NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

ANOMALY DETECTION ON FLOWS AND INCOMING PACKETS WITH GAUSSIAN MIXTURES

by

Tarun Menon

March 2023

Thesis Advisor:	Armon C. Barton
Second Reader:	Gurminder Singh

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** March 2023	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
ANOMALY DETECTION ON FLOWS AND INCOMING PACKETS WITH GAUSSIAN MIXTURES

5. FUNDING NUMBERS

6. AUTHOR(S) Tarun Menon

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)

Firewalls are key for maintaining a secure network, but it cannot be assumed that network traffic that manages to get through one is completely safe. Anomaly detection refers to methods that can be used to discover unique or uncommon occurrences within a particular dataset. Unsupervised machine learning techniques involve machine learning with unlabeled data, and can be utilized in order to perform anomaly detection by ingesting a given set of data and finding instances that diverge from the rest in meaningful ways that may not be obvious to the human eye. In this study we aim to analyze anomalies that are detected in incoming packet and flow network traffic data that successfully passed through a firewall and determine what significance there may be within such anomalies. Considering the vast amount of malicious traffic that exists and gets generated regularly, this study shows that Gaussian Mixtures can be used for discovery of anomalies within network traffic that passed through a firewall to discover potential undesirable or malicious traffic.

14. SUBJECT TERMS
unsupervised learning, anomaly detection, cyber, network traffic analysis, Gaussian Mixtures

15. NUMBER OF PAGES
55

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

ANOMALY DETECTION ON FLOWS AND INCOMING PACKETS WITH GAUSSIAN MIXTURES

Tarun Menon
Civilian, Non US Govt
BS, University of Wisconsin, Madison, 0

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
March 2023

Approved by: Armon C. Barton
Advisor

Gurminder Singh
Second Reader

Gurminder Singh
Chair, Department of Computer Science

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Firewalls are key for maintaining a secure network, but it cannot be assumed that network traffic that manages to get through one is completely safe. Anomaly detection refers to methods that can be used to discover unique or uncommon occurrences within a particular dataset. Unsupervised machine learning techniques involve machine learning with unlabeled data, and can be utilized in order to perform anomaly detection by ingesting a given set of data and finding instances that diverge from the rest in meaningful ways that may not be obvious to the human eye. In this study we aim to analyze anomalies that are detected in incoming packet and flow network traffic data that successfully passed through a firewall and determine what significance there may be within such anomalies. Considering the vast amount of malicious traffic that exists and gets generated regularly, this study shows that Gaussian Mixtures can be used for discovery of anomalies within network traffic that passed through a firewall to discover potential undesirable or malicious traffic.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

Disclaimer

This material is based upon activities supported by the National Science Foundation under Agreement No 1565443. Any opinions, findings, and conclusions or recommendations expressed are those of the author and do not necessarily reflect the views of the National Science Foundation.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

1.1 Introduction

Network security continues to become increasingly challenging and vital when it comes to maintaining functionality and safety of services running across the world. A key component in the pursuit of network defense is the utilization of firewalls. They are primarily known for protecting any given network from external threats that attempt to make their way inside, but they can even be made versatile enough to prevent potential threats from leaving and impacting other networks.

There have been various alternative approaches that have been developed for the purpose of protecting a network from any malicious traffic alongside firewalls, such as Intrusion Detection Systems. Many such approaches have been crafted using supervised machine learning techniques with the hopes that such techniques would pick up on certain features or patterns that might evade the scrutiny of a human. These strategies generally take formatted traffic data and train a model with it for use in determining whether or not any given traffic sample fits into a particular predefined category such as malicious or benign.

This thesis project aims to instead utilize unsupervised machine learning techniques on network traffic which will make discoveries without any human input or restrictions based on predefined categories. Much research in this field aims to craft Intrusion Detection Systems that broadly look at overall traffic traversing a given network, so this work aims to follow a more novel approach of focusing on a subset of traffic travelling through a network. Specifically, the unsupervised machine learning techniques that will be used are meant for performing anomaly detection on traffic that has already passed through a firewall in order to discover any potentially interesting anomalous or even malicious traffic that may be present.

While firewalls are capable of blocking out a great amount of troublesome traffic, they cannot be considered perfect, so we hope that this research will lead to a potential new technique for strengthening networks from unknown dangers.

In this paper, Gaussian Mixtures [1] is used to detect anomalies within incoming packets and traffic flows that succeeded in going through a firewall. The anomalies that are yielded from using Gaussian Mixtures are shown to capture many uncommon or low-occurrence feature values throughout both data sets. Overall the technique is proven to be very successful at anomaly detection, and the technique of focusing solely on incoming traffic shows promising results for use in further applications.

1.2 Thesis Organization

This thesis is organized into the following chapters:

- **Background** where relevant information about the data, techniques used, and adjacent research will be described.
- **Methodology** where the details of how various techniques were employed is described.
- **Results** where the results of anomaly detection will be presented.
- **Discussion and Future Work** where the results are tied together to illustrate their importance, and for listing potential future work.
- **Conclusion** where the discoveries are summarized.

CHAPTER 2:
Background

2.1 Machine Learning

Generally speaking, machine learning (ML) is the practice of providing a computer algorithm with some input data in order for it to be analyzed in a manner that may be too in depth for the capabilities of a human, and then subsequently performing useful tasks using that analysis. There are two main categories when it comes to machine learning styles, one of which is supervised learning: referring specifically to scenarios where the data fed into a machine learning algorithm has been labeled into categories, therefore allowing for tasks such as classification of new data to be performed [1]. Machine learning has proven to be invaluable for performing several key tasks in the computing realm that were otherwise very challenging to engineer solutions for, and for the analysis capability it provides.

2.1.1 Unsupervised Machine Learning

The other main category of ML, unsupervised learning, simply involves using data that has not been labeled into categories. These algorithms can be used in several ways to glean new insights on data in ways such as clustering portions of the data together due to similarity in order to discover potential patterns that were hidden to the naked eye [1]. This is the style of ML that is employed for this research since we are attempting to use the pattern seeking capability for finding traffic that may not actually fit in with the rest.

2.2 Firewalls

In order to maintain a secure network in any given institution, it is standard for a firewall (or potentially multiple firewalls) to be maintained in order to filter traffic that passes through from external locations. The way that firewalls traditionally work is via being configured with various rules that can dictate what traffic to allow or block based on different values such as the port, protocol, or source and destination of the traffic itself. There are several firewalls that also have functionality tied in with an antivirus style system where additional

protection may be accomplished by blocking specific packets with data payloads using signature detection to identify known malicious logic that is present [2].

While it may be the case that firewalls are capable of blocking the majority of threats that attempt to compromise a network or the systems on it, there are possibilities for some novel or unusual attacks to make it through based on how a firewall is configured or simply how outdated it may be. There have been several new technologies created and researched in an attempt to make up for common shortcomings of firewalls, one of which is known as an Intrusion Detection System.

