomous System
src_port	The source port number associated with the flow.
dest_ip	The destination IP address associated with the flow. The feature is also anonymized in the same manner as before.
dest_port	The destination port number associated with the flow
protocol	The protocol number is associated with the flow. For example TCP = 6, UDP = 17, ICMP = 1.
bytes_in	The number of bytes transmitted from source to destination
bytes_out	The number of bytes transmitted from destination to source.
numpktsin	The packet count from source to destination
numpktsout	The packet count from destination to source
entropy	The entropy in bits per byte of the data fields within the flow. This number ranges from 0 to 8.
total_entropy	The total entropy in bytes over all of the bytes in the data fields of the flow
mean_ipt	The mean of the inter-packet arrival times of the flow
time_start	The start time of the flow in seconds since the epoch.
time_end	The end time of the flow in seconds since the epoch
duration	The flow duration time, with microsecond precision

Implementation

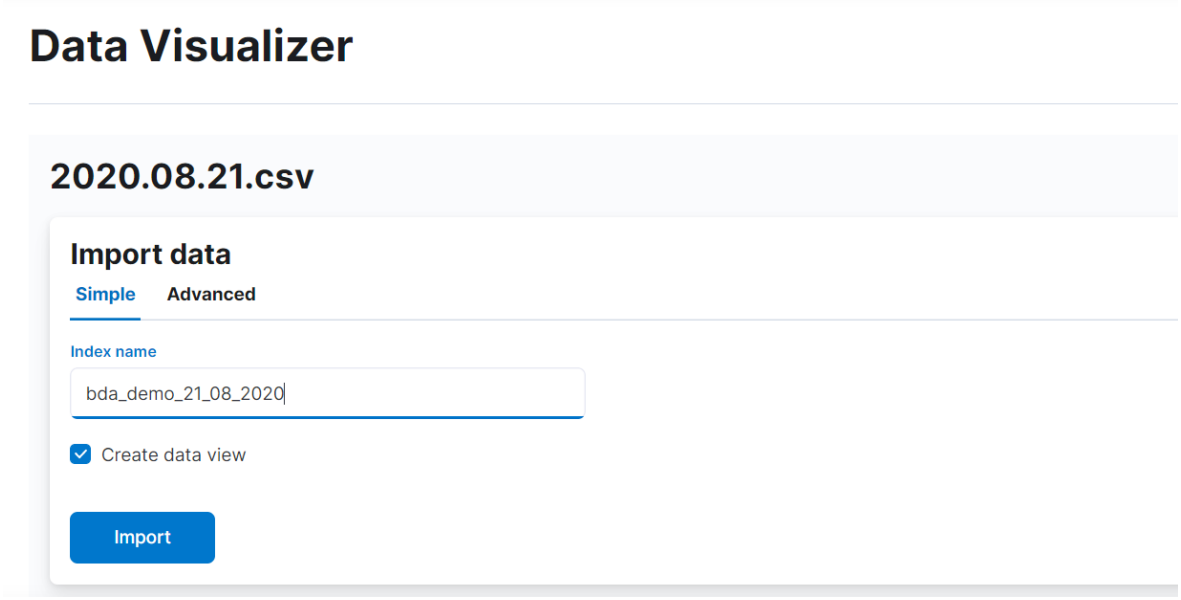
For the purpose of implementing this project, we are using Kibana, a free and open frontend application that sits on top of the Elastic Stack, providing search and data visualization capabilities for data indexed in Elasticsearch. Commonly known as the charting tool for the Elastic Stack. Kibana also acts as the user interface for monitoring, managing, and securing an Elastic Stack cluster — as well as the centralized hub for built-in solutions developed on the Elastic Stack. Consider this dataset is captured from a legitimate organization and we have to analyze the data.

