

Augmented network log anomaly detection

THESIS

Robert J. Gutierrez, Second Lieutenant, USAF

AFIT-ENS-MS-17-M-131

DEPARTMENT OF THE AIR FORCE

AIR UNIVERSITY

###### AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DISTRIBUTION STATEMENT A.

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.AFIT-ENS-MS-17-M-131

Augmented network LOG ANOMALY Detection

THESIS

Presented to the Faculty

Department of Operations Research

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Operations Research

Robert J. Gutierrez, BS

Second Lieutenant, USAF

March 2017

DISTRIBUTION STATEMENT A.  
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT-ENS-MS-17-M-131

Augmented network LOG ANOMALY Detection

Robert J. Gutierrez, BS

Second Lieutenant, USAF

Committee Membership:

Dr. Kenneth Bauer

Chair

Dr. Trevor Bihl

Member

LTC Cade Saie, PhD

Member

AFIT-ENS-MS-17-M-131

# Abstract

This focus of this research is in computer aided human analyst threat detection. The primary motivating question is if time oriented blocks of the data can be used to create meaningful state vectors to suggest subsequent analysis, where state vectors are columns of pre-defined data attributes with representative counts. Using a tabulated vector approach, we exploit changes in the state vectors through graphical and multivariate analysis to identify potentially concentrated anomalies.

# Acknowledgments

First, I must thank my wife and son for their persistent support. I would like to express my sincere appreciation to my faculty advisor, Dr. Kenneth Bauer, and to Dr. Trevor Bihl for their guidance and support throughout the course of this thesis effort. The insight and experience was certainly appreciated. I would also like to thank my sponsor, Lt Col Cade Saie, PhD, from the Army Cyber Command for both the support and latitude provided to me in this endeavor. I also extend my gratitude to Dr. Bradley Boehmke for his expertise and support in the development of the shiny web application.

Robert J. Gutierrez

# Table of Contents

Page

[Abstract iv](#_Toc472581004)

[Acknowledgments v](#_Toc472581005)

[Table of Contents vi](#_Toc472581006)

[List of Figures ix](#_Toc472581007)

[List of Tables x](#_Toc472581008)

[I. Introduction 1](#_Toc472581009)

[1.1 Motivation 1](#_Toc472581010)

[1.2 Research Goals 2](#_Toc472581011)

[1.3 Research Contributions 2](#_Toc472581012)

[1.4 Assumptions/Limitations 3](#_Toc472581013)

[1.5 Organization 3](#_Toc472581014)

[II. Literature Review: Application Domain 4](#_Toc472581015)

[2.1 Chapter Overview 4](#_Toc472581017)

[2.2 Cyber Network 4](#_Toc472581018)

[2.2.1 Army Cyber Data Capture 5](#_Toc472581019)

[2.2.2 Firewalls 6](#_Toc472581020)

[2.2.3 Intrusion Detection/Prevention Systems 6](#_Toc472581021)

[2.3 Previous Cyber Anomaly Detection Methods 10](#_Toc472581022)

[2.3.1 Outlier Detection 10](#_Toc472581023)

[2.3.2 Unsupervised Support Vector Machines 10](#_Toc472581024)

[2.3.3 Event Correlation 10](#_Toc472581025)

[2.3.4 Dynamic Rule Creation 11](#_Toc472581026)

[2.3.5 Principal Components Analysis 11](#_Toc472581027)

[2.4 Feature Vector Creation 12](#_Toc472581028)

[III. Literature Review: Statistical Methods 14](#_Toc472581029)

[3.1 Chapter Overview 14](#_Toc472581031)

[3.2 Statistical Tools 14](#_Toc472581032)

[3.2.1 Mahalanobis Distance 14](#_Toc472581033)

[3.2.2 Dimensionality Assessment 15](#_Toc472581034)

[3.2.3 Factor Analysis 16](#_Toc472581035)

[3.3 Graphical Tools 18](#_Toc472581036)

[3.3.1 Histogram Matrix 19](#_Toc472581037)

[3.3.2 Heatmaps 20](#_Toc472581038)

[3.3.3 Network Graphs 20](#_Toc472581039)

[IV. Methodology 22](#_Toc472581040)

[4.1 Chapter Overview 22](#_Toc472581042)

[4.2 Dataset Description 23](#_Toc472581043)

[4.3 State Vector Creation 23](#_Toc472581044)

[4.3.1 Data Blocks 23](#_Toc472581045)

[4.3.2 State Vectors 24](#_Toc472581046)

[4.4 Multivariate Analysis 25](#_Toc472581047)

[4.5 Shiny Web Application 26](#_Toc472581048)

[V. Analysis and Results 27](#_Toc472581049)

[5.1 Mahalanobis Distance 27](#_Toc472581051)

[5.2 Factor Analysis 28](#_Toc472581052)

[5.3 Example Investigation 31](#_Toc472581053)

[5.3.1 Mahalanobis Column Breakdown 31](#_Toc472581054)

[5.3.2 IP Network Graph 32](#_Toc472581055)

[5.4 Shiny Web App 33](#_Toc472581056)

[VI. Conclusions and Recommendations 35](#_Toc472581057)

[6.1 Conclusions 35](#_Toc472581059)

[6.2 Contributions and Future Research 35](#_Toc472581060)

[Appendix A. Hardware/Software overview 36](#_Toc472581061)

[References 37](#_Toc472581062)

# List of Figures

Page

[Figure 1. Basic Cyber Network 4](#_Toc472581063)

[Figure 2. Data Collection Hierarchy 5](#_Toc472581064)

[Figure 3. Classification of Anomaly Detection Techniques [10] 8](#_Toc472581065)

[Figure 4. Example of Cattell's Scree Plot 15](#_Toc472581066)

[Figure 5. Mail server Histogram Matrix [27] 19](#_Toc472581067)

[Figure 6. Sample Correlation Heatmap 20](#_Toc472581068)

[Figure 7. Sample Network Graph 21](#_Toc472581069)

[Figure 8. State Vector Approach 22](#_Toc472581070)

[Figure 9. Sample of Block Generation 24](#_Toc472581071)

[Figure 10. State Vector Matrix 24](#_Toc472581072)

[Figure 11. Histogram Matrix (All Blocks) 27](#_Toc472581073)

[Figure 12. Histogram Matrix (Top 20 Outliers) 28](#_Toc472581074)

[Figure 13. Heatmap of Rotated Factor Loadings 29](#_Toc472581075)

[Figure 14. Rotated Factor Score Plots 31](#_Toc472581076)

[Figure 15. Block 14: Top 5 Breakdown Distance Columns 32](#_Toc472581077)

[Figure 16. Block 14 IP Network Graph 33](#_Toc472581078)

[Figure 17. Shiny Web App 34](#_Toc472581079)

# List of Tables

Page

[Table 1. Kaiser's Evaluation of the IFS Levels [18] 18](#_Toc472581080)

[Table 2. Dataset Variables 23](#_Toc472581081)

[Table 3. IFS Results 29](#_Toc472581082)

[Table 4. Factor-Variable Breakdown 30](#_Toc472581083)

augmented network log anomaly detection

# I. Introduction

“Just as water retains no constant shape [1],” in cyberspace there are no constant conditions. The United States Army Cyber Command (ARCYBER) is dedicated to securing and defending cyberspace and the information environment. Due to the constantly changing behavior of cyber-attacks, ARCYBER is looking for a reactive approach to efficiently detect and prevent malicious attacks. Currently, the Army network relies on a series of firewalls and intrusion detection/ prevention systems to identify and foil suspicious internet traffic. These devices, when triggered, generate a log file containing details of how it handled each incident, such as the source and destination IP addresses, port numbers, protocols, bytes transferred, etc. Through a tabulated vector approach, this research aims to utilize the information provided in the log file to identify deviations from “normal” behavior that may “suggest the presence of intentionally or unintentionally induced attacks, faults, defects, etc.” that are widely accepted as anomalies [2].

## Motivation

As the number of cyber-attacks continues to grow on a daily basis, so has the delay in threat detection. For instance, in 2015, the Office of Personnel Management (OPM) discovered that approximately 21.5 million individual records of Federal employees and contractors had been stolen [3]. Events such as this are not detected as they occur, but rather days, weeks or even months after computer systems have been compromised. Currently, ARCYBER analysts inspect numerous potential incidents on a daily basis, but have neither the time nor the resources available to perform such a task.

## Research Goals

This research aims to curtail the timeframe in which cyber-attacks go unnoticed and to aid in the discovery of these attacks among the millions of daily logged events, while minimizing the number of false positives and negatives. Using multivariate analysis, the proposed methodology will identify time periods associated with suspected anomalies for further evaluation.

Since attacks are constantly changing, the methodology should be dynamic to account for deviations in “normal” behavior patterns. Specific research goals are:

1. Investigate and examine representative ARCYBER data
2. Develop pattern recognition methods to characterize and detect normal and anomalous activity in ARCYBER data
3. Employ and develop data visualization methods to better visualize real time cyber data
4. Investigate “big data” analytics methodologies as they apply to trend recognition and anomaly detection
5. Provide a web based application that incorporates the above methodology while allowing user input for data visualization

## Research Contributions

This research contributes to the mission of ARCYBER by presenting an approach to detect behavior based signatures within log data while providing a web based application for user driven analysis. With the incorporation of these contributions, this research hopes to aid the analysts in identifying regions of time where a suspected anomaly has occurred.

## Assumptions/Limitations

The effectiveness of the security devices are limited to the policies and configuration of each of those devices which are handled by the Cyber Centers (CCs). In general, the CCs are responsible for configuration, maintenance, tuning, and signature development for all sensors within their region while some are managed by the Network Enterprise Centers (NEC) at each post. Therefore, ARCYBER’s knowledge gap lies in not having access to the current configuration, signatures, and rules for each security device at the global level. The data used for this research is assumed to be a representable sample of ARCYBER data.

