CORRELATION AND COMPARATIVE ANALYSIS OF TRAFFIC ACROSS FIVE NETWORK TELESCOPES

Submitted in partial fulfilment
of the requirements of the degree of

Master of Science

of Rhodes University

Thizwilondi Moses Nkhumeleni

Grahamstown, South Africa
April 8, 2014

Abstract

Monitoring unused IP address space by using network telescopes provides a favourable environment for researchers to study and detect malware, worms, denial of service and scanning activities. Research in the field of network telescopes has progressed over the past decade resulting in the development of an increased number of overlapping datasets. Rhodes University's network of telescope sensors has continued to grow with additional network telescopes being brought online. At the time of writing, Rhodes University has a distributed network of five relatively small /24 network telescopes.

With five network telescope sensors, this research focuses on comparative and correlation analysis of traffic activity across the network of telescope sensors. To aid summarisation and visualisation techniques, time series' representing time-based traffic activity, are constructed.

By employing an iterative experimental process of captured traffic, two natural categories of the five network telescopes are presented. Using the cross- and auto-correlation methods of time series analysis, moderate correlation of traffic activity was achieved between telescope sensors in each category. Weak to moderate correlation was calculated when comparing category A and category B network telescopes' datasets. Results were significantly improved by studying TCP traffic separately. Moderate to strong correlation coefficients in each category were calculated when using TCP traffic only. UDP traffic analysis showed weaker correlation between sensors, however the uniformity of ICMP traffic showed correlation of traffic activity across all sensors. The results confirmed the visual observation of traffic relativity in telescope sensors within the same category and quantitatively analysed the correlation of network telescopes' traffic activity.

Acknowledgements

I would like to acknowledge a number of people who contributed to the success of this research. Firstly, I extend my sincere thanks to Prof. Barry Irwin for the opportunity to complete a Master's programme under his guidance. Thank you for not only supporting me with the datasets but also guiding the project throughout the entire research process. I wish to thank Mr John Richter for his assistance in setting up the working environment. I would like to thank Louella Hastie for her assistance in final proofreading of this document.

To the greater Rhodes University community and the Department of Computer Science thank you for supporting me through my tertiary studies. With regard to this project, I wish to thank the Department for their hospitality in the last few weeks of my research. Having been academically associated with Rhodes University has been a life-changing experience and most of my achievements can be accredited to the institution.

I thank my employer, Anglo American, for sponsoring the course and offering study leave for the required classes. Furthermore, I am grateful for all the support that I have received from the HR department over the last two years.

Personally, I thank my family and friends for all the support: for those who sparked thoughts, encouraged critique or helped to review my thesis. I would also like to thank my partner, Lebone Malele, for her encouragement and assisting with the initial proofreading of this project. Lastly, and most importantly, all glory and honour goes to God.

Contents

List of Figures

List of Tables

List of Code Listings

Chapter 1

Introduction

Research in the area of network telescopes has continued to gain prominence in the information security field. The reason for the growing interest in the field is partially attributed to the changing threat landscape caused by an increase in self-propagating malicious worms. The speed at which these worms are able to traverse the Internet is staggering and the damages they cause costs of their effects can run into billions of dollars. For example, the Code-Red worm infected 359,000 hosts in less than 14 hours and the cost of its impact was reported in excess of $2.6 billion in 2011 terms [35]. With the rapid and devastating impact of self-propagating malware, the need to monitor remote security events by using network telescopes as a first point of detection is well founded. Network telescopes have gained prominence and filled the void left by traditional intrusion detection systems since they allow researchers to monitor unused address space that contains no legitimate traffic. This means that monitored traffic is unwanted and potentially malicious. Furthermore, this proves to be advantageous as it means that researchers and information security experts are not required to overcome the challenge of distinguishing between legitimate and illegitimate traffic.

A network telescope is a monitoring system that passively captures inbound network traffic. In the past decade, by using network telescopes to monitor unexpected traffic flow, researchers were able to identify anomalies in network traffic such as denial-of-service attacks [34] and to study various malicious worm outbreaks [33] [45]. The network telescope's ability to detect random remote events is intrinsically limited by the lens size (monitored address space). Without access to larger network telescopes, it has been proven that relatively small telescope sensors are capable of observing malicious events [14]. By having number of distributed telescopes, researchers are able to capture multiple IP address spaces, consequently increasing the total space monitored. With distinct

distributed network telescopes, the question of similarity in traffic activity becomes pertinent. Similarities of network traffic monitored from distinct network telescopes is a question that has been explored by Rhodes University researchers [18].

1.1 Problem Statement

Organisations such as CAIDA[1] (Cooperative Association for Internet Data Analysis), who have multiple distributed network telescopes to monitor traffic anomalies on a global scale, have contributed to the growing body of network telescope research. Since 2005 in South Africa, Rhodes University has been collecting data for analysis on relatively small (/24[2]) network telescopes [18]. In 2009, another network telescope with a /24 address space was launched at Rhodes. Subsequent to this, three additional /24 telescopes have been launched adding to a distributed network of five telescopes thereby providing multiple sensors for analysis.

With multiple sensors available, researchers have access to increased volumes of concurrent data points or datasets. Increased concurrent datasets decreases the time to detect remote network events [31]. Furthermore, the increased datasets also improve the network telescope's resolution or ability to detect events. This is due to the inherent limitation in network telescope research given that researchers are only able to monitor a portion of the entire address space.

IP address space is increasingly becoming scarce as more devices demand allocation of IP addresses [16]. This imposes a limitation of having large multiple /8 network telescopes, with extensive amount of IP addresses, used purely for monitoring Internet background radiation. In this regard, using smaller distributed network telescopes with large logical distance[3] [20] provides researchers with an increased capability without necessarily requiring access to larger network telescopes.

When using distributed network telescopes to study background radiation, it is important to understand the relativity of traffic activity across the network of telescopes as well as the underlying variables responsible for this activity. Researchers have used various visualisation techniques to show the relation of network telescope activity across multiple network telescope sensors without quantitatively assessing a correlation between different

[1] http://www.caida.org
[2] IP addresses that share the first 24 bits in common and often referred to as 'Class C' network [28].
[3] The logical distance is the numeric difference between two IP addresses represented numerically.

network telescope sensors. It is challenging to quantify the degree of traffic relativity between telescopes without numeric correlation values.

1.2 Research Objectives and Goals

With the current extended network of telescopes, the objectives of the project are to: (1) comparatively analyse traffic observed across five distinct network telescope sensors and (2) investigate the correlation of traffic activity across the different network telescopes' sensors. By determining the correlation across multiple network telescope sensors, the research outcomes would provide greater confidence in forecasting or modelling typical network background radiation.

Therefore the two overarching goals extracted from the objectives above are:

- To comparatively analyse similarities of traffic observed across the network telescope sensors:
  - Determine the degree to which one telescope sensor's traffic is similar to another.
  - Investigate the relationship (if any) and gather insights as to what causes fluctuations or deviations to the relationship between the network telescopes' traffic.
  - Comparatively analyse the different traffic types and their contributions to fluctuations or changes in relativity of traffic.
- To quantitatively analyse the correlation of periodic traffic activity across all sensors:
  - Apart from graphical techniques as representation of traffic activity, assess quantitatively, the correlation of traffic activity across five network telescopes combination.
  - Quantitatively examine whether or not there are repeated periodic patterns of traffic in order to determine if traffic observed over a period is uniform or has repeated cycles.

1.3 Scope

The primary focus of the research project is on network telescope traffic analysis. In this regard, the researcher abstracts away from underlying mechanisms of building network telescopes. The approach is suitable since Rhodes University and other institutions have previously conducted extensive research in setting up and configuring network telescopes' architecture [9] [18] [31].

The research project focuses on five datasets, which have been obtained from respective distinct network telescope sensors on Rhodes University's network. The five datasets were generated from IPv4 network telescopes. Traditional network telescopes, similar to the ones under investigation, are passive since they do not respond to requests to establish connections and simply capture traffic arriving at the network. As no connections are established, the research only focuses on packet header information to infer remote activity without raw packet analysis. Furthermore, packet header information has already been processed into Rhodes University's relational database infrastructure. To this end, the hurdle of processing raw data from telescopes to relational databases is not included in the scope of the project.

1.4 Methodology

To achieve the goals discussed above, the primary methodology will be to conduct incremental and iterative experimental work, analysing datasets and discussing observed trends that are being monitored. The research project will be quantitative and focus on categorising and synthesizing large volumes of data. The methodology adopted will be to extract the underlying variables in order to investigate and understand the dominant variables that are responsible for fluctuations in correlation analysis.

The first approach for the research project will be to fully analyse and understand the five datasets that are available; this includes building a profile of each dataset to understand the outlay of the data. The second approach will examine the datasets' characteristics and make use of summarisation and basic statistics to establish whether there is relativity across the five network telescopes.

Having studied the relativity of traffic using summarisation, the research will proceed to conduct a thorough correlation analysis of the traffic generated by the datasets. Statistical

correlation analysis will require the construction of variables that will model the traffic observed over a period.

1.5 Document Structure

The remaining chapters of this document have been structured as follows:

- **Chapter 2:** highlights the background information relevant to the area of network telescope area. The chapter surveys and analyses existing research on traffic analysis. In related work, the chapter indicates the gaps that previous research projects have not adequately addressed as related to this project's objectives. The chapter discusses the use of distributed network telescopes, summarisation and basic traffic statistical analysis techniques and advanced correlation techniques as part of a time series analysis.
- **Chapter 3:** briefly discusses the network telescope infrastructure with a concise description of each network telescope node used on the research project. An overview of each dataset is discussed. The tools used to conduct the research and the advantages of the platforms selected are also examined.
- **Chapter 4:** focuses on a comparative analysis conducted of the five network telescopes datasets. The chapter implements summarisation and basic statistical techniques as part of traffic analysis.
- **Chapter 5:** examines the implementation of advanced correlation techniques in comparing the traffic activity of network telescopes.
- **Chapter 6:** concludes the research project and revisits the project's objectives. In concluding, the chapter also discusses potential future work and extensions to this project.

Chapter 2

Literature Survey

Network telescopes occupy a range of unused network address space (so called darknets since, seemingly, there is nothing within these networks) [11]. Since the address space is unused, normal traffic is removed therefore making all traffic captured unwanted, potentially harmful or simply malicious.

Network telescopes assist information security researchers in providing an early warning system for worms, denial-of-service and various malware activities [31]. Due to the speed at which worm activities can infect and propagate through the greater network, having a network telescope to monitor unusual traffic is useful as a first point of detection. For example, an analysis of network telescope traffic allows researchers and security experts to understand the environment and consequently develop software that can adapt to the environment.

Network telescopes generate large quantities of data. For example, between 2005 and 2009 a relatively small /24[1] network telescope captured over 40 million packets of data [18]. In order to analyse these large datasets meaningfully, there are a number of techniques that researchers use such as summarisation, correlation analysis and visualisation.

Since the traffic being monitored is captured with a time stamp, time series' can be generated to model traffic activity over a period of time. As more overlapping network telescope datasets are being collected, the correlation analysis of traffic activity across these datasets requires the use of more advanced statistical analysis of time series' such as the cross-correlation and auto-correlation methods.

[1] IPv4 address space with 2^8(256) IP addresses

This chapter begins with discussion of the background of network telescopes including examples of work in analysing malicious activity. The discussion proceeds, providing a motivation for the use of network telescopes and the significance of having distributed network telescope sensors. Summarisation and correlation are examined as well as related research around traffic relativity. Building on basic statistical analysis through summarisation, a discussion around the use of time series' in conducting correlation analysis is provided. The chapter concludes by discussing the limitations of passive monitoring.

2.1 Network Telescope Background

Network telescopes are used to monitor traffic across unused IP address space. Since the IP address space of the telescope sensors are unallocated, no legitimate traffic should be expected [31]. Considering that network telescopes monitor unexpected traffic, they provide a useful mechanism to observe remote security events from malicious activities such as worms and denial of service attacks.

Network telescopes are useful for providing early warning and detection of security events. For example, by monitoring illegitimate traffic and using anomaly detection methods, an early detection system for worms can be developed. In this instance, to monitor illegitimate traffic, a Kalman filter[2] model was used to identify the presence of worms in their early stages [55].

Research using network telescopes is also useful in studying the propagation of malicious worms. For example, a network telescope was used to understand how the Slammer worm managed to achieve rapid growth [33]. Slammer, a computer worm that exploited buffer overflows of Microsoft SQL servers, managed to reach 90% of vulnerable hosts within ten minutes [33]. The speed of scanning rates of Slammer was estimated using network telescopes. Additionally, using network telescopes, researchers were able to detect several of Slammer's flaws with the random number generator that limited the worm's potential impact.

Telescopes can also be used to study denial-of-service attacks. Research has been conducted to infer information about the prevalence of denial-of-service attacks using backscatter analysis . By using a /8 network telescope, researchers were able to observe more than 12,000 denial-of-service attacks in over 5,000 distinct targets [34].

[2]Kalman filter detects the presence of worms by detecting the trend of traffic

Aside from the examples listed above, network telescopes have been used to study the spread of the Code-Red worm [35]. The Code-Red worm exploited a buffer overflow of Microsoft's IIS web servers and caused a widespread outbreak in 2011. Similarly, using network telescopes, researchers studied the Witty worm, which affected the buffer overflow vulnerability of several Internet Security System products such as RealSecure and BlackICE [45]. The detection of malicious remote activity is possible because network telescopes offer researchers the capability to overcome the challenge of collecting global information about worms and provide a significant amount of traffic activity information. For example, a /8 telescope, which the Cooperative Association for Internet Data Analysis (CAIDA) operates, effectively contains 1/256 of all IPv4 addresses [45]. This means that, if a worm propagation is random and unbiased, CAIDA's network telescope would receive roughly one out of every 256 packets sent.

Network telescopes are passive, they do not complete a 3-way TCP handshake to establish a connection and therefore cannot receive TCP payloads. Network telescopes can consequently be used to monitor activity of Internet background radiation. Traffic arriving at a network telescope can be classified in one of the following three categories [18]:

- *Backscatter* - traffic generated from IP addresses within the telescope's range being used for spoofing elsewhere. For example, traffic (i.e. TCP handshake messages) generated by a host responding to a spoofed host wanting to establish a TCP connection.
- *Misconfiguration* - traffic likely emanating from misconfiguration of hosts.
- *Aggressive* - potentially hostile traffic generated through scanning activity, worms and malware.

2.2 Motivation for the Use of Network Telescopes

Research in network telescope traffic has progressed over the decade since the initial work produced by CAIDA in 2002 [31]. In recent years, network telescope research has increased in its prominence among researchers. The increased focus on network telescope traffic is due to the increase in botnet[3]-related activities such as the Conficker worm [4]. Apart from the ability of network telescopes to gather global information about worms

[3]Compromised machines used to send out spam email messages, spread viruses, attack computers and servers - http://www.microsoft.com/en-gb/security/resources/botnet-whatis.aspx

discussed in Section 2.1, there are a number of reasons for the increased interest displayed by security researchers:

- Using production network traffic requires researchers to separate illegitimate traffic from useful and live production traffic [18]. Since network telescope data is unsolicited, researchers can assume that such traffic is illegitimate and is therefore potentially harmful. Using production data could also have negative impact on the functioning of organisations. Furthermore, it is cumbersome to gather information at a global scale using production data.

- An increase in computer processing power aids researchers in processing and classifying rising volumes of data. For example, Rhodes University researchers developed a GPU accelerated packet classification tool to perform fast and accurate classification of packets [36]. The design of a fast classifier was possible due to an accelerated graphic processing unit as well as the leveraging of parallel processing capability available on GPU's which in turn was optimised for parallel classification of packets.

- Added to the increasing processing power is the ability to make sense of large datasets through various visualisation techniques. With visualisation it is easier to view patterns and trends from network packet activity. The InetVis (Internet Visualisation) tool was designed to perform various visual analyses of the network telescope traffic [50]. The InetVis system allowed researchers to use three dimensional plots to display scanning activities thus providing additional insights into scanning activity across different types of traffic. Additionally, researchers at Rhodes University developed a tool to map large IP traffic by using Hilbert Curve fractal mapping [17]. The tool aided analysis of traffic as it could compare data from multiple network telescopes. This is partially achieved by the tool's ability to show sequential relationship between nodes on the produced plot.

2.3 Network Telescope's Size

Before discussing the significance of a network telescope's size in detecting events, addressing terminology is introduced. Internet Protocol version 4 (IPv4) addressing scheme caters for a 32 bit address to identify a host on the network [28]. Therefore an IPv4 address scheme has a potential pool of 2^{32} addresses. Furthermore, originally addresses

were categorised into classes to describe the size of networks. A 'Class A' network has the first 8 bits used by the Network ID and the remaining 24 available for hosts (2^{24} different addresses). Similarly, a 'Class B' network has the first 16 bits occupied by the Network ID and 16 bits remaining for hosts (2^{16} different addresses). The 'Class C' network has the first 24 bits occupied by the Network ID and the remaining 8 bits available for hosts (2^{8} different addresses). 'Class A', 'Class B' and 'Class C' can also be written as /8, /16 or /24 respectively. The latter notation is preferred throughout the paper since it allows one to discern quickly the number of bits used by the Network ID.

Network telescope size (i.e. the address range of the telescope or lens size) is important in the telescope's ability to observe or detect network events. Telescope size will affect the telescope's accuracy and the speed at which it can observe events. There is a relationship between the network telescope's size and its ability to detect events accurately and rapidly. It has been shown that a /1 network telescope is more than twice as good as a /2 given that a /2 takes 2.41 times longer to detect a packet at the same confidence level [31]. The results also showed that the relationship does not scale linearly. In summary, the results can be attributed to the impact of higher resolution that larger network telescopes are able to offer.

Nevertheless, even when using a relatively small /24 network telescope, researchers are still able to observe malware activity as well as perform classification of malware generated traffic [14]. Section 2.2 provides a few examples of research conducted at Rhodes University, which has been primarily based on relatively small /24 network telescopes.

To illustrate the probability of /8 and /24 network telescopes' ability to observe network events, the application of probability theory is considered. If /y represents the network telescope size, then the probability of monitoring a target, which is chosen randomly by a host, is generally given by [31]:

$$p_{(y)} = \frac{1}{2^y} \tag{2.1}$$

By applying probability theory to different network sensors' size, one is able to see that /24 telescope would have a lower probability of observing events relative to a larger /8 network telescope sensor.

$$p_{(8)} = \frac{1}{2^8} = \frac{1}{256} \tag{2.2}$$

$$p_{(24)} = \frac{1}{2^{24}} = \frac{1}{16,777,216} \tag{2.3}$$

It is also important to note that larger network telescopes have a higher probability of detecting events since that an event will occur for a limited time period [31]. The duration of an event therefore plays an important role in terms of whether or not a telescope can detect an event.

2.4 A Case for Distributed Network Telescopes

Network telescopes are limited by their lens size. A reason for this is that a network telescope only monitoring a section of the entire network address range. Without access to a large /8 network telescope, a distributed architecture with multiple sensors improves the ability of smaller network telescopes to detect events quickly.

Multiple distributed network telescopes allow researchers to capture multiple address blocks on the entire address range thereby increasing total address space being monitored. The benefit derived from this is that, by distributing /24 network telescopes, there is a decrease in detection time rate and an improvement in the network telescope's resolution (i.e. ability to detect events) [31]. Additionally, in a case where the network telescope's address blocks are logically spread widely, researchers are also able to minimize targeting bias in monitoring events. There are differences in the distributed model adopted for sensors. In one extreme, distributed network telescopes can take form of a few large contiguous telescope sensors. For example, relatively large /16 telescope sensors were used to develop iSink (Internet sink) by researchers at the University of Wisconsin [54]. In another extreme, distributed network telescopes can take the form of large peer-to-peer networks covering a wide area, however the network telescopes only monitor small to individual IP addresses [43]. Lastly, by distributing network telescopes, resources are spread across different telescopes thereby increasing the network capacity in monitoring events better.

Notwithstanding the advantages of using distributed network telescopes, there are correlation challenges introduced when conducting a comparative analysis of distributed network telescope traffic. Furthermore, although there are an increased number of network telescopes capturing datasets, the anonymisation of sensors is another challenge posed to researchers. These challenges are discussed in detail below.

2.4.1 Correlation Challenges

Using distributed network telescopes introduces an additional challenge of ensuring that the timing of network telescopes' clocks are synchronised. The discussion of correlation becomes important particularly when trying to pinpoint when an event started. The reason for this is that the first packet of an event that is detected by telescope is not necessarily the first packet emitted by the malicious host. The case is the same for the last packet. Using probability statistics, it has been shown that a /8 network telescope has 99% confidence that events happened no earlier than two minutes before the first observation is made in the telescope [31]. It is important to appreciate that the results achieved on a /8 network telescope would not directly translate to a /24 network telescope. In this regard, the question of the correlation of events on smaller telescopes becomes more pertinent.

2.4.2 Sharing Network Telescope Data

The sharing of a telescope's data that is collected with various sensors can prove to be a challenge given the sensitivity of ensuring that portions of the data are preserved. For example, organisations, like CAIDA, have put in place IP anonimisations techniques such as prefix-preserving address anonymisation to protect personally identifying information [32]. The reason behind this prefix-preservation process, apart from political and economic constraints, is simply to protect the integrity of the network telescope in order to avoid potential directed attacks, which would compromise the integrity of the sensor.

2.5 Summarisation and Correlation Analysis

In an effort to manage and understand the immense amount of data that network telescopes generate, researchers use various methods to classify and perform data analysis. There are two main pathways for the analysis of data: summarisation can be used in developing or augmenting concepts and correlation for enhancing understanding and discovering of relations in data [30]. The main goal is to categorise data in a manner that would allow humans to infer knowledge or reach conclusions. Apart from summarisation and correlation, visualisation is simply a technique of presenting results clearly and meaningfully.

2.5.1 Summarisation

Summarisation is responsible for finding patterns through basic techniques such as totals, centrality and the spread of data [30]. For centrality, measures such as mean, median and mode are employed. While, for the spread, the variance and standard deviation can be used. A brief description of these concepts is available in Appendix A. Apart from the basic summarisation techniques, there are more advanced methods that can be used. An example of this is cluster analysis; where data points that are similar to each other are grouped into clusters [53].

2.5.2 Understanding Correlation Analysis

Two factors are correlated when a co-occurrence of a pattern in the values is observed [30]. For example, a multiple of one variable relative to another. Therefore, correlation can be used for both prediction or modelling. Correlation studies of network telescope traffic activity will be useful in predicting network traffic layout in other non-captured network address ranges. Furthermore, correlation studies will allow researchers to have a deeper understanding of background radiation activity (ebbs and flows) and be able to infer greater knowledge.

2.6 A Case for Correlation in Network Telescope Traffic

The correlation of network traffic activity between multiple sensors has been of interest to researchers at Rhodes University, particularly where the network telescope's IP address assignments are not adjacent to each other [20]. In other words, where network telescopes have a logical distance between each other. At Rhodes University, detailed research in this area has not been possible in the past due to the lack of overlapping datasets from different sensors [18]. However this constraint was lifted with the introduction of new telescopes resulting in a total of five network telescopes. The datasets used in this research, collected from the five network telescopes, show the following positive characteristics:

- **Consistent address range size:** the network telescope lens (i.e. the size of IP address range of the network telescope) are all similar /24 networks. The /24 networks have 8 bits available for IP addresses for hosts within the network and therefore monitor 256 IP addresses. This is relatively small compared to a /8 network telescope

operated by CAIDA with 24 bits available for hosts' IP addresses within the network. Therefore, with 24 bits available, the CAIDA network telescope monitors 16,777,216 IP addresses.

- **Moderate logical distance:** the telescopes are not logically located directly adjacent to each other. This allows one to avoid location bias of attacks.
- **Substantial dataset overlap period:** all five network telescopes have been gathering data since May 2011. Since the project's initiation, over a year's worth of datasets were captured.
- **Multiple datasets:** the five datasets allow for multiple comparators with which experiments can be conducted.

2.7 Related Research - Sensor Traffic Relativity

As previously highlighted, comparative analysis for earlier studies was limited by the number of available overlapping datasets. However, recently a paper was published looking at five network telescope datasets [16]. This section explores the work done at Rhodes University and also looks at other research on metrics and basic statistical analysis.

2.7.1 Monitoring Malicious Activity Across Five Sensors

Researchers have recently looked at malicious activity across five distinct network telescopes [16]. The study, containing a 15-month period of observations, demonstrated both the value of having a distributed network of telescopes sensors and the advantage of analysing Internet background radiation across five network telescopes. Although no detailed statistical analysis was conducted (detail analysis was beyond the scope of the paper), graphical results of the study highlighted similarities of the traffic observed across the network telescopes. Research foregrounds the efficient use of combining multiple network telescopes in performing Internet background radiation analysis, thereby saving valuable IP address space.

2.7.2 Basic Statistical Analysis and Metrics

Due to the sheer size of data that is captured by network telescopes, researchers have shown that data summarisation techniques can be used to reduce data into more consumable sections by deriving basic statistical properties such as averages, medians and deviations [7]. Using heuristics and summarisation, researchers were able to detect anomalous activities and also identified a number of malicious attacks such as Conficker and distributed denial of service. Identification of malicious activity was conducted by monitoring the following: the rapid growth in packet count, which showed denial-of-service attacks; changes in packet count rations of top ports, which highlighted presence of Conficker; and variations in packet sizes, which signified the presence of W32.Rinbot[4] [27].

Apart from statistical analysis of network telescopes' data, researchers have also implemented statistical clustering in large networks to aggregate activity types of machines in determining anomalous activity [25]. Machines were clustered into activity groups based on similarities between their activity profiles. Although researchers defined attacks broadly to include information gathering exercises, by using 993 machines capturing data over a month, a total of 27 source IP were determined as attackers.

Metrics allows researchers to derive a consistent way to compare datasets and profile network telescopes. Research conducted has shown that it is possible to use standardised metrics to compare datasets from different network telescopes thereby enhancing information sharing among researchers [19]. The metric-based approach allows researchers to use metrics to conduct analysis without necessarily sharing source data. The research conducted proposed that network telescope metrics could be broadly categorised in two classes:

- Sensor Metrics: Configuration details of network telescope such as lens size, operation mode (passive or with interaction with incoming packets) and meta-data
- Dataset Metrics: Overview of the dataset drawing common traffic metrics (Top host/network observed, destination ports, source ports and protocols)

To the best of the researcher's knowledge, most of the work done on the statistical analysis of traffic activity is mainly focused on single network telescopes. This can be partially attributed to challenges in sharing datasets amongst researchers as there is a need to conceal or protect personally identifying information to prevent poisoning the sensor.