2.3 Intrusion Detection System

One popular type of software that can be used to augment the security of systems on a network beyond a firewall's capabilities is an Intrusion Detection System (IDS). In some ways IDS are similar in design and functionality to firewalls, as they both can be configured with rules and signatures to easily identify traffic that is likely dangerous, however, IDS are fundamentally different from firewalls as they are passive listeners that cannot prevent traffic flow like firewalls can. They are still valuable for performing some automated analysis for cybersecurity specialists to view while monitoring a network, and can even have anomaly detection functionality for identifying less obvious threats [3].

There are also a couple of different classifications for IDS, namely Host-Based IDS which only function on and are meant to protect individual systems, and Network-Based IDS which monitor networks as a whole [3]. Much research in this field attempts to produce new Network-Based IDS, so when the term IDS is mentioned throughout this paper it can be assumed to specifically refer to Network-Based IDS.

2.4 Network Traffic Data Origin

The network traffic used in this thesis project has been captured from the Naval Postgraduate School (NPS) Education and Research Network (ERN) contiguously on September 28, 2022, for approximately 18 hours. There were capture tap points at *p1* and *p2* on the network that captured both on the outside and the inside of the firewall, respectively, as shown in Figure 2.1.

This capture data was originally stored in the form of .pcap files where the snap length was set to capture the first 70 bytes of every packet as the header information alone was meant to be used as input for any ML models [4].

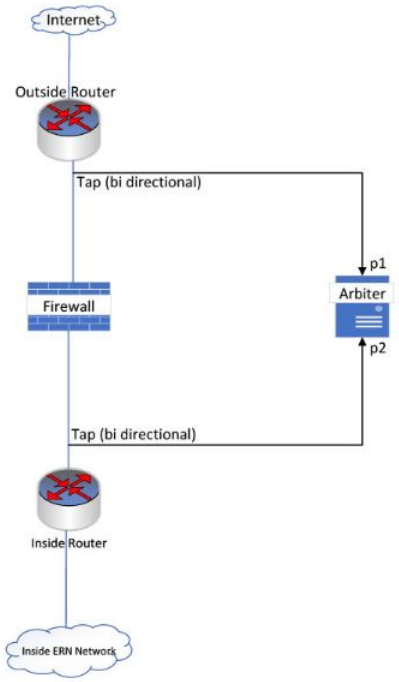

Figure 2.1. Graphic depicting the network traffic capture points on the Naval Postgraduate School's Education Research Network. Source: [4].

The data that was ingested in order to craft the packet header data used for machine learning was from the .pcap files, and was done with the use of *tshark*, then stored in the form of .parquet files. The key elements for the packet headers that *tshark* extracted are listed in Table 2.1. The ERN traffic flow data was then subsequently derived from processing the parsed packet header data, correlating separate packets to an individual flow by comparing the socket pairs, and also stored in the .parquet file format [5].

Table 2.1. Header Fields. Adapted from [5, Table 3].

Header Fields
Timestamp
Source IP
Destination IP
IP Protocol
IP length
TCP source port
TCP destination port
TCP sequence number
TCP acknowledgement number
UDP source port
UDP destination port

2.5 Anomaly Detection

Anomaly detection refers to the process of discovering records that differ significantly from the average record in any collection of data, which are known as anomalies [1]. This research attempts to perform anomaly detection specifically on incoming network traffic that has passed through a firewall including individual packets and aggregated streams. By using anomaly detection techniques on collected traffic that was deemed safe by a firewall, it would be possible to identify any traffic that may be undesirable or otherwise strange that somehow was not blocked by the firewall, which is a goal for this research.

2.5.1 Previous Work

A lot of the research in automated network traffic analysis has been done with the use of various forms of ML, as it fits very well with the problem sets relating to network traffic data. Some early anomaly detection research utilized more traditional ML techniques, such as n-gram models, state machine models, or even older neural networks trained with n-grams.

There are also instances where unique statistical methods were developed such as the one in PHAD where probability calculations were done by using the number of occurrences and distinct values that were seen in particular packet fields [6], or the signal processing approach performed by M. Thottan and C. Ji [7].

Some other rather common traditional approaches to crafting IDS involve supervised ML methods such as Support Vector Machines, Decision Trees, and Random Forests. In order for any of these methods to properly work, they need to be fed labeled network traffic data which is generally done manually by a person. The labels can be used to classify traffic flows as benign or as various other specific attacks such as denial of service. Such ML models tend to perform well on the types of undesirable traffic that were labeled and traffic from the same network from which the original training data was procured [8]. Applying these models toward other networks may not yield great results due to overfitting, and when it comes to other attack types, such models will most likely be entirely unable to discover them without more labeled data input and training.

There has also been research completed on varying unsupervised ML approaches for anomaly detection, such as performing clustering with the well-known k-means algorithm in order to have the ability to detect anomalies that fall outside the range of any given cluster [9]. There also have been some recent approaches involving deep learning using techniques such as Convolutional Neural Networks [10] or Graph Neural Networks [8] to naturally discover patterns for anomaly detection.

One thing that is common across all of these works, regardless of the ML or statistical approaches they have taken, is that they are handling all of the traffic captured from the network. This means that these projects are performing stochastic (statistical properties) or statistical (flow-level/packet-level properties) analysis on all traffic that is going through a particular network [11]. This work intends to focus on flows and specifically incoming network traffic, that of which has passed through a firewall.

2.5.2 Gaussian Mixtures

Gaussian Mixtures is a type of unsupervised machine learning that assumes any given record in a dataset was created by a mixture of many Gaussian distributions that do not have known parameters, where every record generated from a Gaussian distribution makes a cluster. The way that weights are generated is through the Expectation Maximization algorithm where initial parameters are randomly set and then the expectation and maximization steps are repeated. The expectation step assigns a record to a cluster and the maximization step updates each cluster afterwards to reach convergence. This is very similar to K-Means, but differs since it determines size and shape beyond the cluster centers. In order to perform anomaly detection with Gaussian Mixtures, cluster count and a threshold for density must be set, and all records that fall within the threshold in a region that is low enough in density are considered anomalies [1].

CHAPTER 3:
Methodology

3.1 Methodology

Initially, some processing of the captured network traffic packet and flow data will be needed in order to clean out unwanted rows due to issues such as missing values. Afterwards, geolocation information such as the source country and latitude and longitude is appended to every record. This geolocation information will be procured via the utilization of Maxmind's GeoLite databases, which can map IP addresses to such features [12]. Afterwards, Gaussian Mixtures, which is a unsupervised machine learning technique that is capable of anomaly detection, will be applied to the processed data in order to discover anomalies for both data sets, and said anomalies will be analyzed.

3.2 Environment

The environment on which the data processing and ML training took place was the NPS Hamming High Performance Computing (HPC) system. Using this HPC environment, it was possible to harness the processing power of Graphics Processing Units for more rapid performance with running any jobs involving ML. The programming language Python was used for the research with several supporting libraries such as numpy and pandas being utilized for data handling and analysis, with the Scikit-Learn library specifically being used for the Gaussian Mixtures model to perform the anomaly detection.

