Noroff University College

AN EVALUATION OF SCANNING TAXONOMIES AGAINST IBR TRAFFIC

Submitted in partial fulfilment
of the requirements of the degree of

BACHELOR OF CYBER SECURITY

of Noroff University College

Petter Karstensen

Oslo, Norway
April 2023

Declaration

I declare that the work presented for assessment in this submission is my own, that it has not previously been presented for another assessment, and that work completed by others has been appropriately acknowledged.

Name: Petter Karstensen **Date**: April 21, 2023

Abstract

Protection against malicious activities has become increasingly vital in the growing cyberspace. One of the initial steps taken by attackers is the act of network scanning, which highlights the importance of classifying scanning traffic to mitigate such activities. The research field of network scanning taxonomies has seen a decrease in activity over recent years and relies on two pieces of prior research separated by 10 years, namely Barnett and Irwin (2008), and Liu and Fukuda (2018). To evaluate the relevance of previously established taxonomies, it is crucial to have data that reflects global events. Network telescopes offer an efficient means of passive monitoring of an unused IP address space, collecting real-life data. This research project utilises a dataset spanning from December 2020 to March 2021, to assess the effectiveness of two established taxonomies. By examining how the taxonomies categorise data, and comparing the results of the two approaches, this study aims to determine if the taxonomies still contribute to the understanding and processing of current traffic trends.

Keywords: *taxonomy, internet background radiation, IBR, network scanning*

Acknowledgements

I would like to express my gratitude to my supervisor, Barry Irwin, for his valuable guidance, support, and encouragement throughout the duration of this research.

I am also thankful for the support from Mamma, Pappa, Julie, Miadog, and Balder throughout the research.

A special thanks to Elvira for her support during the final stages of the project.

Contents

List of Figures

List of Tables

1

Introduction

Over the past decade, there has been a significant expansion and importance of cyberspace (Cornish *et al.* 2011). The impact of this growth has been global, affecting various sectors such as commerce, governance, and individuals. The reliance on cyberspace brings challenges in terms of the rising occurrence of malicious activities. These activities threaten the security and integrity of information in cyberspace, highlighting the growing hostile nature of the internet.

Protection against malicious activity has become increasingly vital in the current digital landscape. As stated in Anbar *et al.* (2013), network scanning is considered to be the first step taken by attackers trying to gain access. Network scanning is the task of examining a network or Internet-wide services to find active hosts or open ports, eventually performing malicious activities. These scans can generate anomalies in network traffic. Investigating anomalies in network scanning traffic and analyzing the captured data is crucial for improving our understanding of scanning techniques. Such research can contribute to the advancement of security measures such as firewalls and intrusion detection systems. As the security of networks adapts to the knowledge of network scanning attacks, the attackers adapt new techniques to evade detection. It is important for researchers to adopt new methods for the detection and attribution of network scanning (Bou-Harb *et al.* 2014). To achieve this, a clear classification of network scanning traffic, commonly referred to as *Network scanning taxonomies*, is necessary. Having a clear taxonomy of common scanning anomalies will be useful for network administrators to identify network attackers before it is too late (Anbar *et al.* 2013). This allows for the identification of malicious intent and the implementation of preventative measures before an attack can occur.

While the network scanning taxonomies field saw some expansion in the period from 2000 to 2018, the last 5 years have not seen much attention given to the research field. Most of the research referred to in this thesis stems from the beginning of the 21st century. While the research conducted in the 2000s and 2010s is still relevant to the current challenges in the field, it is important to consider newer scanning approaches. Previous research on scanning traffic anomalies may no longer be applicable today and should be revisited. This is not to suggest that earlier work in this area is irrelevant, but rather that an update could be beneficial. The purpose of this bachelor's thesis is to address this gap in the field by examining previously conducted research and focusing on the detection of scanning techniques, providing an evaluation of the value of earlier taxonomies. This will be achieved through the use of packet capture analysis for anomaly classification and a comparison of two established taxonomies.

The remainder of this chapter is structured as follows; In Section 1.1, the problem statement to be addressed in this project is defined. The aims and objectives of the project, including primary and supplementary objectives, will be outlined in section 1.2. The background and related work in the field will be discussed in section 1.3. The approach and methodology adopted for the research will be detailed in section 1.4. The specific requirements and potential risks associated with the project will be presented in sections 1.5 and 1.6.

1.1 Problem Statement

A potentially outdated taxonomy of network scanning techniques is a weakening factor in cyber security. This project is evaluating historical taxonomies against currently observed trends and seeks to confirm their relevance.

1.2 Aims & Objectives

The aim of this research is to revisit taxonomies that can guide cybersecurity specialists toward a better understanding of network traffic. The objective is to recognize the characteristics established in the taxonomies by Barnett and Irwin (2008) and Liu and Fukuda (2018), and evaluate how well these taxonomies can be applied with today's knowledge.

To achieve the objective set for this project, the taxonomies will be tested against Network Telescope data. To simplify the analysis of the two taxonomies, python scripts will be developed to process the data. The quality of the research depends on the success of these scripts. In order to achieve the overall objectives, the following secondary objectives have been identified:

- Identify scanning techniques
- Investigate suitable Python development approaches
- Create scripts in Python for analyzing traffic captures
- Verify the relevance of the taxonomy

1.3 Related Work

To better understand the activity of network scanning, a brief history of scanning is presented. The research by Staniford *et al.* (2002) is among the multiple authors who have researched network scanning in the 1990s and 2000s. Allman *et al.* (2007) provides the history and a deeper understanding of the technical aspects of network scanning. Analyzing scanning traffic requires a dataset consisting of such traffic. As this research uses a dataset from a Network Telescope, a section dedicated to the concept introduced by Moore, Shannon, *et al.* (2004) is provided. The research from Pang *et al.* (2004) and Yegneswaran, Barford, and Plonka (2004) explains how the dataset differs from regular internet activity, supplemented with the newer research by Wustrow *et al.* (2010) and Borgnat *et al.* (2009), which both follows up previous studies with observations of environmental factors.

As this research is based on identifying network scanning, the research by Yegneswaran, Barford, and Ullrich (2003) introduces a scanning technique classification, in this research called *classical classification*, which the taxonomy from Barnett and Irwin (2008) is based. Bou-Harb *et al.* (2014) emphasizes the importance of research and methods for detecting and attributing cyber scanning activities, with a comprehensive review of network scanning techniques.

The main literature that will be utilized in this research is from Barnett and Irwin (2008) and Liu and Fukuda (2018). These introduce the taxonomies for the analysis that is central to the thesis. By comparing the two taxonomies proposed using a common dataset, the earlier-mentioned studies are used as supplementary material to provide an understanding of the results.

1.4 Method

The methodology employed in this project begins with a comprehensive literature review to gain a thorough understanding of the nature of scanning techniques. The literature review serves as the foundation for the main component of the project, which involves the analysis of network traffic by using two existing taxonomies.

The nature of Network Telescopes gathering all incoming traffic, which is used in this research, requires effective analyzing techniques. As mentioned in Barnett and Irwin (2008), Wireshark is unable to perform effective analysis of such a large volume of data. For this research, the taxonomies are implemented in the form of Python scripts, which are subsequently tested with the dataset to assess their relevance in analyzing network scanning. A Comparison between the two taxonomies is also provided to better understand the two proposed approaches.