[4] W32.Rinbot is a worm which exploited Windows Server Service vulnerability which allowed remote code execution without authentication

2.7.3 Handling Statistical Outliers

To conclude this section on related work, an overview of associated research on the handling of statistical outliers is provided. This is essential as network telescope traffic will be susceptible to extreme values. For example, a form of packet flooding as part of a targeted denial of service would affect the packet count estimates (statistic). When conducting a statistical analysis of data it is important to check for outliers (points of data far outside the norm of a figure) since they can lead to distortions of estimates such as mean, standard deviation and variance [37]. The debate around whether or not to remove outliers is beyond the scope of this paper; however, outliers can have adverse effects on correlation measurements. To demonstrate the unfavourable effects of these outliers on correlation, researchers have shown [37] (using a population of 23396 subjects with both weak and strong correlated variables ($r_1 = -0.06$) and ($r_2 = -0.46$)) that cleaned correlation were more accurate (closer to known population correlation). In the experiment under discussion, researchers use a common method of three or more standard deviations from the mean to clean the dataset, so values outside three or more standard deviations were treated as outliers.

2.8 Time Series Analysis

Time series analysis is useful in understanding past events as well as predicting the future [8]. In this regard, the application of most time series analysis is aimed at understanding correlation and fitting models to data in an attempt to forecast future values or generate simulations. Various metrics in economics, science and engineering generally observe data (variables) such as: share price movements, price of commodities, Gross Domestic Product (GDP) per year, exchange rates and weather patterns over a period of time. Similarly, network telescope traffic metrics also measure variables over a period of time such as the daily/hourly packet count. Having variables measured sequentially over a fixed time period allows for the creation of time series [8]. Time series analysis therefore looks at trends, seasonal variations and, most significantly, the correlation (relatedness) of variables over a period of time.

2.9 Advanced Statistical Analysis using Time Series

Section 2.7 explored work around the use of basic statistics and metrics in network traffic analysis. This section develops the discussion of using statistics in traffic analysis further by focusing on more advanced correlation statistical methods using time series'. Although long-range and cross-correlation methods (to be introduced in this section) have been implemented in other fields, a literature survey show little implementation in the field of traffic analysis. This section will introduce the terminology and, in a case where a method has been implemented to analyse repeating patterns, it also discusses the work that was conducted.

To show long-range correlation of a time series, the auto-correlation function method can be used [22]. Long-range correlation analysis tests for correlation of a time series with itself in two different time lags. This is done so that a test of whether or not a series contains repeating patterns might be conducted. In order to compute the correlation of two time series the cross-correlation method is used [8].

2.9.1 Auto-correlation Function

An auto-correlation function calculates the correlation of a series with its own values at different lagged times [8]. With a specified maximum lag, the auto-correlation function would calculate the correlation value of the time series at each consecutive lag until the maximum lag is reached [49]. Correlation results would indicate whether there are periodic patterns in different time lags.

2.9.2 Cross-correlation Function

The simple cross-correlation method is used to quantify the relationship between time series variables. Time series variables may be correlated serially or correlated with a different time lag [8]. The challenge with using the simple cross-correlation method is that it does not cater for shifted series based on time lags. For instance, two time series can be correlated with a lag between them. This is pertinent to network telescope research since the lag could be due to a number of factors relating to network delays of events which are based on the logical or physical distance of the network telescope. In this regard a sample-shifted cross-correlation method is considered.

2.9.3 Related Work using Cross-correlation Method

Understanding the types of information threats and attacks is important for modelling traffic distributions, which could be used to configure information security infrastructure. A related body of work on network traffic correlation analysis makes use of the cross-correlation method to construct a joint distribution of traffic models for dependent and non-identical distributed traffic flows [52]. Researchers observed that due to dependency, a joint distribution platform is best used to model the traffic.

2.10 Limitations with Passive Monitoring

In closing the literature survey, the limitations of passive network telescopes in analysing malicious traffic are discussed.

The first limitation with using passive telescope sensors is that traffic is captured passively as discussed in Section 2.1. Since network telescopes simply capture packet information and drop the packet without establishing a TCP connection, the implementation of passive monitoring does not consider the payload information of packets [16]. With this constraint, researchers are unable to conduct payload analysis. However, with a significant amount of traffic information researchers can still infer knowledge about malicious attacks, such as denial of service, with a high likelihood [34].

Another challenge is the deployment of IPv6. Although the deployment of the IPv6 Internet Protocol has been gradual, network telescope analysis will be affected by the uptake of IPv6 protocol. The reason for this is that IPv6 has larger size of IP addresses relative to IPv4 IP addresses and therefore it makes worm scans less effective or simply infeasible [6]. With a typical /24 network, scanning tools only need to scan 256 addresses to reveal vulnerabilities. On a conservative one second per scan, this would essentially take less than five minutes. A typical IPv6 network will have 64 bits reserved for a host address. Applying the conservative assumption above, this would mean that it would take five billion years to complete a scan. Research on previously deployed IPv6 network telescope sensors showed no traffic observed [12] [16].

Although the large amounts of traffic generated by network telescopes are not a limitation, care should be taken when analysing large datasets. If large time intervals are used, it can be difficult to identify incidents due to dilution and interplay of other incidents [7]. Conversely, tiny time intervals require greater processing in order to observe incidents.

2.11 Summary

In this chapter, a literature review of network telescopes was conducted. At the outset, the literature survey focused on basic concepts to introduce the area of network telescopes. Beyond the introduction and explanation of terminology employed in the chapter, a number of examples of the implementation of network telescope analyses were provided. Network telescopes have been useful in monitoring and studying remote network events such as worms; including Conficker, Slammer, Code-Red, and Witty Worm. The literature survey also demonstrated that network telescopes are useful in providing an early warning system of malicious activity. By employing network telescopes, researchers are also able to infer knowledge about denial-of-service attacks through conducting backscatter analysis.

Additional motivation for the use of network telescopes was also presented. The literature review highlighted the advantages of using network telescopes since they only monitor unexpected or illegitimate traffic and researchers do not have to separate legitimate traffic from the datasets. Increased processing power allows for the processing of larger datasets as well as enabling researchers to use visualisation techniques to present results accurately and meaningfully.

The size of a network telescope is important in the sensor's ability to detecting events. Using probability theory, the chapter highlights the difference between a /8 and /24's probability of observing events. A motivation for using distributed network telescope sensors as means to improve smaller sensors' ability to detect events is provided. Another advantage of using distributed network telescopes is that researchers are able to avoid targeting bias of events by distributing sensors widely in their logical location.

The chapter highlights summarisation techniques and correlation analysis as the two main pathways of data analysis. A survey of work around basic statistical analysis and the use of metrics was also provided. The researcher concludes that, even though there has been previous work around metrics and basic statistical analysis, less work on correlation of multiple sensors has been conducted mainly due to the lack of overlapping datasets in the past.

As well as basic statistical analysis and the use of metrics, the chapter also introduced advanced statistical analysis using time series. Terminology and method for auto- and cross-correlation is discussed. The auto-correlation method is used in detecting repeating patterns in a time series and the cross-correlation method tests for the correlation of two time series.

Having provided background literature, the next chapter focuses on the description and overview of the datasets that have been gathered from the five network telescopes, as well as the tools used to implement the research. Furthermore, the next chapter analyses the logical and physical location of the telescope sensors in exploring similarities of network telescope sensors.

Chapter 3

Datasets and Research Tools

Obtaining appropriate datasets for the research is imperative since third-party datasets are usually anonymised to avoid disclosure of a network telescope's IP address space. The need for multiple comparators is one of the fundamentals of the research project since the study is comparative in nature and focuses on network telescopes' relativity across multiple telescope sensors. Time series' can be employed to model network traffic activity on each sensor.

This chapter discusses details of the datasets, including how data is sourced and the respective data collection methods. It is important to note that there has already been extensive research undertaken at Rhodes University on network telescopes. In this regard, the aim of this chapter is to provide a high-level overview of the network telescope's setup and a brief background on the relational database used in this research. Detailed information on the setup of network telescopes, collection of packets and processing of raw packets can be obtained in various studies done at Rhodes University [16] [18] and other research institutions such as Cymru [9] and CAIDA [31].

The focus of this research is on traffic analysis and as a result, the detailed mechanisms of building network telescopes are avoided. Abstraction from handling raw packet data is made possible by the use of a relational database infrastructure with packet information from different sensors stored in one database. Having one database allows for rapid development of queries and quick access. Additionally, the use of a relational database allows the researcher to work with packet information through relational tables instead of working directly with raw packet dumps. A statistical software package is required to perform statistical analysis and to represent the findings graphically. Since network telescope traffic has been converted to a relational database, querying language is used to

categorise and interpret traffic. Finally, results obtained from the database are used to conduct further analysis using a statistical package.

Understanding the location of the sensors (both physical or logically) is also important in differentiating between network telescope sensors and the effect this may have on traffic. The logical location refers to where the IP address blocks used by the sensors are placed relative to other IP address blocks. Even though the telescopes' sensors are physically located in Africa they may not be logically similar. The chapter concludes by detailing some of the differences and similarities that can be attributed to the physical and logical location of network telescopes.

3.1 Data Source and Collection

Datasets for the research have been obtained from Rhodes University's distributed network of telescope sensors. Rhodes University's research on network telescopes was initiated in 2005 [50] with the launch of a relatively small /24 network telescope sensor. This was the inception of network telescope research at Rhodes University and there have been subsequent telescope sensors added over the years. All network telescope sensors are physically located in South Africa on the TENET[1] network.

To conceal the network telescopes IP addresses, thereby avoiding potential poisoning of the datasets, the researcher has adopted five aliases for the network telescopes. The number represents the higher order IP prefix of the network telescope sensor. To differentiate multiple sensors within one IP prefix (the first 8 bits of the IPv4 address), alphabetic letters ("a", "b" and "c") are used. A brief description of each sensor is provided below:

- 196-a: Launched in August 2005 in the Eastern Cape. This was inception of network telescope research at Rhodes University.
- 146-a: Secondary telescope introduced in August 2009 at Rhodes University.
- 155-a: Launched at the beginning of 2011. The telescope is located in Western Cape.
- 196-b and 196-c: These are similar to 155-a in that both telescopes were launched at the beginning of 2011. They were, however, placed on the 196 IP prefix.

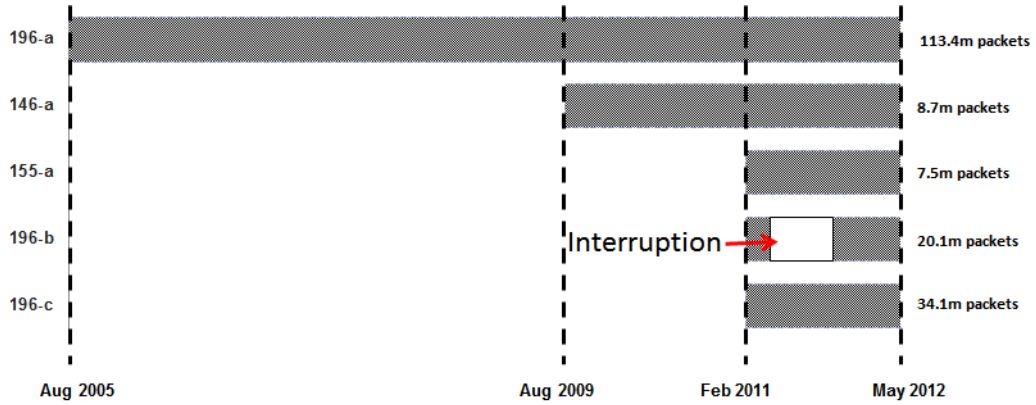


Figure 3.1: Evolution of Rhodes University's network telescope datasets

Figure 3.1 shows the evolution of the datasets using all five network telescopes. The timeline shows the network telescope's life span and the total amount of packets captured with a selected cut-off date of 20 May 2012. The cut-off date was based on the latest available datasets when the project was initiated. As of 20 May 2012, combined telescope data amounted to 183 million packets. Telescope sensor 196-a is the oldest sensor and has been online for over seven years at the time of writing. Sensor 146-a is the second oldest with more than three years worth traffic. Sensors 155-a, 196-b and 196-c are the most recent with over a year's worth of traffic captured.

3.1.1 Selection of Datasets

The first hurdle to overcome was to obtain multiple contiguous and overlapping datasets across different telescope sensors. The dataset needed to be contiguous to cater for the construction of time series'. That is, there should not be prolonged periods where the telescope was offline. It was expected, at an hourly time period, that datasets would experience ad-hoc interruptions due to anomalous network interruptions such as outages emanating from the service provider. However, at a daily time period, elongated outages would render datasets as non-contiguous. The researcher made a conservative assumption that, should a network telescope sensor not receive a packet for an entire hour, there is an outage. Outages analysis will be explored in detail in the upcoming chapters. Although there are advanced techniques to handle short periods of missing data through estimation of the cross-correlation function [46], it is preferable to have continuous and clean datasets.

As shown in Figure 3.1, between 10 February 2011 and 20 May 2012 there were five overlapping telescope sensors online, representing over 464 days (15 months). It is preferable

[1]Tertiary Education and Research Network of South Africa, http://www.tenet.ac.za/

to have a full year of datasets to cater for seasonal changes in activity across telescopes sensors. In this regard, a full year of data collection between the 20th of May 2011 and 20th of May 2012 is selected. Except for 196-b, all network telescopes experienced no outages lasting for a day or more. However, there were hourly outages that will be explored in detail in Section 5.4. Between May 2011 and November 2011, sensor 196-b experienced an interruption of 173 days. Interruptions were caused by the failure of a network interface card [16]. Notwithstanding the six months of interruptions, the research project makes use of a dataset from 196-b however, only for the period where it had been online continuously. The reason for this is the high similarities observed between 196-a, 196-b and 196-c (this will be discussed in detail in Chapter 4). To ensure that there is overlapping and common time period between the network telescopes, whenever comparative analysis of 196-b is conducted only the six month period from November 2011 to May 2012 is considered.

3.1.2 Overview of Data Gathering

There has been extensive research conducted on the Rhodes network telescope datasets in recent years (as discussed in Section 2.7). This section provides a high-level summary of how data was captured on the network telescopes and then processed to the PostgreSQL database. Readers are referred to [18] for detailed information on the database configurations.

Tcpdump[2] (a command-line packet analyser) is used for capturing packets. Files with captured packets are then copied and archived for further analysis [20]. Raw PCAP files are then parsed to a relational database. Raw packet manipulation and analysis is out of the scope of this project as data was already loaded onto the database. Therefore, the research project focuses on the analysis of data stored in relational databases.

3.1.3 Packets Storage - Relational Database

Analysis is conducted on the datasets stored in a relational database. A description of the relational database is provided in this sub-section. As indicated previously, a relational database stores the traffic information (or packet header information) of the telescope sensor. Each network telescope sensor has a defined, separate relational database. Figure

[2]http://www.tcpdump.org/

3.2 shows the Entity Relationship Diagram with table schema containing stored packet information. Each database contained the same database schema (definition of tables, constraints and indexes). This was advantageous in writing queries since similar querying scripts could be executed on different databases by simply changing the database connection. Tables contain the following fields (a definition of each field can be obtained in Appendix B):

- The packets table contains packet header information. This information is obtained from the IP (Internet Protocol) wrapper datagram.
- TCP, UDP and ICMP tables contain protocol-specific header information.

The major IP protocol tables (TCP, UDP and ICMP) were linked to the main packets table with a foreign key constraint enforcing a relationship between the protocol tables and the main packets table containing IP data. Every IP datagram would include information contained in the packets table however, depending on the data encapsulated (TCP, UPD and ICMP data), additional information would be stored. The database schema allows researchers to conduct individual protocol analysis since it offers flexibility to query protocol-specific fields.

3.2 Description of Datasets

As stated in Section 3.1.1, the range of data selected spans a year, from 20 May 2011 to 20 May 2012. During this period the total combined packets gathered from the five telescopes under investigation were 74.3 million. This is equivalent to an average of 203,065 packets per day. Figure 3.3 shows each network telescopes' contribution to the data that was observed. Looking at each sensor packet count, the relative similarities of total counts of 196-a, 196-b and 196-c are notable. These network telescopes form a natural category of telescopes termed as "category A". The lower total packet counts of 146-a and 155-a also allow one to categorise these two network telescopes as "category B". Further detailed analysis to support the categorisation of telescopes is conducted in Section 3.2.1 and Section 3.2.2. In brief, there are a number of factors that contribute to the difference in packet counts between category A and category B such as high-order IP prefix, physical

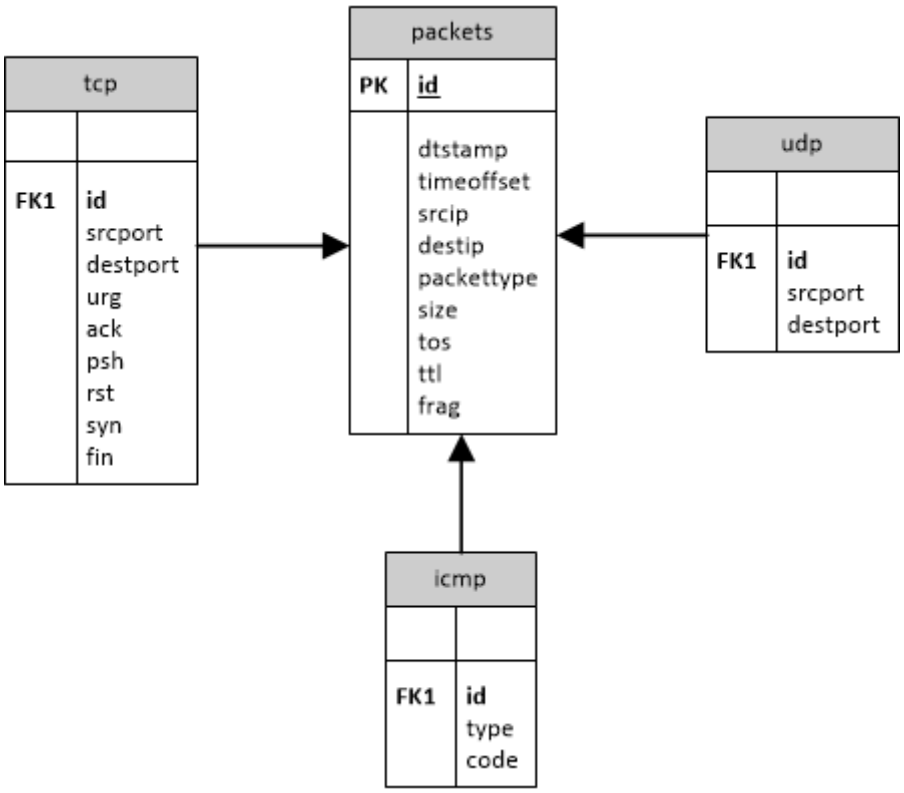


Figure 3.2: Entity relationship diagram of database schema

and logical location and the influence of the Conficker worm. Sensor 196-b has less packet counts compared to other category A telescopes (196-a and 196-c) this is due to the down time caused by a failed network interface card - the 173 days down time occurred between the 20 May 2011 and 9 November 2011 [16]. With only six months considered, between November 2011 and May 2012, 196-b has roughly 50% of packets that have been observed in other category A telescope sensors and, as such, it is considered to be similar to them.

3.2.1 Logical Distance Analysis

Although category A network sensors are part of the same IP prefix (196/8), the IP address ranges of each telescope sensors are non-continuous as there is moderate logical distance between the sensors IP ranges. The logical distance between two IP addresses is defined as the numerical difference between the two addresses [20]. Therefore, to calculate the logical distance between sensors the researcher calculates the numerical differences between the IP addresses representing the sensors.

Logical distance between network telescopes is important as it allows the researcher to test correlation between network telescopes that either are logically far apart or have address

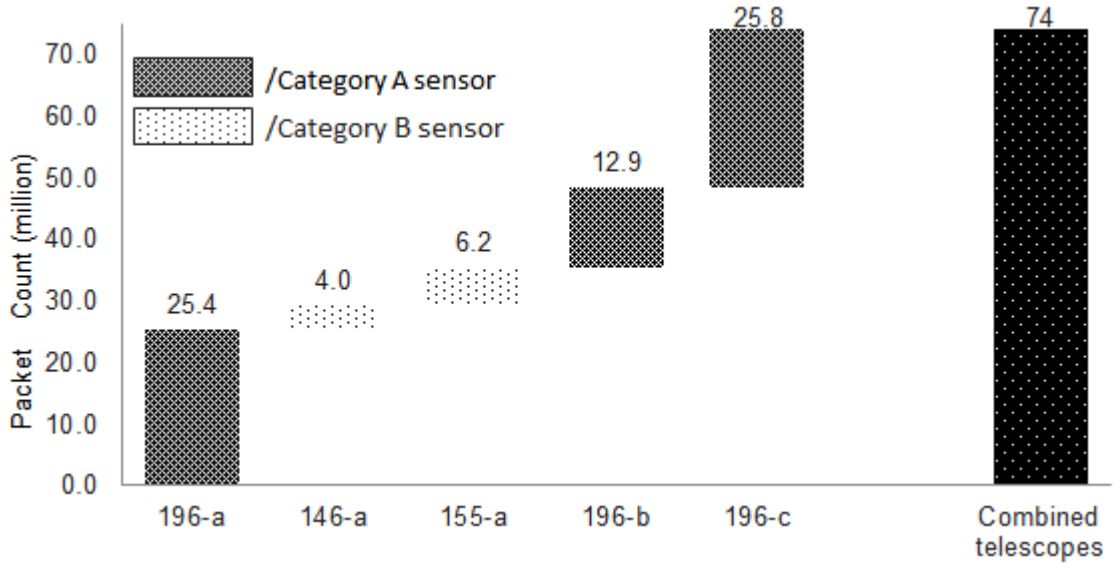


Figure 3.3: Packet count per network telescope sensor between 20 May 2011 and 20 May 2012

ranges that are logically dispersed. In correlation studies ideal network telescope sensors would have a fair degree of logical distance between them (i.e. not logically placed on IP address blocks directly adjacent to each other).

Since the network telescopes are all /24, determining the logical distance is done by simply making use the first three octets of the IP address. For example, with IP address A.B.C.D one simply use the first three octets (A.B.C). The insignificant octet is removed since the sensors' logical distance are being compared not the individual IP address. To convert an IP address to an integer the following approach is used [20]:

$$ConvertINT(A.B.C) = A * 2^{24} + B * 2^{16} + C * 2^{8} \tag{3.1}$$

Equation 3.1 converts IP prefixes (A,B,C) into an integer.

Therefore the logical distance of two IP address prefixes A.B.C and X.Y.Z is given by[3]:

$$ABS\,\{ConvertINT(A,B,C) - ConvertINT(X,Y,Z)\}$$

Having converted the IP address to an integer, the difference between the integer values is the logical distance between the IP addresses.

[3]ABS = absolute value

	146-a	155-a	196-a	196-b	196-c
146-a	0	151,058,944	825,070,592	825,043,968	825,247,744
155-a	151,058,944	0	674,011,648	673,985,024	674,188,800
196-a	825,070,592	674,011,648	0	26,624	177,152
196-b	825,043,968	673,985,024	26,624	0	203,776
196-c	825,247,744	674,188,800	177,152	203,776	0

Table 3.1: Calculated logical distance between network telescope sensors

Using the formula above, a logical distance analysis of the IP address blocks assigned to each network telescope sensor was conducted. Results are detailed in Table 3.1 which shows the logical distance matrix.

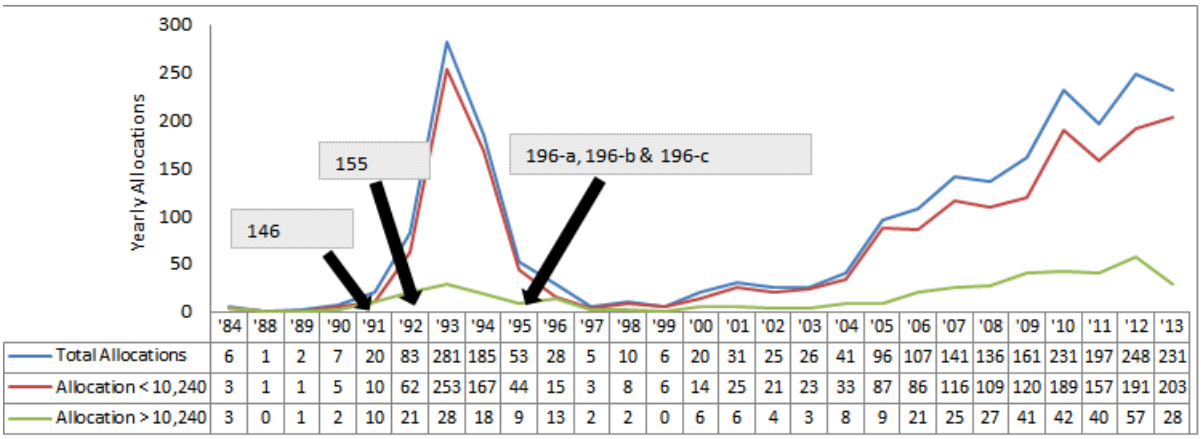
From the results in Table 3.1, the telescopes are generally dispersed however Telescope 196-b and 196-c are the closest. Category A sensors are closer because they are placed on the same high order IP prefix (196/8). Furthermore, category A address blocks are part of one /16 IP assignment. Of interest to the researcher is the relativity between large logically dispersed network telescopes. Previous studies by researchers at Rhodes University demonstrated that there is biasness of traffic emanating from closer address ranges [20]. The reason for this is that scanning techniques are biased to closer IP address.

3.2.2 Telescope Sensor Logical and Physical Location Analysis

Telescope sensors 146-a and 155-a (category B) have other characteristics that allow the sensors to be grouped into one category. The IP addresses of sensors 146-a and 155-a are managed by AFRINIC [15]. AFRINIC, as one of the five Regional Internet Registry (RIR), is responsible for IP address allocations in Africa. Furthermore, high order IP prefix 146 and 155 are legacy assignments of IP addresses initially made by the then Central Internet Registry, before the introduction of Regional Internet Registry [13]. Due to high growth in demand for IP addresses, the Internet Registry was decentralized to regional Internet Registries (ARIN, ARINIC, APNIC, LACNIC and RIPE NCC). The role of the Internet Registry is now carried out by IANA[4] (Internet Assignment Number Authority).

Telescope sensors' 196-a, 196-b and 196-c (category A) IP addresses are also managed by AFRINIC. Although prefix 196 also forms part of a legacy assignment, 196-a, 196-b and 196-c were allocated at a later date and made to relatively smaller institutions.