3.3 Data Processing

The data used for this research was the packet header data of incoming network traffic that had passed through the ERN's firewall, as well as the metadata of all of the traffic flows that passed through the firewall. The way that the packet header data was formatted originally within the .parquet files was with short captures that lasted fractions of a minute each. In order to have more meaningful analysis performed, these files were first aggregated into larger .parquet files that each accounted for a full hour of data capture over the near 18-hour

capture period. The packet header files also originally contained a column that labeled whether or not a packet was incoming or outgoing, so the initial step involved filtering for all of the desired incoming traffic and subsequently dropping the label column. Table 3.1 shows all of the fields of the packet data that was used in the ML after removing unwanted columns and including the desired geolocation data.

Table 3.1. Packet Header Fields.

Packet Header Data
Timestamp
Source IP
Destination IP
IP Protocol
IP Length
Source Port
Destination Port
TCP Sequence Number
TCP Acknowledgement Number
U.S.
EUROPE
COMMONWEALTH
ASIA
MIDDLE EAST
SOUTH AMERICA
OTHER
Latitude
Longitude

Since the data was filtered to only include incoming traffic, the destination IP address in this dataset always referred to a device within the NPS ERN, and the source IP address referred to an external host (the same is true for ports). The latitude and longitude of the source IP address were included as features alongside the source country. As can be seen at the end of Table 3.1, country geolocation information was appended to the data in the form of buckets, and they are one-hot encoded. These buckets encapsulate various countries across the world that traffic originated from to the NPS ERN that were determined when doing initial analysis on the processed packet header data. Table 3.2 depicts all of the countries and what bucket they correspond to.

Table 3.2. Mapping of one-hot encoded buckets for geographic location to countries that are represented within each bucket.

Country Buckets	
Bucket Name	**Countries**
U.S.	United States of America
EUROPE	Ireland, Netherlands, France, Germany, Switzerland, Norway, Sweden, Austria, Finland, Denmark, Italy, Poland, Romania, Bulgaria
COMMONWEALTH	United Kingdom, Australia, Canada
ASIA	Japan, India, Singapore, South Korea, Vietnam, Hong Kong, Malaysia, Taiwan
MIDDLE EAST	United Arab Emirates, Turkey, Qatar
SOUTH AMERICA	Brazil, Chile
OTHER	N/A

There are 6 buckets that encapsulate varying geopolitical areas in the world, with an additional OTHER bucket that accounts for all unspecified countries as well as any packets that failed to have the source IP mapped to a country entirely.

Table 3.3. Flow Metadata aggregated from analyzing packet data.

Flow Metadata
Source IP
Destination IP
IP Protocol
Source Port
Destination Port
Incoming Packets
Incoming Bytes
Outgoing Packets
Outgoing Bytes
Duration
U.S.
EUROPE
COMMONWEALTH
ASIA
MIDDLE EAST
SOUTH AMERICA
OTHER
Latitude
Longitude

Similar to the packet header data, the flow data had the same geolocation data appended to each record, namely latitude, longitude, and the source country buckets as seen in Table 3.2. There are several other similar fields, but with the removal of TCP fields and inclusion of Packet and Byte amounts. However, similar filtering for incoming traffic was not applied for this data as it would not be possible or make sense in the context of analyzing flows, therefore all flows present were analyzed.

In analysis performed by F. Iglesias and T. Zseby [13], ideal features for anomaly detection of network traffic was performed. While there were several features that involved being able to view actual payload of individual packets, effectively all of the features deemed to be highly relevant are within headers. This reinforces the position that using solely header data for performing anomaly detection is worthwhile, alongside the benefit of having to deal with less overhead.

3.3.1 Maxmind

Maxmind has multiple databases available that can be used to locate users on a network for varying purposes such as analytics or fraud prevention and they are regularly updated [12]. The Latitude, Longitude, and source country fields were procured using the Maxmind GeoLite2 Database as they were hypothesized to be valuable features for identifying abnormal traffic. For example, it could simply be the case that network traffic originating from an unexpected location would make for a valid anomaly.

3.4 Applying Gaussian Mixtures

After the data was prepared, Scikit-Learn's Gaussian Mixture class was used as the model for anomaly detection. After setting all of the parameters, the Gaussian Mixture model was fitted on an input of 1 hour of packet header data, and all anomalies that were collected got written to a separate file for each corresponding hour. For the flows, all of the metadata was aggregated and anomaly detection was performed with output to a single file. The number of clusters that was selected was 3 for both data types after testing for an optimal amount with the use of the Bayesian Gaussian Mixture Model [1]. A value of 0.1% was arbitrarily selected to represent packet header anomalies, and 0.5% for flow anomalies. There is not much in the way of definitive research that quantifies the potential amount of undesirable or anomalous traffic that successfully passes through firewalls in general, and this is likely because it would be nearly impossible to properly do so when also accounting for different types of firewalls and networks. Further optimization of these values is left up to future work.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:
Results

4.1 Results

The results will be split into the two separate data sets, first the incoming packet header data, then the flow data. A third section is included for analyzing the latitude and longitude coordinates for both data sets.

4.1.1 Packet Header Data

The total amount of packet headers within the data set was just over 1.6 billion, and the amount of anomalous packets selected was 1,616,977, which corresponds to the chosen percentage of 0.1% for anomalies for the packet data set. The IP addresses alone did not make for particularly valuable analysis, but the geolocation features that were crafted from them did. A few fields for the packet headers were not deemed very pertinent for analysis: namely the timestamp, IP addresses, and TCP sequence or acknowledgement numbers. The timestamp field did not provide valuable insights since the packet header data was split into individual hours, additionally, the data set just accounts for 18 hours of data capture, and patterns based on timestamps would likely be more present when using data spanning multiple days. Since the TCP sequence number and acknowledgment are initially the result of a random selection, they simply could not provide much use as a whole for analysis. As for IP addresses: alone they do not provide any obvious insights, but the geolocation data that was procured from them certainly does, and will be analyzed in their stead.

IP Protocol

All protocols that were present in the data are listed below with their protocol number and name:

- 6: TCP, Transmission Control Protocol
- 17: UDP, User Datagram Protocol
- 1: ICMP, Internet Control Message Protocol

- 50: ESP, Encapsulating Security Payload

As shown in the left pie in Figure 4.1, the vast majority of traffic that was captured used the TCP protocol. For the rest of the traffic, the ESP protocol makes up the majority, as shown within the enlarged portion of Figure. UDP surprisingly makes up far less than ESP, while ICMP barely has a presence in comparison.