1.5 Scope and Limits

The scope of this research is to determine whether the types of scanning traffic identified in the literature review are observable by classifying the traffic using the taxonomies established by Barnett and Irwin (2008) and Liu and Fukuda (2018), or if there are divergent findings from the singular dataset driving from a Network Telescope.

The analysis will exclude Layer 2 scans which is a prevalent factor in Barnett and Irwin (2008), due to limitations in packet capture information.

1.6 Ethical Considerations

There are no ethical considerations of concern with this research. Datasets have been collected by other researchers. Care has been taken in terms of the limiting disclosure of the IP ranges used by Network telescopes.

1.7 Document Structure

The reader has by this time spent around 5 minutes reading the introduction of the research. By this time, 4 thousand scanning traffic packets have been gathered by the network telescope used for the dataset in this research. The remainder of the document is structured as follows:

- The literature review is in chapter 2. Here the history and background of scanning, network telescopes, the reasons for scanning, and the classification of different scan types will be discussed. Additionally, an introduction to the two taxonomies will be provided. The literature review serves as a foundation for the rest of the research, providing the necessary context and understanding of the topic. By the time the literature review has been read another 35 thousand additional packets of scanning traffic have been gathered.

- A discussion of the methodology is presented in chapter 3 follows on from chapter 2 for the comparison will be outlined. This section will detail the methods used to collect and analyze the data, as well as the criteria used to compare the two taxonomies.

- The analysis of the data will then be presented, in chapter 4 where the findings will be discussed and evaluated in relation to the literature review.

- Finally, the thesis will conclude in chapter 5 with a summary of the research and findings from chapter 4. The conclusion will include a discussion of the implications of the research, as well as suggestions for future work. Overall, this research aims to contribute to the study of network scanning classification.

With a reading time of approximately 1 hour and 30 minutes, by this time astounding 80 thousand packets of scanning traffic have been gathered by the network telescope. As later established, the amount of traffic on a Network Telescope is telling of how much scanning activity is present in internet traffic.

2

Literature Review

This chapter presents related works on setting labels on network traffic to further help us to detect unwanted traffic piercing our networks as incessant scanning of hosts by attackers looking for vulnerable servers has become a fact of Internet life Allman *et al.* (2007). The aim is to explore and summarise the basis for the analysis section in this research. The literature review will provide an overview of the history of scanning in section 2.1. The study by Allman *et al.* (2007) serves as the cornerstone of this section, providing valuable insights into the development of scanning over a 12.5-year period. Understanding the history and evolution of scanning is essential in detecting and preventing hostile network traffic in the future.

In section 2.2 the topic of Network Telescopes and their usage in research will be discussed. The aim is to provide an overview from previous studies, including Moore, Shannon, *et al.* (2004), Yegneswaran, Barford, and Ullrich (2003), and Wustrow *et al.* (2010). The purpose is to explain the reason for conducting this research using Network Telescope data and to highlight the various methods used to analyze this data in previous studies.

In section 2.3 on classifications, scanning techniques will be introduced. As an example, a simple and commonly used classification by Yegneswaran, Barford, and Ullrich (2003) will be presented to provide a classical understanding of scans. Furthermore, Bou-Harb *et al.* (2014) will also be included in this section, offering a more comprehensive view of scanning techniques accompanied with key definitions to enhance the reader's knowledge of TCP flags and protocols.

The above-mentioned sections of the literature review serve as the foundation for the two main taxonomies that will be discussed later on. These taxonomies are proposed by Barnett and Irwin (2008) and Liu and Fukuda (2018).

2.1 The history of scanning

Allman *et al.* (2007), stated that the diversity of the internet has made it difficult to categorize various forms of internet activity. It is a great task to untangle the complicated web that is internet activity. One of the classifications of internet activity is called *cyber scanning* in Bou-Harb *et al.* (2014). This activity refers to probing network services to find vulnerabilities. When using the word *find*, it means looking for known or possible vulnerabilities. Staniford *et al.* (2002) mention two general purposes for scanning, a primary and a secondary, for attackers to conduct a scan. The primary purpose of gathering information about the reachability of a network, often including IP and port information for TCP and UDP. The secondary purpose is flooding a network.

In figure 2.1, we can see an overview of the anatomy of attacks. As a first step in an attack, Bou-Harb *et al.* (2014) points to the act of Scanning. In their research, they highlight scanning as a precursor technique to launching various scanning attacks towards running services, network information, intrusion attempts, and malware deployment. This scanning traffic was classified in Staniford *et al.* (2002) as either *horizontal* or *veritical* scanning traffic.

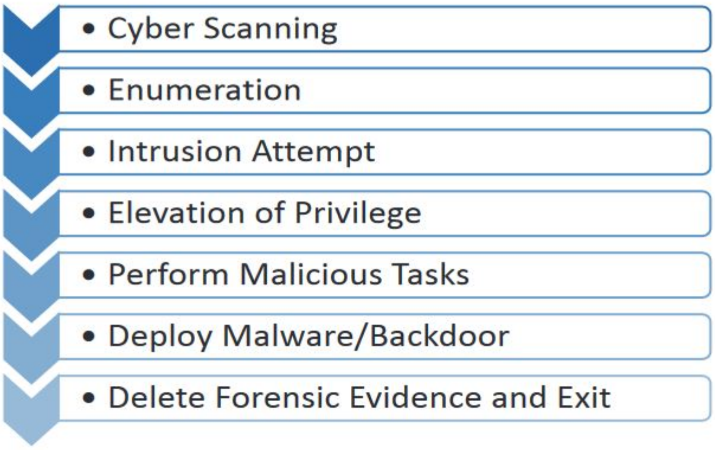

• Cyber Scanning
• Enumeration
• Intrusion Attempt
• Elevation of Privilege
• Perform Malicious Tasks
• Deploy Malware/Backdoor
• Delete Forensic Evidence and Exit

Figure 2.1: The Anatomy of a Cyber Attack (Bou-Harb *et al.* 2014)

As the networks have become more interconnected through the years, the threat of cyber-attacks has become more prevalent. Tools such as *Nmap* have been made public. Nmap enables users to perform detailed scans of the network, including; identifying hosts, open ports, and vulnerabilities. Nmap can also provide detailed information about operating systems and applications, making it easier for attackers to target specific vulnerabilities (Orebaugh and Pinkard 2008).

Allman *et al.* (2007) describes the evolution of scanning activity from the start of their study in 1994 until 2006. For their historical research, the writers only used a single site on a /30 netblock. A /30 netblock allows for 6 usable IP addresses in the netblock which is commonly used for smaller networks. The network traffic logs were collected over 12,5 years from Lawrence Berkeley National Laboratory(LBNL). Throughout their study, they encountered 628 million scans from 2.4 million distinct IP addresses. It is important to note that what they classify as a “scan” or a “scanner” is based on protocols analyzed by the Bro Intrusion detection system [1], Barnett and Irwin (2008) showed that there were flaws in the scan-detection engine of Bro in their research. We must also take into account

[1] Bro has later been renamed Spice

that the Bro Intrusion detection system varied over the course of the research from Allman *et al.* (2007).

To understand the result in Allman *et al.* (2007), we need to understand how they classified *scanners* and *non-scanners*. They used a combination to classify individual connections and all connections from a particular remote host. To classify individual connections, Allman *et al.* (2007) used a summary of connection time, duration, ports, bytes, and protocols. By using the mentioned classifications, their general approach is based on attempts that do not result in established connections representing possible scans. A similar classification is used in Barnett and Irwin (2008), excluding bytes.