[4]http://www.iana.org/about



	'84	'88	'89	'90	'91	'92	'93	'94	'95	'96	'97	'98	'99	'00	'01	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11	'12	'13
Total Allocations	6	1	2	7	20	83	281	185	53	28	5	10	6	20	31	25	26	41	96	107	141	136	161	231	197	248	231
Allocation < 10,240	3	1	1	5	10	62	253	167	44	15	3	8	6	14	25	21	23	33	87	86	116	109	120	189	157	191	203
Allocation > 10,240	3	0	1	2	10	21	28	18	9	13	2	2	0	6	6	4	3	8	9	21	25	27	41	42	40	57	28

Figure 3.4: AFRINIC's evolution of IP address space allocation

Although it is difficult to trace the institutions IP apportions, given that regional registries do not disclose personal registration information such as the name of institution, the researcher conducted further analysis to support the observations that prefix 155 and 146 allocations were made to larger institutions or enterprises. Each Regional Registry has an ftp site[5][6][7][8][9] with monthly reports containing IP address space allocations and assignments. Alternatively, APNIC also provides a link to all the ftp sites with all regional registries monthly reports [1]. By looking at AFRINIC's monthly report[10] it is observed that 146-a and 155-a's allocations of IP address were assigned earlier; in 1991 and 1992 respectively. It is significant to note that 146-a and 155-a assignments were made prior to the sanctions being lifted in South Africa. In 1992, South Africa's participation in the global economy was restricted while at the same time the Internet was also in its infancy.

Figure 3.4 shows the evolution of IP allocations in AFRINIC and the date of allocation for 146-a and 155-a (category B) as well as for 196-a, 196-b and 196-c (category A). The diagram also reveals the apportions of IP address locations above 10,240 IP addresses and also apportions less than 10,240 IP addresses. Moreover, the diagram highlights the rapid increase of small IP address allocations (relative to large allocations) shortly after the 146 and 155 allocations were made. Between 1991 and 1996, the small allocations of IP space tracked the total allocations closely. In 1993, at the peak of allocation, there were a total of 281 allocations, only 28 of which were large allocations. IP prefixes for sensors 196-a, 196-b and 196-c were apportioned on the decline of small IP address block

[5]ARIN - ftp://ftp.arin.net/pub/stats/arin/
[6]AFRINIC - ftp://ftp.afrinic.net/pub/stats/afrinic/
[7]APNIC - ftp://ftp.apnic.net/pub/stats/apnic/
[8]LACNIC - ftp://ftp.lacnic.net/pub/stats/lacnic/
[9]RIPENCC - ftp://ftp.ripe.net/pub/stats/ripencc/
[10]November 2013 Reports are used to conduct the analysis

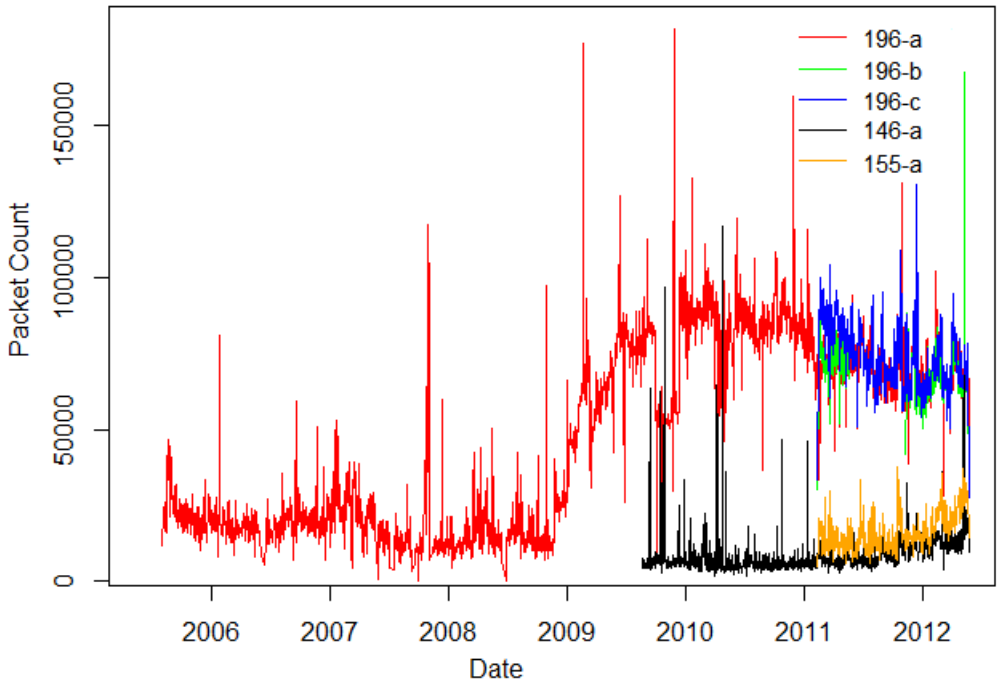


Figure 3.5: Evolution of daily packet count across all network telescopes

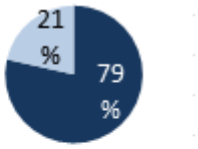
allocations relative to large IP address block allocation. The reduced activity levels (i.e. packet counts across category B as compared to category A) and the legacy allocations do suggest that there is bias for malicious activity towards category A network telescopes as opposed to category B. This occurs because category B allocations were made earlier and more likely made to larger institutions whereas category A allocations were made at a time when the uptake of the Internet were increasing in AFRINIC.

In Figure 3.5, a graph was generated to illustrate the daily packet counts across all five datasets. This graph displays activity over the years of sensors' existence with a cut-off date of May 2012. A lower daily packet count for category B (146-a and 155-a) relative to category A (196-a, 196-b and 195-c) is observed. Of note, is the increase in packet count of 196-a (category A) at the end of 2008. This is due the Conficker outbreak. Conficker and the impact thereof is explored further in Section 4.2. Although slightly higher, prior to 2009 sensor 196-a had a somewhat similar packet count to 146-a and 155-a with significant changes post-Conficker outbreak.

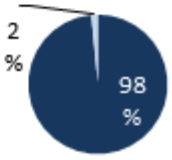
Another element that justifies the grouping of sensors is the neighbouring allocations analysis on each IP prefix with a telescope sensor. A search of AFRINIC's monthly report[11] for IP allocations with prefixes 155 and 146 was conducted. The results showed that in AFRINIC's allocation, prefix 155 has only 65536 block IP allocations with a total

[11]ftp://ftp.afrinic.net/pub/stats/afrinic/

IP Prefix 146	AFRINIC	APNIC	ARIN	LACNIC	RIPENCC	Total
Large allocation > 10240	6	9	138	4	77	**234**
Small allocation < 10240	0	1	0	0	63	**64**



IP Prefix 155	AFRINIC	APNIC	ARIN	LACNIC	RIPENCC	Total
Large allocation > 10 240	10	11	138	1	32	**192**
Small allocation < 10240	0	0	3	0	0	**3**



IP Prefix 196	AFRINIC	APNIC	ARIN	LACNIC	RIPENCC	Total
Large allocation > 10 240	74	0	3	1	0	78
Small allocation < 10240	671	14	35	16	11	747

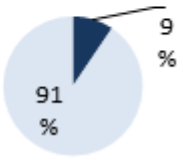


Figure 3.6: Network telescope's prefixes across all Regional Internet Registries

of ten in AFRINIC. Similarly, prefix 146 has only 65536 block IP allocations with a total of six in AFRINIC. Data from all Regional Registries' monthly reports is summarised further in Figure 3.6. The diagram details the five network telescopes prefixes with their respective allocation across all the Regional Internet Registries. To demonstrate the difference in allocations size, in Figure 3.6 the researcher has adopted the following categorisation:

- allocations above 10,240 IP address block grouped as 'large'[12] allocations and
- all other IP address blocks (less than 10,240) are grouped as 'small' allocations.

Category A with prefix 196 (AFRINIC allocations only) has a total of 745 different allocations of varying sizes from 256 to 262144 with 91% of the allocations in the 'small' category and the remaining 9% in the 'large' category. This is different to IP prefix 146 or 155 where there are only 'large' allocations that are contained in AFRINIC. With other Regional Internet Registries, 146 and 155 comprise mainly of 'large' apportions, 79% and 98% respectively.

It has been shown that the neighbouring IP block allocations for category B are generally larger than category A. Therefore, it can be expected that allocations in category

[12]The definition of large is not matched to regional registry's definition of large but used to show the intensity of smaller allocations.

B were made to large corporations or institutions. Moreover, one can infer that larger corporations or institutions would have a higher maturity of security systems. Neighbouring IP block allocations for category A, are generally smaller (less than 10,240 IP addresses). Similarly, it can be expected that category A neighbours are more likely to be small enterprises or those engaged in end-user activities. These users would have limited expertise and lack the capability to maintain information security systems. Additionally, most allocations made in the 146 or 155 prefix are part of ARIN (American Registry for Internet Numbers). ARIN is mainly responsible for the United States, Canada and a few Caribbean Islands [2]. IP prefix 196 has a 'large' number of allocations in AFRINIC. Clearly, allocations to ARIN are largely made to developed regions and AFRINIC allocations are made to developing regions. Classifications of regions can be obtained from the World Bank's website [3]. It is important to note that IP space allocations only represent countries to which the original allocations were made [1]. Consequently, there are limitations to using Regional Internet Registry data because IP address space may be assigned, for example, to multi-nationals with operations in multiple countries and therefore the original location may be inaccurate.

3.3 Tools Used in the Research Project

In this section a discussion of tools that have been used to conduct the research is provided. The rationale behind the selection and the efficiencies realised are also highlighted. For querying and conducting basic data analysis, the researcher made use of PostgreSQL[13]. The majority of statistical analysis and graph generation was conducted using the R Statistical package[14].

3.3.1 Relational Database - PostgreSQL

As discussed in Section 3.1.3, data captured in PCAP files is parsed to a relational database. Since data is stored in a relational database, SQL querying language provides a first step in categorising and performing queries that are used as inputs to further process through R Statistics.

[13] http://www.postgresql.org/
[14] http://www.r-project.org/

The relational database is hosted in a PostgreSQL instance. An advantage to using PostgreSQL is having a cross-platform relational database that operates under different platforms (UNIX,Windows or Mac).

The PostgreSQL graphical user interface (pgAdmin[15]) provides a platform for rapid query development. Having packet information already parsed into an SQL database meant that the researcher could simply abstract from details of parsing data from PCAP files into a database. This enables the researcher to focus primarily on data analysis using PostgreSQL and pgAdmin. This approach, used extensively at Rhodes University, allows researchers to focus on performing experiments rapidly.

3.3.2 Statistical Package - R Statistics

For statistical computing and the construction of graphs, the R Statistics package has been the preferred software package. R is selected because it is an open source software package that can run on multiple platforms (UNIX, Windows or Mac). As it is open source software, R Statistics has good documentation with an extensive online community presence. R Statistics consists of a command-line interface with an intuitive programming language. R Statistics has a built-in capability to produce advanced graphs using simple commands.

3.4 Summary

This chapter provided an overview of the datasets, indicating how they are being accessed, as well as sharing a summary of the processing of data into a SQL database. A period of 12 months between 20 May 2011 and 20 May 2012 was selected based on the available online datasets. The annual period caters for possible seasonal changes.

The chapter provides a brief description of the datasets and showed that telescope sensor 196-a, 196-b and 196-c had roughly the same amount of packets. Similarly, but to a lesser extent, 146-a and 155-b had a related amount of total packets. Empirical data analysis justified the selected categorisation of the five telescope sensors into two categories (category A and category B). Sensors 196-a, 196-b, 196-c all formed part of category A while 146-a and 155-a formed part of category B. The justification of the categories assigned was determined by a combination of factors:

[15] http://www.pgadmin.org/

- Similarities in packet size observed under the period of investigation.
- The logical distance between the network telescopes. Although the sensors are logically dispersed and not placed on adjacent IP address block, as can be expected, 196-a,196-b, 196-c were relatively closer.
- The logical location of the telescope sensors. It was shown that, due to legacy apportions of IP address blocks, category B sensors (146-a and 155-a) and category A (196-a, 196-b, 196-c) sensors allocations are affected by the time at which the IP block apportions were made. The allocations for Category B sensors were made earlier (1991 & 1992) and category A sensors allocations were made later (1995).

The observations made showed that category A sensors are located in more 'end-user' environments which are more susceptible to attacks and malicious activity. This was different to allocations in category B which were larger allocations and likely to have been made to bigger organisations.

In closing, a brief discussion of the relational database tools and the statistical package used was provided. Tools were selected to allow the researcher to abstract away from the hurdle of setting the infrastructure and the system, but rather to focus on analysis and comparative studies of the data.

Given this background, which describes the datasets of all sensors, together with the analysis of the location thereof, the next chapter will focus on the initial finding of the study by comparatively analysing traffic observed across all five sensors.

Chapter 4

Comparative Analysis of Traffic

Chapter 3 focused on providing a basis for the project by profiling the sensors and delivering an analysis of the location thereof. Initial observations of total packet count for the five network telescopes and the logical distance between sensors provided a natural categorisation of network telescopes sensors: sensors 196-a, 196-b and 196-c as category A; and sensors 146-a and 155-a as category B.

The overarching approach in this chapter is to conduct a comparative analysis of network telescopes' sensors by using summarisation and a basic statistical analysis of each dataset. Graphs will be presented to support the observations.

This chapter compares the five network telescopes' datasets highlighting similarities and observed differences. Since the research project follows an iterative process and is based on empirical datasets, the initial results will be studied and a refinement of experiments will be conducted to improve or optimise the results.

4.1 Dataset Comparative Analysis

The initial approach in analysing the relativity of a telescope sensor's traffic activity is to first use summarisation techniques and the graphical representation of trends over time. Summarisation techniques will be used to construct basic metrics. These metrics will be used to investigate packet distributions across the following: major protocols, source and destination ports, and source and destination IP addresses. To investigate underlying trends, generated packet counts are categorised into various period bins (monthly, daily

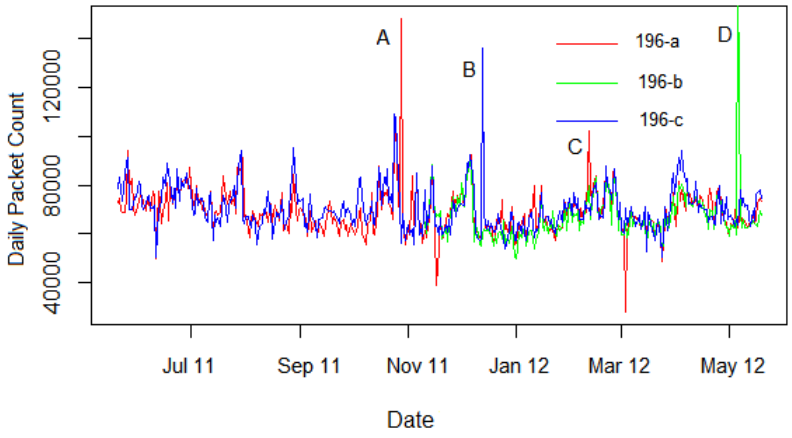


Figure 4.1: Daily packet count of telescope sensors 196-a, 196-b and 196-c between 20 May 2011 and 20 May 2012

and hourly) and placed into various groups ('All packets', TCP, UDP or ICMP). Time series' constructed in this chapter will be used in Chapter 5 for conducting advanced correlation analyses of time series.

4.1.1 Periodic Packet Counts

The first sets of series constructed are general packet count (including all packet types) with various bins (daily, monthly and hourly). These periodic packet counts allow researchers to monitor traffic activity over time and assist in discerning trends and detecting anomalous activity. Having time series plots on same set of axis also allows the researcher to monitor changes of activity in one sensor relative to other sensors.

Constructed series will cover the selected 12 months period (May 2011 to May 2012). As previously stated, when conducting a comparative analysis for sensor 196-b, a six-month period is used to cater for downtime experienced (i.e. only considered the overlapping six-month period across all other telescope sensors). SQL scripts were used to construct the time series by counting the number of packets received on various time bins.

Figure 4.1 shows a daily packet count of sensors 196-a, 196-b and 196-c (classified previously as category A). The plot shows related peaks and troughs of packet counts suggesting relativity of daily traffic activity across the network telescopes. In addition, the general shape and trend of the activity also suggest relativity of traffic activity across category A. Category B (146-a and 155-a) telescopes (shown in Figure 4.2) do show some relativity. These, however, are not as strongly correlated as category A.

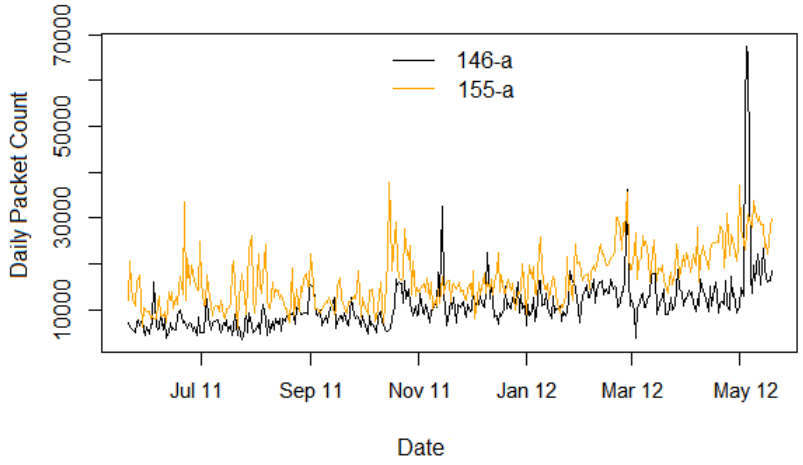


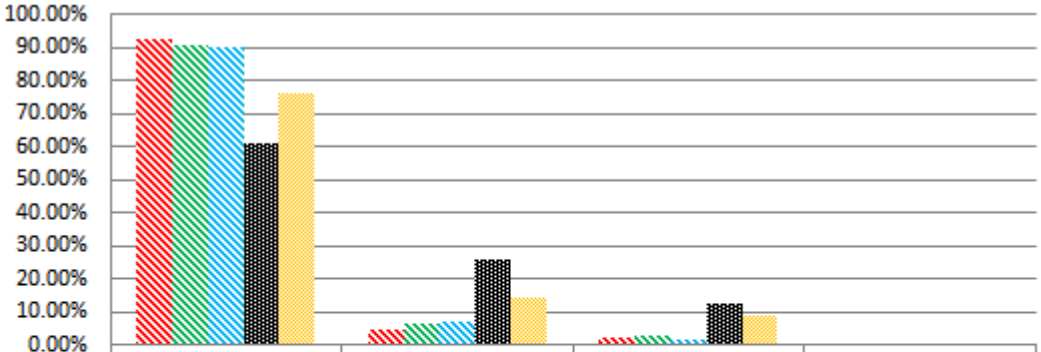
Figure 4.2: Daily packet count of telescope sensors 146-a and 155-a between 20 May 2011 and 20 May 2012

Graphical results show that there is less similarity in inter-category comparisons relative to intra-category comparisons. When examining monthly packet counts, although more summarised, a similarity of trends and the general shape of activity for category A sensors can be observed. Monthly and hourly graphs are included in Appendix C , as well as a combined logarithmic plot of all the telescopes' sensors in one chart.

Visual results observed are aligned to initial observations made regarding relativity on traffic activity in each category. Outliers or traffic spikes are marked with letters "A", "B", and "C" in Figure 4.1. The traffic spikes are uncoordinated across the relative sensors and, therefore, it can be expected that these spikes would cause distortion in relativity analysis. This rapid build-up of traffic will be explored in detail later on in the chapter.

4.1.2 Packet Type Analysis

Before analysing each traffic type separately, an analysis of packet distribution across the major protocol (TCP, UDP and ICMP) is conducted. In this experiment, traffic has been summarised by the percentage share of each major protocol across the five network telescope sensors using the selected annual period. Figure 4.3 shows the percentage of packets of each protocol across the five sensors. The results show that category A network telescopes (196-a, 196-b and 196-c) have a similar distribution of packets across the major protocols (TCP, UDP and ICMP). TCP, UDP and ICMP traffic have a total combined contribution of more than 99,9% across all datasets. Although sensor 196-b only has a six-month overlap period, the telescope sensor has an almost identical distribution of



	TCP	UDP	ICMP	Other
196-a	92.49%	5.17%	2.33%	0.011%
196-b	90.66%	6.56%	2.77%	0.021%
196-c	90.45%	7.42%	2.12%	0.011%
146-a	61.34%	25.82%	12.84%	0.004%
155-a	76.34%	14.79%	8.86%	0.004%

Figure 4.3: Packet type per network telescope

packets across the major protocols relative to other category A sensors. TCP traffic is dominant in category A telescope sensors, accounting for 90% of all traffic. UDP traffic and ICMP traffic have a lesser share of the traffic in category A.

It is evident that category B network telescopes (146-a and 155-a) have a slightly reduced level of TCP dominance, while UDP and ICMP proportion is higher in category B relative to category A telescopes. One reason for this difference in TCP traffic is the prevalence of Conficker-related traffic in category A telescopes (Conficker worm is discussed in detail in Section 4.2).

The results show similarities in the distribution of packets across the major protocols for category A telescope sensors. Category B telescope sensors also have a similar distribution of packets across the major protocols. Additionally, these results are comparable to the results observed in previous work [16]. Interestingly, results obtained by a different study [38], with a week's trace of data between 28 April 2004 and 5 May 2004, shows ICMP traffic as the second-largest contributor of traffic as opposed to UDP. Notwithstanding the short period of analysis (April 28 to May 5), the lower UDP packet count was caused by filtering of port 1434/udp which is associated with Slammer worm. Although the oldest sensor in this research (sensor 196-a) starts its collection of data from 2005, it can be seen in Figure 4.4[1] that UDP traffic has, for the majority, been the second-largest contributor of traffic on 196-a. However, a brief period (shown with the callout box in the diagram) between August and October in 2006 where ICMP traffic was second-largest contributor

[1]Double bars on the diagram indicate that the extreme values exceed 40,000 daily packet count

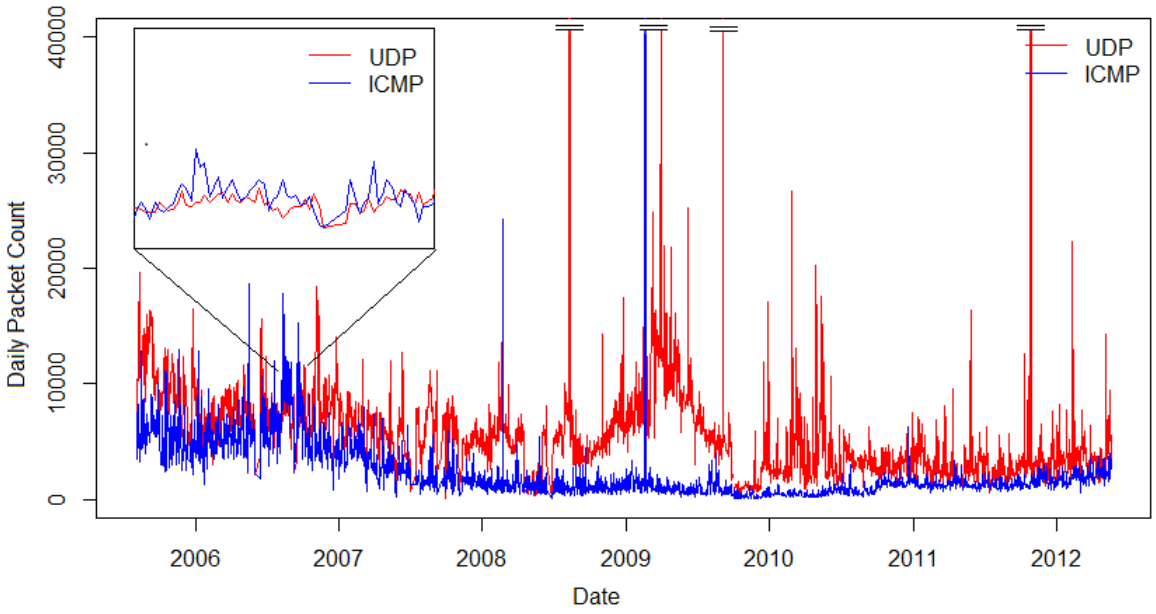


Figure 4.4: UDP and ICMP daily packet count of telescope sensor 196-a between 3 August 2005 and 20 May 2012

of traffic is observed. ICMP packet spikes are also observed resulting in a larger ICMP packet count.

As an example of category A traffic composition, by examining sensor 196-c's daily packet count broken down into the three major protocols (shown in Figure 4.5), it is observed that daily the TCP packet count is a dominating indicator and tracks the daily 'All packet types' series. Sensor 196-c's daily packet count plot shows the TCP packet count series being distant from the UDP and ICMP's packet count series. That is, there is a higher magnitude of TCP packet counts than UDP or ICMP. It is observed that minimal deviations are caused by the underlying UDP or ICMP packet activity. Although the y-axis uses a logarithmic scale of 10, it is evident that only high levels of UDP fluctuations have an intermittent influence on 'All packet types' series. Related points where UDP activity caused fluctuations are highlighted on the chart (letters "A", "B" and "C").

Figure 4.6 shows the packet count across major protocols (TCP, UDP and ICMP) for sensor 146-a as an example of a category B telescope. The dominance of TCP is weaker compared to the 'All packet types' series. In contrast to sensor 196-a, results show that UDP and ICMP counts are closer to TCP counts and therefore the UDP variable has a greater influence on the 'All packet types' variable. To this end, the daily packet count variable does not closely track the TCP Only packet count variable. This means that, as opposed to category A, the other traffic types (UDP and ICMP) have a greater influence on the fluctuations of the daily traffic activity.

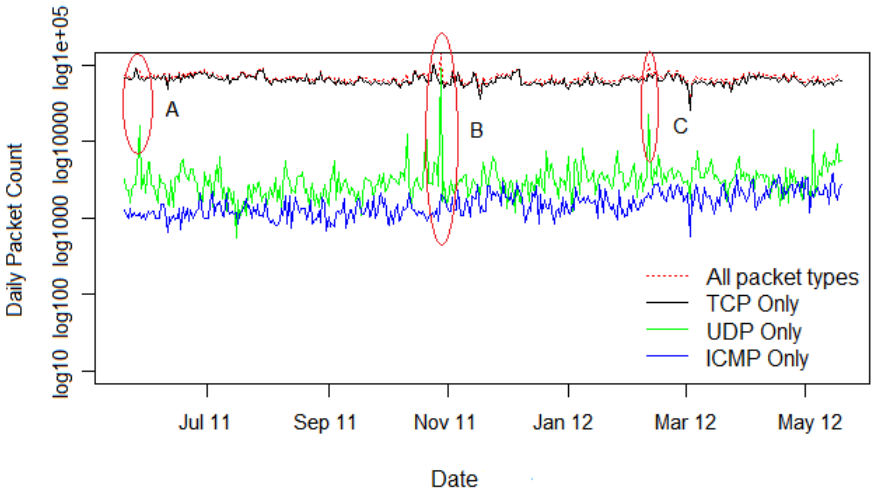


Figure 4.5: Daily packet count of telescope sensor 196-c between 20 May 2011 and 20 May 2012 - protocol overview

Packet count contribution of traffic outside of TCP, UDP or ICMP is extremely small (between 0.004% and 0.02%) across all network telescope nodes. Hence, the researcher did not analyse other packet types, as network telescope traffic activity is considered insignificant.