## Organization

In the second chapter, we present a literature review related to cyber networks, the Army’s data collection process and anomaly detection techniques. Chapter 3 provides a comprehensive literature review of the statistical methods to be utilized in this research, Mahalanobis distance and factor analysis, as well as background on the graphical methods to display the analysis. Chapter 4 details the methodology for this research including a description of the dataset, dataset reduction and the multivariate analysis. Chapter 5 presents the analysis of the results with an in-depth example. Chapter 6 offers conclusions and recommendations for further research on anomaly detection in log files.

# II. Literature Review: Application Domain



## Chapter Overview

As cyber-attacks become more difficult to identify, so has the need for more advanced detection methods. Problematically, normal behavior for cyber networks is generally not well defined and changes over time, resulting in high false positive rates [4]. Most related research has focused on anomaly detection at the device/software level, with little exploration into anomaly detection in the log files generated from the preexisting devices/software. This chapter will first provide a brief background on the devices used to protect security networks. Next is a brief overview of the Army’s data collection process. Finally, a literature review of previous research on the application of anomaly detection methods is presented.

## Cyber Network

To understand network security devices, one must first visualize a cyber network. Figure 1 conceptualizes a basic cyber network where a user is protected by an Intrusion Detection System (IDS) / Intrusion Prevention System (IPS) and a firewall. Each security device plays a crucial role in protecting the user’s computer from outside and inside threats.

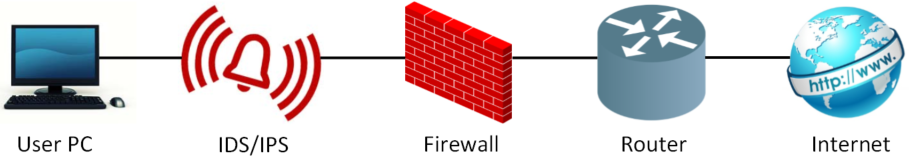


Figure 1. Basic Cyber Network

### Army Cyber Data Capture

With several types of security sensors spread across army networks worldwide, all flagged traffic is recorded into log files. These raw logs are first normalized into a structured data file by a connector, a stand-alone device or software that forwards data and sometimes converts from one format to another, and are then forwarded to Cyber Centers (CC). In Van Vleet’s [5] interview with Army Lieutenant Colonel Griffin, Griffin describes CCs as “organizations that provide LandWarNet service to their theaters while simultaneously defending the network from cyber threats and they operate around the clock”. At the CCs, a regional Security Information and Event Management (SIEM) device aggregates, correlates, monitors and generates alerts from the received data. Next, a second connector forwards the data to a global SIEM known as the Integration Center (IC). After the data is reprocessed at the IC, it is then uploaded into a big data platform. Figure2 provides a visualization of the data collection process.

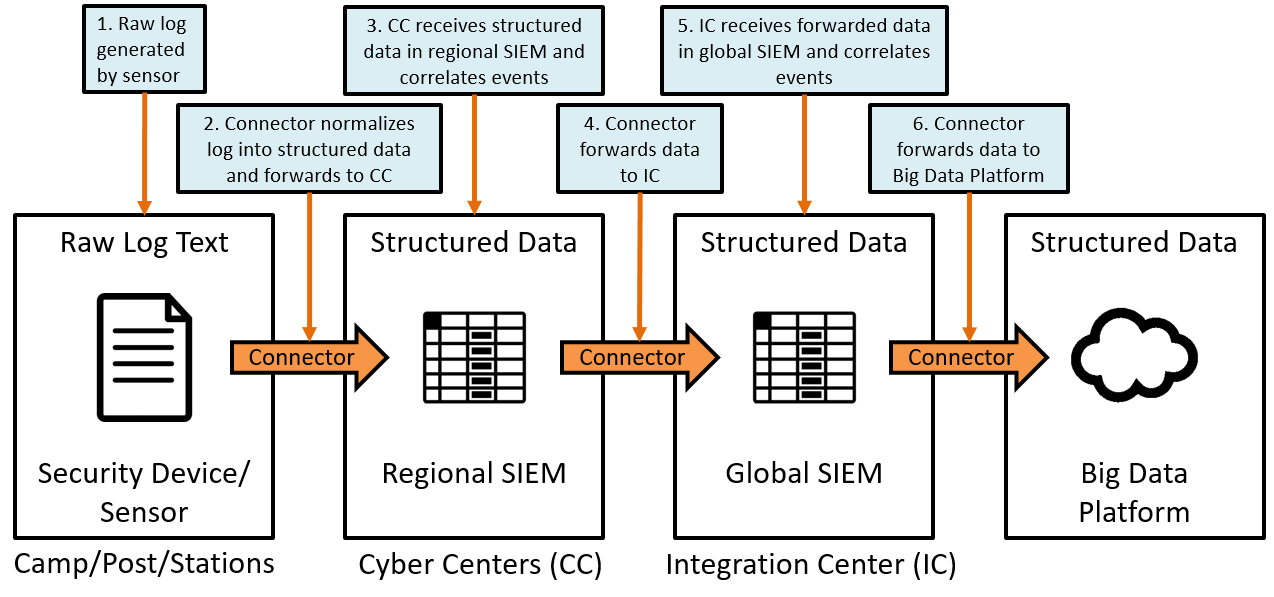


Figure 2. Data Collection Hierarchy

### Firewalls

First, there are firewalls which act as a blockade between an internal and external network. Current firewalls include Palo Alto Networks, Cisco ASA, Mcafee, WinGate, and Norton 360. Firewalls prevent risks, including (1) an internal host system’s exposure to inherently insecure Internet protocols and services, and (2) probes and attacks launched from hosts on the Internet [6].

Firewalls are of three general types, per Wu and Irwin [6]: 1) packet filtering (PF), 2) proxy gateway and 3) circuit level inspection. Packet filtering firewalls consider each incoming and outgoing packet, apply a predefined set of rules to decide whether to forward it or not based on the information available in the packet’s header [6]. Proxy gateways or servers perform as a security filter by providing the appearance of connecting directly to the outside network [6]. Circuit level inspection firewalls use a Socket Secure (SOCKS) proxy server that utilizes an access control list (ACL) to determine if a request is permitted.

### Intrusion Detection/Prevention Systems

Intrusion detection is the process of monitoring network traffic to detect imminent threats such as computer security policy violations and acceptable use policy violations and creates an alert if a threat is detected [7]. Intrusion prevention goes one step further by attempting to stop the detected incidents. Therefore, intrusion detection and prevention systems (IDPS) must identify possible incidents, log information about them, attempt to stop them, and report them to security administrators [7]. These systems monitor internet traffic according to a set of rules similar to firewalls or look for known malicious signatures. A signature is a pattern of information that can be used to recognize an attack on a network or host based device” [8]. IDPSs can be setup in two ways, host based (HIDS) or network based (NIDS), where the former is deployed on each individual computer while the latter is positioned along the network. IDPSs are then divided into two classifications: signature based and anomaly/behavior based. Signature based IDPSs rely on previously identified attack signatures. Anomaly based IDPSs “generate the normal behavior/pattern of the protected system and deliver an anomaly/outlier alarm if the observed behavior at an instant does not conform to expected behavior” [6, p. 852]. Originally proposed by Denning [9] in 1987, anomaly detection techniques in cyber have been a key interest, as seen in [4]. Shown in Figure 3, there are three detection techniques used by anomaly based IDPSs: statistical based, knowledge based and machine learning based. Within each detection technique, there are several analysis methods that can be applied, data dependent.

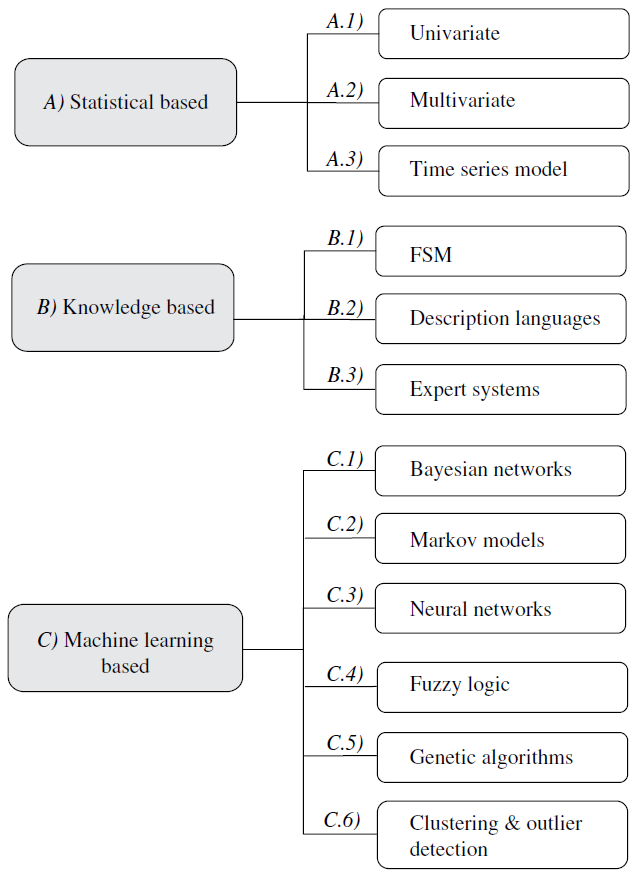


Figure 3. Classification of Anomaly Detection Techniques [10]

The first behavior based technique, statistical based, creates a profile representing the system’s stochastic behavior based on numerical attributes such as traffic rate, count of packets for each protocol, number of unique IP addresses, and etc. These numeric values define a statistical profile that can be compared to a distribution or previously defined profile. As more network traffic flows into the IDPS, the “normal” profile is updated, thus improving the detection accuracy. The benefits of statistical approaches include not requiring prior knowledge about the normal profile and the ability to track malicious activities over long periods of time [10]. However, one drawback is that this method can be misled to consider network traffic generated during an attack as normal [10].