From the start of the study in 1994, Allman *et al.* (2007) describes the beginning of the study as virtually no scanning activity. They later observed an increase in 1998 but with significant differences on a month-to-month basis. It was not until October 1999 that they observed scans of 30 scans per day in a month, which only happened once prior to this date. This suggests that prior to this time period, the number of scans being performed was relatively low and sporadic.

However, in the study, they noticed a significant shift in 2001. They observed that more connections transitioned from being legitimate to being part of scanning activity. The shift in 2001 correlated with the emergence of the *Code Red* and *Nimda* worms outbreaks. These worms were able to spread rapidly due to vulnerabilities in the Microsoft Windows operating system and caused widespread damage to computer systems and networks. The two worms were closely linked to the routing instabilities observed in 2001, degrading the end-to-end functions of the global internet (Cowie *et al.* 2001). The impact of these worms on the internet infrastructure and the increased activity of scanning traffic indicates that the worms played a major role in the shift observed in the study by Allman *et al.* (2007).

As we can see in figure 2.2 there is a relatively low volume of scanning traffic before a significant spike in 2001, related to the *Code Red* and *Nimda* worms, as mentioned earlier, the spike is marked A in the figure. Once again, in 2004, there is a major increase in scanning hosts, marked as point B in the figure. This visual representation of the data further supports the idea that the emergence of these worms in 2001 had a significant impact on the increase in scanning activity observed in the study. We can also see a consistently increasing rate of Non-Scanners, Allman explains this by highlighting the growth of general network traffic. If we look a Pang *et al.* (2004), they illuminate our understanding further of all the traffic that occurs in Allman *et al.* (2007) research.

Pang *et al.* (2004) characterizes traffic that is “up to no good” as *nonproductive*. The *nonproductive* traffic consists of hostile reconnaissance scans, backscatter, spam or exploit attempts. The traffic is characterized by its failure to establish a connection. They term this traffic *background radiation*. In the mentioned traffic logs from LBNL, Pang *et al.* (2004) shows that *background radiation* concurred to more than double the site’s successfully established incoming connections. Wustrow *et al.* (2010) found that in their research that they had an increase of roughly 100% each year of *nonproductive* traffic from 2007-2010 when observing an unused IP address space. While *productive* traffic had an increase of 50% each year over the same period when observing commercial internet traffic in the research by Labovitz *et al.* (2010).

In the research from Borgnat *et al.* (2009) conducted over a 7-year period from 2001-2008, they found that traffic always had a large number and variety of anomalies. The traffic analyzed in that research is

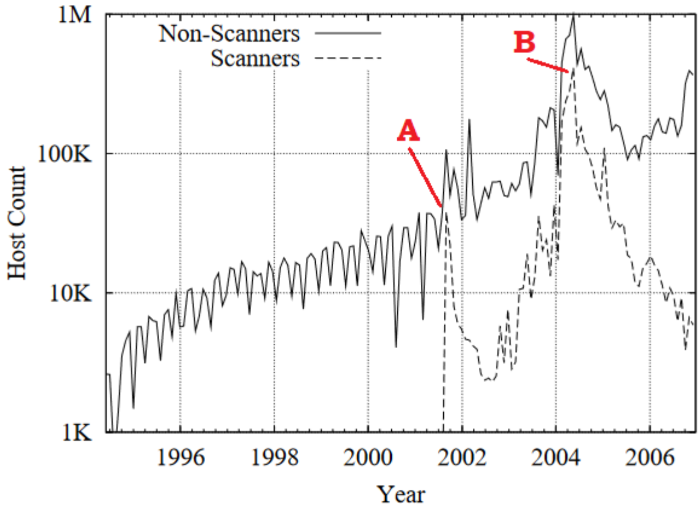


Figure 2.2: Host-Level summary of incoming traffic (Allman *et al.* 2007)

comparable with the traffic in the research by Allman *et al.* (2007). Borgnat *et al.* (2009) ask questions about what *normal* and *regular* traffic actually is, when observing commercial internet traffic. To avoid that question, this research will utilize data obtained from Network Telescopes, as it is considered to be more accurate for finding scanning connections, as shown in the research by Moore, Shannon, *et al.* (2004) and Yegneswaran, Barford, and Ullrich (2003). The next section of the thesis will explore the concept of Network Telescopes and the type of data they provide.

2.2 Network Telescope

In order to establish taxonomies for network scanning, it is essential to have access to a comprehensive dataset of scanning traffic. The classification and analysis of this data require a sample that accurately reflects real-world scenarios of scanning events. Collecting enough real-world data to analyze is a formidable task, The impractical nature of collecting data that could be either legitimate or not is a time and storage-consuming exercise. Borgnat *et al.* (2009) raises the question of what *normal* or *regular* traffic actually is. In the study, they used a *Measurement and Analysis on the WIDE Internet*(MAWI) dataset, which consists of 15 minutes of traffic captures per day, captured from a trans-Pacific link between Japan and the United States (Fontugne *et al.* 2010). Borgnat *et al.* (2009) found that through this period, each data set had a large variety of anomalies, making it difficult to pinpoint *normal* data.

For the purpose of analyzing illegitimate traffic, the data set should consist mostly of such traffic, avoiding the time wasted on analyzing *normal* data. One method of obtaining this type of data is through the use of Network Telescopes. As defined by Moore, Shannon, *et al.* (2004), in the research of Network Telescopes, they are a section of the IP address space where little to no legitimate traffic is exist. The use of Network telescopes emerged as the main mechanism for gathering a great sample size of illegitimate traffic at the start of the 21st century. Wustrow *et al.* (2010) mentions a variety of names, including *network sinks* as used in Yegneswaran, Barford, and Plonka (2004). In this research *Network Telescopes*, as suggested by Moore, Shannon, *et al.* 2004 is used.

The term *Network Telescopes* derives from the study of astronomy, where astronomy telescopes are widely used for studying portions of relevant data, which in astronomy would be specific regions of the universe. Pointing the telescope towards the appropriate region quickly captures data that can be easily processed (Moore, Shannon, *et al.* 2004). In networking terms, a telescope would be set up at an address space to provide samples of internet traffic that, in theory, would capture events such as; Denial-of-Service attacks (DoS) and network scanning. Moore, Shannon, *et al.* (2004) states that Network Telescopes attract data that could be used for analyzing and characterizing globally occurring events based on *unsolicited traffic*, unsolicited traffic referring to network scans, and "spam" traffic. This is confirmed by Yegneswaran, Barford, and Plonka (2004) who states that packets destined for an unused address are almost always malicious.

The fact that a network telescope is monitoring an unused IP space, is a useful mechanism for measuring and understanding internet attack behavior (Yegneswaran, Barford, and Ullrich 2003). Yegneswaran, Barford, and Ullrich (2003) used traffic logs of rejected data from a Network Intrusion Detection System (NIDS) to examine potential characterizing patterns of the earlier mentioned *Code Red* and *Nimda* worms' behavior as non-worm scans. They confirm that Network Telescopes would be a more effective way to gather and analyze more relevant data than NIDS in their section on Network Telescopes, stating that Network Telescopes are a useful mechanism for understanding internet attack behavior. Yegneswaran, Barford, and Plonka (2004) proposes the use of network telescopes as an optimized way to expand knowledge of abuse activity without problems associated with NIDS.