Analyzing the dataset:

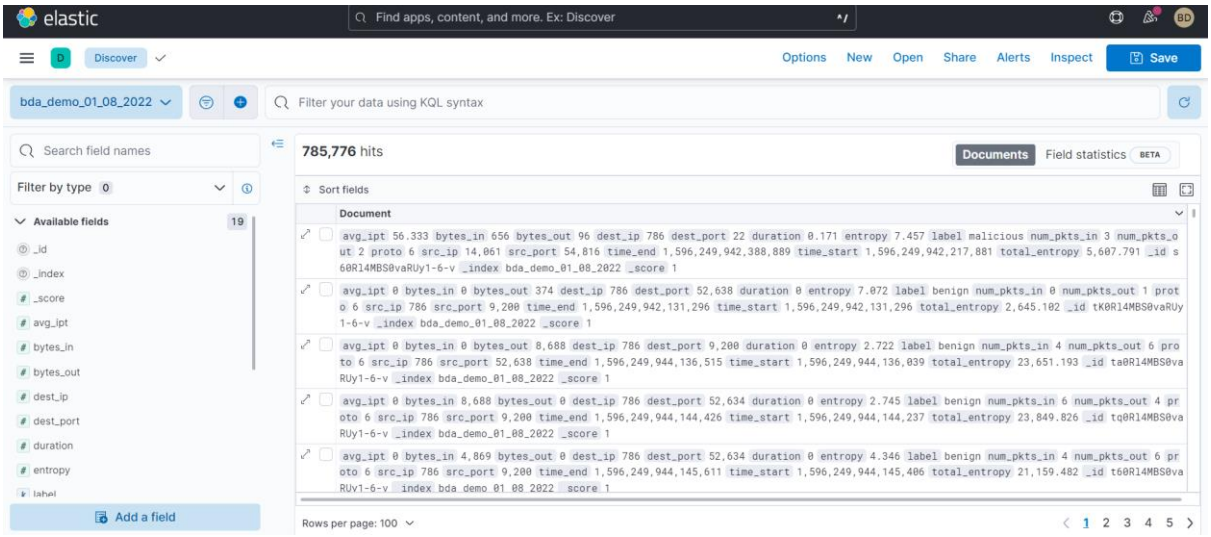
1. Importing the CSV file onto Kibana:
   First, we log in to the ElasticSearch Cloud portal, navigate to the tab where we have to upload the data for visualization. Here we could have done some transformations of data for certain attributes such as protocol, source_ip, dest_ip, etc., but unfortunately, we were unable to find this type of function in Kibana.