The second technique, knowledge based, summarizes the “normal” behavior from the data and incorporates expert knowledge to look for anomalies [6, p. 853]. This technique is time consuming and labor intensive yet it yields a low false alarm rate and has the potential to detect zero-day and mutated attacks.

The final technique for anomaly based IDPSs is machine learning based, which seeks to determine the underlying structure of the data and becomes more accurate as new data is provided [6]. While many machine learning methods exist, common methods include Bayesian networks, Markov models, neural networks, genetic algorithms, clustering, outlier detection, and spectral anomaly detection. Machine learning methods provide flexibility in discovering new attacks as well as adaptability when given new data. A comparison of these techniques can be found in [2] [10].

Conversely, these methods tend to generate a large false positive rate which is why they are not commonplace in commercial use. Recent work by Walter [11] addresses the class imbalance problem that plagues the accuracy of machine learning methods. In his research, he not only provided an assessment on four machine learning methods handling class imbalance, but also on the effectiveness of such techniques in feature selection.

## Previous Cyber Anomaly Detection Methods

### Outlier Detection

While there are many variations in the definition of an outlier, it can simplified to an observation that stands out among other observations. Security devices can employ two types of outlier detection, supervised and unsupervised. Supervised detection relies on training data to either be attack-free or contain attacks that are labeled as such. Unsupervised detection identifies attacks by determining if an event has a statistical deviation from the rest of the data. While many NIDS currently rely on supervised anomaly detection, Zhang and Zulkernine [12] proposed the incorporation of random forests in unsupervised anomaly detection due to its applicability in related fields such as pattern analysis and bioinformatics. Instead of distinguishing between an attack and normal activity, random forests categorize the data by building a pattern for each unique class. An outlier can then be determined by its proximity to a class according to a specified threshold [12].

### Unsupervised Support Vector Machines

Lazarevic et al. [2] discussed that unsupervised support vector machines (SVM) differ from standard supervised SVMs by “different values for SM parameters (variance parameter of radial basis functions (RBFs), expected outlier rate), the models with different complexity may be built”.

### Event Correlation

Previous work in log-based intrusion detection has attributed the large number of false positives to research being limited to a single source rather than correlating the activities of all involved components [13]. Referring back to Figure 1, a cyber network is made up of interconnected security devices, where each device produces its own log file. Abad et al. [13] remark that because an intrusion generally leaves multiple signs of its presence, an anomaly detected in one log warrants examination in the other logs during the same time period. Therefore, an intrusion correlation system should be able to discover the relationships between two or more events to shed light on complex distributed attacks.

Due to the sheer volume of log data across multiple security devices, conducting log analysis had become more difficult without an organized aggregation technique. Grimalia et al. [14] found that distributed event correlation outperforms centralized techniques by overcoming the limitations of network bandwidth utilization and database query efficiency. Streamlining the process of event correlation enables security devices to identify potential intrusions in a timely manner [14].

### Dynamic Rule Creation

Brier and Branišová [15] developed an approach for log analysis that constructs an anomaly profile based on log data split into blocks, where a block is characterized by the starting time, session duration and type of a service.

### Principal Components Analysis

Due to the large number of events detected by IDPSs, Garcia-Teodoro et al. [10] suggested using principal components analysis (PCA) as a technique to reduce the number of variables down to a set of uncorrelated components, representative of the amount of variance they explain. The main advantage of PCA is its independence of a distributional assumption. While the use of major components (those that explain ~ 50% of the total variation) has been used for outlier detection, Shyu et al. [16] proposed the additional usage of the minor components (those whose variance is less than 0.20), claiming that their method can distinguish whether an outlier is an extreme value or it does not have the same correlation structure as the “normal” data. Then in 2006, Wang and Battiti [17] developed the identification matrix, which applied PCA to NIDS. This matrix was determined by the squared Euclidean distance between an observation’s deviation from the average and the observation’s distance between its reconstruction onto the subspace of the normal behavior.

## Feature Vector Creation

Winding e. t al. [16] proposed that machine learning techniques can best be applied to log files by aggregating the records into a set of feature vectors, which can utilize boxplot outlier analysis and clustering. A feature vector is a count of occurrences for the unique values in a set of variables. The authors provided an example of feature vector development that began with amassing the log file records into the following set of feature candidates:

* Repeated attempts of access by a single IP
* Number of source IPs per destination IP
* Number of destination IPs per source IP
* Number of destination ports on a given source/destination IP pair
* Unique IPs
* Maximum activity from a single IP
* Failed and successful connections from the same IP
* Attempts to access invalid IPs
* Inbound/Outbound bytes per unit time.

Then using the 9 feature candidates, the feature vectors below were derived:

* Source IP address, number of destination IP Addresses
* Destination IP, number of failed access attempts
* Source IP, destination IP
* Destination Perspective Vector (destination IP, count of Source IPs, number of successful accesses, number of failed accesses, count of destination ports, number of bytes transferred[inbound], number of bytes transferred [outbound])

# III. Literature Review: Statistical Methods



## Chapter Overview

The purpose of this chapter is to provide an introduction to the statistical and graphical tools that will be utilized in the methodology of this research. It begins with a discussion on outlier detection using Mahalanobis distance. Then, a brief overview on the methods for choosing the number of factors to retain. Next is a discussion of factor analysis, a multivariate analysis technique that relates a set of variables to corresponding factors, followed by an objective assessment of the factor analysis solution. Finally, this chapter focuses on the graphical tools used to display the aforementioned analysis.

## Statistical Tools

This section will provide details on the statistical methods applied in this research, beginning with Mahalanobis distance and a breakdown distance metric. Then, a thorough discussion on dimensionality reduction, followed by an explanation of factor analysis.

### Mahalanobis Distance

Mahalanobis distance [17] is a multivariate method for outlier detection which calculates the distance between a data point and the overall mean, independent of the scale between variables Let *MD* be the Mahalanobis distance defined as

(1)

Where:

= vector of observations =

= mean vector of data =

= inverse covariance matrix.

Next, let *BD* in denote the breakdown distance matrix, which computes a measure of the relative contribution of each variable, *xi*, to the MD,

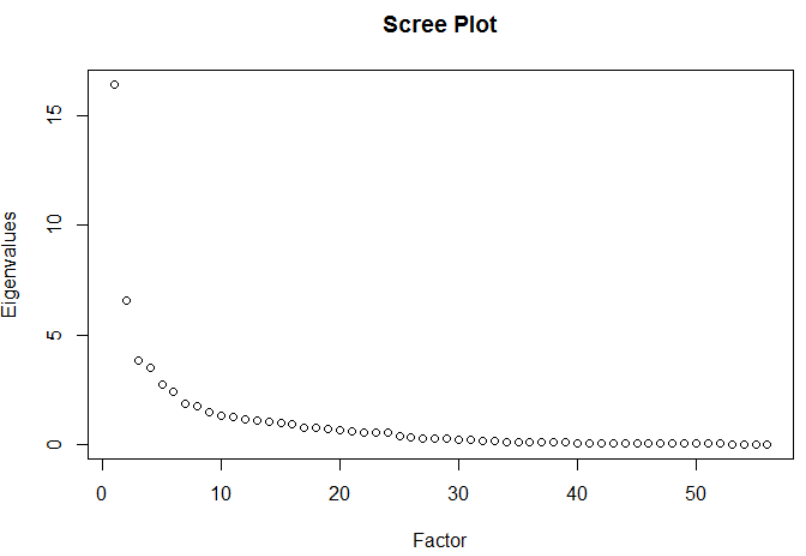
(2)

Where:

= the variance of *xi.*

### Dimensionality Assessment

There are three generally accepted methods of determining the dimensionality for correlation matrix inputs. The first and most commonly used is Kaiser’s Criterion [18] which advises to retain those factors whose eigenvalues are greater than 1.0. Second is Cattell’s Scree Test [18] which involves graphing the eigenvalues and retaining those that form the steep curve. In Figure 4, we look for the values that form a steep curve and those that look like a straight line. The retained factors are those that lead up to the flat line.



Look for steep curve

Identify flat line

Figure 4. Example of Cattell's Scree Plot

The third method is a modified scree test called Horn’s Parallel Analysis (i.e. Horn’s Curve) [19]. Let a *n x p* matrix equal the size of the data of interest, where *n* is the number of rows and *p* is the number of columns. This method utilizes a Monte Carlo simulation to randomly generate a large number of data matrices of size *n x p* where each element is drawn from a standard normal probability distribution. Next, we extract the eigenvalues from the *p x p* correlation matrix. By plotting the average eigenvalues in descending order, the eigenvalues form a logical bound that provides a recommendation for the number of factors to keep. For this research the dimensionality is determined by comparing the eigenvalues generated from Horn’s Parallel Analysis to the sorted eigenvalues from the data of interest and retaining only those factors whose eigenvalues are greater than or equal to Horn’s eigenvalues.

### Factor Analysis

The purpose of factor analysis is to explain the structure of the data by relating the correlations between variables through a set of factors [21]. The factors are first determined through the eigen-decomposition of the correlation matrix. Then, using the dimensionality assessment noted above, the number of factors *r* to retain is determined. Let Λ denote the factor loadings matrix from what is known as the principal components solution for factor analysis, the factor loadings are estimated using [24].

(3)

Where:

λ*i* = eigenvalue for each factor

e*i* = eigenvector for each factor

*i* = 1…*r*

Factor loadings can range from -1 to 1, which indicate how much that factor affects each variable. Values close to 0 imply a weak effect on the variable.

To simplify the results for interpretation, the factor loadings can undergo an orthogonal or oblique rotation [23]. Orthogonal rotations assume independence between the factors while oblique rotations allow the factors to correlate. For this research, we utilize the most common rotation option known as varimax. Varimax rotates the factors orthogonally to maximize the variance of the squared factor loadings which forces large factors to increase and small ones to decrease, providing easier interpretation.