A Network telescope can be implemented in a number of ways. Irwin (2012) proposed an explanation of different modes of operations for Network Telescopes, ranging from passive to live systems. A passive network telescope offers a complete logging feature, logging all traffic and not giving any response. At the other end of the scale, the live systems are endpoints offering the same capabilities as a public Internet site. The nature of most network telescopes not establishing a connection makes it possible to categorize the monitored traffic as scanning traffic Allman *et al.* (2007). But as established by Barnett and Irwin (2008), Pang *et al.* (2004) observed that this traffic also reflects background radiation which could be other than scanning traffic.

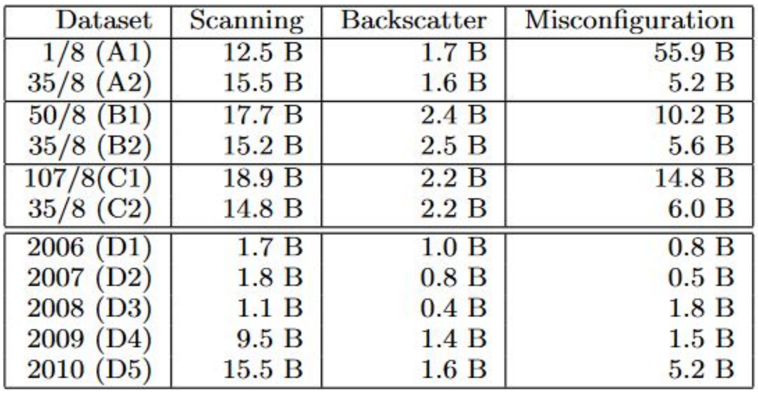
2.2.1 Internet Background Radiation

To understand why the use of Network Telescopes is beneficial for research of mostly illegitimate traffic, there is a need to understand the nature of the traffic that occurs on these unused internet address blocks. The approach had heavy use before, but until the study from Pang *et al.* (2004) there was not a detailed characterization of internet background radiation proposed (Wustrow *et al.* 2010). The most striking aspect of their analysis was the widespread presence of *Internet Background Radiation* (IBR), its size, the diversity of targeted services, and the highly dynamic nature of many aspects of the observed traffic (Wustrow *et al.* 2010). As mentioned earlier, they classify this traffic as *nonproductive*. The study from Pang *et al.* (2004) considers some of this traffic to be prevalent and widespread, hence they use the term IBR to describe it. Pang *et al.* (2004) states that *nonproductive* traffic is either directed towards non-existent addresses, inactive servers, or servers that do not receive it. It could also be the result of a hostile reconnaissance scan, *backscatter* from a flood attack, spam, or an attempt to exploit a vulnerability. In their research, they listen to traffic from "thousand to millions of IP addresses" (Pang *et al.* 2004), from several unused network blocks. They found that nearly 30,000 packets per second on a Class A network were IBR. In their research, they filtered the excessive amount of traffic by using *iSink*, an LBNL-Sink set up with Honeyd, which was introduced by Yegneswaran, Barford, and Plonka (2004). The different configurations of these Network Telescopes, lead to widely-ranging results. The algorithmic filtering from Pang *et al.* (2004) is a factor problematized by Wustrow *et al.* (2010). In their research, they showed that the differences in result observed in Pang *et al.* (2004) was due to environmental factors, not algorithmic factors.

Background radiation can be classified into three main groups (Wustrow *et al.* 2010). *scanning*, *backscatter*, and *misconfigurations*. In Pang *et al.* (2004) they noted that TCP protocol packets dominated all observed network blocks, mainly TCP packets consisting of TCP SYN flags (up to 95%). ICMP packets also contributed to a large number of packets, while UDP packets represented a smaller amount. This traffic is classified as scanning traffic. In figure 2.1 we see that scanning traffic represents a big part of the overall data in Wustrow *et al.* (2010). In Pang *et al.* (2004), *backscatter* is referred to as unsolicited traffic responding to spoofed attacks. This is often achieved by manipulating the source IP address of the packets to match that of a trusted host or network. *Backscatter* traffic varied in representation throughout their research. Moore, Voelker, *et al.* (2001) studied denial-of-service attacks after a series of high-visibility sites were attacked. They found that *unsolicited* traffic is mostly responses to spoofed attacks, but it is not possible to validate that all traffic is.

Misconfiguration is a result of software and hardware errors (Wustrow *et al.* 2010). In their research, they found that misconfigurations had the most fluctuating result between the observed network, as seen in table 2.1. This traffic is also termed as *benign* traffic, benign being legitimate IBR traffic that is produced as a result of network misconfiguration or as a response to DDoS attacks stemming from other sources (Liu and Fukuda 2018).

Table 2.1: Billions of packets received per week for each pollution type (Wustrow *et al.* 2010)



Dataset	Scanning	Backscatter	Misconfiguration
1/8 (A1)	12.5 B	1.7 B	55.9 B
35/8 (A2)	15.5 B	1.6 B	5.2 B
50/8 (B1)	17.7 B	2.4 B	10.2 B
35/8 (B2)	15.2 B	2.5 B	5.6 B
107/8(C1)	18.9 B	2.2 B	14.8 B
35/8 (C2)	14.8 B	2.2 B	6.0 B
2006 (D1)	1.7 B	1.0 B	0.8 B
2007 (D2)	1.8 B	0.8 B	0.5 B
2008 (D3)	1.1 B	0.4 B	1.8 B
2009 (D4)	9.5 B	1.4 B	1.5 B
2010 (D5)	15.5 B	1.6 B	5.2 B

2.3 Scanning Classification

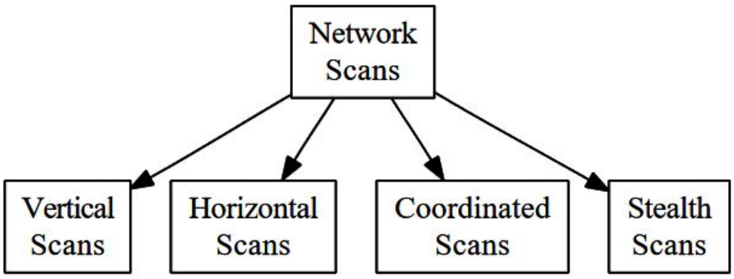
This classification offers a different view on scanning from the taxonomies that we later will introduce. The classic scanning technique classification proposed by Yegneswaran, Barford, and Ullrich (2003) offers a broad understanding of scanning techniques. The classifications are based on four areas. In section 2.4, Bou-Harb *et al.* (2014) will be presented for a more in-depth classification of different scanning techniques.

2.3.1 Classic Classification

This classification of scans was introduced in Staniford *et al.* (2002). It is *horizontal*, *veritical*, and as a combination of the two, *block* scans. Yegneswaran, Barford, and Ullrich (2003) revisited these categorizations and adapted them to a broad categorization of scans including four well-known scan types, the two first mentioned plus *coordinated* and *stealth* scans. This classification can be seen in figure 2.3.