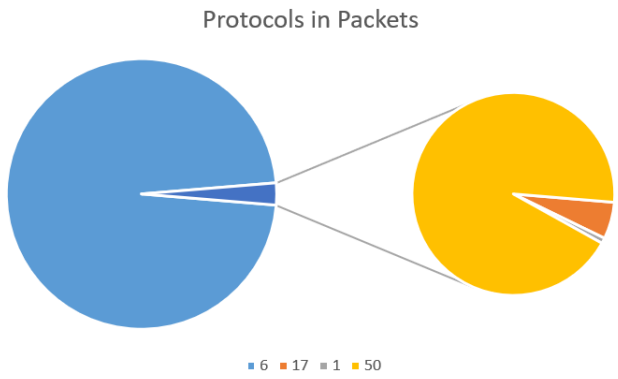

Figure 4.1. Pie chart displaying the distribution of IP protocols within the packet header data set.

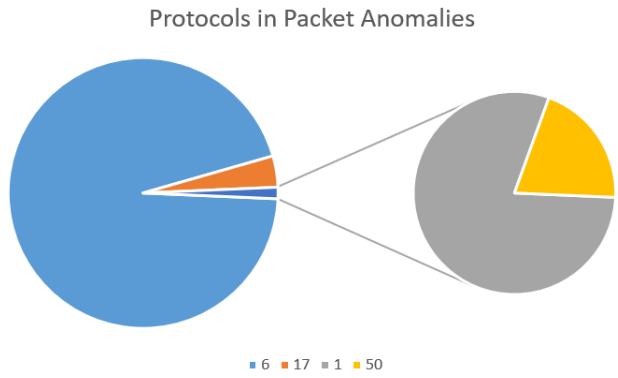

Figure 4.2. Pie chart displaying the distribution of IP protocols within the packet header anomalies.

TCP is still the majority for anomalies, but a far larger portion is the UDP traffic in comparison. The enlarged portion of Figure 4.2 shows that much more ICMP traffic was accounted for, with ESP making up the least of the data instead. ESP is meant to be used for supporting the functionality of IPsec, which is generally used for the operation of Virtual Private Networks (VPN) [14]. This could explain why it made up such a significant portion of the total traffic, because tunneled VPN users would have any and all traffic use the protocol and be captured this way when the VPN gateway is behind the firewall.

Packet Length

The size of packets is quite uniform with the vast majority matching with the Ethernet MTU size of 1500 bytes. The anomalous packets were able to capture a good amount of the less common smaller packets as seen in the far smaller 25th percentile in Figure 4.3, as well as a slightly smaller value for the 50th percentile.

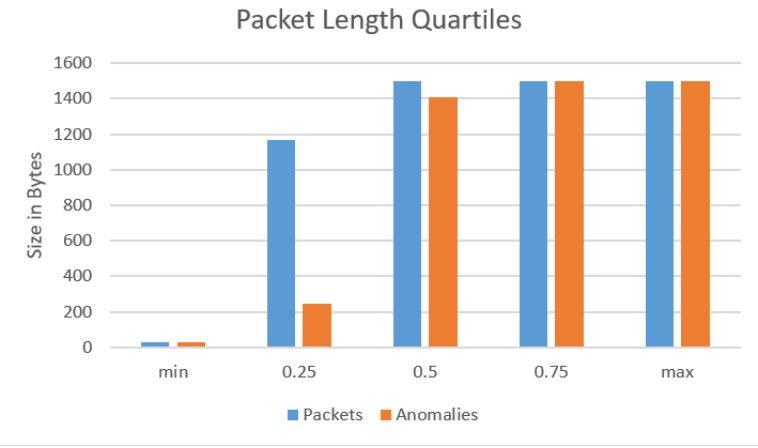

Figure 4.3. Bar graph comparing packet length between overall packets and anomalous packets.

Packets with smaller length could be caused by varying reasons such as differing operating systems or protocol used. The anomalies consisting of smaller sized packets accounts for such unique instances.

Port Numbers

An overwhelming amount of the source ports from external hosts indicate servers hosting HTTP or HTTPS websites that users on the NPS network reached out to, close to 80% for both overall and anomalous packets. Many ports are used overall in the data, but some other well-known ports seen in the anomalous traffic are as follows:

Source Port (External Servers):

- 20/21: FTP data and control
- 22: SSH
- 25: SMTP
- 43: WHOIS
- 993: IMAP
- 995: POP3

Destination Port (NPS Servers):

- 22: SSH
- 25: SMTP
- 53: DNS
- 500: IKE

Source Country

As expected, nearly all incoming packets going to NPS originated from the U.S., roughly 99%. In Figure 4.4, the rest of the country buckets make up too small a percentage to easily discern the ratios, but Europe was larger than Asia and the Commonwealth which were comparable. The Middle East and South America were mere fractions of a percentage and compare to each other in the enlarged section on the right of Figure 4.4. The anomalies on the other hand have a significant difference in ratio for countries. In Figure 4.5, both the U.S. and Asia are nearly equal around 40%, with Europe and the Commonwealth having far more presence as well. The Middle East and South America remain effectively the same, and while the anomaly detection collected all the packets from those locations, the amount of packets was still too small to be a significant portion of the anomalous packets.

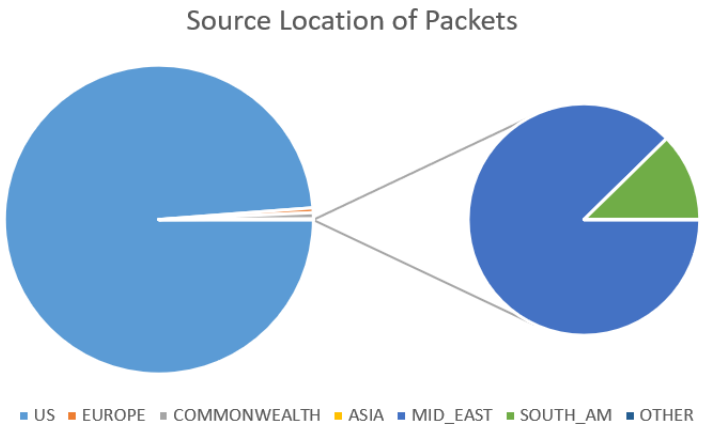

Figure 4.4. Pie chart displaying the distribution ratio of source country buckets across the packet header data set.

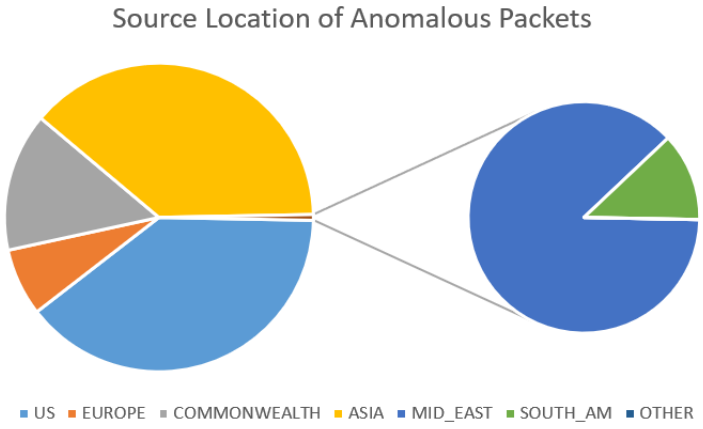

Figure 4.5. Pie chart displaying the distribution ratio of source country buckets across the packet header anomalies.