Factor scores are used to examine the behavior of the observations relative to each factor. This research will plot factor scores against one another as a method for outlier detection. The scores are calculated by using Equation 4 below.

(4)

Where:

Xs = standardized observations

R-1 = inverse of the correlation matrix

Λ = factor loadings matrix

To assess the quality of a factor analysis solution, Kaiser [15] proposed the Index of Factorial Simplicity (IFS) that measures the tendency towards unifactoriality for both a given row and the entire matrix as a whole. Beginning with a loadings matrix *C* (not to be confused with the covariance matrix), Kaiser first considers Carroll’s criterion shown below.

(5)

Where:

j = row index

s, t = column index

*v*js = value in the loadings matrix

By defining the overall squared IFS as seen in Equation 6 below, the value can vary between 0 and 1.

(6)

Substituting the value of *C* and performing a little algebra, the IFS value is then simplified to the solution seen in Equation 7.

(7)

Where:

q = the number of factors

j = row index

s = column index

*v*js = value in the loadings matrix

After taking the square root of the value from Equation 7, we compare the value to Kaiser’s proposed evaluations of the IFS value shown in Table 1.

Table 1. Kaiser's Evaluation of the IFS Levels [18]

|  |  |
| --- | --- |
| **IFS Level** | **Evaluation** |
| In the .90s | Marvelous |
| In the .80s | Meritorious |
| In the .70s | Middling |
| In the .60s | Mediocre |
| In the .50s | Miserable |
| Below .50s | Unacceptable |

## Graphical Tools

This section will describe the graphical tools applied to the statistical methods outlined in Section 3.2, along with a visual example. This research will first use a histogram matrix to display the MD and BD, followed by a heatmap to visualize the factor loadings, and conclude with a network graph to understand the source and destination IP pairings.

### Histogram Matrix

A Histogram Matrix (HMAT) is an innovative visualization technique developed by Frei and Rennhard [27], combines graphical and statistical techniques to aid security administrators in efficiently identifying anomalies. HMAT is designed to scan large log files that show steady normal behavior and examine the messages displayed for each observation. Instead of seeing bars as in a typical histogram, the data is represented through a series of circles (where the radius directly corresponds to the height of the bar) [27]. Figure 5 below shows a histogram matrix, where one day is divided into twenty four slots of one hour.



Figure 5. Mail server Histogram Matrix [27]

While the size of the circle provides valuable information, Frei and Rennhard [27] also add color to the circles which serves as indication to the relative likelihood of the time slot. The authors determined the color of a circle by comparing the distribution of the sizes of the circles in its column with previous columns. In Figure 5, the large red circle indicates an unusually large amount of messages, greater than 5 standard deviations from the norm. HMAT also provides user interaction, where an administrator can click on one of the circles to reveal all the log messages that define that circle.

### Heatmaps

Heatmaps are graphical representations of a data matrix through the use of a color scale. Developed in 1991 by Kinney, heatmaps were designed as a visualization tool for stock market analysis [19, p. 219]. In statistics, the most common use for heatmaps are for correlation matrices, illustrating the relationship between variables ranging from -1 to 1. See Figure 6 as an example.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| -0.667 | 0.095 | 0.547 | -0.656 | 0.314 | 0.268 | 0.113 |
| 0.848 | -0.234 | 0.397 | 0.937 | -0.699 | 0.378 | 0.826 |
| 0.283 | 0.573 | -0.862 | -0.485 | 0.874 | 0.318 | -0.444 |
| -0.990 | -0.263 | 0.845 | 0.800 | 0.637 | -0.214 | 0.552 |
| 0.413 | 0.038 | -0.507 | 0.731 | 0.157 | -0.652 | 0.933 |
| -0.497 | 0.353 | 0.154 | 0.235 | 0.354 | 0.478 | -0.334 |
| 0.896 | -0.472 | 0.968 | 0.024 | -0.546 | 0.563 | 0.693 |

Figure 6. Sample Correlation Heatmap

### Network Graphs

Network graphs are graphical models that depict a relationship between two or more nodes, connected by edges [29]. This research employs network graphs to illustrate the interaction between source and destination IP addresses, allowing for effortless focal node detection. Thus, each unique IP address is a node. An edge represents the interaction between sources and destinations with its thickness denoting the frequency of interactions between the nodes. See Figure 7 for an example.

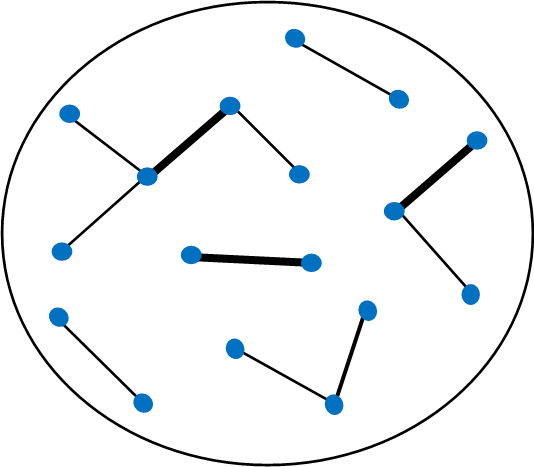


Figure 7. Sample Network Graph

# IV. Methodology



## Chapter Overview

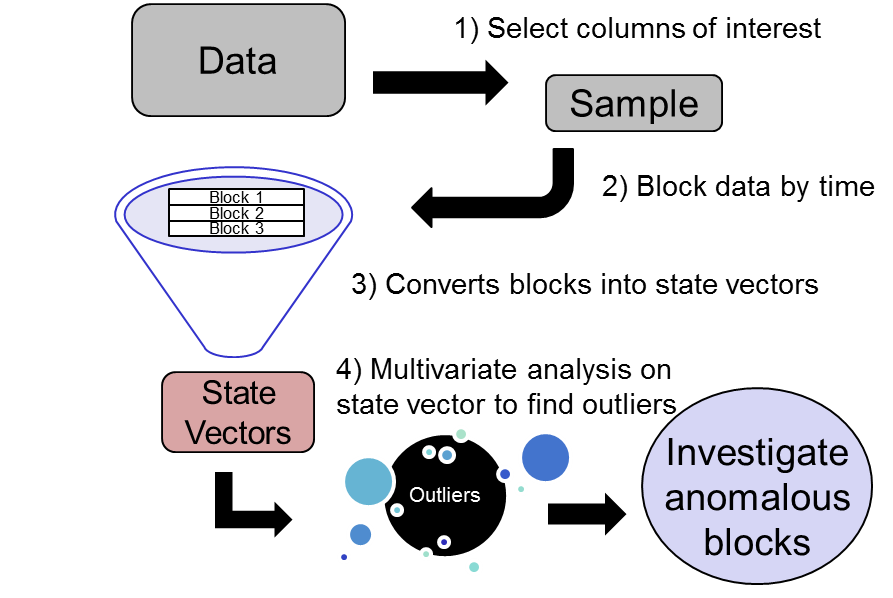


Figure 8. State Vector Approach

Detailed in Figure 8, a tabulated vector approach takes pre-defined data attributes and transforms them into representative counts using descriptive statistics which in turn generates a state vector for a given period of time. These vectors are then aggregated into a state vector matrix. Using a representative dataset, this research aims to provide a proof of concept for exploiting changes in the state vector matrix, allowing for rapid graphical and statistical analysis of potential concentrated anomalies within blocks of time.

## Dataset Description

As discussed in Section 2.2.1, the data used in this research had been collected from sensors located around the world and uploaded into a big data platform. Though there are over 400 data fields, this research focuses on the fields shown in Table 2 below.

Table 2. Dataset Variables

|  |  |
| --- | --- |
| **Field Name** | **Description** |
| Device Vendor | Company who made the device |
| Device Product | Name of the security device |
| Source Address | IP address of the source |
| Destination Address | IP address of the destination |
| Transport Protocol | Transport protocol used |
| Bytes In | Number of bytes transferred in |
| Bytes Out | Number of bytes transferred out |
| Category Outcome | Action taken by the device |
| ad.SCN | Country of the source IP address |

## State Vector Creation

### Data Blocks

After reducing the dataset to the desired variables in Table 2, we then create our data blocks. The 39,304 observations are chronologically separated into 136 data blocks, each containing 289 observations. As seen in Figure 9, categorical variables are labeled as factors while numerical variables are labeled as numeric. The number of levels associated with a categorical variable denote the number of unique entries.

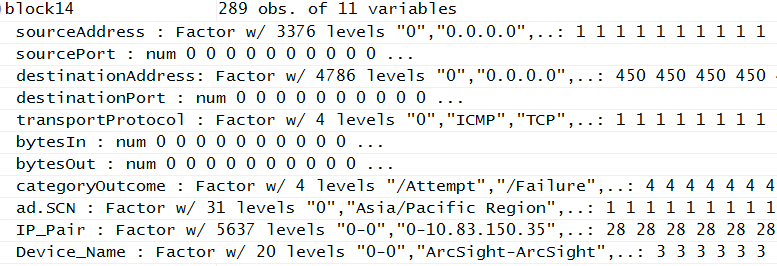


Figure 9. Sample of Block Generation

### State Vectors

As each block is generated, the categorical fields are separated by their levels and a count of occurrences for each level are recorded into a vector. All numerical fields, such as bytes in and bytes out, are recorded as a summation within the block. Due to the large number of levels associated with IP addresses, only the top ten source and destination IP address counts are recorded. These vectors are then aggregated into a single matrix, known as the state vector matrix, as seen in Figure 10.