The taxonomy later introduced from Barnett and Irwin (2008) claims to be a simplistic taxonomy, based on all four categorizations suggested in the next paragraphs. However, the taxonomy from Barnett and Irwin (2008) offers more flexibility by adding attributes to different scans. Their intent for the research was to use the upcoming classification as a baseline, although, they found that the four classifications overlapped each other in their own taxonomy. The classification introduced in this section is a great baseline for understanding the taxonomies from Barnett and Irwin (2008).

Figure 2.3: Classic Classification (Barnett and Irwin 2008)



2.3.2 Vertical Scans

A *Vertical scan* is defined by Staniford *et al.* (2002) as scans on some or all ports on a single host. The reason for conducting *vertical scans* is the attacker's interest in a particular host to gather information about running vulnerable services or finding vulnerabilities from subject information gathering. Yegneswaran, Barford, and Ullrich 2003 defines *vertical scans* as a sequential or random scan of multiple (5 or more) ports of a single IP address, they also declare that a *vertical scan* takes place during an hour period. Barnett and Irwin (2008) mentions that Yegneswaran, Barford, and Ullrich (2003) indicates regular use of automated tools to perform *vertical scans*. Yegneswaran, Barford, and Ullrich (2003) describes these as "*strobe*" scans, based on an original script-kiddie tool.

2.3.3 Horizontal Scans

Horizontal scans are concentrated on a particular service. A service can for example be Domain Name Service (DNS) hosted on port 53/tcp. *Horizontal scans* consist of scans from a single host to multiple IP addresses targeted on a single port. These kinds of scans are usually performed to identify hosts with a specific vulnerability Barnett and Irwin (2008). Yegneswaran, Barford, and Ullrich (2003) classifies a horizontal scan as a scan from a single source of several hosts (5 or more) on a netblock, aimed at the same target port. They further explain that these scans could be a recruiting peers procedure for larger DoS attacks.

2.3.4 Coordinated Scans

Coordinated scans, also referred to as *distributed scans* by Barnett and Irwin (2008), is a scan type that utilizes multiple sources. All sources are aimed at particular destination ports of a netblock. These scans usually stem from more aggressive scanners. An aggressive scan often involves a comprehensive scan of the target in a short amount of time. The scans are usually aimed at overwhelming the target with traffic, rendering it unavailable for legitimate users. These scans can be a result of multiple hosts collaborating to achieve a goal(Green *et al.* 1999), in some cases botnets (Bou-Harb *et al.* 2014).

2.3.5 Stealth Scans

Stealth scans are both *horizontal* and *vertical* scans performed with the intention of evading detection. The key parameter for a stealth scan according to Yegneswaran, Barford, and Ullrich 2003 is a maximum threshold of one hour, meaning that each connection from the host is no further apart than one hour. Scans from one host exceeding the threshold are considered unrelated.

2.4 Scanning Techniques

Scanning techniques differ from the classic classification in section 2.3.1. In this section, a classification of scanning techniques presented in terms of scanning abilities and exchange messages is provided. Bou-Harb *et al.* (2014) offers a review of known scanning techniques. We display a summary of the relevant depth, of use in this research, of the techniques. The Scanning Techniques presented here offer five branches; Open Scan, Half Open scan, Stealth scan, sweep scan, and miscellaneous scans.

2.4.1 TCP flags

Most of the techniques presented in the upcoming sections use the Transmission Control Protocol (TCP). TCP uses flags to indicate the state of a connection (Taylor 2005). It can also be used to pinpoint anomalies in a connection (Fukushima and Goto 1999).

SYN & ACK

Synchronize (SYN) flags are utilized to synchronize the sequence numbers during the establishment of a new connection between two hosts. Normally, both parties send and receive a pair of packets with SYN flags (Fukushima and Goto 1999). SYN packets occur often in scanning traffic, Pang *et al.* (2004) clearly shows that TCP dominates all three networks in their study. They also found that up to 95% of scanning traffic included TCP SYN flags.

Acknowledgement (ACK) flags are used as an acknowledgement from the receiving side of a connection. The flag comes with an acknowledgement sequence number. ACK flags are typically included in all packets, except the first (Fukushima and Goto 1999).

Three-way handshake

As TCP is connection-oriented, the conversation between a client and a host must establish a connection before communicating. The initial connection reserves resources at both ends for the remainder of the conversation. For TCP this procedure is called a *three-way-handshake* stemming from the establishment of a connection consisting of three packets. This process utilises SYN and ACK flags. The three-way handshake is executed by two communicators validating themselves by interchanging three sequential, equally sized TCP packets (Kuhrer *et al.* 2014). This procedure is depicted in figure 2.4. The first packet of the handshake includes a Sequence number (random number), SYN flag, Max segment size (to avoid fragmentation), and window size (buffer capacity). The second packet answers with an acknowledgement number (sequence number +1) and ACK flag in addition to the mentioned fields in the first packet. The third packet consists of the sequence number, acknowledgement number, and ACK flag. The data transfer can now begin (Bou-Harb *et al.* 2014).

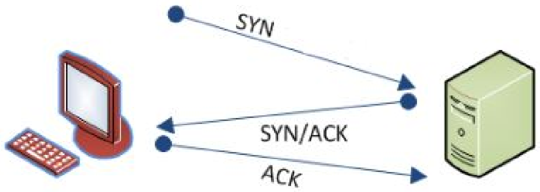


Figure 2.4: TCP Three-way-handshake

FIN, RST & PSH, URG

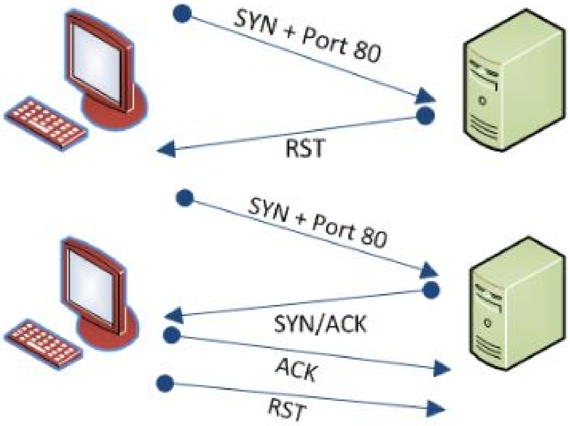
FIN (Finish) and RST (Reset) are similar in the way that they terminate a connection. FIN flags are transmitted when the data transfer is finished. There is normally no data loss. Both ends of a connection usually exchange FIN flags at the end of a connection (Fukushima and Goto 1999), if not, the receiver of the FIN flag can continue the communication. The RST flag is used to abort a connection. The receiver of the flag stops communication immediately, data can be lost when using the RST flag.

Data with URG (Urgent) flags are prioritized and sent to the application layer, even if there is more data waiting. The URG stands for the urgency of the data, and the flag alerts that the receiver is asked to process these packets before any other. The URG flags are rarely used (Fukushima and Goto 1999). The PSH (Push) flag is seen more often. The nature of the Transport layer is to fill up the buffer size, (established in the three-way handshake), before sending data. For some applications, this behavior is not desirable. By sending the PSH flag, the packets are processed when received (Fukushima and Goto 1999). The main difference between the two is the fact that for PSH flags packets are delivered in sequence, while for URG flags, packets are delivered out of sequence. The flags presented here will later aid our understanding of scanning techniques

2.4.2 Open Scan

Open scan (or as reffed to in Bou-Harb *et al.* (2014), *vanilla scan*), is the most basic of scanning techniques. The name, *vanilla*, originated from the fact that *open scans* follows the TCP three-way handshake, which is a standard procedure for every TCP-based connection. The Scanning technique is simple in the way that there are mainly two possible outcomes. If an attacker tries to access a port on a host, the host replies with either an ACK flag or an RST flag. The ACK flag means the port is open, RST flag means the port is closed. The outcomes are depicted in 2.5. Open scans are easy to detect and are not preferred due to their inefficiency and the high chance of triggering security measures.