4.2 TCP Analysis - Destination Port

As already indicated, TCP traffic is the most common protocol across all the sensors. TCP protocol offers a reliable communication service by establishing connection between hosts and implementing flow control, as well as acknowledgment techniques [10]. This section, as part of an separate analysis of TCP traffic data, will investigate the top ports that account for a significant amount of TCP traffic.

Table 4.1 shows TCP's top 20 ports for all telescope sensors as a proportion of all TCP packets. Results show that between 68.3% and 74.1% of traffic routes to the top five ports in each network telescope, with the exception of sensor 155-a. Sensor 155-a has a flatter distribution of packets across the top 20 ports because the top 20 ports accounts for 56% of all TCP traffic. The dominance of port 445/tcp in category A network telescopes is quite evident with roughly 67% of the traffic routing to port 445/tcp.

A brief description of the transport protocol port numbers and the service can be obtained from IANA[2]. A description of the top four ports is provided below:

[2]http://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xml

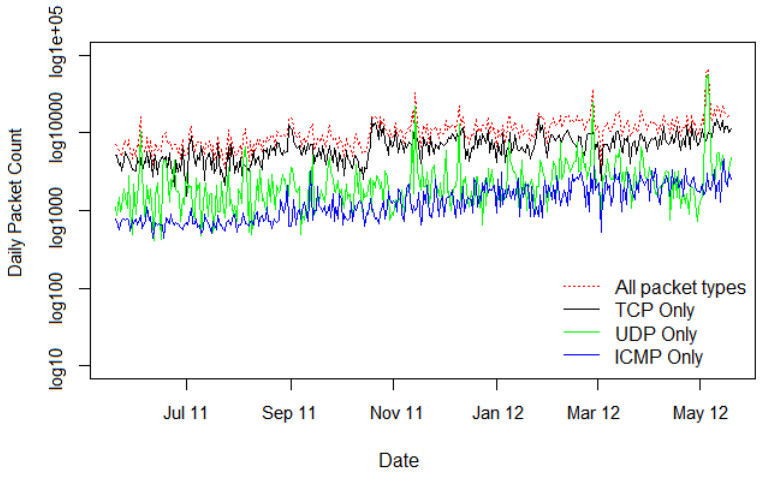


Figure 4.6: Daily packet count of telescope sensor 146-a between 20 May 2011 and 20 May 2012 - protocol overview

Table 4.1: Top 20 TCP destination ports across all telescope sensors

Rank	196-a		146-a		155-a		196-b		196-c	
	Port	%	Port	%	Port	%	Port	%	Port	%
1	445	66,9%	445	21,9%	3389	9,4%	445	64,1%	445	66,6%
2	22	2,2%	3389	17,6%	1433	7,3%	80	2,6%	22	2,1%
3	80	2,0%	80	12,2%	80	7,0%	22	2,4%	1433	2,0%
4	3389	1,9%	1433	10,8%	445	6,7%	3389	2,0%	80	1,9%
5	1433	1,7%	8080	5,8%	57471	4,8%	1433	1,6%	3389	1,5%
6	49787	1,0%	22	5,7%	22	4,0%	23	1,3%	10300	1,0%
7	23	1,0%	23	4,5%	8080	3,1%	8080	1,1%	8080	0,9%
8	8080	0,9%	139	4,0%	23	2,6%	1234	0,7%	135	0,9%
9	135	0,8%	135	1,8%	1234	1,8%	135	0,7%	23	0,8%
10	5900	0,6%	3306	1,6%	1024	1,5%	443	0,6%	5900	0,6%
11	443	0,5%	443	1,3%	3072	1,4%	3306	0,5%	3306	0,5%
12	25	0,4%	5900	1,0%	139	1,2%	25	0,5%	443	0,5%
13	1234	0,4%	25	1,0%	3306	1,0%	39459	0,3%	25	0,4%
14	139	0,3%	4899	1,0%	443	1,0%	5900	0,3%	1234	0,4%
15	3072	0,3%	9415	0,9%	135	0,8%	1024	0,3%	3072	0,3%
16	1024	0,3%	8909	0,8%	5900	0,8%	3072	0,3%	1024	0,3%
17	3306	0,3%	21	0,6%	21	0,8%	110	0,3%	46904	0,3%
18	110	0,2%	210	0,4%	25	0,7%	15215	0,2%	139	0,3%
19	3128	0,2%	3128	0,4%	0	0,5%	139	0,2%	3128	0,2%
20	6881	0,2%	5631	0,3%	8909	0,5%	34643	0,2%	21	0,2%
Others		17,9%		6,4%		43,25%		19,9%		18,3%

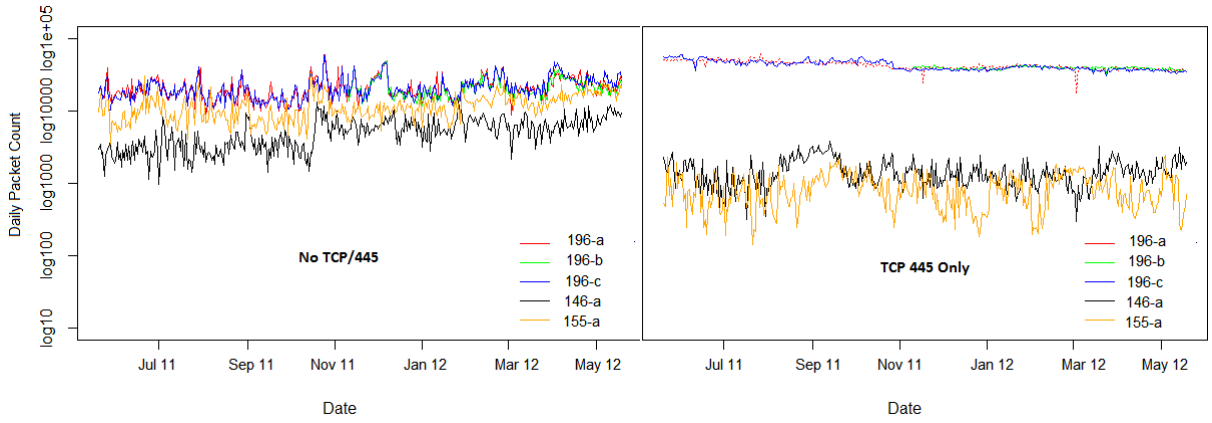


Figure 4.7: Port 445/tcp vs. no 445/tcp - daily packet count of all telescope sensors between 20 May 2011 and 20 May 2012

- Port 445/tcp runs Microsoft's active directory service. Conficker worm uses a specially crafted remote procedure call over port 445/tcp [39]. Therefore, the dominance of 445/tcp across sensors 196-a, 196-b and 196-c is due to Conficker and related scanning. Additional detailed analysis of Conficker will be conducted shortly.
- Port 80/tcp is responsible for HTTP traffic and the port's high ranking is expected since the port listens to web traffic.
- Port 22/tcp and port 3389/tcp are responsible for secure socket shell (SSH) and remote desktop protocol respectively. The high ranking of these ports indicates substantial SSH and RDP scanning.

Given the large presence of Conficker worm, additional analysis of the worm is carried forward. Conficker is a self-propagating worm which appeared in November 2008. Conficker exploit uses a specially crafted remote procedure call causing infected Microsoft machines (2000, XP, 2003 Server and XP) to run instructions without authentication [39]. Traces of Conficker date back to September 2008, however, by the end of October 2008 Microsoft had released a patch [27]. The outbreak of Conficker only came online in mid-November 2008. The existence of Conficker, at the time of writing, is an indication of the lack of uptake in patch management given that a patch was released since in October 2008. As discussed in Section 3.2, the high order IP prefix 196 generally has smaller IP space assignments and is located in developing countries with relatively less maturity in information security. In this regard, it is likely that patch management is not prioritised given the more 'end-user' environment present. These factors would result in a relatively more vulnerable IP space that is attractive to malicious attacks. This can be seen as a

contributing factor to increased scanning in these environments. In Section 3.2 it was shown that larger IP apportions were generally made for 146 and 155 IP prefixes. Furthermore, apportions in 146 and 155 prefixes were mostly made to developed countries and, as such, it can be expected that there is increased knowledge of information security and application of patches.

Apart from the differing logical and physical location of sensors, Conficker has, however, a bug in its random number generator. The number generator is used to create IP addresses used to identify hosts. A Windows random generator provides 15-bit random numbers. In order to create a 32-bit IP address, Conficker uses two 15-bit random number generators [39]. This effectively means that there are two bits (first bit of Octet 2 and 4) that are not catered for in the final generated IP address resulting in excluded IP space. The excluded range of IP addresses falls between [5]:

$$x.128.x.x - x.255.x.x \tag{4.1}$$

and

$$x.x.x.128 - x.x.x.255 \tag{4.2}$$

Another reason for the natural categorization of category A and category B sensors is the comparison of IP address for category A sensors relative to those of category B. Given the range in Equation 4.2 and looking at the full IP address, sensors 146-a and 155-a (category B) fall outside Conficker's target range, while 196-a, 196-b and 196-c (category A) are inside Conficker's target range [16]. To show the impact of this differentiation, the researcher constructed a time series with all TCP packets excluding port 445/tcp. A second set of series are constructed with TCP packets that only route to port 445/tcp. Figure 4.7 shows quite clearly that 'non-TCP' traffic has relatively similar peaks and troughs across all five network telescopes. Looking at 445/tcp only, it is observed that there is a disjunction between category A and category B - with category A experiencing higher packet counts than category B.

Table 4.2: Top UDP destination ports across all telescope sensors

Rank	196-a		146-a		155-a		196-b		196-c	
	Port	%	Port	%	Port	%	Port	%	Port	%
1	5060	34,98%	5060	19,87%	5060	22,53%	5060	32,38%	5060	52,61%
2	**[19416]**	6,74%	**[24003]**	14,61%	1434	6,24%	**[21566]**	12,34%	**[22549]**	8,61%
3	1434	4,22%	1434	5,34%	6257	2,41%	1434	5,14%	1434	6,58%
4	137	2,15%	6257	2,02%	137	1,99%	137	1,66%	41560	5,07%
5	6257	1,66%	**[18261]**	1,87%	53	1,99%	6257	1,62%	137	3,13%
6	**[473]**	1,63%	**[41511]**	1,80%	**[6568]**	1,76%	**[1046]**	0,94%	**[41559]**	2,73%
7	**[38834]**	1,43%	**[4375]**	1,78%	**[60505]**	1,57%	**[48170]**	0,84%	6257	2,51%
8	**[6655]**	0,88%	137	1,60%	**[43815]**	1,20%	**[1288]**	0,52%	**[64578]**	2,14%
9	**[26848]**	0,83%	**[5159]**	1,20%	**[32737]**	0,98%	53	0,52%	53	2,12%
10	53	0,82%	**[21284]**	1,06%	39455	0,95%	39455	0,51%	54311	1,87%
Others		44,67%		48,85%		58,37%		43,52%		61,57%

4.3 UDP Analysis - Destination Port

Continuing from the previous TCP section, this section focuses on UDP as the next common protocol. UDP is a lightweight (i.e. has less packet overhead compared to TCP) and connectionless protocol [40]. Unlike TCP, UDP does not guarantee delivery packets or the protection against duplicates. Table 4.2 shows the top 10 UDP ports for all network telescopes based on the proportion of traffic routing to a particular port. Port 5060/udp ranks the highest across all the sensors. Port 5060 is widely used for SIP[3] (Session Initiation Protocol) traffic. SIP is responsible for multimedia communication including voice, video and voice over IP [21]. Results therefore show the prevalence of SIP scanning in UDP traffic.

An observation worth taking into account is the high ranking of port 1434/udp across all sensors, signifying the presence of SQL Slammer worm. Slammer worm exploits buffer overflow vulnerability on computers running Microsoft SQL Server [33]. At the time of writing, Slammer has been present for over a decade since its outbreak on the 25 January 2003. Following the outbreak, Microsoft released a patch on 31 January 2003 [29]. The presence of Slammer from its outbreak in January 2003, up to the time of this study, again reiterates the lack of the information security awareness in implementation patches.

Of interest is the unusually high 24003/udp ranking in the top 10 ranking. This is significant because 24003/udp only appears in sensor 146-a (highlighted in bold on the

[3]port list available at http://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xml

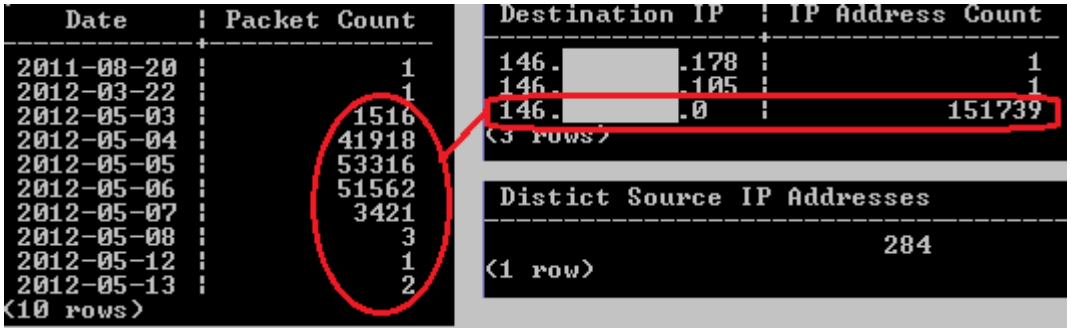


Figure 4.8: Anomalous spike - sensor 146-a on port 24003/udp

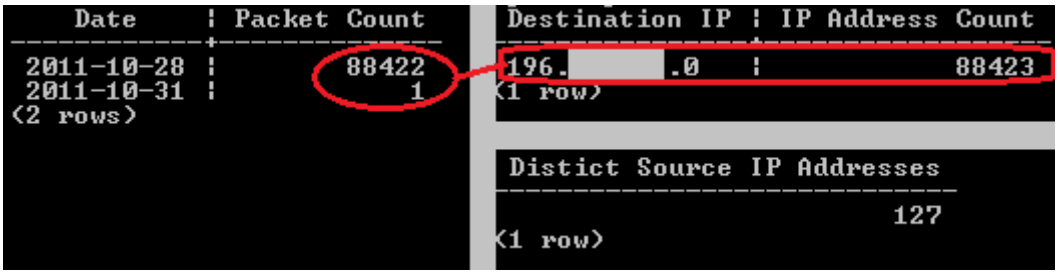


Figure 4.9: Anomalous spike - sensor 196-a on port 19416/udp

port table). While investigating activity on 24003/udp, the daily packet count shown in Figure 4.8 highlights very limited to no activity for sensor 146-a for the whole year except a significant build-up between the 3 and 7 May 2012. In addition, results indicate that the target was on a single IP address within the sensor address space. The source IP address count also shows that there were roughly 284 different IP addresses routing to 24003/udp. Hourly packet count analysis shows a build-up over the weekend, from Friday morning until Monday morning (shown in Figure 4.10).

Likewise, an investigation of sensor 196-a also indicated that 19416/udp (also uncommon to all other telescopes) experienced a similar spike of traffic on 28 October 2011. The results of 19416/udp are captured in Figure 4.9. Results show no packet count for an entire year except for 28 October 2011 and 31 October 2011. There were 127 different and distinct source IPs routing traffic to one destination IP (196.x.x.0). By manually examining the top three ports, the researcher also identified anomalies (i.e. rapid packet build-up) on:

- port 21566/udp for sensor 196-b (results shown in Figure 4.11); and
- port 22549/udp for sensor 196-c (results shown in Figure 4.12).

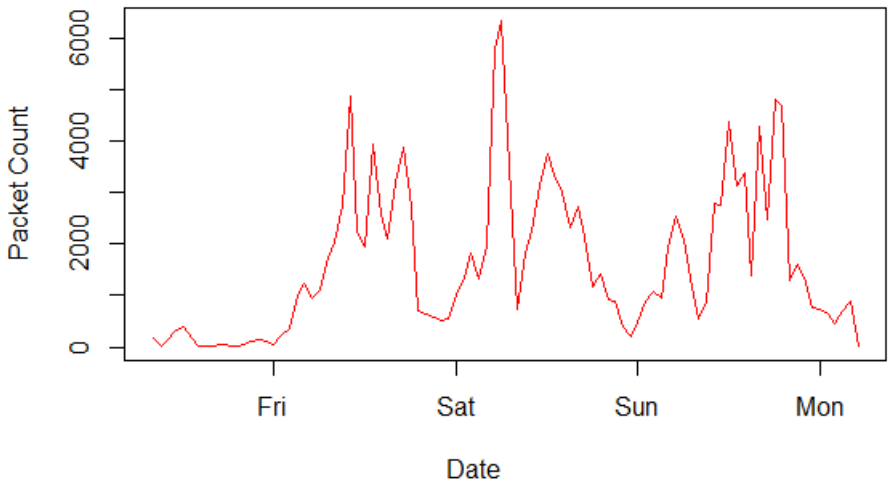


Figure 4.10: Hourly series of an anomalous spike - sensor 146-a on port 24003/udp

Date	Packet Count
2012-05-06	104079
2012-05-19	7

(2 rows)

Destination IP	IP Address Count
196.[redacted].0	104086

(1 row)

Distict Source IP Addresses
280

(1 row)

Figure 4.11: Anomalous spike - sensor 196-b on port 21566/udp

Date	Packet Count
2011-09-29	1
2011-12-13	72560
2012-01-03	1
2012-01-07	1
2012-01-08	1
2012-01-21	1
2012-01-22	2
2012-01-23	1
2012-01-24	1
2012-01-30	1
2012-02-01	1
2012-02-10	1
2012-02-14	1
2012-02-15	1
2012-02-21	1
2012-02-27	1
2012-03-05	1

(17 rows)

Destination IP	IP Address Count
196.24.70.27	1
196.24.70.0	72576

(2 rows)

Distict Source IP Addresses
265

(1 row)

Figure 4.12: Anomalous spike - sensor 196-c on port 22549/udp

When analysing UDP traffic, the term 'anomalous traffic spike' is used to classify events that resemble that of a denial-of-service attack. Detailed analysis of denial-of-service attacks is beyond the scope of this paper and therefore the researcher refrains from carrying out further investigations to ascertain whether these anomalous spikes are in fact denial-of-service attacks. Anomalous spikes are classified according to the following criteria:

- Sudden rapid build-up of UDP packets to a specific port.
- Distributed machines being used to launch an attack. Although IP spoofing (generating random source IP address) can be achieved rather easily, multiple source IP addresses are an indication of a potential distributed attack.
- Traffic routing to a single destination IP through a specific port.

Anomalous traffic spikes cause deviations on correlation with network telescope traffic activity. This occurs because these anomalous spikes focus on a single machine or IP address and are not replicated on other network telescopes. Furthermore, the packet counts would form outliers due to the rapid packet build-up.

The initial observations of anomalous spikes were the result of the researchers noticing unique ports in the top 10 UDP ports. Following in the same thought process, a method of looking for unique ports with relatively large packet count is presented. The researcher constructs a matrix, denoted as X and represented by Equation 4.3, with top $1 \ldots n$ ports across all telescopes (a, b, c, d, e). Ports are ranked based on the proportion of traffic routing to a particular port with a cutoff value of (n) for the rankings. A threshold value (p) is selected to signify the lowest ranking of ports that is considered. Another matrix, denoted as Matrix Y and represented by Equation 4.4, with ports ranking from $n \ldots p$ is developed. A search for unique values (port numbers) in Matrix X is then conducted. If a unique value was found in Matrix X, unique values are compared to all the values in Matrix Y. Matrix Y is constructed to ensure that values that are unique in Matrix X are not merely unique because of the cutoff imposed by the height (n) of Matrix X. If unique values were not found in the second matrix, the port number is considered as an anomaly requiring additional investigation.

$$\begin{matrix} a_1 & b_1 & c_1 & d_1 & e_1 \\ a_2 & b_2 & c_2 & d_2 & e_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_n & b_n & c_n & d_n & e_n \end{matrix} \tag{4.3}$$

and

$$\begin{matrix} a_{n+1} & b_{n+1} & c_{n+1} & d_{n+1} & e_{n+1} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n+p} & b_{n+p} & c_{n+p} & d_{n+p} & e_{n+p} \end{matrix} \tag{4.4}$$

Table 4.3: Anomalous spike investigation results using UDP traffic

Sensor	Port	Duration & Packets[4]	destination IP	diff SrcIPs	Anom. spike?
196-a	473	multiple days	"196.x.x.109"	699	Unclear
196-a	38834	1 day btw 14:00 and 16:00 (18763 packets)	"196.x.x.14"	5824	Yes
196-a	6655	1 day (11488 packets)	"196.x.x.0"	79	Yes
196-a	26848	1 day at 17:00 (10835 packets)	"196.x.x.140"	5	Yes
146-a	18261	2 days (15:00 to 3:00-next day)	"146.x.x.0"	191	Yes
146-a	41511	multiple days	multiple	11128	Unlikely
146-a	4375	1 day btw 4 and 5 am (18506 packets)	"146.x.x.110"	3450	Yes
146-a	5159	multiple days	"146.x.x.99 & 114"	7509	Unclear
146-a	21284	1 day btw 06:00 and 07:00 (1096 packets)	"146.x.x.0"	94	Yes
155-a	6568	1 day at 03:00 (16095 packets)	"155.x.x.120"	1	Yes
155-a	60505	multiple days	"155.x.x.165"	652	Unclear
155-a	43815	multiple days	multiple	6472	Unlikely
155-a	32737	multiple days	multiple	178	Unlikely
196-b	1046	1 day btw 16:00 and 17:00 (7888 packets)	"196.x.x.120"	1499	Yes
196-b	48170	1 day btw 07:00 and 11:00 (7079 packets)	"196.x.x.175"	2635	Yes
196-b	1288	1 day btw 04:00 and 05:00 (4415 packets)	"196.x.x.120"	1270	Yes
196-c	41559	multiple days	"196.x.x.143"	954	Unclear
196-c	64578	multiple days	"196.x.x.72"	328	Unclear

To implement this technique, an experiment was conducted looking at top 10 ports (i.e. $n = 10$) and selecting a threshold of 10 ($p = 10$). Using this technique, a total of 22 UDP port numbers were identified spread across all five telescopes. Manual analysis of the 22 ports was used to look at the duration of traffic routing to the specific UDP port, as well as the destination and source IP addresses. Table 4.3 shows the results of manual analysis[5]. Of the 22 ports numbers identified:

- 14[6] ports numbers that were identified exhibited behaviors that allows for them to be classified as anomalous spikes.
- 5 required additional analysis to categorise.

[5]The other 4 ports analysed above are not included on the table
[6]ten plus the four previously analysed

- 3 were categorised as unlikely to be anomalous spikes.

4.4 ICMP Analysis

ICMP (Internet Control Message Protocol) is a relatively simple protocol however; there is limited awareness of the security issues that are associated with the protocol [44]. ICMP is used to send messages about problems in the network [41]. Therefore, ICMP offers error reporting and allows users to investigate traffic issues. Attackers also use the ICMP protocol in fingerprinting and scanning. This is achieved by using the 'ping' and the 'traceroute' command to detect online hosts and to determine the path towards the target [44].

ICMP messages can be grouped into ICMP error messages and ICMP query messages. The two ICMP fields, which are stored in the experimental database, are ICMP Type (the message type) and the ICMP Code (which provides further details of ICMP message type). Table 4.4 shows the top ranking ICMP types observed across the five sensors. Below is the definition of ICMP types and their respective message type numbers [42][26]:

- Echo Reply (0) & Echo Request (8) are used to test for network connectivity through the 'ping' command.
- Destination Unreachable (3) message is sent when the destination is unreachable.
- Time Exceeded (11) message is sent when a host drops the packet due to time to live (TTL) being exceeded[7].

Results obtained in Section 4.1.2 showed that the percentage of ICMP packets was relatively small, ranging between 2.1% and 2.7%. Comparatively, sensors 146-a and 155-a have 12.8% and 8.9% of ICMP packets respectively. The difference between category A and category B's percentage share of the total packets is due to the larger TCP traffic observed in category A sensors. Results in Table 4.4 show that the composition of the ICMP packet types observed across all network telescope sensors is relatively similar, with a small difference observed with sensor 196-b. The similarity in ICMP packet counts, as well as the portion of each packet type, provides initial evidence of similarities of ICMP traffic observed in all sensors.

[7] Alternatively the time to live field will be zero when it is discarded

Table 4.4: ICMP traffic distribution across sensors

ICMP Type	196-a		146-a		155-a		196-b		196-c	
	Packets	%	Packets	%	Packets	%	Packets	%	Packets	%
Echo Request	387154	65.6%	357603	69.2%	380592	69.4%	207254	58.3%	355285	64.9%
Dest Unreachable	125340	21.3%	108461	21.0%	112396	20.5%	71800	25.8%	121762	22.3%
Time Exceeded	62535	10.6%	36778	7.1%	42809	7.8%	45185	12.7%	54777	10.0%
Echo Reply	10777	1.8%	12186	2.4%	11085	2.0%	9504	2.7%	10794	2.0%
Others	3926	0.7%	1603	0.3%	1304	0.2%	1877	0.5%	4556	0.8%
Total	**589732**		**516631**		**548186**		**335620**		**547174**	

Table 4.5: Central tendency and variation results using daily and hourly packet counts

	196-a		146-a		155-a		196-b		196-c	
	Daily	Hourly	Daily	Hourly	Daily	Hourly	Daily	Hourly	Daily	Hourly
Mean	69300	2898	10990	460	16900	705	66980	2791	70590	2941
Median	68460	2821	9973	321	15800	590	65230	2683	69060	2833
SDev	9574	998	6227	513	6057	566	10533	1271	9248	983

4.5 Basic Statistical Analysis

The mean and variance statistics are important to understand the outlay of datasets as they summarise the centrality and spread of data. In an effort to measure the central tendency of periodic packet counts, the mean and the median calculations are considered. An explanation of these statistical methods is contained in Appendix A. The median is also calculated to avoid the influence of excessive outliers on the mean.

Table 4.5 details the mean, median and standard deviation of each sensor at different period groupings of packets (daily and hourly). Results show that the mean of packet counts either daily or hourly, is slightly higher than the median. The difference between the mean and the median shows the influence of outliers. Similarity of mean values is observed between sensors 196-a, 196-b and 196-c. As visually observed in Section 4.1, sensors 196-a, 196-b and 196-c have a similar trend in traffic activity. The lower mean and median for sensor 146-a and 155-a are caused by lower total packet counts.

To measure how data is dispersed, variance and standard deviation calculations are used. Dispersion methods such as variance and standard deviation are important as they show how widely spread values can be from the mean. Similar to previous views on central tendency with the mean, sensors 196-a, 196-b and 196-c have similar standard deviation.