4.1.2 Flow Metadata

The total amount of flow metadata consisted of 23,138,472 records, and the amount of anomalous flows selected was 115,693, which corresponds to the chosen percentage of 0.5% for anomalies for the flow data set. Just like the packet data, IP addresses alone did not make for particularly valuable analysis, but the geolocation features that were crafted from them did. Similar to the packet header analysis, the IP address fields will not be investigated, but all other fields from Table 3.3 will be.

IP Protocol

The overall protocols in the flow metadata essentially had a ratio of 66.6% TCP to 33.3% UDP. A graph for comparison was not created for this because while there were ICMP and EDS flows present, the percentages of 0.02% and 0.000017% for each respectively was too small to be remotely seen.

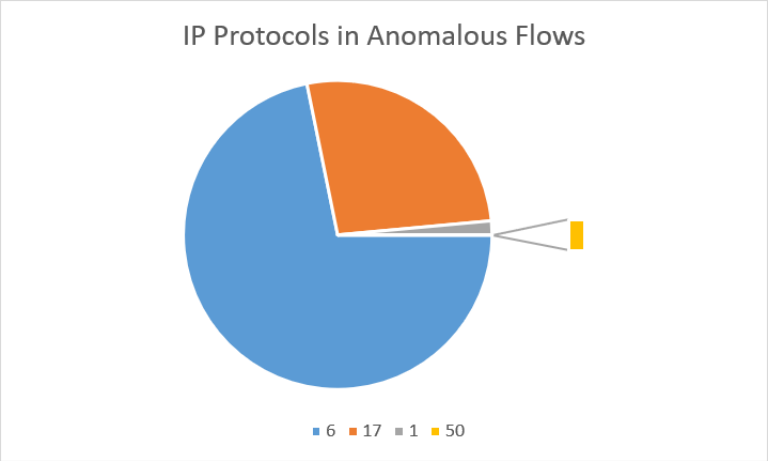

Figure 4.6. Pie chart displaying the distribution ratio of source country buckets across the flow metadata data set.

With anomalous flows, the ratio of TCP interestingly increased slightly, with both ICMP and EDS becoming a sizable enough portion of the total. EDS flows only consisted of 0.0035% of the anomalous flows, but was a significantly larger proportion alongside ICMP in Figure 4.6 compared to the overall total.

Port Numbers

For flows, when looking at destination ports (NPS Servers), a majority of roughly 80% seem to be for handling DNS traffic, with roughly 18% handling HTTPS website traffic. Approximately 415 well-defined ports were present in the overall source ports (External Servers), and not much difference in terms of the NPS server ports from the packets. Some other interesting ports for external servers were seen in anomalies:

Source Port (External Hosts):

- 7: Echo protocol
- 123: Network Time Protocol
- 388: Local Data Manager
- 497: Retrospect

The fact that there is traffic coming from the Echo protocol might be considered suspect due to the fact that it could be used for activities such as network mapping. Retrospect also seems to be a rather uncommon protocol for backups, so keeping an eye on such traffic may be worthwhile for security purposes.

Packet and Byte Count

Overall, the amount of incoming packets is rather small, but there are some exceptional cases that the anomaly detection worked fairly well for based on the disparity in the 75th percentile in Figure 4.7. The maximum amount was excluded from the bar graph as it is an extreme outlier in comparison, precisely 63789370 packets.

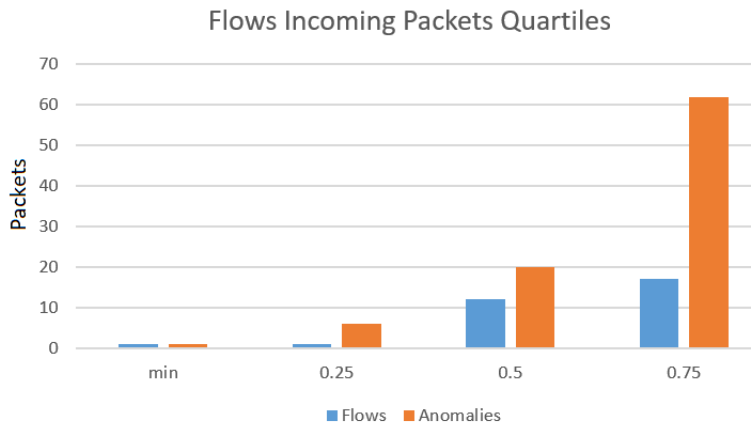

Figure 4.7. Bar graph displaying difference between incoming packet count quartiles for overall and anomalous flows.

As would be expected based on the incoming bytes graph, the anomalies captured the more extreme instances of higher bytes as seen in Figure 4.8. The maximum byte count was a whopping 95.5 Gigabytes worth of incoming data for one flow, and was also excluded due to being an extreme outlier. Flows that represent large amounts of incoming data are potentially more valuable to focus on for analysis since they could contain large amounts of interesting data being brought onto the NPS network.

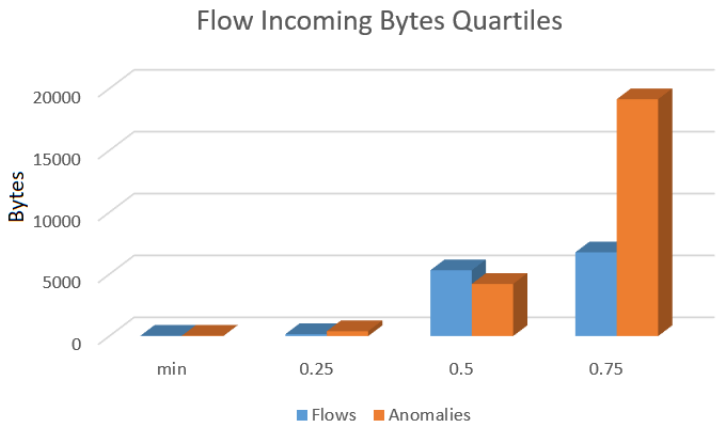

Figure 4.8. Bar graph displaying difference between incoming bytes amount quartiles for overall and anomalous flows.

Rather similar to the incoming packets graph, the outgoing packets quartiles in Figure 4.9 show that the anomalies were more likely to have higher amounts of data flow, with the outlier maximum value in this case being 21645650 packets. Flows that represent large amounts of outgoing data may be interesting for analysis since they could potentially involve exfiltration of data.