2. Upload the CSV file and give a proper index name

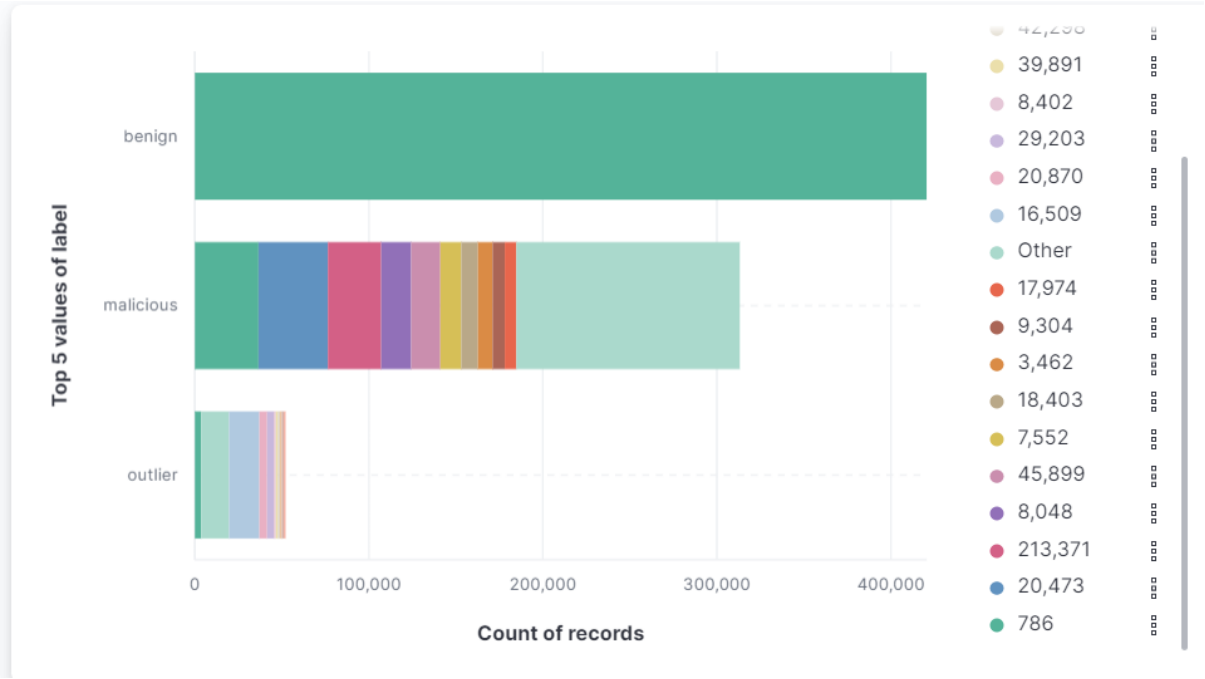


3. Go to Discover and then we can analyze the data uploaded



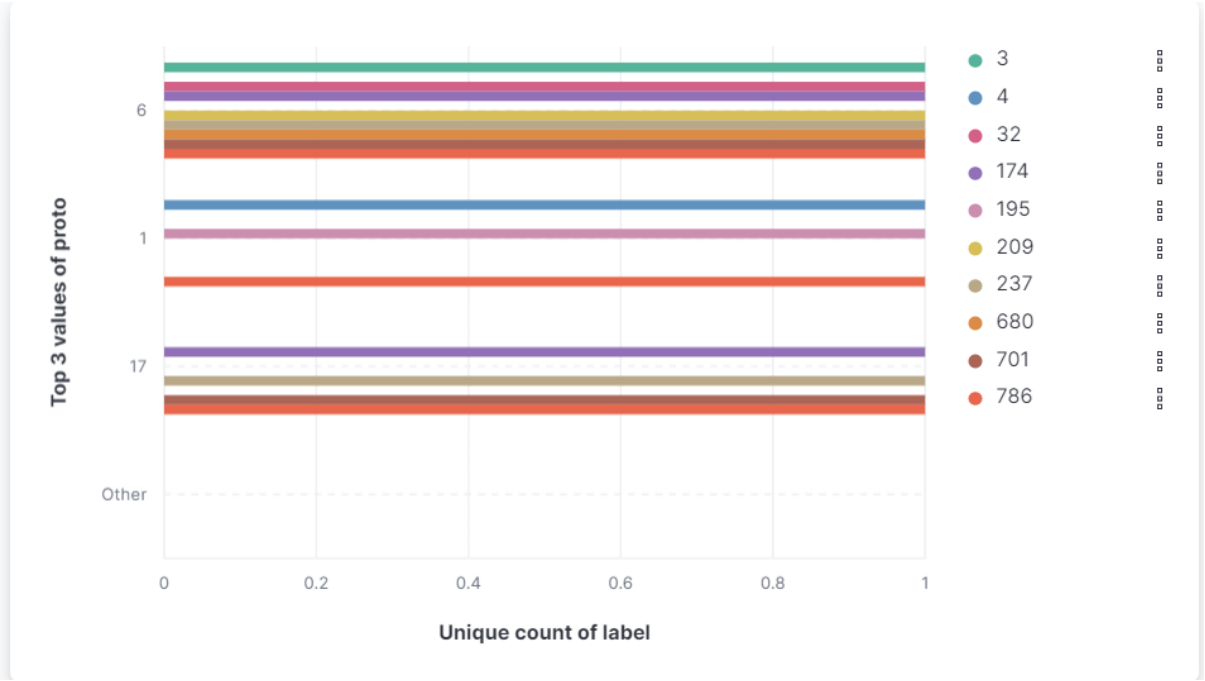
4. Here we can query the data as needed.
   For example, if we need to query the data which is labelled as “malicious”, we can do so by using the KQL syntax. After the query is returned, we can see what the source_ip is, what IP is being targeted, whether the malicious IP has pivoted towards a different network or not and so on.

5. Graph – I: Horizontal Bar Plot
   The following is the bar plot for the type of traffic which is labelled in the dataset along with the source_ip.