The box and whisker plots are considered to summarise the degree of spread of the daily

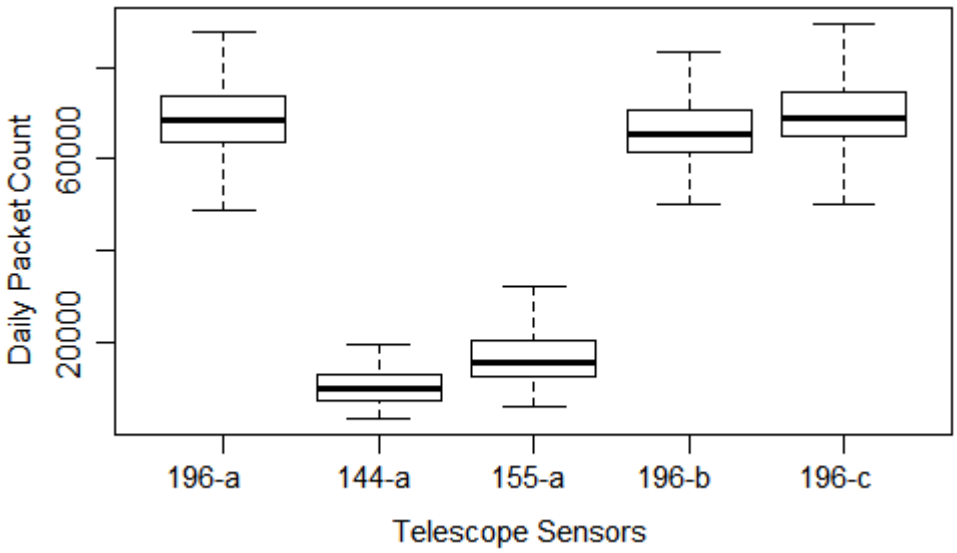


Figure 4.13: Box plot of daily packet count for all sensors

packet count parameter. Figure 4.13 shows the five box plots in one set of coordinates. The box plot, constructed using daily packet counts, displays an overview of the distribution of daily counts. The box plot also had outliers removed to suppress distortion caused by excessive outliers. The thicker black line shows the median and the 25th and 75th percentile are represented by the top and bottom of the "box" [30]. Maximum and minimum values (excluding outliers) are displayed by the horizontal lines above the "box" (or alternatively the whisker). Based on the box plot's position and spreads there is a clear distinction between category A sensors (196-a, 196-b and 196-c) and category B sensors (146-a and 155-a). The results obtained from the box-plots support the categories into which the sensors have been placed.

4.6 Source IP Address Analysis

In this section, an analysis of the source IP address, relative to the total packet counts, is provided with results captured in Table 4.6. A total of 74.3 million different packets were captured by the five network telescopes' sensors over the period of one year. The captured packets were sourced from 10.3 million different IP addresses. When analysing each network telescope node separately, it was observed that category A telescopes have similar packets per IP of 6.7 with a higher total packet count. However sensors 146-a and 155-a have 10.3 and 15.1 packets per IP respectively. The weakness of this metric is that IPs can be spoofed relatively easily and thus, this is not necessarily an accurate

Table 4.6: Distinct source IP addresses

	196-a	146-a	155-a	196-b	196-c	Total
Distinct source IP	3 813 944	392 077	409 762	1 908 204	3 815 196	10 339 183
Total packet count	25 362 068	4 023 238	6 184 900	12 861 095	25 835 932	74 267 233
Packet count / Distinct source IP	6,65	10,26	15,09	6,74	6,77	

representation of the actual distinct source IP address. Nonetheless, the average packet per IP measure and the observed impact of Conficker does provide grounds that category A telescopes are placed on a more hostile segment of the network space than category B ones.

4.7 Summary of Findings

This section summarises the findings of the chapter. Its main aim is to examine related findings highlighted within the chapter.

A number of periodic time series' were constructed to show network telescope traffic activity. Periodic packet count plots showed that category A sensors (196-a, 196-b and 196-c) had similarities of traffic activity. Category B sensors also showed similarities in traffic activity however less so when compared to category A results.

When analysing the major protocols, results showed that category A had a similar distribution of packets in major protocols (TCP, UDP and ICMP). TCP protocol was observed as dominant in category A with 90% of the traffic being TCP traffic. Results also showed similarities in the distribution of traffic across the major protocols in category B (146-a and 155-a). TCP traffic dominance was slightly reduced in category B sensors with 61% and 76% of all traffic being TCP traffic for sensor 146-a and sensor 155-a respectively.

In category A, TCP traffic's daily packet counts tracked the total packet counts (i.e. daily traffic activity) showing the dominance of TCP traffic. In category B, unlike category A, UDP and ICMP traffic caused fluctuations on the relativity of traffic activity showing the significance of UDP and ICMP when considering the activity of all traffic. Other protocols, apart from UDP, TCP and ICMP, accounted for an insignificant proportion (less than 0.02%) of traffic and, therefore, were not explored further.

Analysis of TCP traffic showed the following:

- Port 445/tcp is dominant in category A sensors accounting for more than 60% of all TCP traffic. Category B sensors had a reduced level of 445/tcp traffic with 22% for 146-a and 9% for 155-a.
- There was a significant prevalence of Conficker worm. It was observed that, due to algorithmic errors in the random number generator of the worm, there were IP address ranges that were unreachable. Category B sensors fell under the unreachable address range. The pervasiveness of Conficker was seen as a contributing factor to the large difference in packet counts between category A and category B sensors.

With UDP traffic as the second-largest contributor of traffic, the following results were obtained:

- Conducting port analysis showed the prevalence of SIP scanning and SQL Slammer.
- UDP traffic contained a number of rapid packet build-up (termed as anomalous spikes) which resembled distributed denial-of-service. Anomalous spikes were uncoordinated and directed towards a single sensor at a time.
- UDP port analysis showed that the top 10 ports observed in all sensors had 22 ports that were unique. These ports were suspected to contain anomalous spikes. Through manual investigation of the 22 ports: 14 were categorised as anomalous spikes, 5 required additional analysis and 3 were unlikely to be anomalous spikes. Anomalous spikes were uncoordinated and therefore caused deviations in relativity of UDP traffic's activity.

Analysis of ICMP traffic showed the following results:

- ICMP traffic was uniformly distributed across all the sensors (i.e. similar packet counts across all ICMP types).
- The prevalence of fingerprinting was clear due to the number of Echo Requests observed.

Basic statistical methods such as mean, median and standard deviations were also used to comparatively analyse the datasets. Daily and hourly traffic was analysed and it was observed that the median was generally lower than the mean. This is due to the influence of outliers. Comparatively, it was shown that category A sensors had similar values for

mean, median and standard deviation. Although there were similarities in the lower values being obtained in category B, there were slight disparities between the values of statistics generated. The box plots generated showed a clear disjunction between category A and category B sensors.

4.8 Summary

The approach followed in this chapter was to first conduct a comparative analysis of the traffic generated by all sensors. It focused on using summarisation and basic statistics techniques with results being presented in graphs and tables. Generated packet count series' plots showed similarities in traffic activity on category A telescope sensors. Similarities of traffic activity on category B telescopes were also observed but were not as correlated as category A. Conficker accounted for a significant proportion of packets in the category A sensors however, less so for category B sensors. UDP traffic analysis showed a significant amount of anomalous spikes that would cause deviations in relativity of UDP traffic. ICMP traffic analysis showed that traffic was distributed uniformly in different types of ICMP messages and across all sensors.

Initial comparative analysis provides a foundation for more advanced cross-correlation analysis experiments. The next chapter uses time series' to quantitatively analyse correlation of traffic activity.

Chapter 5

Advanced Correlation Analysis - Time Series

The results shown in Section 4.1 support the basis on which the five network telescopes sensors have been categorised. Through summarisation and basic numeric statistics techniques, it has been observed that there are similarities in sensor traffic activity between category A telescope sensors (196-a, 196-b and 196-c). Likewise, though to a lesser extent, there are also similarities between category B telescope sensors (146 and 155), but these are distinctly different from category A.

Time series can be constructed by using variables with a fixed time period between observations. For example, daily temperatures can be used to construct a time series with a fixed 24-hours time period. Time series are used in the field of Economics to model parameters such as Gross Domestic Product, Consumption and Gross National Investment over a period of time [47]. Time series can also be used to forecast economic indicators. This can be achieved by looking at a leading indicator or variable that precedes and correlates to another variable being investigated. With a higher correlation between the leading indicator and the investigated variable, better accuracy in forecasting the investigated variable is achieved.

Traffic captured using network telescope sensors can also be represented in a time series. The reason for this is that packets are associated with a datetime stamp. The datetime stamp field, found in the database schema, records the time at which a packet is received. Having a time-based packet capture system allow one to attach observations (packet counts) to a specific period (either daily, monthly or hourly). Constructed time series

allows the researcher to conduct advanced correlation analysis between multiple variables or simply between the same variable at different time lags. Activity plots (daily and hourly packet counts) used previously were generated by graphing the time series that were generated from network telescope's traffic.

5.1 Interpreting the Results

In conducting a correlation analysis of network telescope activity, the correlation coefficient for auto-correlation and cross-correlation functions was used. As discussed in Section 2.9, auto-correlation and cross-correlation methods are used to test for correlations in a generated time series. The correlogram plots showing correlation coefficients at different time lags are also used to interpret correlation results. Correlation coefficients are used to quantitatively assess and compare the degree of relativity between two generated time series variables.

5.1.1 Cross-correlation Coefficient and the Auto-correlation Coefficient

The coefficient (r) of both auto-correlation and cross-correlation functions is a calculated value bound between -1 and 1. The value of $r = \pm 1$ represents a perfect correlation, that is, either a positive or a negative relationship. If the coefficient is zero, it indicates that there is no correlation or relationship between the variables. Because there are different categorisations of correlation coefficients, it can be difficult to interpret and categorise correlation coefficient values. However, coefficients can generally be categorised as follows: $r \leq 0.35$ is usually considered weak or low correlation; $0.36 \leq r \leq 0.67$ is considered moderate correlation; and $0.68 \leq r \leq 1$ is strong or high correlation [48]. In this regard, Table 5.1 shows full categorisation of correlation coefficients that were adopted in this research project.

Although not ideal, one can employ a technique that computes cross-correlation and auto-correlation of two variables with missing points, such as simply passing over the missing values [24]. The challenge with passing over values, especially with regard to the auto-correlation of time series, is that time intervals become distorted with missing values. Depending on the level of granularity of the time series, short disruptions such as missing hourly packet capture can be expected. Disruptions can be caused by various network

Table 5.1: Categories of correlation coefficients

Category	Correlation coefficient values
weak positive (negative) relationship	0 to 0.35 (0 to -0.35)
moderate positive (negative) relationship	0.36 to 0.67 (-0.36 to -0.67)
strong positive (negative) relationship	0.68 to 1.0 (-0.68 to -1.0)

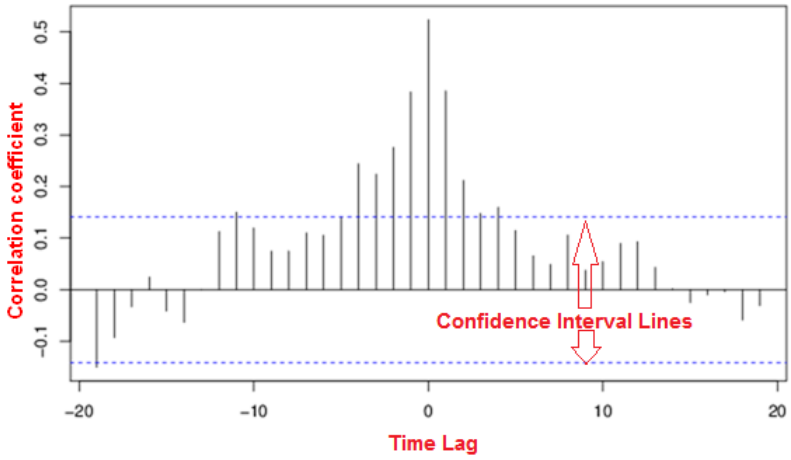


Figure 5.1: Correlogram example

interruptions such as outages borne by the service provider. In this regard, in analysing hourly time series, a passing of missing values method is used. At a daily interval, a full-set of data points is available for the period under investigation (20 May 2011 to 20 May 2012) across the four telescopes. Moreover, as noted previously, the researcher has six months of data points for telescope 196-b.

5.1.2 The Correlogram

Part of correlation analysis requires an understanding of the correlogram used to plot both the auto-correlation function and cross-correlation function. The correlogram shows the distribution of the correlation coefficient within the maximum specified time lag.

The correlogram plot (example shown in Figure 5.1) shows the time lag on the x-axis and the correlation coefficient values on the y-axis. For the purposes of this research, the lag is the time period between series values and it can either be daily or hourly lag based on the time period under investigation. For example, when analysing the daily packet count, the lag will be a day. While similarly, for an hourly packet count the lag will be an hour. The correlogram is constructed with a 95% confidence interval. The confidence interval is

Table 5.2: Confidence intervals for hourly and daily correlograms

	daily time series	hourly time series
Year	[-0.1;0.1]	[-0.02;0.02].
Six Months	[-0.14;0.14]	[-0.023;0.023]

shown by dotted lines on the correlogram and highlights the reliability of the coefficient's estimate. In the correlogram, if the correlation coefficient lies outside the dotted lines then the null hypothesis of the coefficient being zero can be rejected [8]. If the correlation coefficient is zero then, the correlogram is normally distributed with a mean of $-\frac{1}{n}$ and variance of $\frac{1}{n}$. The 95% confidence interval of a normal distribution is given by the mean ± 2 standard deviations [8]. In this regard, the 95% confidence interval is calculated using the following formula:

$$\frac{1}{n} \pm \frac{2}{\sqrt{n}} \tag{5.1}$$

Using Equation 5.1, the confidence intervals (the lower confidence limit interval and upper confidence interval limit) are presented in Table 5.2 for various time periods (n). To calculate the confidence interval using a daily packet count time series, n will equal 365 days to account for a full year. Looking at hourly packet count time series, the value of n will be 8760. Confidence intervals for a six-month dataset are shown to accommodate telescope sensor 196-b which contains six months' worth of data. The confidence intervals are represented by dotted lines on the correlogram.

For auto-correlation, lag 0 is always 1 since auto-correlation analysis investigates the correlation of a variable with itself in different time lags. At time lag 0 the function auto-correlates with itself.

5.2 Long-range Correlation Analysis

Before investigating the correlation between two time series, it is important to first establish whether there are repeating patterns on the time series constructed from the sample dataset. Therefore, to investigate long-range (auto) correlation (i.e. whether the time series cross-correlates with itself in two different points in time), the researcher makes use of the auto-correlation function.

5.2.1 Auto-correlation Analysis

The auto-correlation function is extended from a covariance function. Covariance functions are used to study correlation between two variables. Given two variables (x and y), the sample covariance is defined as follows [23]:

$$Cov(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \{\bar{y}\})}{n - 1} \tag{5.2}$$

with $\bar{x}$and $\bar{y}$ defined as the sample mean:

$$\bar{x} = \frac{\sum x_i}{n}; \bar{y} = \frac{\sum y_i}{n} \tag{5.3}$$

Following from this, the sample correlation coefficient (ρ) is defined as follows:

$$\rho = \frac{Cov(x, y)}{sd(x)sd(y)} \tag{5.4}$$

The sample standard deviation is the square root of the variance and, as such, sd (x) and sd (y) are defined as follows:

$$sd(x) = \frac{\sum(x_i - \bar{x})^2}{n - 1}; sd(y) = \sum(y_i - \bar{y})^2 \tag{5.5}$$

Therefore, to calculate the auto-correlation covariance of a time series (x) with (t) as the time period of the series and (k) as the lag between the series, the following function is used [8]:

$$c_k = \frac{1}{n}\sum_{t=1}^{t-k}(x_t - \bar{x})(x_{t+k} - \bar{x}) \tag{5.6}$$

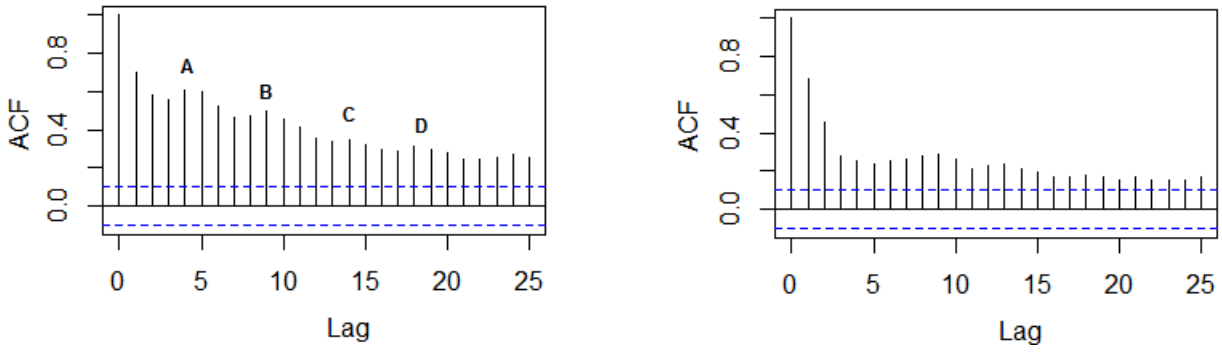
By normalising c_k, researchers can compute the auto-correlation coefficient with bound values between -1 and 1. The normalising function is simply c_0 (cross correlation covariance at 0 time lag) [8]. The auto-correlation coefficient is defined as:

$$\rho = \frac{c_k}{c_0} \tag{5.7}$$

Listing 1 Auto-correlation function definition

```
acf(x, lag.max = NULL, type, plot = TRUE, na.action  ...)

x          -> univariate or multivariate variable
lag.max    -> maximum lag to test for cross correlation
plot       -> if set to true, a plot is produced
na.action -> function to handle missing values (na.fail/na.pass)
```



155-a 146-a

Figure 5.2: Auto-correlation correlograms of sensors 155-a and 146-a using daily packet counts

To determine the auto-correlation function and produce the plots, R Statistic's "acf" function is used. The "acf" function is described in Listing 1[1] [51] .

The correlogram is used to check if the series is correlated with itself in different time lags. The auto-correlation coefficient is plotted on the y-axis and the x-axis represents the time lags. Essentially, if the series is has no repeating patterns, the auto-correlation coefficient at time lags greater than 0 should be close to zero.

By using the auto-correlation method to analyse the generated series, results showed that the daily packet counts time series across category A do not show repeating patterns beyond a lag period of more than 1, as observed auto-correlation coefficients are less than 0.36 at time lag greater than 1. The auto-correlation correlograms of 196-a, 196-b and 196-c are contained in Appendix D. Sensor 155-a has a particularly interesting correlogram (shown in Figure 5.2); as it displays a gradual decay with moderate correlation coefficient and repeated peaks marked with alphabets ("A", "B", "C" and "D"). The moderate correlation coefficient (until lag 11 for sensor 155-a) and the gradual decay are an indication of repeated patterns [8]. Sensor 146-a (Figure 5.2) does show repeated peaks,

[1]Definition of parameters that are used in this research

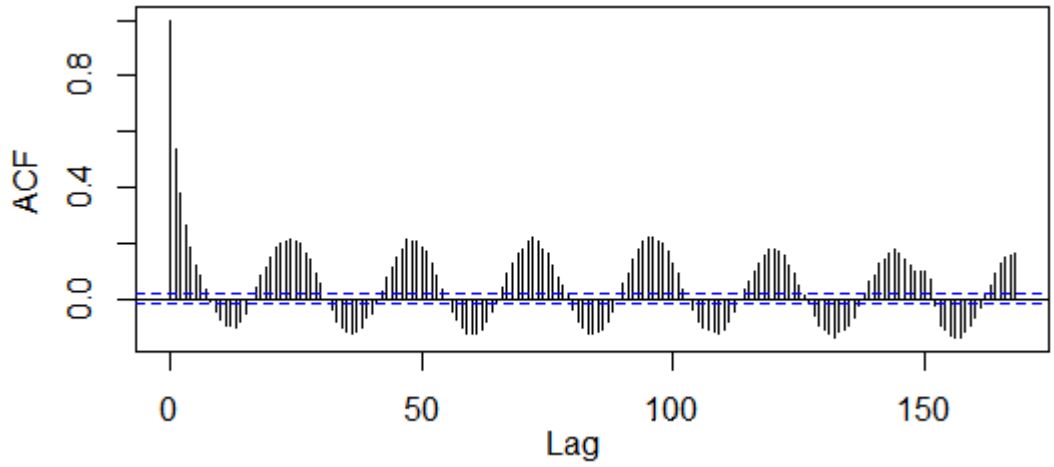
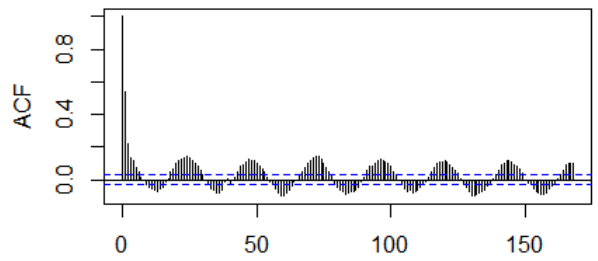


Figure 5.3: Auto-correlation correlogram of sensor 196-a using hourly packet count

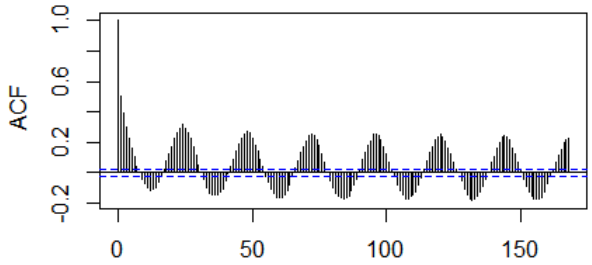
however, beyond the time lag of 2 the correlation coefficient is weak.

When conducting an analysis of the hourly packet count time series, although weak correlation coefficients are obtained, it is observed that there are repeating and diminishing patterns every 24 hours with correlation coefficients declining as the time lag increases. This is illustrated by Figure 5.3 which shows a correlogram with a maximum lag of 168 hours (this represents a week's worth of lag). Results presented in Figure 5.4 show that the 24-hourly pattern is observed across category A telescopes sensors thereby providing evidence of 24-hour cyclical network telescope traffic activity.

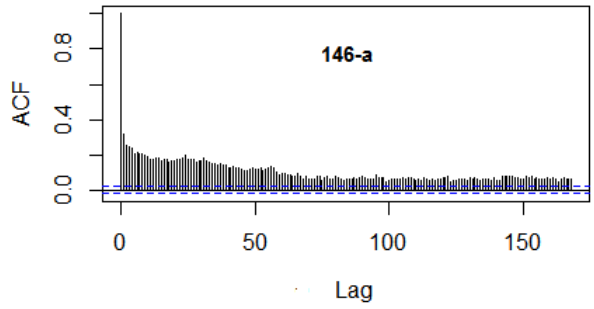
Category B sensors simply demonstrate a decline of correlation coefficients with no 24-hourly cycles. Category B sensors, particularly sensor 155-a, show evidence of repeated daily traffic patterns. However, category A sensors show repeated 24-hour cycles with weak correlation coefficient observed. The 24-hourly cycles are related to Conficker worm scanning as shown in Figure 5.5. The correlogram shows Conficker-related traffic only (i.e. port 445/tcp) using the hourly packet count of sensor 196-a. Other category A sensors (196-b and 196-c) show similar results and can be viewed in Appendix E. The correlation coefficients at the peaks of the correlogram are in the strong category showing that the time series auto-correlates in 24-hourly intervals. The actual auto-correlation coefficients are also contained in Appendix E.



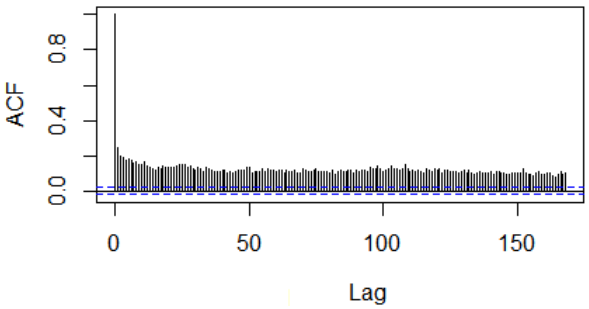
196-b



196-c



146-a



155-a

Figure 5.4: Auto-correlation correlogram of sensors 196-b, 196-c, 146- and 155-a using hourly packet count

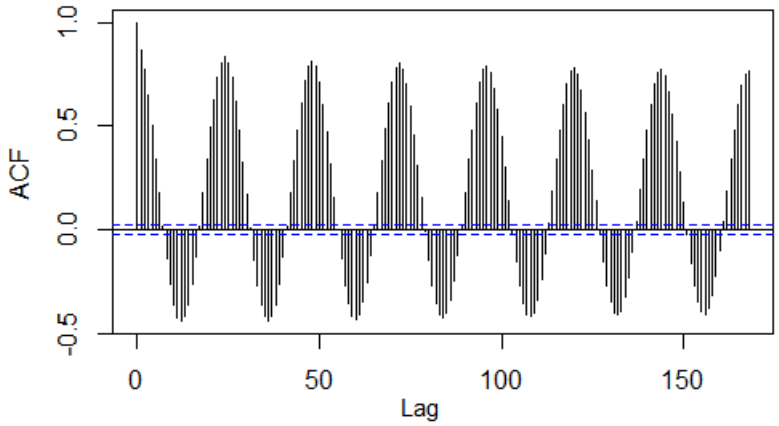


Figure 5.5: Auto-correlation correlogram of sensor 196-a using hourly packet counts on port 445/tcp

5.3 Cross - Correlation Analysis

Having multiple telescope sensors, allows the researcher to access multiple series of data captured across the same period of time. As highlighted in Section 2.9, a cross-correlation method is used to investigate correlation (i.e. estimate the degree to which two time series are related) across multiple variables. To conduct a cross-correlation analysis, a number of parallel time series across the telescope sensors datasets are constructed.

The cross-correlation function can be seen as an extension of the auto-correlation function 5.6 in Section 5.2. Instead of having the same variable with different time lags; now there are two time series variables. Therefore, given two time series (x and y), the cross-correlation covariance function can be defined as follows [8]:

$$c_k(x, y) = \frac{1}{n} \sum_{t=1}^{n-k} (x_{t+k} - \bar{x})(y_t - \bar{y}) \tag{5.8}$$

Time lag (denoted as k on Equation 5.8) is important when examining logically distant network telescopes, taking into account the network delays and the proximity of the nodes to source machines. This ensures that a test for cross-correlation can be conducted even with lags between the sensors.