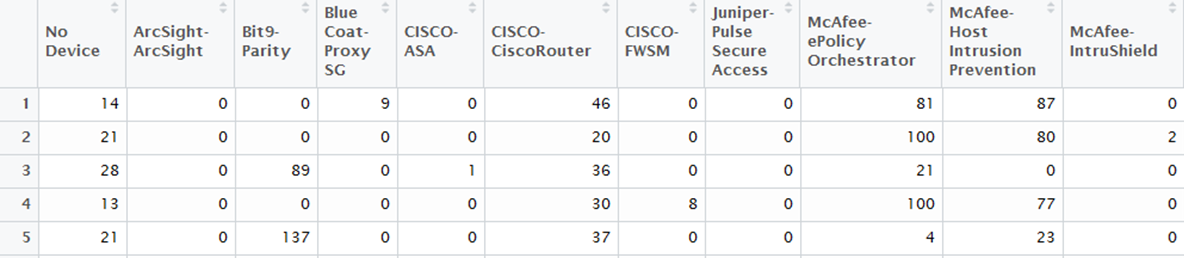


Figure 10. State Vector Matrix

Before the statistical tools, mentioned in Chapter III, are applied to the state vector matrix, the columns of the state vector matrix must meet three criteria: 1) the variance within a column must be greater than 0.1 to avoid matrix singularity; 2) the columns must be linearly independent to avoid computational errors associated with rank deficiency; 3) the values of the correlation matrix cannot exceed a threshold of ±0.95. The first criterion is addressed by removing the columns whose variance is less than or equal 0.1. The second criterion makes use of the *findLinearCombos* function developed by Max Kuhn [need citation], which iteratively removes the columns of a rank deficient matrix to isolate the sets of columns involved in the dependencies, leading to the removal of the columns found to counter the linear combinations. The final criterion searches for columns with correlation values greater than or equal to 0.95. The identified columns are removed and the reduced state vector matrix is ready for multivariate analysis.

## Multivariate Analysis

This research first utilizes the squared Mahalanobis distance as an outlier detection metric to determine the color of each time block in the histogram matrix. The breakdown distance enhances the histogram matrix by adjusting the size of the circles according to its normalized value for each variable.

To begin using factor analysis, the dimensions of the reduced state vector matrix are first passed to the Horn’s curve function to find the recommended set of eigenvalues. Next, the reduced state vector matrix and the eigenvalues generated from Horn’s curve are passed to the factor analysis function. The dimensionality is determined by finding the eigenvalues of the correlation matrix of the state vector matrix and retaining only those factors whose eigenvalues are greater than or equal to those produced by Horn’s curve. Then, the factor analysis function generates two sets of factor scores and factor loadings, unrotated and rotated. Next, we find the IFS values to assess the quality of our solutions and select the set of scores and loadings associated with the larger IFS value.

## Shiny Web Application

Shiny is a web application framework developed for the R software platform. This enables us to display the multivariate analysis noted above in an interactive environment that can be operated by users with little to no experience in this type of research.

# V. Analysis and Results



## Mahalanobis Distance

Figure 11 shows the histogram matrix for the entire dataset, where each row refers to a time block and each column represents a variable. Using the Mahalanobis distance, the outlier time blocks are distinguished by the darker shade of blue. Then, we normalized the breakdown distance for each column to single out the outlier variables within each time block by the size of the circle.

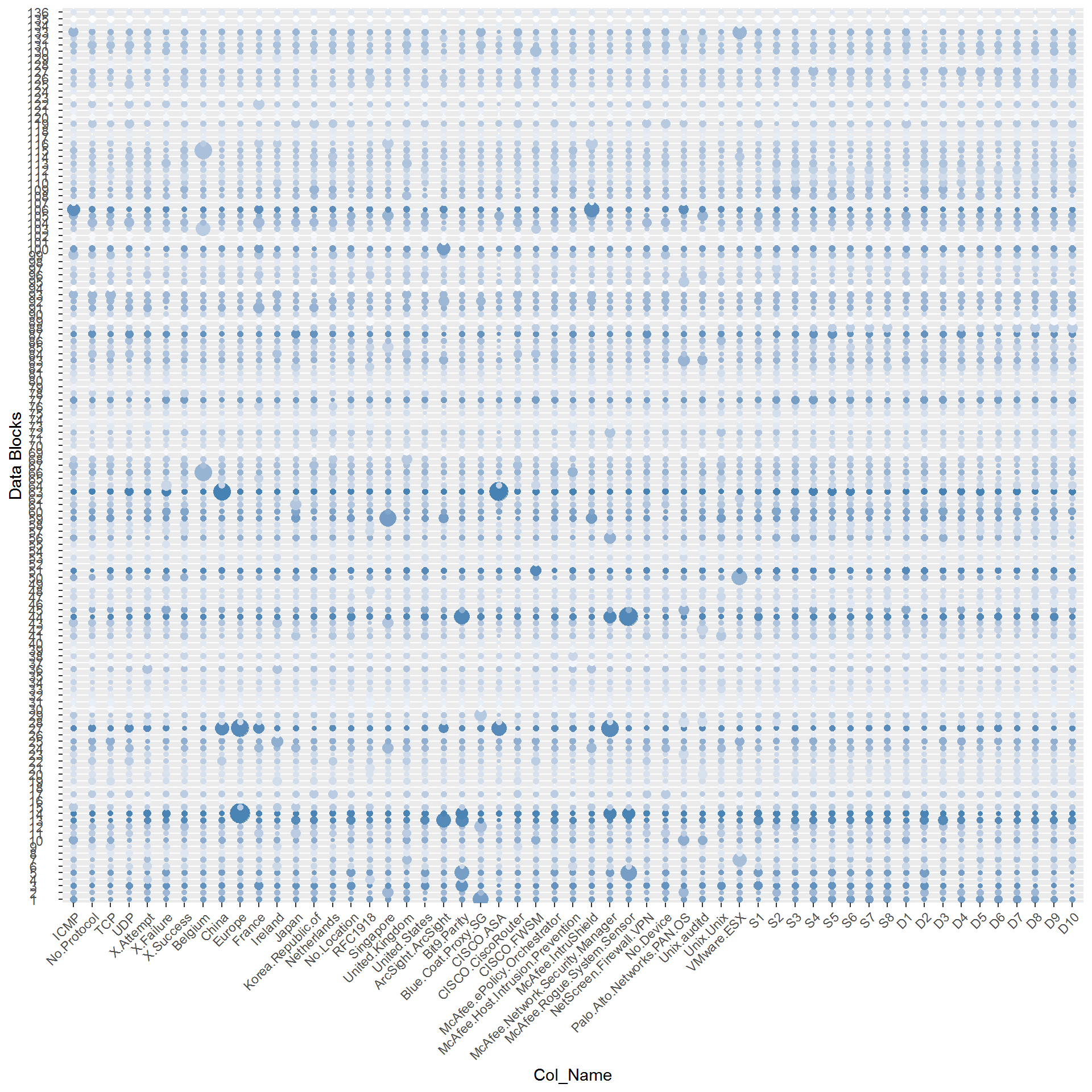


Figure 11. Histogram Matrix (All Blocks)

The rows that are shaded darker imply that they are outliers relative to their MD. Then the columns that have larger circle indicate the variables that are driving the MD for that particular block. The significance of this figure is displaying 2 dimensional analysis.

Figure 12 shows the histogram matrix for the twenty rows with the highest Mahalanobis distance. This enables us to see the temporal aspect of the outlier blocks, providing insight into how the observed anomalies may be correlated with one another.

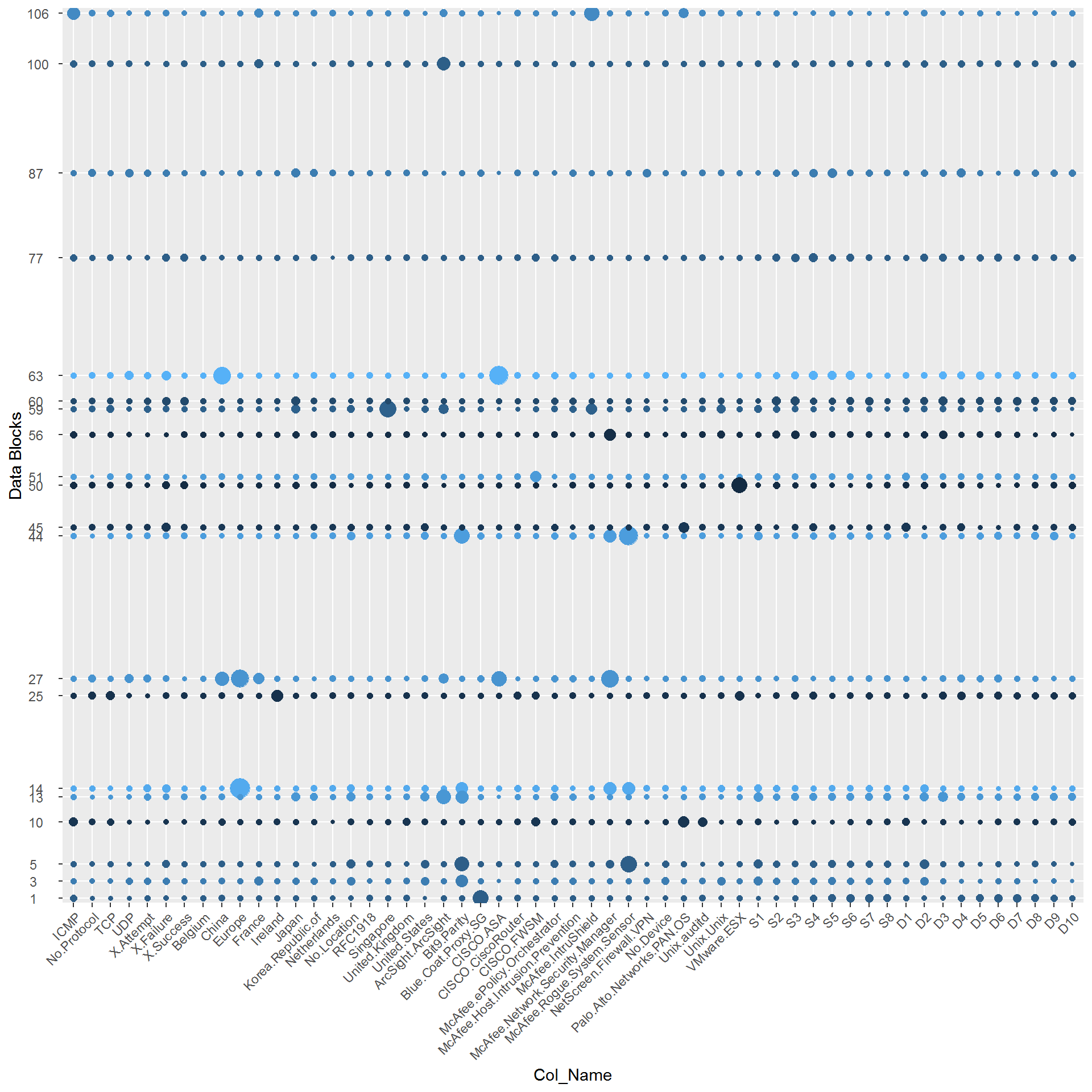


Figure 12. Histogram Matrix (Top 20 Outliers)

## Factor Analysis

The IFS levels are presented in Table 3. The rotated IFS level is higher than the original IFS level, serving as rationale for using the rotated factor loadings and scores in the subsequent analysis. According to Table 1, a value of 0.6125 is deemed as mediocre.