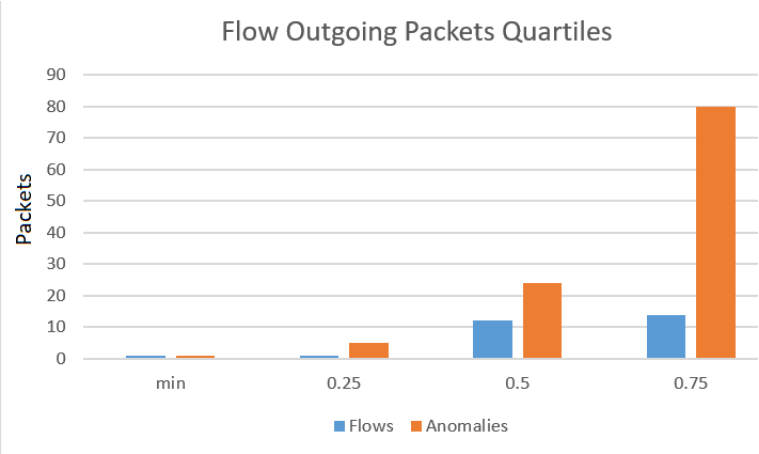

Figure 4.9. Bar graph displaying difference between outgoing packet count quartiles for overall and anomalous flows.

Unlike the outgoing packets, the outgoing bytes in Figure 4.10 are relatively even with each other. The values were very high, so it was scaled using natural logarithm for producing a better graph. For context of the actual value of bytes between the two, Table 4.1 is provided.

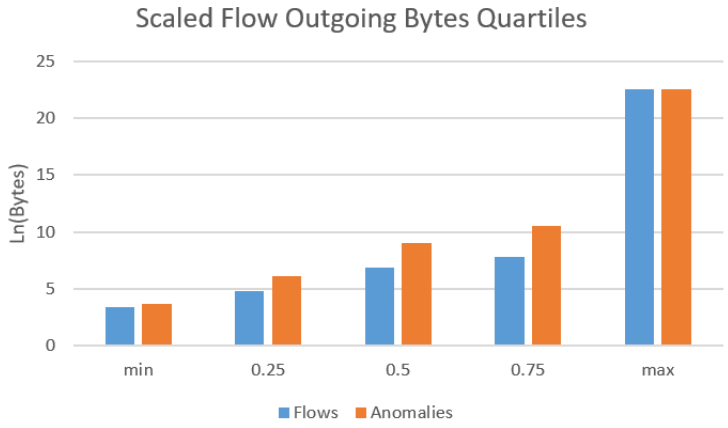

Figure 4.10. Bar graph displaying difference between outgoing bytes amount quartiles for overall and anomalous flows.

Table 4.1. Mean and Standard Deviation of Flow and Anomalous Outgoing Bytes.

Packet Length Statistics		
	Overall	**Anomalies**
Mean	43850.26	3666516
Standard Deviation	4412471	61619340
Max	6150538000	6150538000

Duration

Similar to the previous graph of outgoing bytes, the Figure 4.11 graph was scaled using natural logarithm due to high values. A majority of flows had a duration close to 0 seconds long, and the anomalies heavily leaned towards the longer flows. This relates quite directly with the fact that the flow anomalies corresponded to higher byte and packet counts as well. Table 4.2 is provided for context of the actual value for duration across flows.

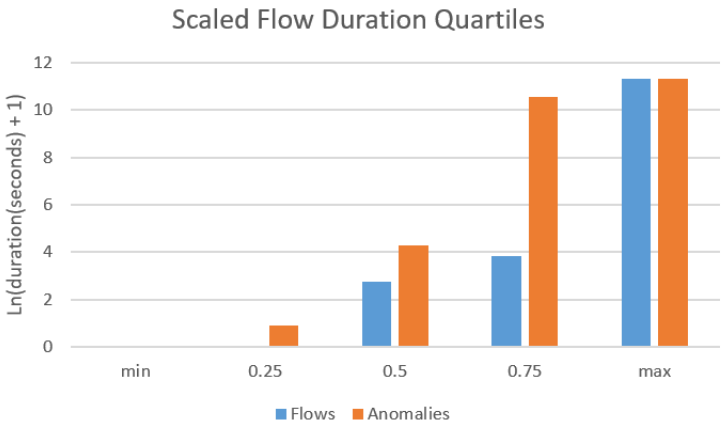

Figure 4.11. Bar graph displaying difference between duration quartiles for overall and anomalous flows.

Table 4.2. Mean and Standard Deviation of Flow and Anomalous Durations.

Packet Length Statistics		
	Overall	**Anomalies**
Mean	2557.465	16382.35499
Standard Deviation	10228.03	25044.01674
Max	6150538000	6150538000

Source Country

It is immediately obvious that the flows have more presence of foreign countries in Figure 4.12 than the packets in Figure 4.4. This is likely because of the sheer quantity of packets in comparison to flows, since several packets can exist per flow. Close to 93% of flows originated from the U.S., with 4% coming from Europe, 1.5% from Asia, 1.2% from the

Commonwealth, and 0.07% from the Middle East. Something strange is the fact that South America does not appear whatsoever in the flows, despite the fact that it can be seen in some packets like in Figure 4.4. This implies that despite the fact that there was some incoming traffic from South America that made it through the firewall, none of it ever received responses from within NPS in order to be counted as a flow. It is reasonable to state that this actually indicates some undesirable traffic since a packet that does not receive a response generally indicates that it did not reach for a valid host or service and was simply ignored despite making it through the firewall. Since such traffic made it through the firewall, it is likely that there was no serious malicious logic embedded, but could have been some form of automatic probing activity.

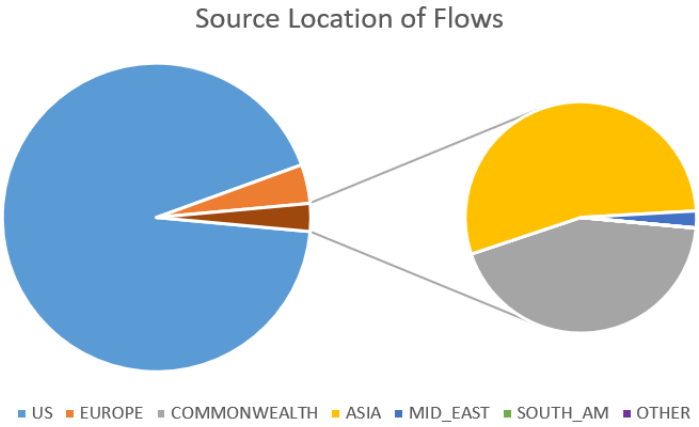

Figure 4.12. Pie chart displaying the distribution ratio of source country buckets across the flow metadata data set.

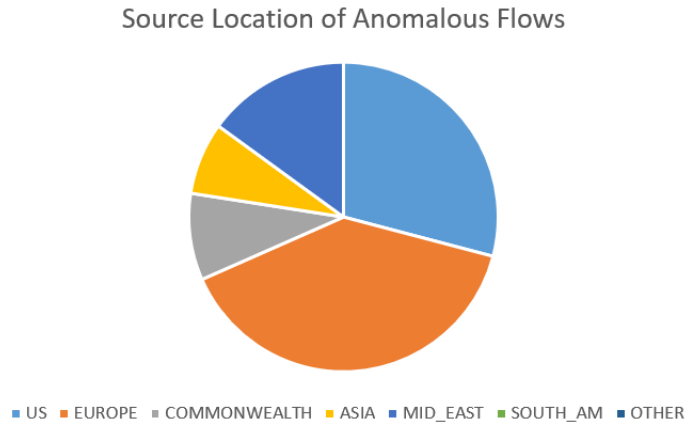

Figure 4.13. Pie chart displaying the distribution ratio of source country buckets across the flow metadata anomalies.