Figure 2.5: Open Scan outcomes (Bou-Harb *et al.* 2014)



2.4.3 Half Open Scan

The name *Half-Open scan*, tells us the nature of this scan. The scan aims to gather information about ports, if they are open or not, without completing a full TCP three-way handshake. The reason for utilizing a *half-open scan* has two main advantages. Half-open connections (SYN-SYN/ACK-RST) are not logged by destination applications, which means that in theory, a half-open scan would be stealthier than an open scan. However, the drawback is that the use of TCP flags means the connections can be logged by a firewall (Bhuyan *et al.* 2011). The second advantage stems from that the connection to the application is not established, it spares the system resources on the host side.

In Bou-Harb *et al.* (2014), *Version detection scans* are said to utilize *Half-Open scans* for finding running services, before running a *version detection scan*, which typically implies using a three-way handshake and banner grabbing (Shaikh *et al.* 2008). Shaikh *et al.* (2008) states that *Version detection scans* as part of a vulnerability assessment, is a key reconnaissance activity to find possible exploits.

2.4.4 Stealth Scans

As the name implies, the goal is to evade detection. As we have seen from the two previously mentioned scans, the use of SYN flags makes the scans easy to detect. As stated in Bou-Harb *et al.* (2014), *stealth scans* attempt to bypass filtering devices by using different flags instead of SYN to appear as valid traffic. *Stealth scan* techniques do not complete the standard three-way handshake (de Vivo *et al.* 1999).

There are multiple scanning techniques under *Stealth scans*. SYN-ACK scans initially receive an RST flag for an open port and no response for a closed port. This is similar to a *half-open scan*, but with an added ACK flag that corrupts the three-way handshake. This technique is similar to *IDLE scans*, except *IDLE scans*, uses a *zombie* host to hide the source IP of the attack.

FIN, Xmas Tree, and Null Scans are grouped together in this classification on the basis that they are similar techniques. They are all based on FIN scan behaviour (de Vivo *et al.* 1999). FIN segments are dropped by listening ports or receive RST flags for closed ports (de Vivo *et al.* 1999). In figure 2.6, we see an example of these scans, here as a Xmas Tree technique. Since no TCP sessions are created for any of these scans. none of these should appear in any application logs (Bou-Harb *et al.* 2014). It is possible to make these scans even harder to detect by utilizing fragmentation (de Vivo *et al.* 1999).

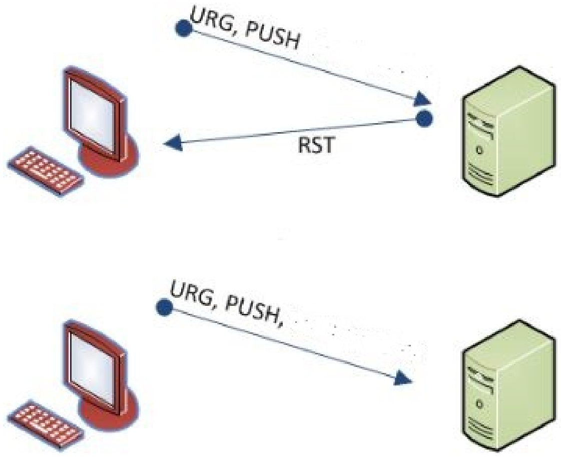


Figure 2.6: Xmas Tree scan outcomes (Bou-Harb *et al.* 2014)

TCP Fragmentation scans are as explained in Oberheide and Karir (2006) and stressed in Bou-Harb *et al.* (2014) not a scanning technique in itself, but rather the process of executing a scan. In Oberheide and Karir (2006), they state that IP packet fragments are trivial to craft and can easily bypass firewall rules. In their text on Honeyd, they show examples of how IP fragmentation is often incorrectly reassembled in software, making it hard to catch. By sending smaller fragments of a malicious packet, the packet can bypass most intrusion detection systems (de Vivo *et al.* 1999).

2.4.5 Sweep Scans

Sweep scans are different from the previously mentioned scans in the way that they are not scans to identify ports, but rather to identify hosts (Bou-Harb *et al.* 2014). The main goal of *sweep scans* is to identify active hosts. One way to identify active hosts is by *ICMP echo Request scan*. The simple idea behind this type of scan is to send a *ping* (ICMP type 8) packet to a host, if the host answers (ICMP type 0), the target is active. If there is no answer, the target is offline Arkin 1999. The *ICMP echo scans* is mentioned by Mirkovic and Reiher (2004), they characterized the scan by packet count.

An interesting addition to this category is *TCP-SYN scans*. Similar to *half-open scans* in the fact that an SYN packet is sent to a host, the TCP-SYN scan is not looking for open ports, but rather active systems. If the SYN packet gets any answer, the system is active (Bou-Harb *et al.* 2014), and the session is topped

2.4.6 Miscellaneous Scans

Miscellaneous scans include scans with various protocols. The most notable scan from this category is *UDP scans*. As UDP connections are not dependent on a handshake to establish a connection, it leaves *UDP scans* amplified in effectiveness (Bou-Harb *et al.* 2014). UDP protocol use can generate a multitude of attacks. Still, when it comes to scanning, the nature of UDP being a connectionless service, relevant information may not be given to the scanners (Bhuyan *et al.* 2011). Bhuyan *et al.* (2011) mentions in their research that UDP ports is easily blocked and if used, *UDP scans* are mostly used to find open ports. This can be done by a host sending a packet to a victim and the victim replies with ICMP type 3 if the port is unreachable or UDP data if the port is open, shown in figure 2.7.