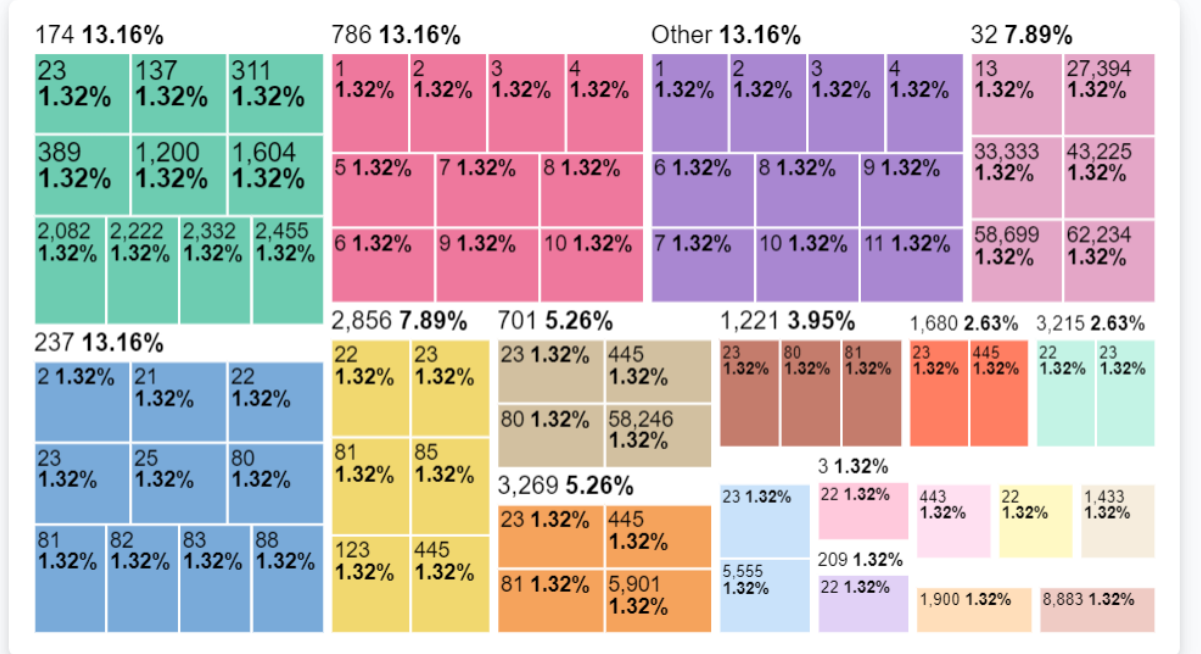
6. Graph – II: Horizontal Bar

Protocol wise distribution of the traffic



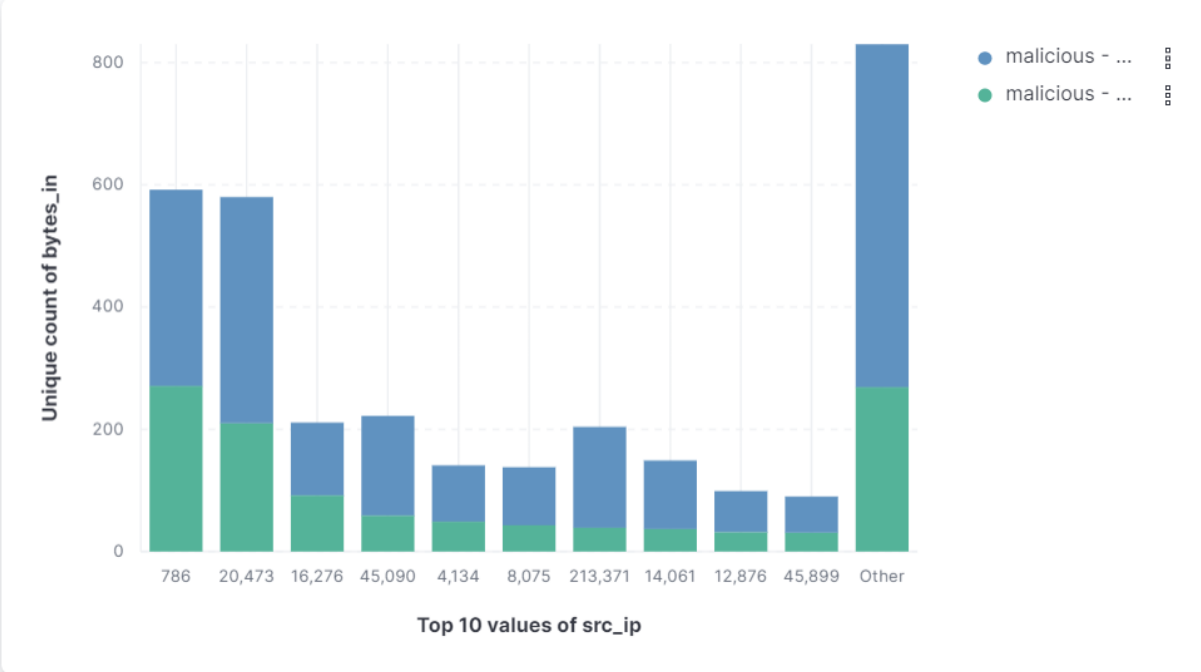
7. Graph – III: Treemap

Now we what ports are the malicious IPs targeting.



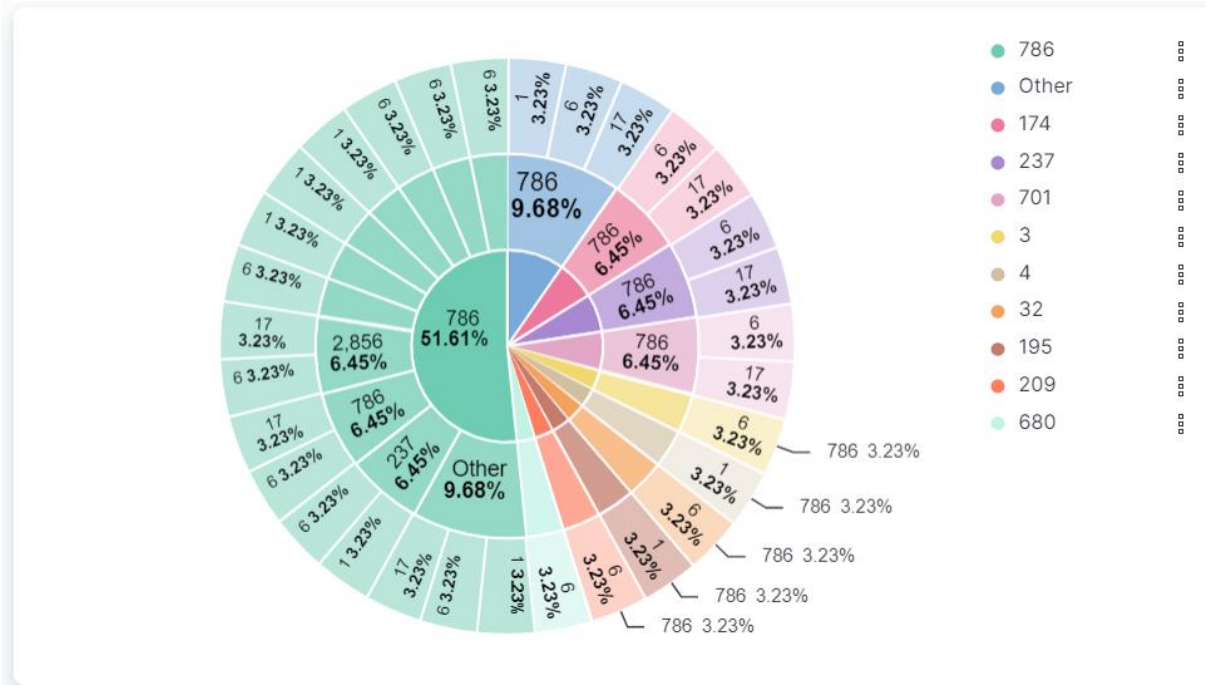
8. Graph – IV: Vertical Bar

Transfer of data in bytes from source to destination and vice-versa for the IPs which are flagged as malicious.



9. Graph – V: Pie Chart

Communication between the malicious src_ip and dest_ip along with the protocols.



Results and Conclusion

The following can be concluded after analyzing the above mentioned graphs.

1. Graph – I:

   According to the given dataset, and after analyzing the bar graph, we can say that most of the traffic which is labelled benign is originating from src_ip = 786, and thus we can say that this IP is a legitimate IP present inside a network of an organization.