Similar to auto-correlation, the cross-correlation coefficient is calculated by normalising the auto-correlation covariance $c_k(x, y)$ such that the coefficient is bound between -1 and 1 (similar to auto-correlation covariance). The normalising function is defined as follows:

$$z = \sqrt{c_0(x, x) c_0(y, y)} \tag{5.9}$$

Therefore, cross-correlation coefficient is defined as follows:

$$\rho = \frac{c_k(x, y)}{z} \tag{5.10}$$

R Statistical package has an implementation of the cross-correlation function ("ccf"), which is used to calculate the cross-correlation coefficient of two variables. R Statistics' implementation of cross-correlation is described in Listing 2 [51]:

Listing 2 Cross-correlation function definition

```
ccf (x, y, plot, na.action,...)

x,y        -> Univariate variable (numeric vector or matrix)
plot       -> If set to true, a plot is produced
na.action -> Function to handle missing values (na.fail/na.pass)
```

5.3.1 Daily Packet Counts - Time Series

Two time series are constructed to cater for variables X_1 and X_2. Since the daily packet counts' time series does not contain missing values, the 'na.action' is left with default values such that the computation halts in the case of missing values. The experiment tests cross-correlation against a combination of telescope sensors' traffic. Results in cross-correlation analysis will also be presented in a cross-correlation matrix. The cross-correlation matrix will be presented in the following syntax:

$$(+|-)coefficient(l = x) \star \{\star\} \tag{5.11}$$

Where:

- +/- sign indicate whether or not there is a positive relationship (positive or negative coefficient)
- x is the maximum coefficient achieved
- rating can either be:
  - $\star$ for weak correlation;
  - $\star\star$ for moderate correlation;
  - $\star\star\star$ for strong correlation; or
  - blank for values below 95% confidence interval.

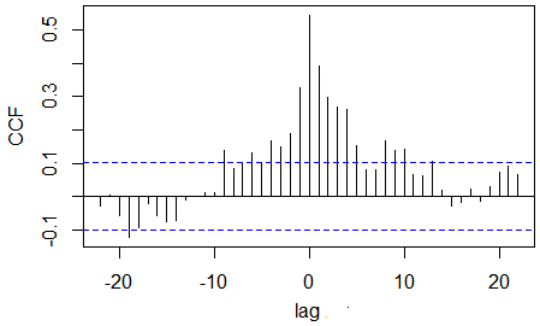
The cross-correlation results contained in Table 5.3 supports the previously observed results, achieved through summarisation and basic statistics, in addition to the graphical traffic activity plots. Results show that sensors 196-a, 196-b and 196-c (defined as category A) have a moderate correlation coefficient ranging from 0.45 to 0.55. There is relativity

Table 5.3: Cross-correlation matrix - daily packet count

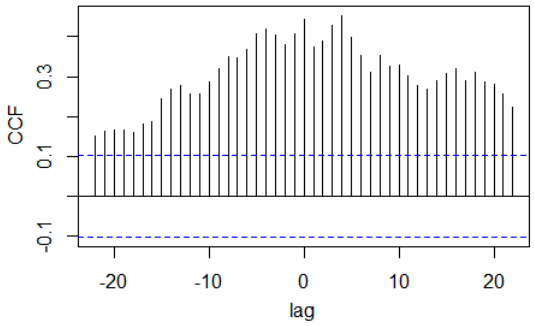
	196-a	146-a	155-a	196-b
146-a	- 0.132 ($l = 5$) $\star$	——	——	——
155-a	+0.200 ($l = 0$) $\star$	+0.454 ($l = 4$) $\star\star$	——	——
196-b	+0.523 ($l = 0$) $\star\star$	+0.407 ($l = 0$) $\star\star$	+0.274 ($l = 13$) $\star$	——
196-c	+0.546 ($l = 0$) $\star\star$	- 0.126 ($l = 13$) $\star$	+0.233 ($l = 0$) $\star$	+0.450 ($l = 0$) $\star\star$

between sensors 146-a and 155-a (defined as category B) with a cross-correlation coefficient of 0.44. However, the cross-correlation coefficient between telescope nodes in category A and category B is considered weak for all experiments except for the correlation between 146-a and 196-b, which demonstrates a moderate relationship.

Figure 5.6 illustrates correlograms for category A comparison (196-a vs. 196-c) as well as category B comparison (146-a vs. 155-a). For a comparison of sensor 196-a vs. sensor 196-c, the diagram shows a distinct peak at time lag of 0. In contrast, the comparison of sensor 146-a vs. sensor 155-a shows gradual peaks, not a distinct peak, with two slightly close peaks at time lag of 0 and time lag of 4. At time lag 0, the cross-correlation coefficient is 0.444 and at time lag 5 the coefficient is 0.454 (slightly higher). Between the time lag of -6 and +5, all coefficients are in the moderate category. This indicates that the two series, at specified time lags, correlate uniformly.



196-a vs. 196-c



146-a vs. 155-a

Figure 5.6: Cross-correlation correlogram of sensor 196-a vs. 196-c (A) and sensor 146-a vs. 155-a (B) using daily packet count

Considering the amount of interplay of events that occurs because of sensors and general Internet clutter, the results achieved provide evidence of relativity of traffic in both

categories. In addition to the time series correlation analysis, when profiling gathered datasets in Section 3.2, the researcher observed significant similarities between category A sensors in the following areas:

- Total packet counts for the 12-month period were identical.
- Although logically distant, all three telescope sensors were placed in the same prefix.
- The neighboring IP address blocks were smaller and, predominantly, in developing countries.

These factors do seem to provide additional evidence for the cross-correlation of sensors in category A. Category B's similarities can be attributed to the following (also highlighted in Section 3.2):

- Similar and lower (when compared to category A) total packet counts across the 12-month period.
- Larger neighboring IP block assignments on the same high-order prefix as category B's sensors.
- Earlier provision of IP addresses (compared to category A) with larger apportions of IP space.
- Telescopes IP ranges outside Conficker's range due to a fault in the random number generator.

Apart from examining all traffic without distinction, further correlation analysis of different types of traffic is required. This will ensure that correlation tests can be conducted based on major traffic types (TCP, UDP and ICMP) and, therefore, one can test to see which traffic types achieve better correlation between sensors.

5.3.2 TCP Traffic Time Series

In a previous analysis, the first series constructed was the daily packet count by simply using a bin size of 24 hours and counting the number of packets received in each bin. To study the underlying variables, the daily packet count was filtered to consider TCP packets only. Other traffic types will be examined in the sections that follow. The newly generated

Table 5.4: Cross-correlation matrix - daily TCP packet count

	196-a	146-a	155-a	196-b
146-a	-0.224 ($l = 9$) $\star$	——	——	——
155-a	+0.287 ($l = 0$) $\star$	+0.443 ($l = 4$) $\star\star$	——	——
196-b	+0.820 ($l = 0$) $\star\star\star$	+0.340 ($l = 5$) $\star$	+0.278 ($l = 16$) $\star$	——
196-c	+0.760 ($l = 0$) $\star\star\star$	-0.242 ($l = -10$) $\star$	+0.321 ($l = 0$) $\star$	+0.860 ($l = 0$) $\star\star\star$

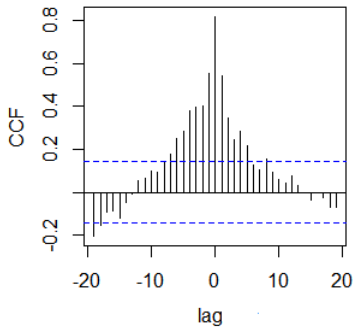
datasets of only TCP packets went through similar cross-correlation experiments. The results are presented in Table 5.4.

The first thing to observe is the movement from moderate to strong cross-correlation coefficients with a maximum value of 0.86 between sensor 196-b and sensor 196-c. In Section 4.1.2, results showed that sensors 196-a, 196-b and 196-c TCP packet proportions were above 90%. By removing less than 10% of the traffic, the cross-correlation results have improved significantly. Looking at category B (146-a and 155-a), it is observed that the sensor's relativity is still categorised as moderate with a minute decline of the coefficient: from 0.454 with all traffic to 0.443 with only TCP traffic.

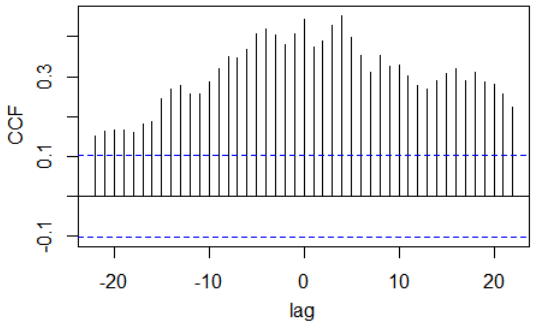
Figure 5.7, shows the correlogram of sensor 196-a vs. sensor 196-b as well as sensor 146-a vs. sensors 155-a. An initial observation of sensor 196-a vs. sensor 196-b reveals that the correlogram resembles a normally distributed function. There is a clear distinct peak at the time lag of zero. The correlogram of 146-a vs. 155-a shows a gradual decline as the time lag increases with two high peaks at lag 0 and lag 4. The double peaks are similar to the experiment conducted earlier that showed all traffic types. Appendix G contains other comparisons of the sensor combination.

Table 5.5: Cross-correlation matrix - daily TCP packet count with SYN flag on

	196-a	146-a	155-a	196-b
146-a	-0.237($l = 9$) $\star\uparrow$	——	——	——
155-a	+0.262 ($l = 0$) $\star\downarrow$	+0.448 ($l = 4$) $\star\star\uparrow$	——	——
196-b	+0.813 ($l = 0$) $\star\star\star\downarrow$	+0.136($l = 5$) $\star\downarrow$	+0.272 ($l = 16$) $\star\downarrow$	——
196-c	+0.754 ($l = 0$) $\star\star\star\downarrow$	-0.250 ($l = -10$) $\star\uparrow$	+0.293 ($l = 0$) $\star\downarrow$	+0.858 ($l = 0$) $\star\star\star\downarrow$



196-a vs. 196-b



146-a vs. 155-a

Figure 5.7: Cross-correlation correlograms of sensor 196-a vs. 196-b (A) and 146-a vs. 155-a (B) using daily TCP packet count

With improved correlation results obtained by only analysing TCP traffic, the next experiment focuses on TCP packets that have a SYN flag set. Packets with SYN flag set are classified as active, meaning that a response would be required from the network telescope [16]. Therefore, the aim of the experiment is to test if there are further improvements to correlation by refining TCP traffic. Table 5.5 shows the cross-correlation matrix of TCP packets with the TCP SYN flag set. Arrows are used to indicate whether results have improved when compared to previous results illustrated in Table 5.4 (all TCP traffic). Results highlight that cross-correlation coefficients using active TCP packets are marginally affected. For example, the cross-correlation coefficient for the comparison of sensor 196-b vs. sensor 196-c has decreased from 0.8600 to 0.858. The reason for this is that inactive TCP packets' total share of all TCP packet count is small (1.5% for all traffic observed in all five sensors) and, therefore, a lesser influence on the total daily packet count (TCP only) variable. Although it is important to consider active packets when analysing malicious activity, by removing inactive traffic the results show only a marginal effect on correlation of traffic.

Table 5.6: Cross-correlation matrix - daily UDP packet count

	196-a	146-a	155-a	196-b
146-a	+0.187 ($l = -17$) ⋆	———	———	———
155-a	+0.098 ($l = 0$)	+0.336 ($l = 4$) ⋆	———	———
196-b	+0.348 ($l = -1$) ⋆	+0.570 ($l = -1$) ⋆⋆	+0.276 ($l = -5$) ⋆	———
196-c	+0.053 ($l = -19$)	+0.103 ($l = -3$) ⋆	+0.105 ($l = 0$) ⋆	+0.062 ($l = -4$)

Table 5.7: Cross-correlation matrix - daily UDP packet count

	196-a	**146-a**	**155-a**	**196-b**
146-a	+0.609 ($l = 0$) ⋆⋆	———	———	———
155-a	+0.653 ($l = 0$) ⋆⋆	+0.699 ($l = 0$ ⋆⋆⋆	———	———
196-b	+0.818 ($l = 0$) ⋆⋆⋆	+0.330 ($l = 0$) ⋆	+0.422 ($l = 0$) ⋆⋆	———
196-c	+0.853 ($l = 0$) ⋆⋆⋆	+0.676 ($l = 0$) ⋆⋆⋆	+0.711 ($l = 0$) ⋆⋆⋆	+0.810 ($l = 0$) ⋆⋆⋆

5.3.3 UDP Traffic Time Series

Section 4.1.2 demonstrated that UDP accounts for less than 10% of total packet share across category A. For category B, it is noted that UDP accounts for a larger share with 15% and 25% for sensor 155-a and sensor 146-a respectively.

Table 5.6 shows the results of cross-correlation experiments with UDP packets constructed using daily packet counts. Results demonstrate that there is a weak correlation between sensors in category A. Correlation between sensor 146-a and sensor 155-a, in category B, is also considered weak. As an anomaly, the cross-correlation coefficient of sensor 196-b vs. sensor 146-a falls in the moderate category. Looking at a correlogram in Figure 5.8, which shows the cross-correlation function between sensor 146-a and sensor 155-a, it is observed that the correlogram has a number of peaks and troughs. Results suggest that UDP traffic is less relative as compared to the previous experiment on TCP traffic.

During the UDP port analysis section the presence of uncoordinated anomalous spikes was examined in detail. It evident that there was a significant presence of these spike attacks to uncommon UDP ports with significant packet build-up in a short space of time that targeted a single sensor. These random sensor attacks were uncoordinated across the telescopes and are responsible for the lower correlation results achieved in this experiment.

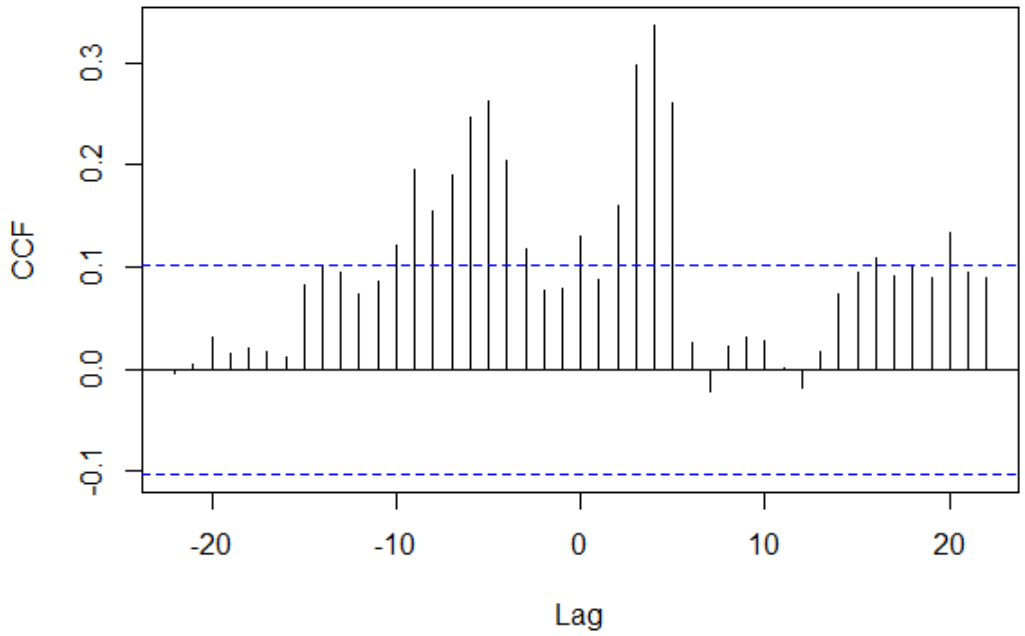
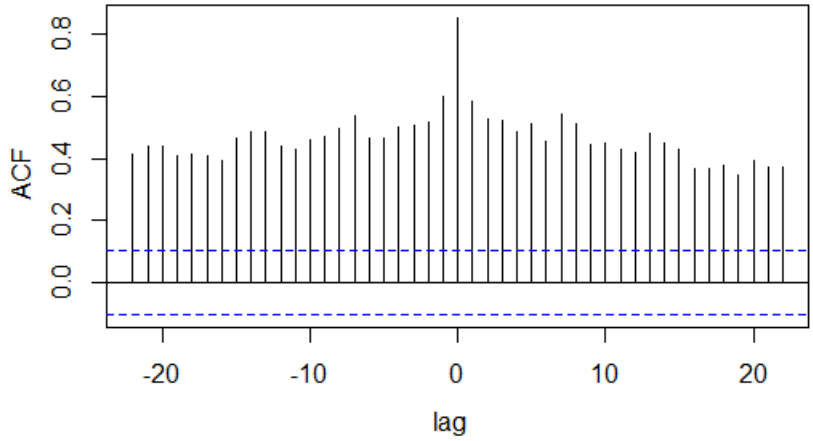


Figure 5.8: Cross-correlation correlograms of sensor 146-a vs. 155-a using daily UDP packet count

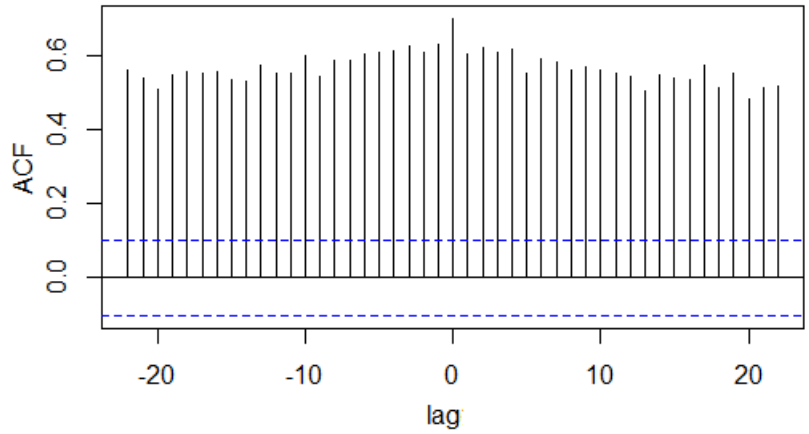
5.3.4 ICMP Traffic Time Series

Figure 5.7 contains the cross-correlation analysis of ICMP traffic The results show moderate to strong correlation across all network telescope sensors' combinations except for sensor 146-a vs. sensor 196-b. These results were consistently obtained regardless whether the analysis occurred in a single category or across categories. A peak at the time lag of zero can be observed, however, the correlograms of all sensors demonstrated uniformity across the different time lags. Figure 5.9 shows the cross-correlation function of sensor 196-a vs. 196-c (category A) and sensor 146-a vs. 155-a (category B). Both correlograms shows the highest coefficient at time lag 0 but the adjacent lags are all in the moderate category. Appendix I contains correlograms of other comparisons.

The shape of the correlograms constructed provide evidence of uniform traffic being observed across the sensors. Similarly, the moderate to high correlation coefficient observed indicates that, irrespective of category, ICMP traffic shows that there are similarities across the sensors.



196-a vs 196-c



146-a vs 155-a

Figure 5.9: Cross-correlation correlograms of sensor 196-a vs. 196-c and 196-a vs. 146-a using daily ICMP packet count

Table 5.8: Cross-correlation matrix - daily packet count for traffic without port 445/tcp

	196-a	**146-a**	**155-a**	**196-b**
146-a	+0.336 ($l = -5$) ⋆ ↑	———	———	———
155-a	+0.541 ($l = 0$) ⋆⋆ ↑	+0.446 ($l = 0$ ⋆⋆ ↑	———	———
196-b	+0.872 ($l = 0$) ⋆⋆⋆ ↑	+0.185 ($l = 5$) ⋆↓	+0.298 ($l = 0$) ⋆ ↑	———
196-c	+0.843 ($l = 0$) ⋆⋆⋆ ↑	+0.381 ($l = 5$) ⋆⋆ ↑	+0.586 ($l = 0$) ⋆⋆ ↑	+0.872 ($l = 0$) ⋆⋆⋆ ↑

5.3.5 Non-445/TCP Time Series

Section 4.2 demonstrated the influence of port 445/tcp (used by Conficker) and its impact on sensor traffic activity. In previous plots showing traffic without TCP traffic, relativity was visually observed across all sensors. This section conducts a cross-correlation analysis of non-Conficker traffic. This is achieved by removing all traffic targeting port 445/tcp and then constructing a time series. This experiment was conducted with a view to study how relative traffic would be without the influence of Conficker and related 445/tcp scanning.

The results of cross-correlation calculations are captured in Table 5.8. Arrows are used to show the comparison between results obtained in this experiment and the initial results captured in Figure 5.3. When comparing the initial results obtained, with all traffic types and traffic without port 445/tcp traffic, improved correlation results across all telescope sensor combination were obtained, with the exception of sensor 146-a vs. sensor 196-b. Category A sensors achieved the highest correlation coefficient ranging between 0.843 and 0.872. Although results show three weak correlation comparisons, it is also important to note that the other seven comparisons are in medium to strong correlation irrespective of category. It is also significant that, even with weak categorisations, the coefficients are not as low as the previous analysis. Sensor 146-a and 155-a (category B) are outside Conficker's range and the results of this experiment affirm this. In this regard, results show only a minimal improvement of the cross-correlation coefficient from 0.444 (obtained in previous results) to 0.446 for sensor 146-a vs. sensor 144-a comparison.

5.3.6 Time Series - Destination IP Address

Previous work on 256 IP address block (small /24 telescope sensors) conducted has shown a clear disjunction on packet count per IP for destination IP's below x.x.x.127 and those greater than x.x.x.127 [16]. Similarly, using annual datasets between the selected period (20 May 2011 to 20 May 2012) in Appendix H, plots show the distribution of packets across IP addresses.

Table 5.9: Cross-correlation matrix - traffic with destination IP address above x.x.x.127

	196-a	146-a	155-a	196-b
146-a	+0.348 ($l = 0$) ⋆	——	——	——
155-a	+0.460 ($l = 0$) ⋆⋆	+0.556 ($l = 0$) ⋆⋆	——	——
196-b	+0.698 ($l = 0$) ⋆⋆⋆	+0.141 ($l = 5$) ⋆	+0.254 ($l = 11$) ⋆	——
196-c	+0.686 ($l = 0$) ⋆⋆⋆	+0.321 ($l = 0$) ⋆	+0.419 ($l = 0$) ⋆⋆	+0.810 ($l = 0$) ⋆⋆⋆

This experiment extracts the upper half of the /24 IP space (looking for IP addresses between x.x.x.127 and x.x.x.255) and looks at daily packet counts across network telescope sensors. Since it is known that the selected destination IP addresses are outside of Conficker's automated scanning range, it is possible to remove the influence of automated scans [18]. This is different from the experiment in Section 5.3.5 as the unreachable addresses were simply excluded while still observing other traffic routing to 445/tcp. This allowed the researcher to look at all traffic across all ports and packet types while focusing on a limited destination address range.

The results of the cross-correlation experiment are shown in Figure 5.9. Results show a strong correlation between category A sensors. Although correlation between sensors in category B is moderate, improvements can be observed with a cross-correlation coefficient of 0.556. These results indicate that, with Conficker's automated scans removed, relativity of sensors in each category is observed.

5.4 Hourly Packet Count Analysis

Thus far research has been conducted using daily bin sizes for correlation analysis. In this section the researcher aims to reduce granularity and look at hourly bin sizes to conduct cross-correlation calculations. The challenge with hourly bin sizes is that there are missing values and, as previously indicated, cross-correlation analysis is sensitive to missing values thus one needs to treat missing values.

Table 5.10 shows the hourly outages for the entire selected year period[2]. Due to the vast amount of packets that sensors receive daily, a conservative assumption was made: should the network telescope sensor not receive a packets for an entire hour; it is assumed that there was an outage. These minimal hourly outages can be attributed to a number of factors such as regional outages with service providers. No telescope sensor (apart from

[2]Considering six months period for Sensor 196-b

Table 5.10: Telescope sensor's hourly downtime for the 12 month period

Telescope	Total downtime (hours)	Up-time
196-a	31 hours	99.65%
146-a	39 hours	99.57%
155-a	12 hours	99.86%
196-b	0 hours	100%
196-c	1 hour	99.99%

Table 5.11: Cross-correlation matrix - hourly packet count for all packet types

	196-a	146-a	155-a	196-b
146-a	+0.029 ($l = 5$) $\star$	——	——	——
155-a	+0.089 ($l = 0$) $\star$	+0.127 ($l = 0$) $\star$	——	——
196-b	+0.364 ($l = 0$) $\star\star$	+0.219($l = -32$) $\star$	+0.067 ($l = 0$) $\star$	——
196-c	+0.320 ($l = 0$) $\star$	- 0.027 ($l = 10$) $\star$	+0.089 ($l = 1$) $\star$	+0.320 ($l = 0$) $\star$

196-b) had a full day outage. The hours reported on the table are not contiguous but rather spread across the entire period. In a full year (with 8785 hours), the highest outage experienced in one sensor was 39 hours. A total of 83 hours of downtime in all sensors was observed.

By default, missing values are not allowed in computing cross-correlation coefficients. However, it is possible to simply pass through missing values by setting "na.pass" on the 'na.action' input parameter. Table 5.11 shows the hourly cross-correlation analysis of the five sensors. The first observation is that all sensors, except 196-a vs. 196-c, have weak correlation results. Furthermore, when comparing hourly packet counts with daily packet counts of all packet types, the cross-correlation coefficients were significantly lower. Although the number of missing values does play a role on correlation coefficient, these results obtained show that increasing the resolution and looking at hourly traffic activity does not yield better correlation results.

5.5 Summary of Findings

This chapter focuses on using advanced correlation analysis to test for relativity in sensor activity. Having traffic captured with related time stamps allowed the researcher to construct a number of time series. Auto-correlation and cross-correlation of time series

analysis were used. The results of the correlation analysis were presented with correlograms and their respective cross- or auto-correlation coefficients.

5.5.1 Auto-correlation Results

Using the auto-correlation technique to determine whether a time series correlates with itself in different time lags (i.e. repeated patterns), it was observed that category A sensors have weak auto-correlation (i.e. no repeating patterns) using daily packet counts. Sensor 155-a showed a particularly interesting repeated decaying pattern with strong to moderate auto-correlation coefficient until a time lag of 11. Sensor 155-a also had a similar decaying pattern, however, the cross-correlation coefficients were much lower and at a time lag beyond 2 they were weak. Similarly, looking at an hourly packet count, the correlograms for category A showed repeated and diminishing cycles every 24 hours, although the correlation coefficient beyond time lag of 2 was weak. Looking at port 445/tcp, 24-hourly cycles were observed with strong auto-correlation coefficients every 24 hours thereby showing evidence of Conficker's cyclical scanning patterns.

5.5.2 Cross-correlation Results for All Traffic

Following from auto-correlation experiments, the cross-correlation method was used to test for correlation of two variables (i.e. two time series). Initial results, looking at packet counts of all packet types, affirmed the results shown in Section 4.1 and indicated the relativity of network traffic activity across category A sensor's combination. Similarly, category B sensors showed relativity. The cross-correlation coefficient fell in the moderate bracket in this experiment for comparisons in each category. Although these results did not initially contain coefficients greater than 0.67 (i.e. strong correlation), given the amount of interplay and anomalous spikes, the results were encouraging and allowed the researcher to study the underlying variables further. When comparing combinations of category A sensors with category B sensors, results showed weak cross-correlation.