Anomaly detection heavily leaned towards the source country features, and mixed up the distribution of anomalies to make the U.S. actually have 10% less presence than Europe.

4.1.3 Latitude and Longitude

The source country for network traffic has been displayed, but in order to analyze latitude and longitude an approach specific to the U.S. was taken. Both packet header and flow metadata data sets were filtered for source traffic from the U.S., and then subsequently filtered for being within the geographic region depicted in Figure 4.14.

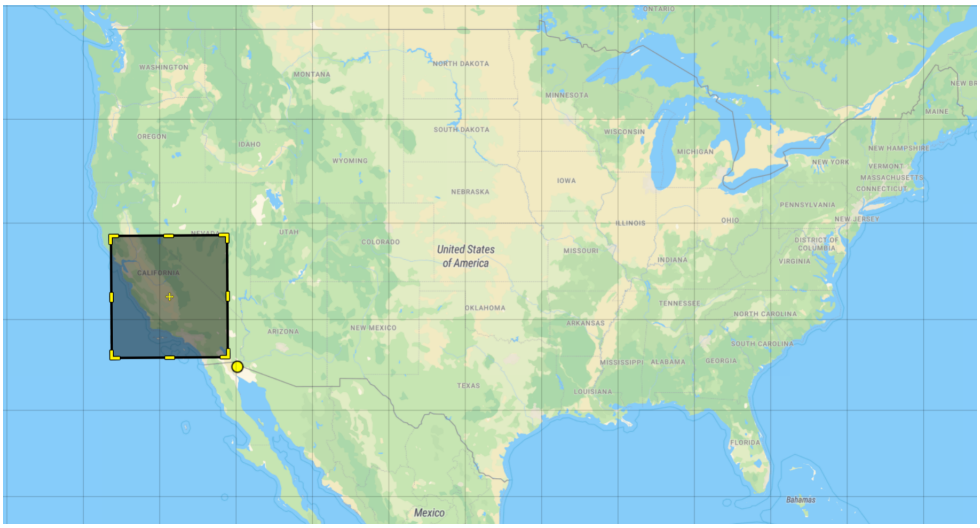

Figure 4.14. Map of the United States with blocked out region depicting area that was filtered for. Source: [15].

The coordinate corners of the dark region on the map are the following:

(LONG, LAT):

(-123.18,32.91),(-115.53,39.39)

The ratio of U.S. traffic within southern California and outside of southern California was graphed in Figure 4.15 for both flows, packets, and anomalies for each.

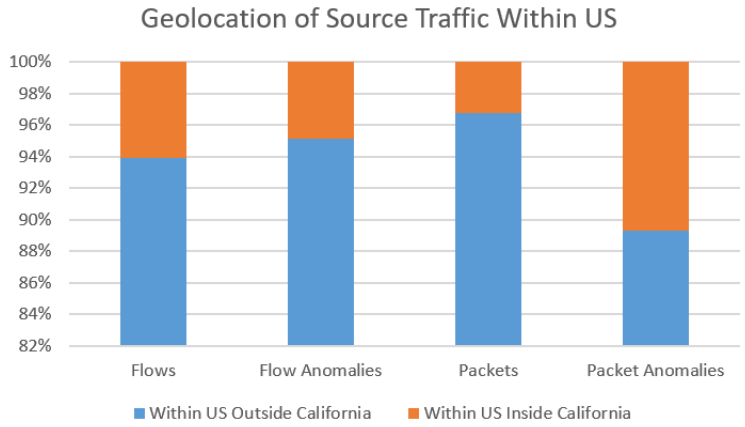

Figure 4.15. Bar graph displaying the ratio of traffic within and outside of Southern California in the U.S. for both flows, packets, and corresponding anomalies.

The amount of overall network traffic that was determined to be from outside the general physical location of NPS was surprising, but upon further inspection it seemed as though this was because the GeoLite2 database was unable to specifically pinpoint the latitude and longitude of many IP addresses within the U.S., and generalized all of them to be roughly in the middle of the country within the state of Kansas. There does not seem to be any clear pattern to glean from the result in Figure 4.15, and a likely reason would be due to the inaccuracy of the GeoLite2 database for coordinates within the U.S.

CHAPTER 5:
Discussion and Future Work

5.1 Discussion

In this research the Gaussian Mixtures unsupervised machine learning technique was used on network traffic data that was not labeled into any predefined category. One of the data sets used was packet headers, specifically ingress packets that were not blocked by the firewall. The other data set was for flow metadata of network traffic flows between hosts on either side of the NPS firewall.

A primary goal for anomaly detection is determining records that contain feature values that are uncommon compared to most other records in the data, and Gaussian Mixtures detected anomalies that met this criteria quite well. Despite making up a minuscule portion of overall traffic, a good portion of anomalies accounted for UDP and ICMP traffic (especially in packets). Since ICMP is used for functionality such as pinging or other error providing codes, and UDP can be used without establishing a clear flow, they both would be good packets for a network admin to be aware of as potential danger.

The MTU for Ethernet packets is 1500 bytes, so it makes sense that an overwhelming majority of packets attempt to make full use of this limit in order to efficiently transfer data with as few packets as possible. Unlike the overall traffic, roughly half of the anomalous packets have packet lengths smaller than 1500 bytes. Packets with smaller lengths could mean something innocuous such as being the final packet in a series of data, but can be interesting for network admins as well if there are oddly specific packet sizes going through a network that may represent something akin to a covert channel.

While the well-known service ports seen from the NPS side were fairly straightforward services such as HTTPS, SMTP, and DNS, there were few more noteworthy ones on external servers. FTP was seen, which is worth monitoring by network admins as novel or obfuscated malware could be brought into the network through the firewall. The Echo protocol was also found which would be valuable to monitor due to the possibility of another

external user performing IP spoofing to make the Echo server seem like the original sender of some traffic.

Based on all of the flow statistics, it is clear that most flows involved a rather brief back and forth between hosts. In every case anomaly detection heavily leaned towards the less common flows since it accounted for flows with more incoming packets and bytes as well as more outgoing packets. While the amount of traffic exchanged in a flow alone would not indicate something suspicious, correlating it with the service that is attached to it could shine light on strange behaviors such as long DNS TCP conversations.

Both data sets had over 90% of the traffic originate from within the U.S., but the anomalies captured a great amount of the foreign sourced traffic with roughly 65% of anomalies being foreign source traffic across the data. It is fairly straightforward as to why the source country of any given traffic would be valuable to a network admin, as any countries that are known to be adversarial could be given greater scrutiny.

Another interesting finding from the work is the fact that the combination of analyzing both incoming packets and flows yielded some interesting findings. Namely, the fact that there were ingress packets captured that were sourced from South America, but that there were no flows originating from South America. Such correlations between the two data sets could be a key indicator for network admins regarding potentially unwanted probing activity, especially if the source country is unusual.