   Whereas, if we see the traffic labelled as malicious, we can see that there are multiple source IPs who are not in the network which are targeting the organization.
   From this, we can further analyze there are certain IPs which stand out that are taking most of the space in the bar graph. So, we can conclude that there might be certain specific IPs which can be a part of APTs and are determined to target this organization.
   Also, we can see that src_ip = 786 is also flagged as malicious. This could be because there might be some malicious users who have already pivoted through the network or it might just be a false positive.

2. Graph – II:

   Now we analyze the protocols where most of the traffic is concentrated.
   As it is mentioned in the dataset,
   6 = TCP
   1 = ICMP
   17 = UDP

   So we can infer from here that most of the malicious traffic which is coming is from the TCP (Transmission Control Protocol) and then we can further see the ports which are targeted.

   There is less concentration in the UDP protocol, so we can infer that there is no DDoS happening currently in the UDP side which usually uses the UDP Flood attack.

   Also, no ICMP Flood attack is happening because there is less concentration in the ICMP protocol.

3. Graph – III:

   Here we are visualizing what ports are exactly being targeted by the malicious source IPs.
   Hence, we can infer from the available data that most of the malicious IPs are targeting port numbers:
   a. 22 - SSH
   b. 23 - Telnet
   c. 21 – FTP
   d. 80/443 – HTTP/S
   e. 445 – Microsoft AD

   All the above-mentioned ports are running critical services on internal servers and we can infer that these IPs are really targeting the important services which hold a lot of sensitive information/data regarding the employees, clients as well as the organization.

   It is often recommended to shut down services like Telnet and FTP, but as we can see that this particular organization has not followed the recommendation, these services are a prime target for the malicious IPs.

4. Graph – IV

   After determining what ports are being targeted by the IPs, we can check whether these are True Positives or False Positives by looking at the data sent and received between the source IP and the destination IP which is in bytes.

   Here, src_ip=786 is also flagged because of the reason mentioned in the Graph I. But there might also be a case of Malicious Insider, where the employee of the organization is attempting to steal data or exploit some services,

   Other than the above mentioned IP, we can observe that the bytes_in (bytes travelling from source to destination) are less, but bytes_out (data travelling from destination to source) are more. This is a typical case of gaining sensitive information after exploitation is done in a server. The malicious user typically will send a small query to retrieve the data, but this query is quite dangerous as it will return some huge amounts of data from the server. This can be observed for most of the Malicious source IPs and we can infer that the observed IPs are dangerous and this can be turned as a training data in the future for predicting future threats.

5. Graph – V:

   From the pie chart, we can see the communication happening between the malicious source IPs and the target destination IPs.
   We can observe that most of the source IPs which are flagged malicious are targeting the internal architecture of the organization and are using different protocols for doing so.

   Although, we removed the possibility of DDoS or DoS attacks for the UDP, there are a number of services which are open on UDP. Namely, the SNMP service is open widely throughout organizations. It is recommended to shut down this service, but from the data we can infer that this service is active for this organization and the malicious users have found that vulnerability and are trying to exploit the same.

   It can also be inferred that mostly the TCP services are being targeted by the malicious users which supports our previous analysis of Graph – II.

   We can also observe what exact IPs are being targeted here from the Pie chart. And as we can also determine who the source IPs are and from the previous analysis of the graphs, we can flag those IPs as a threat towards the organization.

Future Work

We can further implement prediction models using Machine Learning to predict a potential source of threat. Further using advanced analysis it will also be possible to flag an insider threat by analyzing the ports and IPs they are targeting along with total packets which are being transferred, how much data in bytes is being exchanged and the duration of a particular Source IP communicating with a Destination IP. Automation of all these processes is also achievable as these systems are already developed and maintained by commercial organizations, but they can be improved by analyzing the data more carefully and to support the analysis, Deep Learning and AI can also be implemented for more accurate prediction and prevention of future potential threats.

References

1. https://www.kaggle.com/datasets/mryanm/luflow-network-intrusion-detection-data-set

2. https://ayushmaskey.github.io/projects/network_traffic_analysis

3. https://www.ijitee.org/wp-content/uploads/papers/v8i6s/F60190486S19.pdf

4. https://www.researchgate.net/publication/326499385_Network_Traffic_Monitoring_System_Based_on_Big_Data_Technology

5. https://github.com/simondelarue/Network-traffic_analysis_with_ELK