Table 3. IFS Results

|  |  |  |
| --- | --- | --- |
|  | **IFS Level** | **Evaluation** |
| Original | 0.5674 | Miserable |
| Rotated | 0.6125 | Mediocre |

The heatmap in Figure 13 shows the relationship between the columns of the reduced state vector to the rotated factor loadings.

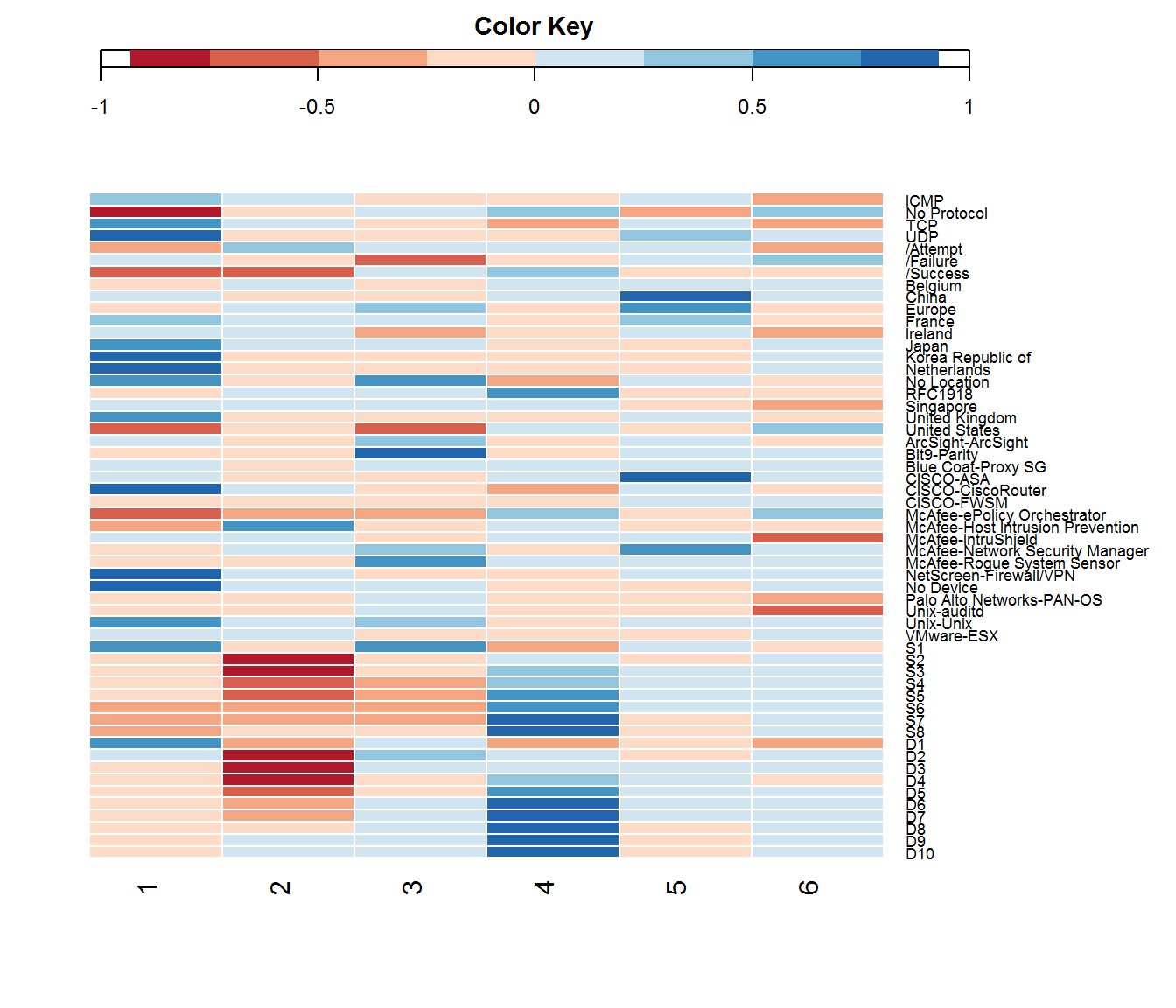


Figure 13. Heatmap of Rotated Factor Loadings

Table 4 clarifies the high and low variables by factor correlation for each factor. The factor loading breakdown can provide insight into the relationships between variables. For example, in factor 5 we see that the two devices, CISCO-ASA and McAfee-Network Security Manager are directly related to the geographic variables China and Europe. While the true relationship between these variables are unknown, we may presume that these devices are set up to capture signatures from those locations.

Table 4. Factor-Variable Breakdown



The rotated factor score plots are shown in Figure 14. Based on these plots, we can clearly see the outlier time blocks, such as blocks 27, 63 and 14.

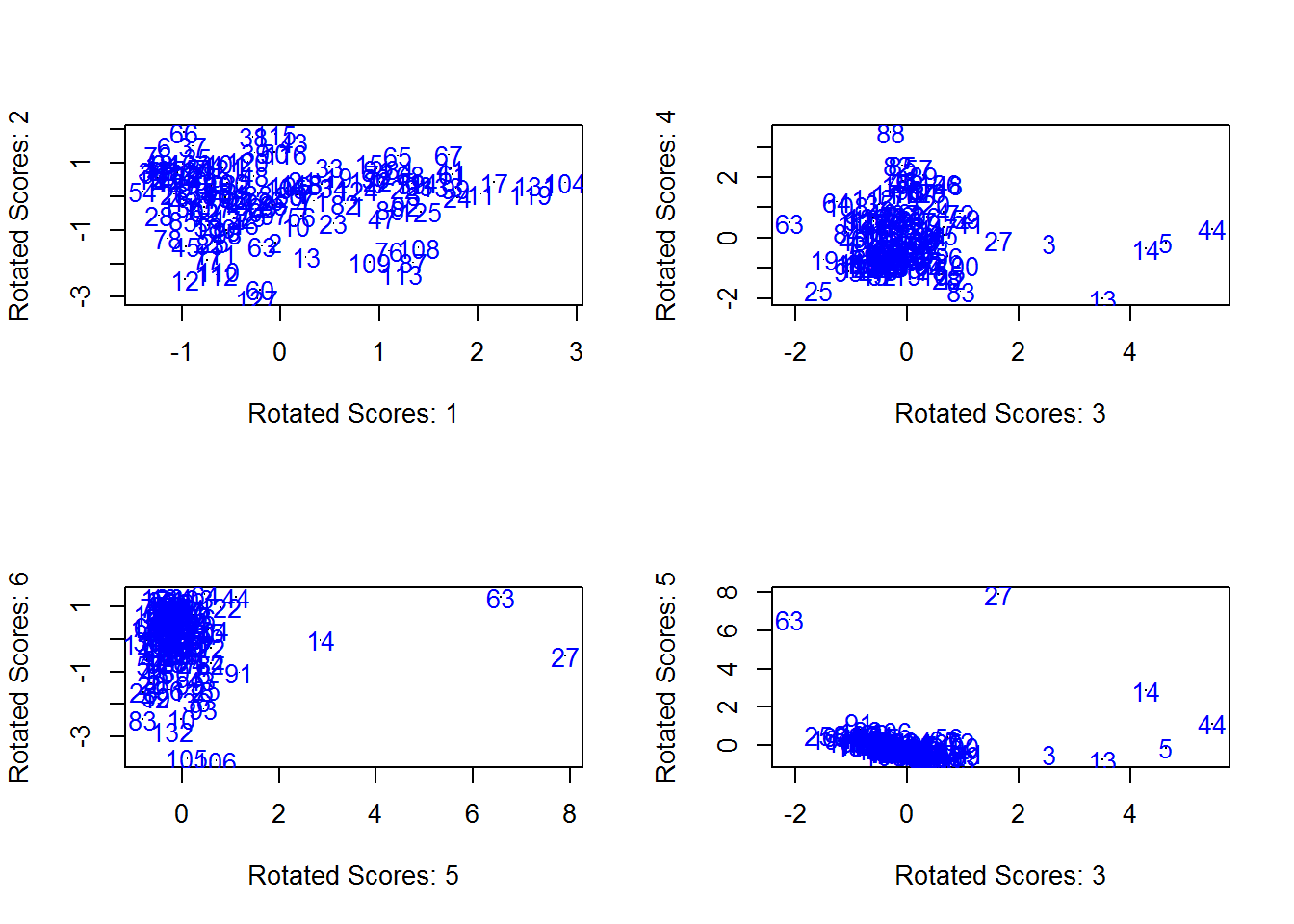


Figure 14. Rotated Factor Score Plots

## Example Investigation

Based on the analysis in Sections 5.1 and 5.2, block 14 is a clear outlier and is used as our sample point for further investigative analysis.