5.5.3 Cross-correlation Results for Major Protocols

Having statistically demonstrated evidence of cross-correlation across category A and category B sensors, an investigation of the underlying similarities was conducted by looking at different types of traffic separately.

Looking at TCP traffic only, category A TCP cross-correlation analysis results were in the strong bracket with the cross-correlation coefficient between 0.76 to 0.86. Given the significant amount of traffic being analysed and the fact that category A sensors are not adjacent to each other on the 196 prefix, the results achieved showed high relativity. Since TCP traffic accounts for above 90% of category A's traffic, by simply removing 10% of the traffic the cross-correlation coefficient moved from moderate to strong. Correlograms of category A comparisons resembled a normal distribution with a clear distinct peak at time lag of 0. Category B sensors however remained in a moderate category achieving only 0.4 cross-correlation coefficient between the sensors. Looking at active TCP traffic, results showed that there was a slight, negligible decrease of the correlation coefficients. This was due to the small amounts of inactive packets observed.

Results obtained in looking at UDP traffic results were mixed and uncategorised. Although most of category A cross-correlations were weak (some even insignificant), a few moderate correlations were reached, even across categories. The poor results achieved were mainly caused by the anomalous spikes investigated in Section 4.1. The anomalous spikes were uncoordinated and focused on one sensor at a time.

ICMP traffic results showed that traffic activity across all sensors (irrespective of category) showed moderate to strong correlation. Interestingly, by analysing the correlograms it was shown that ICMP traffic was uniform. Although at time lag of 0 there were clear peaks, neighbouring time lags also showed relatively similar but slight lower cross-correlation values.

5.6 Summary

This chapter implemented auto- and cross-correlation methods to conduct advanced correlation analysis using time series. By using the auto-correlation function, experiments were conducted to test if each of the generated time series auto-correlates. Category A sensors showed evidence of repeated 24-houly cycles due to the Conficker worm.

The second part of the chapter focused on implementing the cross correlation method to test sensor combinations relativity. The cross-correlation function showed moderate correlation in each category when looking at all traffic. Results were improved significantly when looking at TCP traffic and strong correlation in category A's comparisons were calculated. Category B's comparisons remained in the moderate category for TCP traffic. Implementing the cross-correlation function using UDP traffic showed mixed and

uncategorised results but mainly results were in the weak category. The cross-correlation function showed that coefficients of ICMP traffic were uniform irrespective of category.

This chapter quantitatively analysed the correlation of traffic activity for all the sensors. The next chapter provides a conclusion to the research project by providing a brief overview of the work done and examining the project's objectives.

Chapter 6

Conclusion

This chapter summarises the research project by providing an overview of the work conducted and the outcomes thereof. The chapter also revisits the objectives and goals that were set for the project at the outset. When revising the objectives, the extent to which the project's goals were met is examined. The last section discusses the potential future extension of the research project.

6.1 Overview of Research Project

This research project focuses on correlation analysis of traffic activity observed in five /24 network telescopes. To develop a foundation, the project focused on comparative analysis of traffic generated by the telescopes. Building on this foundation, an advanced correlation analysis of traffic activity across the five sensors was conducted.

6.1.1 Preliminary Analysis of Datasets and Location of Sensors

The research project was introduced by discussing datasets and conducting analysis of the logical and physical location of network telescopes. With a selected period of a year, results obtained showed that datasets generated by category A sensors (196-a, 196-b and 196-c) had similar higher packet counts when compared to the lower packet counts observed in category B (146-a and 155-a) datasets. Consequently, the researcher proceeded to conduct a detailed analysis of the logical location. It was revealed that, although sensors

were relatively logically distant (not logically next to each other); sensors in category A were closer to each other. The reason for this is that Category A sensors are placed on the same high-order IP prefix (prefix 196).

The in-depth analysis of IP prefix location (both physical and logical) using the Regional Internet Registry, provided further insight into the legacy of IP assignments that aided the researcher in the categorisation of the sensors. The high-order IP prefix, in which category A sensors were placed, had most apportions in Africa. Conversely, apportions in high-order IP prefixes, in which category B sensors were placed, were mainly in the United States. Results showed that category A sensors were placed in a high-order IP prefix with predominantly smaller IP address blocks whereas category B's high-order prefixes had larger IP address blocks. Larger IP address block assignments are more likely to belong to large enterprises while smaller assignments are more likely to be found in an "end-user" environment, which tends to be more susceptible to malicious activity. Moreover, larger corporations are likely to have higher information security maturity.

6.1.2 Comparative Analysis Using Summarisation

Having built a profile of datasets (logical and physical location, size, legacy IP assignments), the research project proceeded to conduct a comparative analysis of traffic observed across the five sensors. Iteratively, through summarisation techniques and the use of graphical plots, a number of similarities were observed in each category. Time series' of traffic activity, as well as their respective plots, were generated. Generated plots with all traffic types, showed similarities in traffic activity in each category of sensors.

The results from the graphically observed data were encouraging, which led the researcher to continue to investigate different types of traffic separately. Apart from similarities in the distribution of packets across the major protocols in each category of telescope sensors, further category-specific similarities were observed. In analysing TCP traffic, the influence of Conficker worm was observed as a contributor, among other sensor location factors, in the categorisation of sensors. Random number generator errors in Conficker meant that category B sensors were unreachable by the worm's automated scanning. The examination of UDP traffic, by looking at top ports, revealed that the traffic contained a number of anomalous spikes which resembled denial-of-service attacks. Because these anomalous spikes were uncoordinated, they caused the fluctuation in the relativity of UDP traffic activity when comparing different sensors. Comparative analysis of ICMP traffic showed evidence of similarities of traffic across all sensor combinations irrespective of the category.

Basic statistical methods also supported the initial observations of the categorisation of sensors presented, by showing similar spreads and centrality in each category.

6.1.3 Implementing Advanced Time Series Correlation Methods

Subsequent to the comparative analysis, that used summarisation and basic statistics, tests for correlation in traffic activity using advanced correlation analysis were conducted. At first instances, tests on whether there are repeating patterns in each generated time series (representing traffic activity of each sensor) were conducted. Time series were generated using a daily time period. Results of the daily time series showed that the series do not auto-correlate because of the resulting weak correlation coefficient calculated. At an hourly time interval, although the correlation coefficient was weak, examination of the correlogram showed evidence of diminishing cycles in every 24 hours of traffic. The next step focused on testing for correlation across all sensor combinations using cross-correlation method Time series were generated by firstly, looking at all traffic types and, secondly, refining the time series and studying different major TCP, UDP and ICMP traffic separately. Time series excluding the 445/tcp port were also generated, as well as studies of IP ranges outside Conficker's reach. The following results were achieved when using the cross-correlation method:

- Looking at all traffic types, the cross-correlation coefficient was in the moderate bracket for category A sensors. Similarly, when testing combinations of category B sensors the correlation coefficients calculated were also in the moderate bracket.
- Looking at TCP traffic, the cross-correlation coefficient for category A sensors was in the strong bracket, while Category B sensor combinations remained in the moderate bracket.
- Looking at UDP traffic cross-correlation coefficient results were mixed and uncategorised.
- Looking at ICMP traffic, cross-correlation coefficient results showed relativity of traffic activity with a moderate to strong correlation bracket across all sensor combinations.
- Looking at traffic without the influence of Conficker (i.e. excluding 445/tcp), results were significantly improved when compared to all traffic. Category A comparisons were in the strong bracket and Category B comparisons were in the moderate category.

6.2 Project Objectives and Goals

The two main objectives of the research project were: (1) to comparatively analyse traffic observed by telescope sensors; and (2) to quantitatively investigate the correlation of traffic activity across the five telescope sensors. From these objectives, a number of research goals were identified and in this section the goals are revisited as follows:

- In comparatively analysing the similarities of traffic, Chapter 4 used graphical plots and basic statistics (means, median and standard deviation) to show relativity of traffic in each category of telescope sensors. The categorisation of sensors was introduced in Section 3.2 following iterative analysis of where the sensors' IP prefixes were located (logically and physically) as well as the influence of the Conficker worm. The categorisation exercise formed part of an investigation into the differences in traffic observed by sensors and further highlighted the impact of legacy issues related to the IP address assignment.

- In analysing the different types of traffic, TCP traffic's dominance across category A sensors was observed due to 445/tcp port traffic. The influence of Conficker proved to be a main distinguishing element between category A and B sensors due to Conficker's inability to reach Category B sensors' IP range. Comparative analysis of UDP traffic, in Section 4.3, led to the discovery of a number of anomalous uncoordinated traffic spikes. ICMP traffic analysis conducted in Section 4.4 showed ICMP traffic as uniform across all sensors.

- In using statistical methods to quantitatively analyse traffic activity across all sensors, Chapter 5 focused on the implementation of time series. Time series were implemented to model traffic activity represented over a period of time. The implementation of the auto-correlation method showed that constructed time series representing daily traffic activity had no repeated patterns, however, hourly analysis revealed evidence of 24-hour cycles.

The fulfillment of the each goal contributes towards the main project's objectives. The research project has conducted comparative analysis by using basic statistics and summarisation techniques. Additionally, comparative analyses of major traffic types were conducted separately. In concluding the analysis, the project focused on quantifying the cross-correlation coefficient to determine the correlation of traffic activity in all the sensor combinations.

6.3 Future Work

Given the time limit imposed on research work, there are suggested areas of research that would aid further correlation studies of network telescope traffic activity. This research project is based on datasets that were collected at Rhodes University and, as larger and different datasets become available, further experiments can be carried out.

6.3.1 Multivariate Time Series Correlation Analysis

The research project directed efforts towards a univariate correlation analysis of time series focusing mainly on the packet count variable. Using multivariate correlation analysis, researchers would be able to conduct correlation analysis of traffic activity using more than one variable at a time. For example, researchers could examine the combination of daily packet count variable with the average daily packet size to create a bivariate time series and use multivariate analysis of time series to test for relativity in traffic activity. This would further enhance the research by examining the relationship of multiple variables.

6.3.2 Test Correlation with Additional Third Party Dataset

The research was conducted with five network telescope nodes spread across different network segments of the address range with a relatively logical distance. It would be interesting to further support the results achieved by using datasets captured in other regions or segments and conduct a correlation analysis of the same time period with the datasets used in this research. Furthermore, it would be interesting to conduct similar experiments on a larger network telescope or compare a larger network telescope's relativity with the small network telescope. Provided the periods under analysis are the same, cross-correlation of different sized telescope sensors could be conducted since the time series of larger telescope sensor would simply have higher packet count variable.

6.3.3 Smoothing Techniques

Network telescope traffic flow, similar to many time-substantial time series such as the economics series [8], are subjected to significant amount of noise. As an extension to this project, a technique of time series smoothing can be applied to extract the underlying

trend. After analysing UDP traffic, the researcher discovered a number of anomalous spikes of traffic and these were responsible for extreme packet counts that were uncoordinated across all sensors. If one were to examine the time series with smoothing techniques, one could investigate the correlation of traffic activity with a smooth time series.

6.3.4 Automated Metrics and Dashboards for Analysis

Chapter 4 makes use of summarisation and basic statistical techniques to analyse traffic of the five datasets. Conducted experiments were iterative in nature and were based on observations or trends of the initial results. The refinement of experiments was also employed to arrive at a set of plots and graphs used to compare the datasets. As an extension, by using generated graphs and plots as a basis; there is an opportunity to create dashboards and metrics that can be automatically generated to compare multiple datasets. Readers are referred to the work in network telescope metrics showing that standardised metrics can be used to share information among researchers [19].

References

[1] APNIC. Resource ranges by RIR, November 2013. Available from http://www.apnic.net/publications/research-and-insights/ip-address-trends/by-rir; accessed 10 October 2013.

[2] ARIN. Number Resources, 2013. Available from http://www.iana.org/numbers; accessed 6 September 2013.

[3] World Bank. Countries and Economies, 2013. Available from http://data.worldbank.org/country/; accessed 2 November 2013.

[4] Nevil Brownlee. One-way traffic monitoring with iatmon. In *Passive and Active Measurement*, pages 179–188, Vienna, Austria, 2012. Springer.

[5] Carnivore. Conficker does not like me?, November 2009. Available from http://carnivore.it/2009/11/03/conficker does not like me; accessed 10 October 2013.

[6] T. Chown. IPv6 Implications for Network Scanning. RFC 5157, March 2008. Available from http://www.ietf.org/rfc/rfc5157.txt; accessed 10 September 2013.

[7] Bradely Cowie and Barry Irwin. A Baseline Numeric Analysis of Network Telescope Data for Network Incident Discovery. In *Southern African Telecommunications Networks and Applications Conference (SATNAC)*, pages 339–344, Spier Estate, South Africa, 2010.

[8] Paul SP Cowpertwait and Andrew V Metcalfe. *Introductory time series with R*. Springer, 2009.

[9] Cymru. The darknet project, 2013. Available from http://www.team-cymru.org/Services/darknets.html; accessed 10 September 2013.

[10] Marina del Rey. Transmission Control Protocol, RFC 793, September 1981. Available from https://www.ietf.org/rfc/rfc793.txt; accessed 10 September 2013.

[11] G. Armitage F. Baker, W. Harrop. IPv4 and IPv6 Greynets, RFC 6018, September 2010. Available from http://tools.ietf.org/html/rfc6018; accessed 10 Devember 2013.

[12] M. Ford, J. Stevens, and J. Ronan. Initial Results from an IPv6 Darknet13. In *International Conference on Internet Surveillance and Protection*, pages 13–13, Cote d'Azur, 2006. IEEE.

[13] E. Gerich. Guidelines for Management of IP Address Space, May 1993. Available from http://tools.ietf.org/html/rfc1466; accessed 2 November 2013.

[14] Uli Harder, Matthew Johnson, Jeremy T. Bradley, and William J. Knottenbelt. Observing Internet Worm and Virus Attacks with a Small Network Telescope. In *PASM 2005, 2nd International Workshop on the Practical Application of Stochastic Modelling*, volume 151 of *Electronic Notes in Theoretical Computer Science*, pages 47–59, May 2006. The data has now been released in anonymised form at http://www.doc.ic.ac.uk/ uh/network-telescope/.

[15] IANA. IANA IPv4 Address Space Registry, May 2013. Available from http://www.iana.org/assignments/ipv4-address-space/ipv4-address-space.xml; accessed 01 November 2013.

[16] B. Irwin. A baseline study of potentially malicious activity across five network telescopes. In *Cyber Conflict (CyCon), 2013 5th International Conference on*, pages 1–17, 2013.

[17] B. Irwin and N. Pilkington. High Level Internet Scale Traffic Visualization Using Hilbert Curve Mapping. In JohnR. Goodall, Gregory Conti, and Kwan-Liu Ma, editors, *VizSEC 2007*, Mathematics and Visualization, pages 147–158. Springer Berlin Heidelberg, 2008.

[18] Barry Irwin. *A framework for the application of network telescope sensors in a global IP network*. PhD thesis, Rhodes University, 2011.

[19] Barry Irwin. Network Telescope Metrics. In *Southern African Telecommunications Networks and Applications Conference (SATNAC)*, pages 339–344, Fancourt, George, Western Cape, 2012.

[20] Barry Irwin and Richard J Barnett. An Analysis of Logical Network Distance on Observed Packet Counts for Network Telescope Data. In *Southern African Telecommunications Networks and Applications Conference (SATNAC)*, volume 31, Royal Swazi Spa, 2009.

[21] G. Camarillo A. Johnston J. Peterson R. Sparks M. Handley E. Schooler J. Rosenberg, H. Schulzrinne. SIP: Session Initiation Protocol, RFC 3261, June 2002. Available from http://www.ietf.org/rfc/rfc3261.txt; accessed 6 November 2013.

[22] Jan W Kantelhardt, Eva Koscielny-Bunde, Henio HA Rego, Shlomo Havlin, and Armin Bunde. Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 295(3):441–454, 2001.

[23] Eric D Kolaczyk. *Statistical analysis of network data*. Springer, 2009.

[24] CBG Limonard. Missing values in time series and the implications on autocorrelation analysis. *Analytica Chimica Acta*, 103(2):133–140, 1978.

[25] David J Marchette. A Statistical Method for Profiling Network Traffic. In *Workshop on Intrusion Detection and Network Monitoring*, pages 119–128, Santa Clara, California, USA, 1999.

[26] Microsoft. Internet Control Message Protocol (ICMP) Basics, February 2007. Available from http://support.microsoft.com/kb/170292; accessed 23 October 2013.

[27] Microsoft. Microsoft Security Bulletin MS08-067 - Critical, October 2008. Available from http://technet.microsoft.com/en-us/security/bulletin/ms08-067; accessed 8 November 2013.

[28] Microsoft. IPv4 Addressing, January 2009. Available from http://technet.microsoft.com/en-us/library/dd379547(v=ws.10).aspx; accessed 10 October 2013.

[29] Microsoft. Buffer Overruns in SQL Server 2000 Resolution Service Could Enable Code Execution (Q323875), January 2013. Available from http://technet.microsoft.com/en-us/security/bulletin/ms02-039; accessed 16 October 2013.

[30] Boris Mirkin. *Core concepts in data analysis: summarization, correlation and visualization*. Springer, 2011.

[31] D. Moore, C. Shannon, G. Voelker, and S. Savage. Network Telescopes: Technical Report. Technical report, Cooperative Association for Internet Data Analysis (CAIDA), Jul 2004.

[32] David Moore. Summary of Anonymization Best Practice Techniques, September 2008. Available from http://www.caida.org/projects/predict/anonymization/; accessed 10 November 2013.

[33] David Moore, Vern Paxson, Stefan Savage, Colleen Shannon, Stuart Staniford, and Nicholas Weaver. Inside the slammer worm. *Security & Privacy, IEEE*, 1(4):33–39, 2003.

[34] David Moore, Colleen Shannon, Douglas J Brown, Geoffrey M Voelker, and Stefan Savage. Inferring internet denial-of-service activity. *ACM Transactions on Computer Systems (TOCS)*, 24(2):115–139, 2006.

[35] David Moore, Colleen Shannon, and k claffy. Code-Red: A Case Study on the Spread and Victims of an Internet Worm. In *Proceedings of the 2Nd ACM SIGCOMM Workshop on Internet Measurment*, IMW '02, pages 273–284, New York, NY, USA, 2002. ACM.

[36] Alastair T Nottingham and Barry Irwin. gPF: A GPU Accelerated Packet Classification Tool. In *Southern African Telecommunications Networks and Applications Conference*, pages 339–344, Royal Swazi Spa, 2009.

[37] Jason W Osborne and Amy Overbay. The power of outliers (and why researchers should always check for them). *Practical assessment, research & evaluation*, 9(6):1–12, 2004.

[38] Ruoming Pang, Vinod Yegneswaran, Paul Barford, Vern Paxson, and Larry Peterson. Characteristics of Internet Background Radiation. In *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, IMC '04, pages 27–40, New York, NY, USA, 2004. ACM.

[39] Hassen Saidi Phillip Porras and Vinod Yegneswaran. An Analysis of Conficker's Logic and Rendezvous Points, March 2009. Available from http://mtc.sri.com/Conficker/; accessed 10 October 2013.

[40] J. Postel. User Datagram Protocol, RFC 768, August 1980. Available from http://www.ietf.org/rfc/rfc768.txt; accessed 10 September 2013.

[41] J. Postel. Internet Control Message Protocol, RFC 792, September 1981. Available from http://tools.ietf.org/html/rfc792; accessed 6 November 2013.

[42] J. Reynolds & J. Postel. Assigned Numbers, October 1994. Available from http://www.ietf.org/rfc/rfc1700.txt; accessed 18 September 2013.

[43] Joel Sandin. P2P Systems for Worm Detection, September 2003. Available from www.icir.org/vern/dimacs-large-attacks/sandin.ppt; accessed 7 August 2013.

[44] SANS. ICMP Attacks Illustrated, 2001. Available from https://www.sans.org/reading-room/whitepapers/threats/icmp-attacks-illustrated-477; accessed 10 October 2013.

[45] Colleen Shannon and David Moore. The spread of the witty worm. *Security & Privacy, IEEE*, 2(4):46–50, 2004.

[46] DM Simpson, AFC Infantosi, and DA Botero Rosas. Estimation and significance testing of cross-correlation between cerebral blood flow velocity and background electro-encephalograph activity in signals with missing samples. *Medical and Biological Engineering and Computing*, 39(4):428–433, 2001.

[47] James H Stock and Mark W Watson. Variable trends in economic time series. *The Journal of Economic Perspectives*, 2(3):147–174, 1988.

[48] Richard Taylor. Interpretation of the correlation coefficient: a basic review. *Journal of diagnostic medical sonography*, 6(1):35–39, 1990.

[49] Lisa Tompson. Advanced Time Series Analysis, 2012. Available from http://www.ucl.ac.uk/jdi/events/int-CIA-conf/ICIAC11_Slides/ICIAC11_1E_LTompson; accessed 23 August 2013.

[50] Jean-Pierre van Riel and Barry Irwin. InetVis, a Visual Tool for Network Telescope Traffic Analysis. In *Proceedings of the 4th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*, AFRIGRAPH '06, pages 85–89, New York, NY, USA, 2006. ACM.

[51] W. N. Venables and B. D. Ripley. Auto- and Cross- Covariance and -Correlation Function Estimation, 2002. Available from http://stat.ethz.ch/R-manual/R-patched/library/stats/html/acf.html; accessed 10 October 2013.

[52] Du Xu and Haishan Zhong. Cross correlation analysis and construction of joint distribution traffic model. In *Communications, Circuits and Systems, 2008. ICCCAS 2008. International Conference on*, pages 538–542, 2008.

[53] Rui Xu, Donald Wunsch, et al. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.

[54] Vinod Yegneswaran, Paul Barford, and Dave Plonka. On the design and use of Internet sinks for network abuse monitoring. In *Recent Advances in Intrusion Detection*, pages 146–165, Sophia Antipolis, French Riviera, France, 2004. Springer.

[55] Cliff C. Zou, Weibo Gong, Don Towsley, and Lixin Gao. The monitoring and early detection of internet worms. *IEEE/ACM Trans. Netw.*, 13(5):961–974, October 2005.

Appendix A

Overview of Basic Statistical Methods Used

These appendices provide a brief description of the basic statistical methods that were used for evaluating centrality, spread and correlation [30].

This appendices provides a brief description of basic statical methods that are used for evaluating centrality, spread and correlation [30].

1. Mean , Median and Mode

   (a) The sample mean is calculated by summing up the values observed and dividing by the total number of the sample. The mean calculation is sensitive to outliers and, therefore, extreme values can have impact on the mean. Given variable (x) and (n) observations, Equation A.1 is the formula for sample mean.

$$\bar{x} = \frac{\sum x_i}{n} \tag{A.1}$$

   (b) The median value is the middle value observed when sorting the numbers from smallest to largest. If there is an even number of observations, the median is the average of the two middle values on an ordered dataset.

2. Variance

   (a) Sample variance is defined as the average squared deviations of values from their sample mean.

3. Standard deviation

   (a) The standard deviation is a square root of the sample variance. This is the average deviation from the mean. Equation A.2 shows the calculation of a standard deviation given variable (x) with (n) observations.

$$sd(x) = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{A.2}$$

Appendix B

Packet Header Information

This section contains the table fields for the relational database that were used as part of the research. A high-level overview of the fields used in this research is provided. For a complete list, the reader is referred to the following RFCs[41] [40] [10].

Packets Table:

- datetime stamp: time stamp of the packet
- source IP: 32-bits source address
- destination IP: 32-bits destination address
- packet type: protocol identifier
- size: packet size
- time to live: maximum time of the datagram in the Internet system

TCP Table:

- source port: 16-bit source port number
- destination port: 16-bit destination port number
- control bits:
  - URG: Urgent Pointer field significant

  - ACK: Acknowledgment field significant
  - PSH: Push Function
  - SYN: Synchronise sequence number

UDP Table:

- source port: 16-bit source port number
- destination port: 16-bit destination port number

ICMP Table:

- type: identifier for ICMP message type
- code: 0 = net unreachable; 1 = host unreachable; 2 = protocol unreachable; 3 = port unreachable; 4 = fragmentation needed and DF set; 5 = source route failed.

Appendix C

Daily, Hourly and Monthly Packet Counts

Figure C.1 shows the disparity of traffic relativity between inter-category comparisons. Sensors in category A (196-a, 196-b and 196-c) have similar peaks and troughs. Similarly, category B sensors (146-a and 196-b) also shows similar peaks and troughs but these are not as strongly correlated as compared to category A. There is a clear disjunction between category A comparisons and category B comparisons.

Figure C.2 shows the hourly packet counts of all network telescope sensors. Higher packet counts are observed in Category A sensors relative to Category B sensors.

Figure C.3 and C.4 show the monthly packet counts for category A and category B's sensors respectively. The diagrams clearly demonstrate that monthly packet counts for

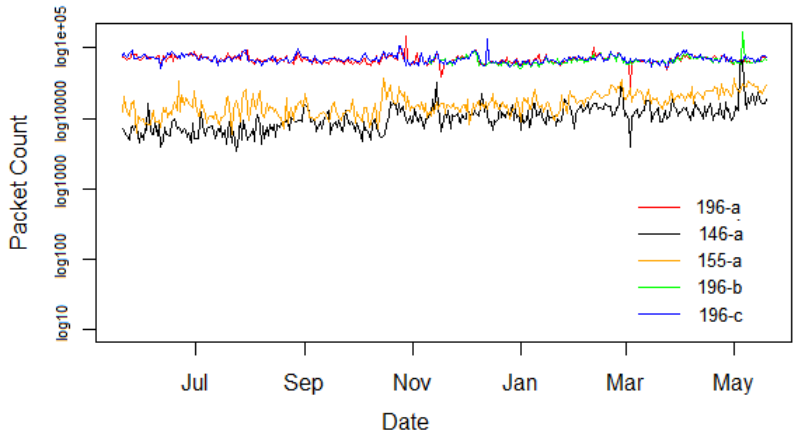


Figure C.1: Daily packet counts (category A vs. category B)

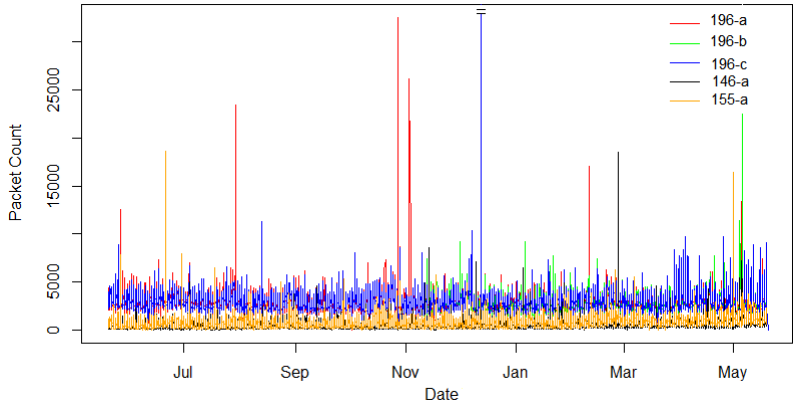


Figure C.2: Hourly packet count (category A vs. category B)

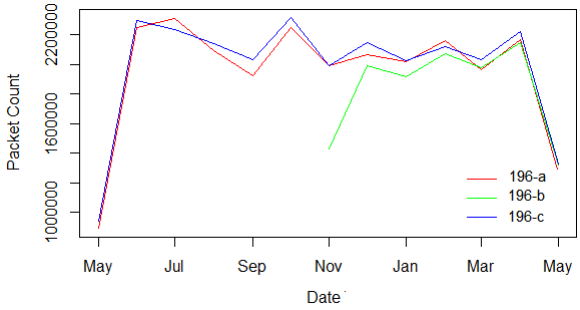


Figure C.3: Monthly packet count of telescope sensor 196-a, 196-b and 196-c (category A)

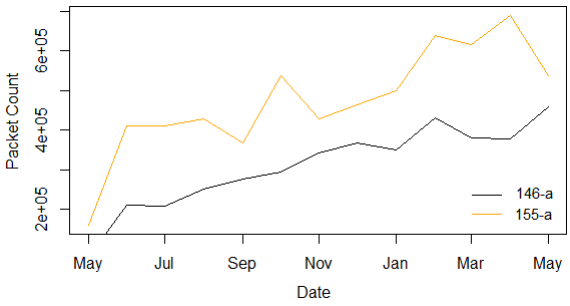


Figure C.4: Monthly packet count of telescope sensor 146-a, 155-a (category B)

category A have similar peaks. Category B's monthly packet counts show a similar increasing trend.