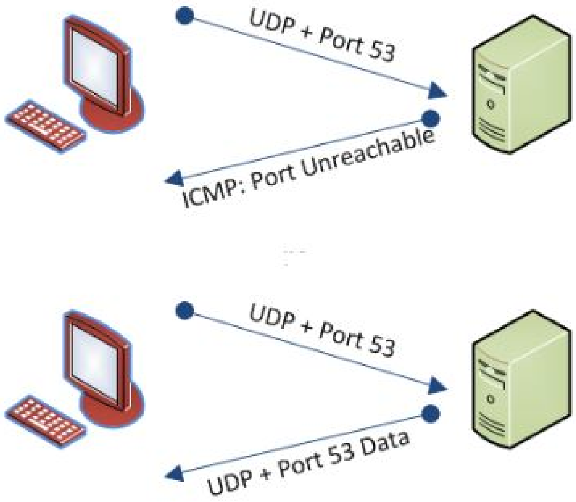


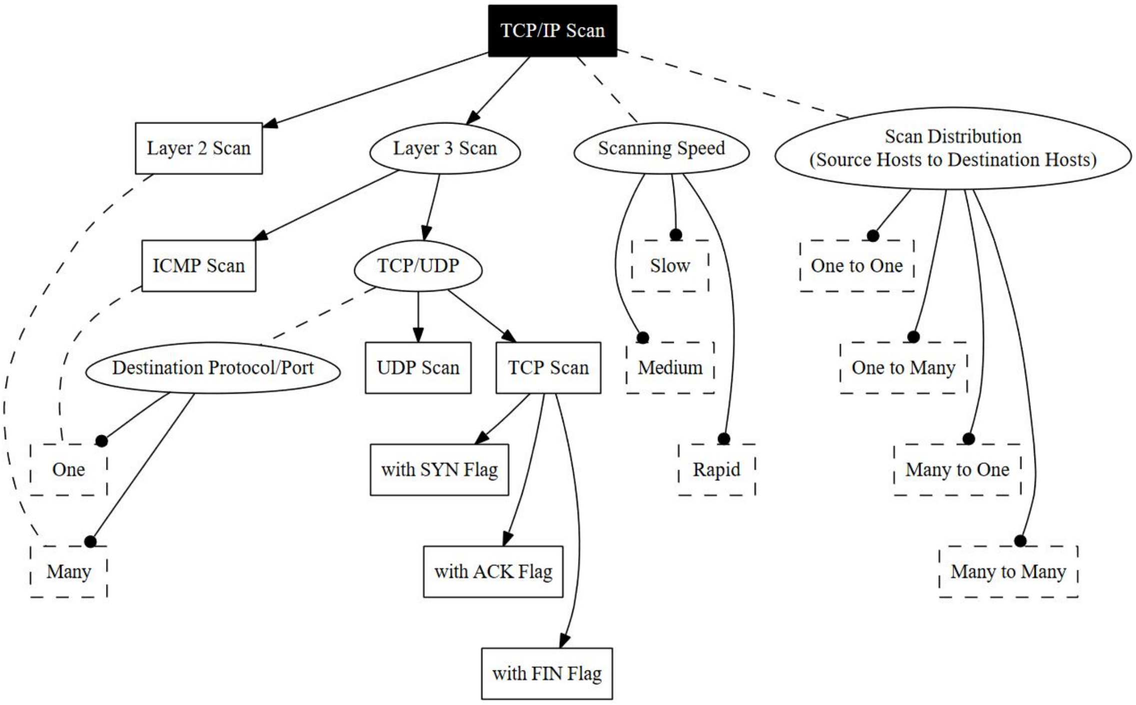
Figure 2.7: UDP scan outcomes (Bou-Harb *et al.* 2014)

IP protocol scans are also a part of this category. The goal of such a scan is to potentially find IP protocols in use in the victim's system, by asking the host for a protocol, the attacker can determine if a host uses that protocol if there is an answer in the form of an RST packet, then yes, If there is no answer, then no (Bou-Harb *et al.* 2014).

2.5 Barnett & Irwin (2008)

Barnett and Irwin (2008) base their constructed taxonomy on the classic classification presented in section 2.3.1. The proposed taxonomy, in their own words, encompasses all the features of the classification presented in figure 2.3. Additionally, Barnett and Irwin (2008) formed their taxonomy on observations of data from a Network Telescope. They conclude their research by saying that their proposed taxonomy is a more flexible version of what is explained in the section 2.3.1. however, they do not presume their taxonomy is the most complete. In their taxonomies, there are seven main classifications with different attributes. These are highlighted by square blocks in figure 2.8, depicting the taxonomy.

Figure 2.8: Taxonomy of Network scanning Techniques (Barnett and Irwin 2008)



2.5.1 Attributes

For the taxonomy, groups of attributes are proposed, these attribute groups include *Scanning Speed*, *Scan Distribution*, and *Destination protocols*. A total of nine total attributes are proposed. Because scans differ in speed to potentially avoid detection, the *scanning speed* is a measure attributable to all types of Scan types, speeds range from slow to rapid, with medium also being represented. For *Scan Distribution*, there are a total of four attributes seen on the far right side of figure 2.8, they include two options, *one* and *many*. They are respectively one IP or many IPs. The *destination protocol/Port* attribute group represents two different uses in scan categorization, depending on the scan type. Of the three groups proposed, the *destination protocol/port group* is stated to be attributed in a complex manner (Barnett and Irwin 2008).

2.5.2 Scan types

The taxonomy consists of three main scan types, respective Layer Two, ICMP, UDP, and TCP scans. As we can see in figure 2.8, they use the TCP/IP model. Layer two, in that context, refers to the internet layer. Layer two scans differ from ICMP, TCP, and UDP scans, due to their nature of targeting different Layer three protocols. This scan is a recognition activity for finding which protocols the host would respond to on the target network at Layer Three, the Transport Layer. An IP-protocol scan, which is mentioned in Bou-Harb *et al.* (2014), is a typical example of a *layer two scans*.

Under the category: Layer 3 scan, we find *ICMP scan*, where Echo request (ping) packets are sent to a host. ICMP scanning is different from other scan types, as they are only classified as a scan if one instance of a ping occurs. Bou-Harb *et al.* (2014) shows that *ICMP scans* as categorized under *Sweep scans*, these scans try to identify active hosts, not open ports.

UDP scans represent scans using UDP. These scans are attributable to speed, distribution, and destination port. In the research by Barnett and Irwin (2008), it is said that the most common *UDP* scan involves multiple scanning hosts targeting a single destination port, or single scanning hosts targeting multiple destination ports. *UDP scans* are mentioned in Bou-Harb *et al.* (2014) under miscellaneous scans.

TCP scans are the most common scan type, they also offer the most complexity. All *TCP scans* are attributable in the same way as *UDP scans*. An extra option for *TCP scans* is added caused of the Three-way-handshake for TCP connections, the option flags. Although there are more than SYN, ACK, and FIN flags, these flags were chosen on the basis that they were identified as being set when crafting packets. As we have seen in Bou-Harb *et al.* (2014), there is a multitude of scanning attacks that uses other flags than SYN, ACK, and FIN.

2.6 Liu & Fukuda (2018)

Liu and Fukuda (2018) mentions in their introduction that there is a need for a simple and effective taxonomy of anomalies from Network Telescope traffic. They state that examining longitudinal data traffic without obtaining packet inter-arrival time (IAT), is more efficient. To do this there is a need for pinpoint anomalies of the traffic in the datasets. They propose five main types of anomalies; *scanning*, *one flow*, *backscatter*, *IP fragment*, and *small activities*. They claim a 94% definition rate, leaving only a 6% unlabelled source rate. They showed in their research that most of the scanning traffic is placed in either one or two anomaly-group. They also include *Other* events for each Protocole. This anomaly group is categorized as an anomaly group because it includes all traffic not picked up by any other categorization.

2.6.1 Parameter Dependency of the Taxonomy

In their taxonomy research, Liu and Fukuda (2018) analyzed the impact of various parameters on their classification system. They assessed the effects of three parameters: *N* (Number of hosts), *R* (Ratio of scan occurrence), and *M* (Many hosts). They evaluated each parameter by changing one at a time and determining its impact on the results using real data. Their findings showed that the *R* parameter had a significant impact on the number of major anomalies, whereas the *M* parameter acted as the boundary between heavy and light scans, with a setting of 3 in the final iteration of the taxonomy. The range of *R* was between 25 and 90, while *M* was between 2 and 4 in their experiments.