### Mahalanobis Column Breakdown

The histogram shown in Figure 15 displays the frequency of the top five columns with the largest *BD* value for block 14, relative to the values for neighboring blocks. The purpose is twofold: 1) observe the columns that cause the block to be an outlier and 2) take note of the temporal relationship between the top five columns and the blocks.

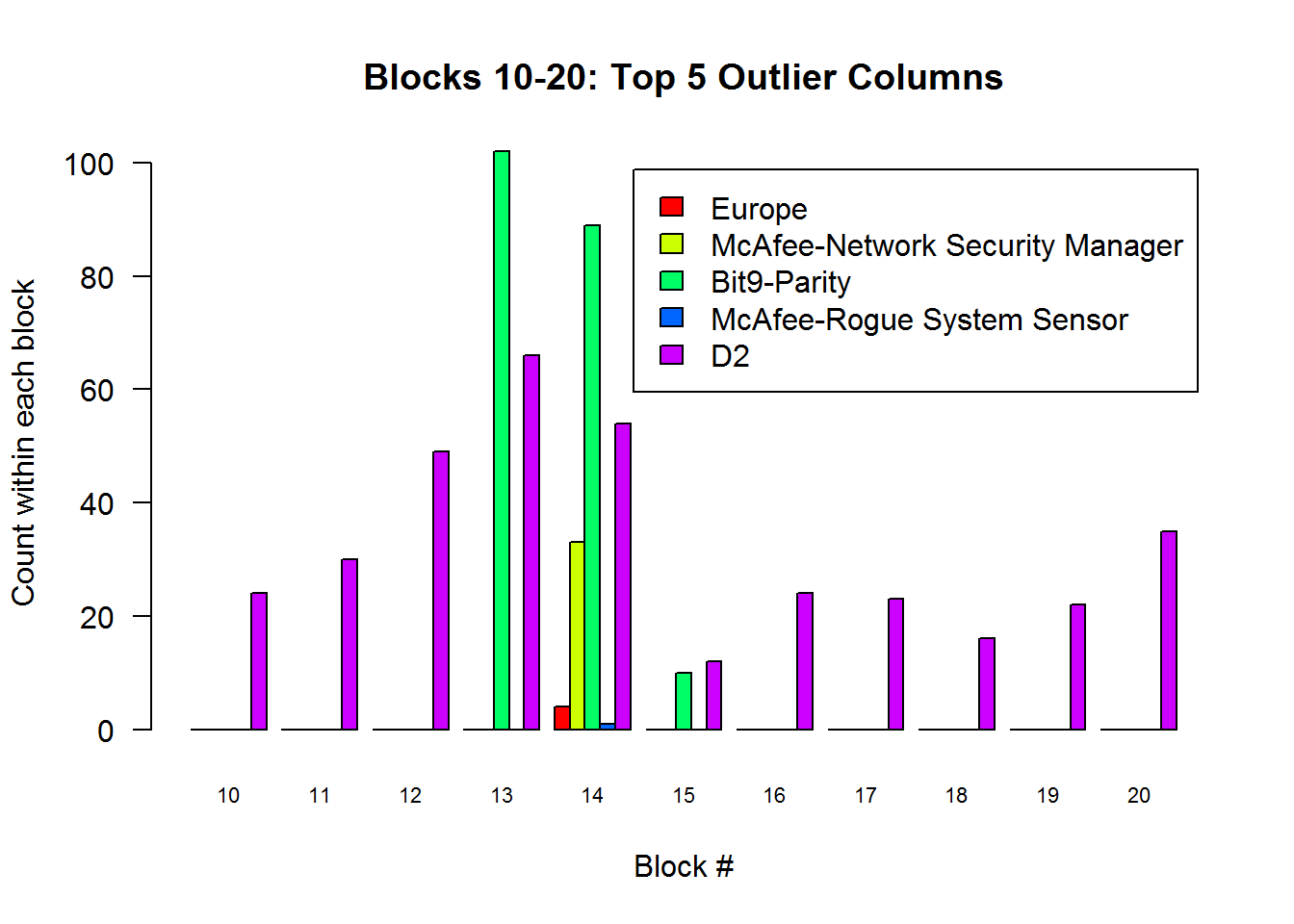


Figure 15. Block 14: Top 5 Breakdown Distance Columns

Based on Figure 15, we

### IP Network Graph

In Figure 16, we display the network graph for block 14. Taking a closer look at a couple of the larger clusters, where multiple nodes are connected to a single node, we can readily identify the destination IP addresses that look to be the target of multiple source IP addresses.

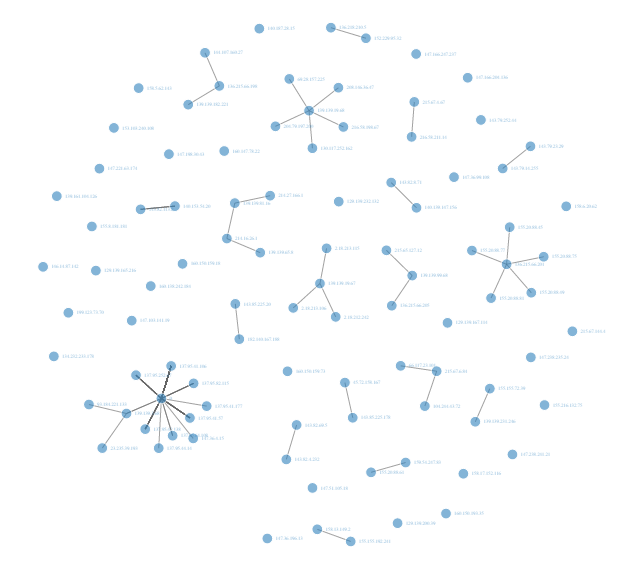


Figure 16. Block 14 IP Network Graph

## Shiny Web App

The above analysis has been conducted in a standard script file for use in R. However, to use this research in the big data platform for the army, the tools have been developed into a Shiny web application. Below is an image of the application homepage.

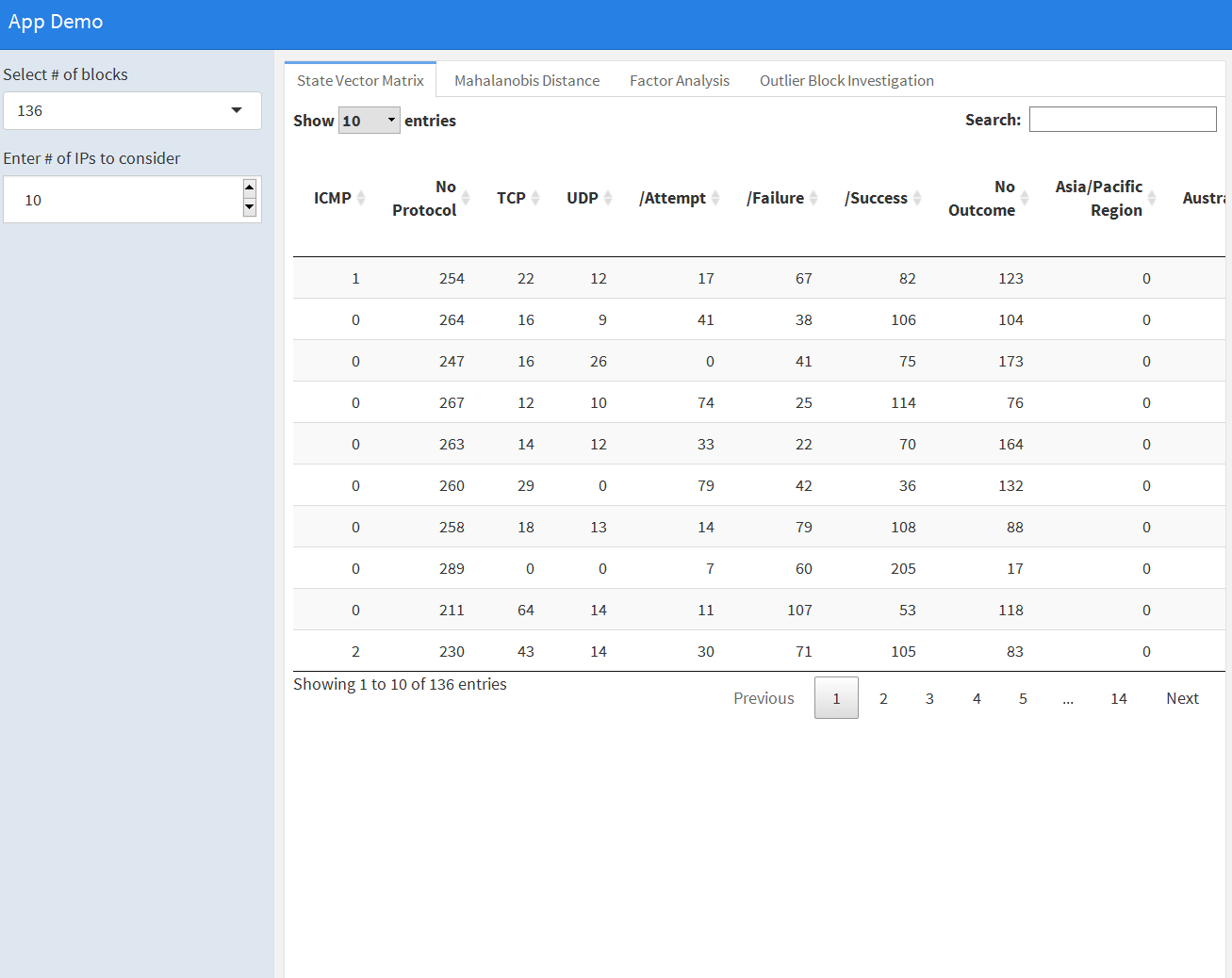


Figure 17. Shiny Web App

# VI. Conclusions and Recommendations



## Conclusions

Text

## Contributions and Future Research

This research contributes to the field of network intrusion detection by incorporating multivariate methods to narrow the search window for log file analysis.