When it comes to implementations, in a similar fashion to previous works, this can be used as part of an IDS system as a way to inform network admins of traffic that could be causing issues. Theoretically it could be done by running Gaussian Mixtures on an already captured data set in order to generate clusters, and then proceeding to determine if any given packet or flow falls outside of any of those clusters based on some threshold in order to be considered an anomaly.

Another potentially valuable implementation would be for use as a form of forensics. After an attack occurred and the recovery process has completed, any traffic that was captured around the time that compromise was suspected to have occurred can have the Gaussian Mixture anomaly detection technique performed on it, and analysts could streamline their search for the offending traffic towards some of the anomalies that were detected.

Limitations

A big limitation to consider for this study would be that there is no traffic within the data sets analyzed that was confirmed to be malicious or even undesirable. Since all of the traffic analyzed was accepted by a firewall, it is entirely reasonable to imagine that none of the data is undesirable at all. There is no feasible method for determining what traffic if any is actually undesirable due to the sheer size of the data sets as well as the lack of payload.

While the maxmind GeoLite2 databases were able to successfully tie every IP address to a country, the actual latitude and longitude coordinates showed signs of not being accurate. When looking deeper into flows, only 6% of them were considered to be from within the southern California region. About 31% of the coordinates were located in Kansas and 26% were close to the border between Oregon and Washington. It is possible that a good amount of the traffic geolocated at the Oregon and Washington border were actually from California, but the fact that the majority of traffic was supposedly sourced from Kansas is a big red flag. The likely explanation of this is that the coordinate geolocation from the GeoLite2 City database is quite coarse or simply did not have enough information on west coast Internet Service Providers.

Another consideration is that the network data used was from the Naval Postgraduate School, which is a fairly unique institution when considering its combined military and education purposes. The threat actors targeting NPS as well as the security posture of the NPS network may be uncommon compared to other ASNs. While it is very reasonable for similar research to be replicated in other environments, it should be expected that results would vary based on the different conditions in each environment.

5.2 Future Work

There are various possibilities for future work from this research that could lean towards Machine Learning, or to the data itself.

- Other unsupervised machine learning anomaly detection methods can be tested with a similar approach.
- Instead of incoming packets, a data set could be created for exclusively outgoing packets allowed through a firewall in order to perform analysis on the types of traffic

internal users are trusted to send externally.

- Use or create a data set that comprises of known malicious or otherwise undesirable traffic that actually does get through firewalls in the real world. One method for accomplishing this could be by artificially generating the data of stealthy attacks for use in anomaly detection.
- Using a more accurate geolocation database for determining latitude and longitude coordinates in order to perform better analysis on traffic from specific locations within the country, relative to where data was captured.
- Instead of grouping by individual hours for packets, other efficient aggregation and computation techniques can be used for performing machine learning on larger portions of data.

CHAPTER 6:
Conclusion

6.1 Conclusion

In this study, we determined that Gaussian Mixtures perform well for detecting anomalies with network traffic data, both packets and flows. The strength of geolocation features with network traffic for machine learning models was displayed, and using even more accurate databases would yield greater results. The anomalies managed to capture many unique values for varying features that would be useful for narrowing down more suspicious traffic. Similar to other research in the field, this strategy can be deployed with software such as an Intrusion Detection System, or even for a forensics style program that analyzes traffic that entered a network for potential covert threats after an attack had occurred.

This study demonstrates a solution for additional network level alerts that can be used across different network types with relative ease, and shows that there are more arising potential solutions to the increasing threats present in the digital world.

THIS PAGE INTENTIONALLY LEFT BLANK

List of References

[1] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. Sebastopol, CA, USA: O'Reilly Media, 2019.

[2] Cisco Systems, Inc., "What is a firewall?." Accessed Feb. 1, 2023 [Online]. Available: https://www.cisco.com/c/en/us/products/security/firewalls/what-is-a-firewall.html

[3] Check Point Software Technologies Ltd., "What is an intrusion detection system (IDS)?." Accessed Feb. 5, 2023 [Online]. Available: https://www.checkpoint.com/cyber-hub/network-security/what-is-an-intrusion-detection-system-ids/

[4] M. Calnan, "Multi-dimensional profiling of cyber threats for large-scale networks," M.S. thesis, Dept. of Natl. Sec. Aff., NPS, Monterey, CA, USA, 2022 [Online]. Available: https://hdl.handle.net/10945/71108

[5] B. Allen, "Netcom 2023 network traffic analysis (draft)," unpublished.

[6] M. V. Mahoney and P. K. Chan, "Phad: packet header anomaly detection for identifying hostile network traffic," Melbourne, FL, Florida Institute of Technology., 2013 [Online]. Available: https://repository.lib.fit.edu/handle/11141/94

[7] M. Thottan and C. Ji, "Anomaly detection in ip networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 8, pp. 2191–2204, 2003.

[8] D. Pujol-Perich, J. Suarez-Varela, A. Cabellos-Aparicio, and P. Barlet-Ros, "Unveiling the potential of graph neural networks for robust intrusion detection," *SIGMETRICS Perform. Eval. Rev.*, vol. 49, no. 4, p. 111–117, Jun 2022. Available: https://doi.org/10.1145/3543146.3543171

[9] R. Kumari, Sheetanshu, M. K. Singh, R. Jha, and N. Singh, "Anomaly detection in network traffic using K-mean clustering," in *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*, 2016 [Online]., pp. 387–393. Available: https://ieeexplore.ieee.org/document/7507933

[10] R.-H. Hwang, M.-C. Peng, C.-W. Huang, P.-C. Lin, and V.-L. Nguyen, "An unsupervised deep learning model for early network traffic anomaly detection," *IEEE Access*, vol. 8, pp. 30 387–30 399, 2020.

[11] S. Valenti, D. Rossi, A. Dainotti, A. Pescapè, A. Finamore, and M. Mellia, "Reviewing traffic classification," in *Data Traffic Monitoring and Analysis*, E. Biersack,

C. Callegari, and M. Matijasevic, Eds. Berlin, Heidelberg: Springer, 2013, pp. 123–124.

[12] Maxmind, Inc., "Geolite2 free geolocation data," Jan. 23, 2023 [Online]. Available: https://dev.maxmind.com/geoip/geolite2-free-geolocation-data?lang=en

[13] F. Iglesias and T. Zseby, "Analysis of network traffic features for anomaly detection," *Machine Learning*, vol. 101, Dec. 2014 [Online]. doi: https://doi.org/10.1007/s10994-014-5473-9.

[14] HYPR Corp, "Encapsulating security payload (ESP)." Accessed Feb. 22, 2023 [Online]. Available: https://www.hypr.com/security-encyclopedia/encapsulating-security-payload-esp

[15] Klokan Technologies, "Bounding box." Accessed Feb. 26, 2023 [Online]. Available: https://boundingbox.klokantech.com/

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California