Appendix D

Auto-correlation Correlograms - All Traffic

Figure D.1, Figure D.2 and Figure D.3 show category A's auto-correlation correlograms which demonstrates weak auto-correlation coefficients. The correlograms show that most coefficients are closer or below the confidence interval lines and, therefore, one cannot reject the null hypothesis that the correlation is zero for values below the confidence intervals.

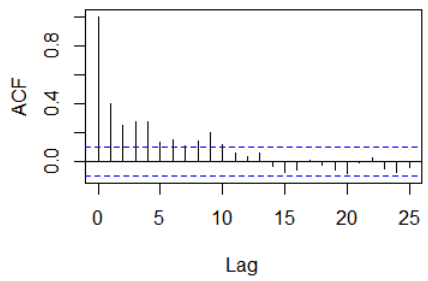


Figure D.1: Auto-correlation correlogram for sensor 196-a using daily packet count

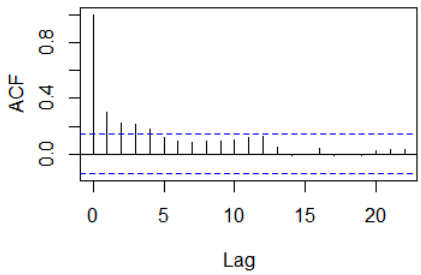


Figure D.2: Auto-correlation correlogram for sensor 196-b using daily packet count

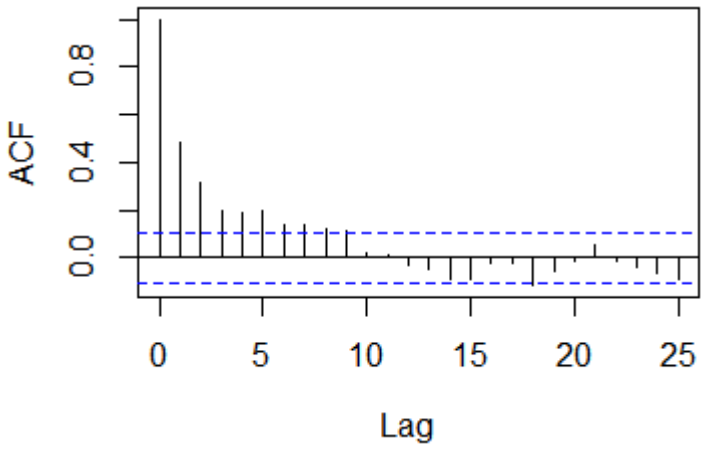


Figure D.3: Auto-correlation correlogram for sensor 196-c using daily packet count

Appendix E

Hourly Auto-correlation Correlograms - Port 445/tcp

The correlograms of category A comparisons showed evidence of repeating patterns when looking at Conficker related traffic. Remaining category A correlograms (not included in main body) are shown in this appendix.

E.1 Auto-correlation Correlograms for Sensors 196-b and 196-c

Figure E.1 and Figure E.2 show auto-correlation correlograms for sensor 196-b and 196-c respectively (category A). Both diagrams show strong auto-correlation coefficients at 24-hourly cycles.

E.2 Correlogram Auto-correlation Coefficients

As an example, auto-correlation coefficients of Sensor 196-c are shown in Figure E.3. The diagram shows distinct peaks, circled in red, at every 24-hour cycle. The strong auto-correlation coefficients are highlighted in yellow.

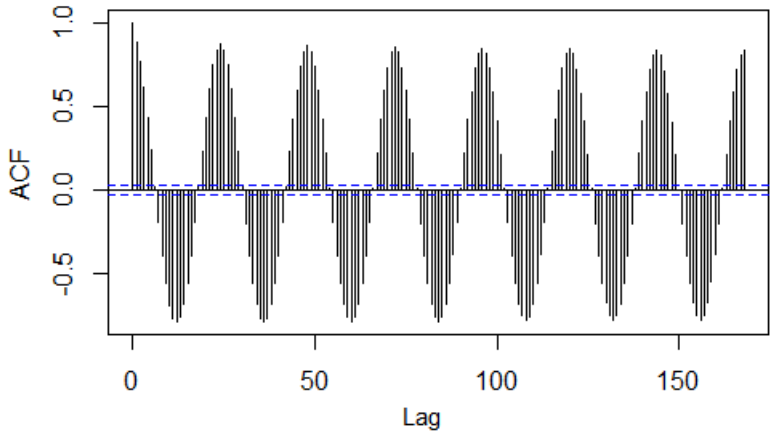


Figure E.1: Auto-correlation correlogram for sensor 196-b using hourly packet count

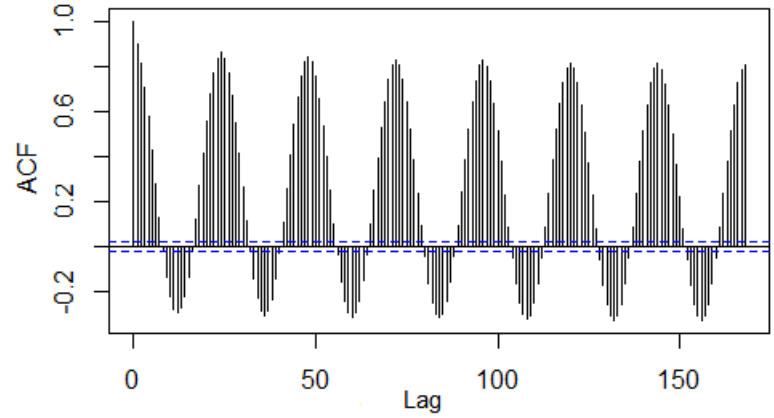


Figure E.2: Auto-correlation correlogram for sensor 196-b using hourly packet count

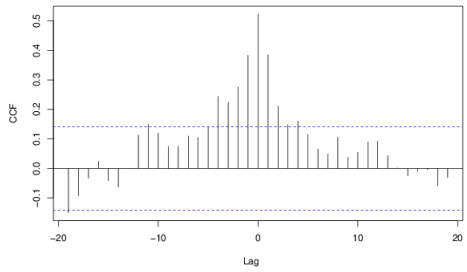
0	1	2	3	4	5	6	7	8	9	10	11	12	13
1.000	0.900	0.815	0.706	0.577	0.432	0.279	0.130	-0.011	-0.130	-0.219	-0.273	-0.289	-0.271
14	15	16	17	18	19	20	21	22	23	24	25	26	27
-0.221	-0.134	-0.018	0.120	0.269	0.417	0.555	0.676	0.772	0.837	0.861	0.836	0.770	0.673
28	29	30	31	32	33	34	35	36	37	38	39	40	41
0.549	0.412	0.265	0.118	-0.021	-0.140	-0.228	-0.281	-0.301	-0.284	-0.231	-0.144	-0.027	0.112
42	43	44	45	46	47	48	49	50	51	52	53	54	55
0.258	0.406	0.542	0.663	0.759	0.822	0.844	0.821	0.755	0.657	0.536	0.397	0.253	0.103
56	57	58	59	60	61	62	63	64	65	66	67	68	69
-0.034	-0.151	-0.240	-0.292	-0.310	-0.292	-0.239	-0.150	-0.035	0.102	0.248	0.390	0.526	0.644
70	71	72	73	74	75	76	77	78	79	80	81	82	83
0.741	0.803	0.826	0.802	0.739	0.641	0.522	0.386	0.239	0.093	-0.043	-0.159	-0.244	-0.296
84	85	86	87	88	89	90	91	92	93	94	95	96	97
-0.314	-0.297	-0.243	-0.158	-0.041	0.095	0.241	0.387	0.522	0.644	0.738	0.802	0.825	0.799
98	99	100	101	102	103	104	105	106	107	108	109	110	111
0.734	0.638	0.515	0.377	0.232	0.087	-0.048	-0.163	-0.248	-0.300	-0.318	-0.301	-0.250	-0.160
112	113	114	115	116	117	118	119	120	121	122	123	124	125
-0.045	0.090	0.236	0.383	0.519	0.638	0.731	0.793	0.815	0.789	0.726	0.628	0.506	0.370
126	127	128	129	130	131	132	133	134	135	136	137	138	139
0.227	0.079	-0.056	-0.170	-0.254	-0.307	-0.326	-0.305	-0.252	-0.164	-0.050	0.087	0.233	0.378
140	141	142	143	144	145	146	147	148	149	150	151	152	153
0.511	0.629	0.727	0.790	0.811	0.786	0.722	0.625	0.502	0.367	0.223	0.078	-0.057	-0.172
154	155	156	157	158	159	160	161	162	163	164	165	166	167
-0.256	-0.307	-0.324	-0.304	-0.251	-0.165	-0.047	0.089	0.233	0.380	0.513	0.631	0.726	0.787
168													
0.807													

Figure E.3: Auto-correlation coefficients for sensor 196-c

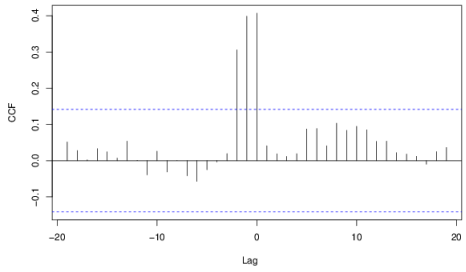
Appendix F

Cross-correlation Correlograms - All Traffic

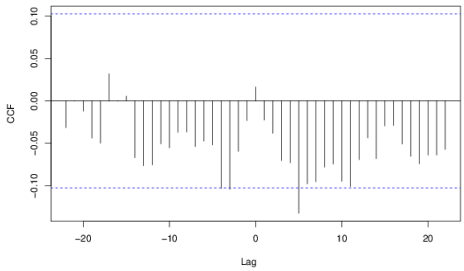
A number of correlograms were constructed to investigate the relationship between each network telescope sensor's activity. Figures F.1 and Figure F.2 contain the combination of correlograms for cross-correlation results looking at all traffic.



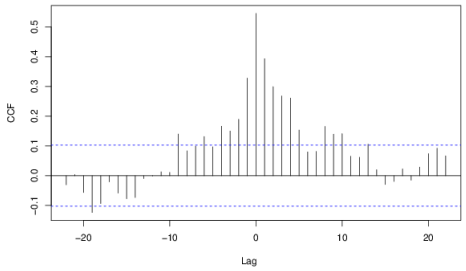
Cross-correlation correlogram of sensor 196-a vs. 196-b using daily packet count



Cross-correlation correlogram of sensor 146-a vs. 196-b using daily packet count

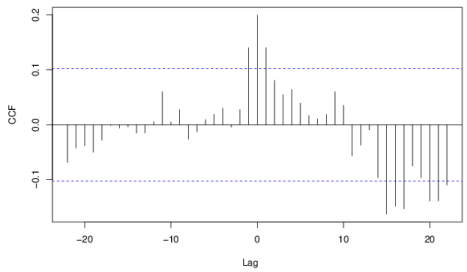


Cross-correlation correlogram of sensor 196-a vs. 146-a using daily packet count

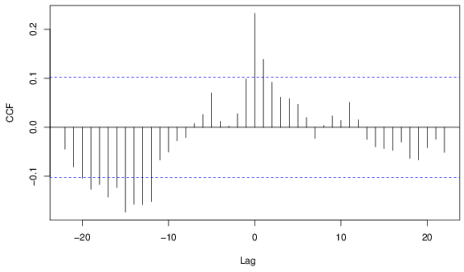


Cross-correlation correlogram of sensor 196-a vs. 196-c using daily packet count

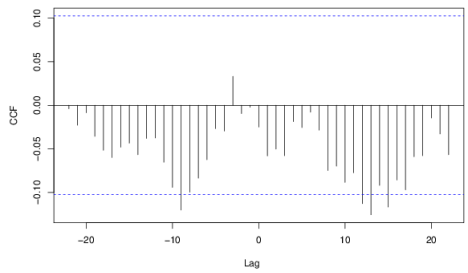
Figure F.1: Cross-correlation correlograms set A



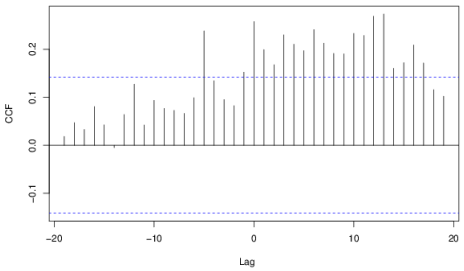
Cross-correlation correlogram of sensor 196-a vs. 155-a using daily packet count



Cross-correlation correlogram of sensor 155-a vs. 196-c using daily packet count



Cross-correlation correlogram of sensor 146-a vs. 196-c using daily packet count



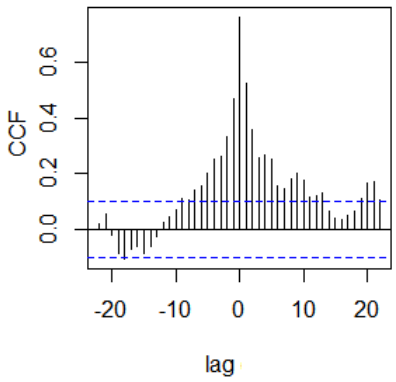
Cross-correlation correlogram of sensor 155-a vs. 196-b using daily packet count

Figure F.2: Cross-correlation correlograms set B

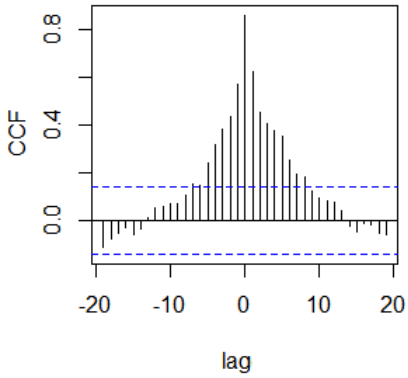
Appendix G

Cross-correlation Correlograms - TCP Traffic

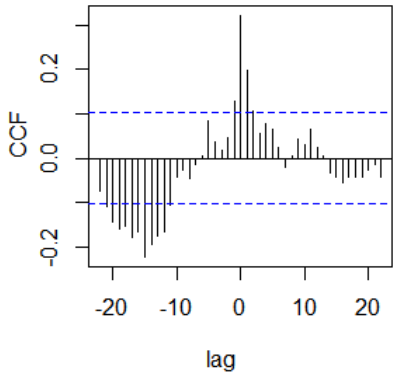
Figure G.1 and Figure G.2 show the cross-correlation correlograms of TCP traffic. Higher cross-correlation coefficients were calculated across category A telescope sensors and, similarly, the correlograms show a distinct peak at time lag 0.



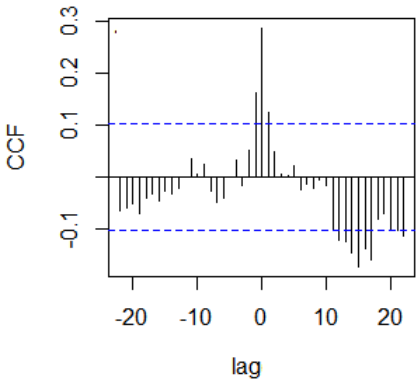
Cross-correlation correlogram of sensor 196-a vs. 196-c using daily packet count



Cross-correlation correlogram of sensor 196-b vs. 196-c using daily packet count

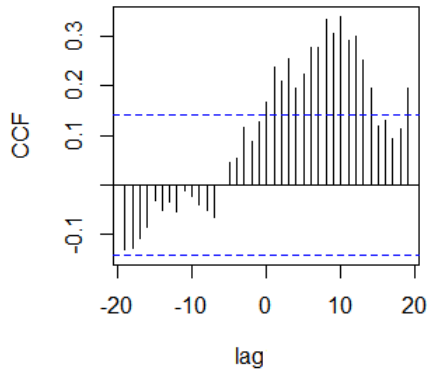


Cross-correlation correlogram of sensor 155-a vs. 196-c using daily packet count

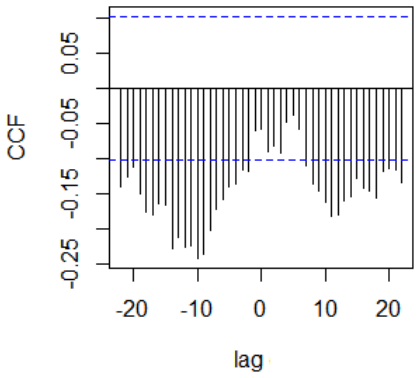


Cross-correlation correlogram of sensor 196-a vs. 155-a using daily packet count

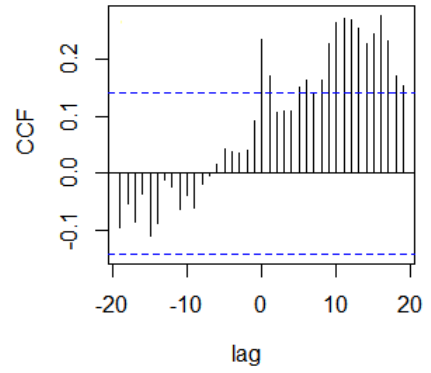
Figure G.1: Cross-correlation correlograms using daily TCP packet count - set A



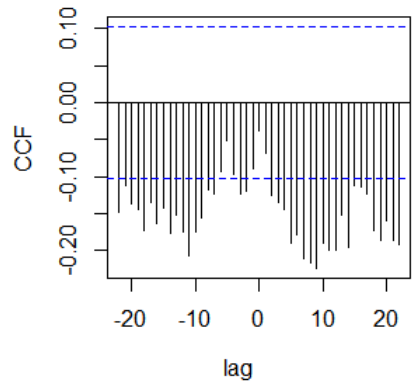
Cross-correlation correlogram of sensor 146-a vs. 196-b using daily packet count



Cross-correlation correlogram of sensor 146-a vs. 196-c using daily packet count



Cross-correlation correlogram of sensor 155-a vs. 196-b using daily packet count



Cross-correlation correlogram of sensor 196-a vs. 146-a using daily packet count

Figure G.2: Cross-correlation correlograms using daily TCP packet count - set A

Appendix H

Packet Distribution Across Destination IP

Figure H.1 shows the packet count as per destination IP of all five network telescope sensors. Similar results have been obtained by researchers at Rhodes University [16]. Results illustrate the reduced packet counts on traffic outside Conficker's reach. At destination IP x.x.x.128 there is a sharp reduction on packet counts for category A sensors (196-a,196-b and 196-c). This is marked with letter A on the diagram. Results also show that sensors 146-a and 155-a gradually see a reduction of packet counts in IP addresses greater than x.x.x.128.

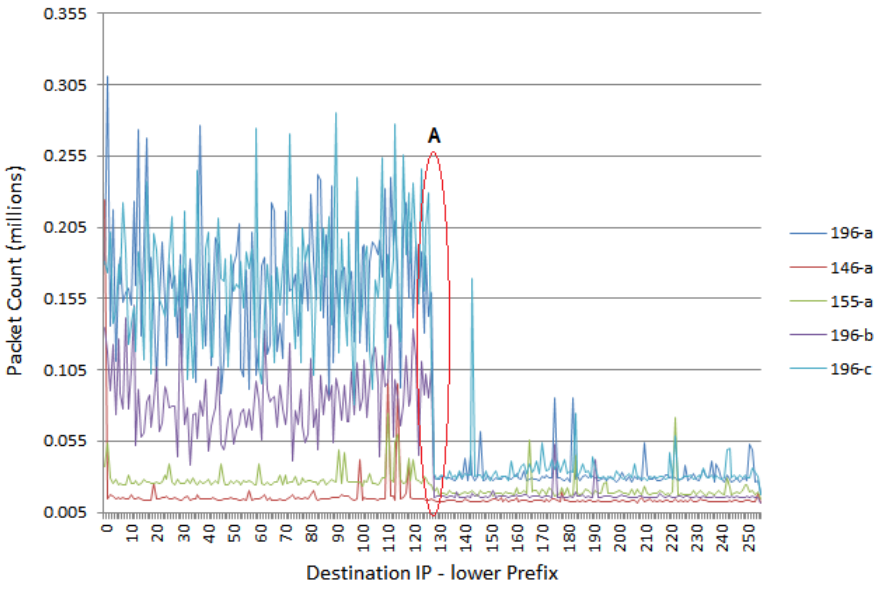
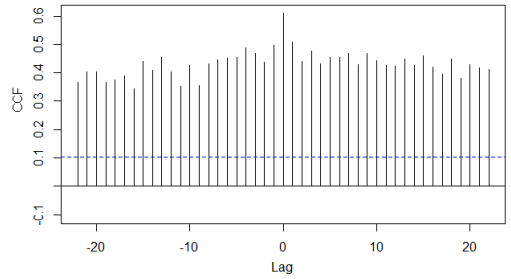


Figure H.1: Packet distribution across destination IP for all network telescope sensors

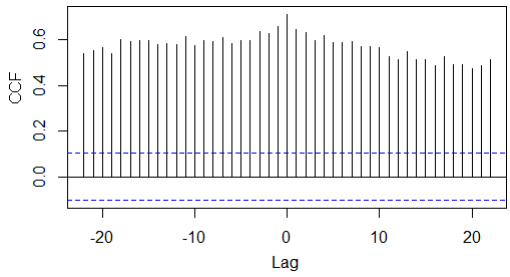
Appendix I

Cross-correlation Correlograms - ICMP Traffic

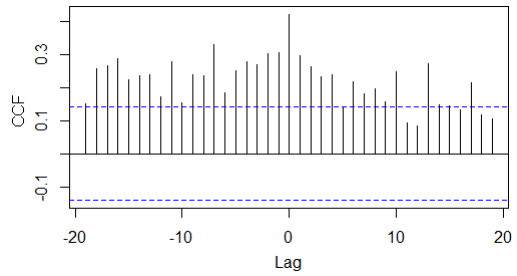
Figure I.1 and Figure I.2 shows the cross-correlation correlograms for ICMP traffic. A peak at time lag 0 is observed with a gradual decline of cross-correlation coefficients as the lag increases.



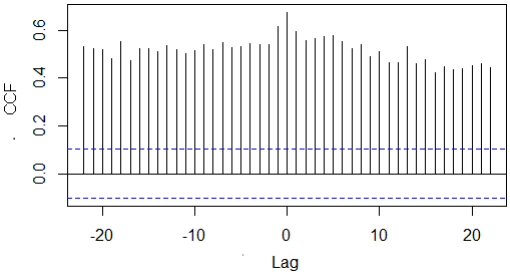
Cross-correlation correlogram of sensor 196-a vs. 146-a using daily ICMP packet count



Cross-correlation correlogram of sensor 155-a vs. 196-c using daily ICMP packet count

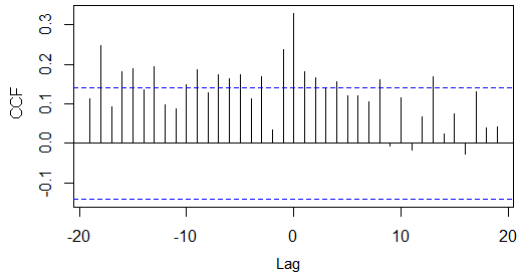


Cross-correlation correlogram of sensor 155-a vs. 196-b using daily ICMP packet count

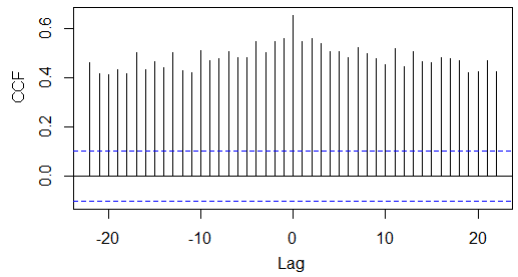


Cross-correlation correlogram of sensor 146-a vs. 196-c using daily ICMP packet count

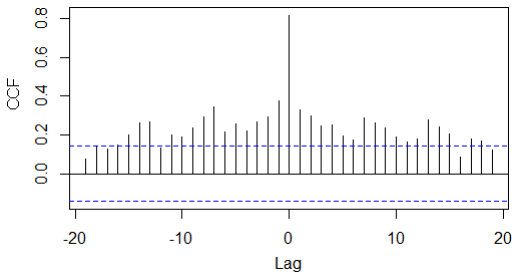
Figure I.1: Cross-correlation correlograms using ICMP traffic - Set A



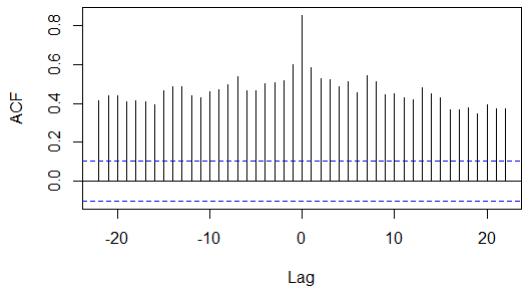
Cross-correlation correlogram of sensor 146-a vs. 196-b using daily ICMP packet count



Cross-correlation correlogram of sensor 155-a vs. 196-b using daily ICMP packet count



Cross-correlation correlogram of sensor 196-a vs. 196-b using daily ICMP packet count



Cross-correlation correlogram of sensor 196-b vs. 196-c using daily ICMP packet count

Figure I.2: Cross-correlation correlograms using ICMP traffic - Set B