For future research, we recommend three topics for further investigation. The first topic involves proper variable selection from the over 400 variables aggregated in the big data platform. In conjunction with the variable selection, a study should be conducted on the rotation methods used in factor analysis, specifically the comparison between oblique rotations and orthogonal rotations. The oblique rotation allows for factor correlation which is directly affected by the variables chosen for the analysis. Finally, we recommend exploration into the appropriate block size for time series analysis.

# Appendix A. Hardware/Software overview

The entirety of this research is conducted within the R statistical computing environment.

Look at Maj Walter’s thesis and create something similar

# References

|  |  |
| --- | --- |
| [1] | S. Tzu, The Art of War, New York: Oxford Univeristy, 1963. |
| [2] | A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur and J. Srivastava, "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection," SDM, 2003. |
| [3] | "Cybersecurity Resource Center: Cybersecurity Incidents," [Online]. Available: https://www.opm.gov/cybersecurity/cybersecurity-incidents/. [Accessed 27 September 2016]. |
| [4] | M. Zamani and M. Mahnush, "Machine Learning Techniques for Intrusion Detection," *arXiv preprint arXiv:1312.2177 ,* 2013. |
| [5] | G. Van Vleet, "2d Regional Cyber Center opens," 30 October 2013. [Online]. Available: https://www.army.mil/article/114105/. [Accessed 13 September 2016]. |
| [6] | C.-H. J. Wu and J. D. Irwin, Introduction to Computer Networks and Cybersecurity, Boca Raton: CRC Press/Taylor & Francis Group, LLC, 2013. |
| [7] | K. Scarfone and P. Mell, *Guide to Intrusion Detection and Prevention Systems (IDPS),* Gaithersburg, MD: NIST special publication, 2007. |
| [8] | "Symantec Attack Signatures," [Online]. Available: https://www.symantec.com/security\_response/attacksignatures/. [Accessed 12 September 2016]. |
| [9] | D. E. Denning, "An Intrusion Detection Model," *IEEE Transactions on Software Engineering,* vol. 2, pp. 222-232, 1987. |
| [10] | P. García-Teodoro, J. E. Díaz-Verdejo, G. Macía-Fernández and E. Vásquez, "Anomaly-based network intrustion detection: Techniques, systems and challenges," *Computers and Security,* vol. 28, no. 1, pp. 18-28, 2009. |
| [11] | R. W. Walter, *Methods to Address Extreme Class Imbalance in Machine Learning Based Network Intrusion Detection Systems,* Air Force Institute of Technology, 2016. |
| [12] | J. Zhang and M. Zulkernine, "Anomaly Based Network Intrusion Detection with Unsupervised Outlier Detection," in *IEEE International Conference on Communications*, 2006. |
| [13] | C. Abad, J. Taylor, C. Sengul, W. Yurcik, Y. Zhou and K. Rowe, "Log correlation for intrusion detection: A proof of concept," in *Computer Security Applications Conference*, 2003. |
| [14] | M. R. Grimalia, J. Myers, R. F. Mills and G. Peterson, "Design and Analysis of a Dynamically Configured Log-based Distributed Security Event Detection Methodology," in *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, 2011. |
| [15] | J. Breier and J. Branišová, "Anomaly Detection from Log Files Using Data Mining Techniques," in *Information Science and Applications*, 2015. |
| [16] | M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn and L. Chang, "A Novel Detection Scheme Based on Principal Component Classifier," Miami University Dept of Electrical and Computer Engineering, Coral Gables, 2003. |
| [17] | W. Wang and R. Battiti, "Identifying Intrusions in Computer Networks with Principal Component Analysis," in *First International Conference on Availability, Reliability and Security*, 2006. |
| [18] | R. Winding, T. Wright and M. Chapple, "System Anomaly Detection: Mining Firewall Logs," in *Securecomm and Workshops*, 2006. |
| [19] | P. C. Mahalanobis, "On the generalized distance in statistics," in *Proceedings of the National Institute of Sciences*, Calcutta, 1936. |
| [20] | H. F. Kaiser, "The Application of Electronic Computers to Factor Analysis," Educational and Psychological Measurement, 1960. |
| [21] | R. B. Cattell, "The scree test for the number of factors," *Multivariate behavioral research,* vol. 1, no. 2, pp. 245-276, 1966. |
| [22] | J. L. Horn, "A rationale and test for the number of factors in factor analysis," *Psychometrika,* vol. 30, no. 2, pp. 179-185, 1965. |
| [23] | C. Spearman, "The abilities of man," 1927. |
| [24] | K. W. Bauer, Class Lecture, OPER 685, Applied Multivariate Analysis. School of Engineering and Managment, Air Force Institute of Technology, Wright-Patterson AFB OH, Spring Quarter 2016. |
| [25] | J. W. Osborne and A. B. Costello, "Best Practices In Exploratory Factor Analysis: Four Recommendations For Getting The Most From Your Analysis," *Pan-Pacific Management Review,* vol. 12, no. 2, pp. 131-146, 2009. |
| [26] | H. F. Kaiser, "An index of factorial simplicity," *Psychometrika,* vol. 39, no. 1, pp. 31-36, 1974. |
| [27] | A. Frei and M. Rennhard, "Histogram Matrix: Log file visualization for anomaly detection," in *Third International Conference on Availability, Reliability and Security*, 2008. |
| [28] | R. Silhavy, R. Senkerik, Z. K. Oplatkova, P. Silhavy and Z. Prokopova, Software Engineering Perspectives and Application in Intelligent Systems: Proceedings of the 5th Computer Science On-line Conference 2016 (CSOC2016), vol. 2, Switzerland: Springer, 2016. |
| [29] | S. Hojsgaard, D. Edwards and S. Lauritzen, Graphical Models with R, New York: Springer Science & Business Media, 2012. |
| [30] | CISCO, "What Is a Firewall?," [Online]. Available: http://www.cisco.com/c/en/us/products/security/firewalls/what-is-a-firewall.html. [Accessed 12 July 2016]. |
| [31] | B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola and R. C. Williamson, "Estimating the Support of a High-Dimensional Distribution," *Neural Computation,* vol. 13, no. 7, pp. 1443-1471, 2001. |
| [32] | S. Ramaswamy, R. Rastogi and K. Shim, "Efficient Algorithms for Mining Outliers from Large Data Sets," *ACM SIGMOD Record,* vol. 29, no. 2, 2000. |
| [33] | M. D. Rich, "Evaluating Machine Learning Classifiers for Hybrid Network Intrusion Detection Systems," Wright-Patterson AFB, 2015. |
| [34] | E. M. Knox and R. T. Ng, "Algorithms for Mining Distance-Based Outliers in Large Datasets," in *Proceedings of the International Conference on Very Large Data Bases*, 1998. |
| [35] | D. M. Hawkins, Identification of Outliers, vol. 11, London: Chapman and Hall, 1980. |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| REPORT DOCUMENTATION PAGE | | | | | | | | Form Approved  OMB No. 0704-0188 | | |
| Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.** | | | | | | | | | | |
| 1. REPORT DATE *(DD-MM-YYYY)*  23-03-2016 | | 2. REPORT TYPE  Master’s Thesis | | | | | | 3. DATES COVERED *(From - To)*  Sept 2015 – March 2016 | | |
| 4. TITLE AND SUBTITLE | | | | | | | | 5a. CONTRACT NUMBER | | |
|  | | | | | | | | 5b. GRANT NUMBER | | |
|  | | | | | | | | 5c. PROGRAM ELEMENT NUMBER | | |
| 6. AUTHOR(S) | | | | | | | | 5d. PROJECT NUMBER | | |
| Gutierrez, Robert J, 2nd LT | | | | | | | | 5e. TASK NUMBER | | |
|  | | | | | | | | 5f. WORK UNIT NUMBER | | |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)  Air Force Institute of Technology  AND ADDRESS(ES) | | | | | | | | 8. PERFORMING ORGANIZATION REPORT  NUMBER | | |
| Graduate School of Engineering and Management (AFIT/EN)  2950 Hobson Way  Wright-Patterson AFB OH 45433-7765 | | | | | | |  | AFIT-ENS-17-M-131 | | |
| 9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) | | | | | | | | 10. SPONSOR/MONITOR’S ACRONYM(S) | | |
| Army Cyber Command | | | |  | | | | ARCYBER | | |
| LTC Cade Saie | | | |  | | | |  | | |
|  | | | |  | | | | 11. SPONSOR/MONITOR’S REPORT | | |
|  | | | |  | | | | NUMBER(S) | | |
| cade.m.saie.mil@mail.mil | | | |  | | | |  | | |
| 12. DISTRIBUTION / AVAILABILITY STATEMENT    Distribution Statement A. Approved for Public Release; Distribution Unlimited | | | | | | | | | | |
| 13. SUPPLEMENTARY NOTES  This work is declared a work of the U.S. Government and is not subject to copyright protection in the United States. | | | | | | | | | | |
| 14. ABSTRACT | | | | | | | | | | |
| 15. SUBJECT TERMS | | | | | | | | | | |
| 16. SECURITY CLASSIFICATION OF: | | | | | 17. LIMITATION  OF ABSTRACT | 18. NUMBER OF PAGES | | | 19a. NAME OF RESPONSIBLE PERSON  Dr. Kenneth W. Bauer, AFIT/ENS | |
| a. REPORT  U | b. ABSTRACT  U | | c. THIS PAGE  U | | UU | 41 | | | 19b. TELEPHONE NUMBER *(include area code)*  (937)255-6565 x4328  kenneth.bauer@afit.edu | |
|  | | | | | | | | | | Standard Form 298 (Rev. 8-98)  Prescribed by ANSI Std. Z39.18 |