The parameters *N1*, *N2*, and *N3* had a more significant effect on labelling major anomalies. Changing *N1* and *N2* in the range of 3-7 had a significant impact on SMALL SYN and LIGHT TCP NETWORK SCAN, while the group SMALL UDP remained unaffected. They ultimately set *N1* = *N2* = 5. Similarly, varying *N3* in the range of 10-20 affected SMALL SYN but had no impact on LIGHT TCP NETWORK SCAN, with a final value of *N3* = 15 in the taxonomy.

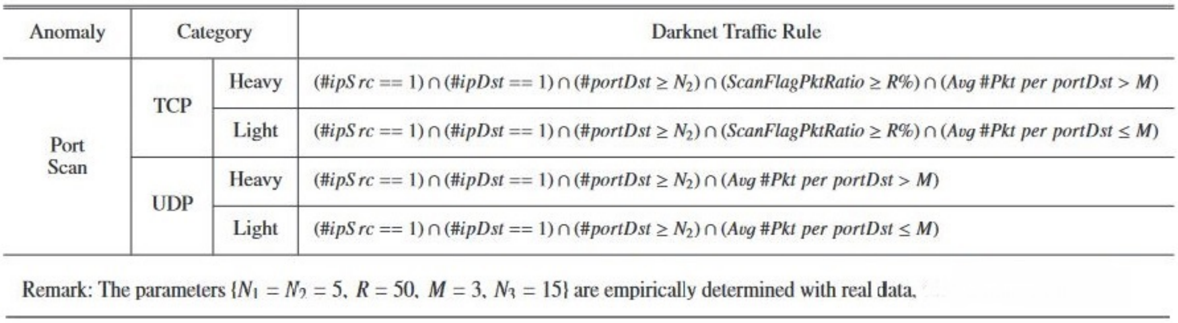
2.6.2 Port Scan

PORT SCANS is referring to what we saw earlier in Bou-Harb *et al.* (2014) as the most common cyber scanning technique, especially TCP PORT SCANS. A PORT SCAN event is raised when a number of destination ports for a *flow* (source IP - destination IP connection) exceeds the threshold of *N2*. This applies to both TCP and UDP PORT SCANS.

Additionally, for TCP PORT SCANS, flags are taken into account. The proportion of packets with scan flags must be larger than the threshold ratio *R* in a flow. The inclusion of flags for TCP is as stated in Fukushima and Goto (1999) an important addition for pinpointing anomalies, as explained in 2.4.1. The scanning flags they take into account are “SYN”, “FIN”, “FIN-ACK”, and “NULL”. The addition of flags is important to know if the attacker is looking for active destination ports, so the activity can be categorized as port scanning. These flag combinations were discussed in Bou-Harb *et al.* (2014), although, we saw several different approaches of port scanning flag combinations than what this taxonomy offers. The scan flags presented by Liu and Fukuda (2018), stem from the research by Yegneswaran, Barford, and Ullrich (2003) who classified Network Telescope traffic based on TCP flags, the findings were later confirmed by Wustrow *et al.* (2010).

For port scanning, there are subcategories for each main category, respectively heavy and light traffic. This is measured by the Average packet per Destination port to be more than *M*=3 or equal or less than *M*.

Table 2.2: Port Scan Darknet Traffic Rules (Liu and Fukuda 2018)



Anomaly	Category		Darknet Traffic Rule
Port Scan	TCP	Heavy	$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst \geq N_2) \cap (ScanFlagPktRatio \geq R\%) \cap (Avg\ \#Pkt\ per\ portDst > M)$
		Light	$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst \geq N_2) \cap (ScanFlagPktRatio \geq R\%) \cap (Avg\ \#Pkt\ per\ portDst \leq M)$
	UDP	Heavy	$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst \geq N_2) \cap (Avg\ \#Pkt\ per\ portDst > M)$
		Light	$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst \geq N_2) \cap (Avg\ \#Pkt\ per\ portDst \leq M)$

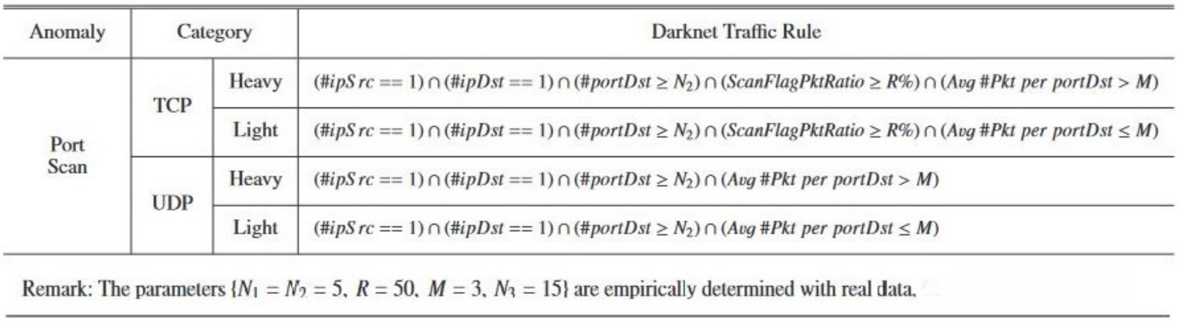
Remark: The parameters $\{N_1 = N_2 = 5,\ R = 50,\ M = 3,\ N_3 = 15\}$ are empirically determined with real data.

2.6.3 Network Scan

In network scans there are three categories; TCP and UDP, as in Port Scans, and in addition, ICMP. Also for Network scans, there are subcategories of heavy and light, as we saw in the port scan anomaly group. Network scans are unlike port scans attempting to find peers (Zombies) for DDoS attacks or find victims for an exploit. The difference between network and port scans for UDP and TCP in the taxonomy is that the Destination port equals 1, and destination IP $\geq$ *N1*. This derives from the theory that potential DDoS attacks usually stem from multiple agent machines Mirkovic and Reiher (2004).

The new addition to the anomaly group, ICMP, must by nature of the protocol have different rules to be categorized. Under *Sweep Scans* in Bou-Harb *et al.* (2014), we can see *ICMP scans* mentioned. ICMP echo request scans (ping scans) are used for finding active hosts. For ICMP network scans, only echo requests are considered.

Table 2.3: Network Scan Darknet Traffic Rules (Liu and Fukuda 2018)



Anomaly	Category		Darknet Traffic Rule
Port Scan	TCP	Heavy	$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst \geq N_2) \cap (ScanFlagPktRatio \geq R\%) \cap (Avg\ \#Pkt\ per\ portDst > M)$
		Light	$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst \geq N_2) \cap (ScanFlagPktRatio \geq R\%) \cap (Avg\ \#Pkt\ per\ portDst \leq M)$
	UDP	Heavy	$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst \geq N_2) \cap (Avg\ \#Pkt\ per\ portDst > M)$
		Light	$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst \geq N_2) \cap (Avg\ \#Pkt\ per\ portDst \leq M)$

Remark: The parameters $\{N_1 = N_2 = 5,\ R = 50,\ M = 3,\ N_3 = 15\}$ are empirically determined with real data.

2.6.4 One Flow

One flow is the category hosting large traffic flows. They are defined by conversations with more than 15 packets for a flow. Liu and Fukuda (2018) explains that this anomaly could be a result of network misconfiguration, which was characterized by Wustrow *et al.* (2010) as a result of hardware and software errors, shown in table 2.1 at an increasing level, but in Liu and Fukuda (2018) as decreasing.

2.6.5 Backscatter

Backscatter is mentioned in the section 2.2 on Network Telescopes. In the taxonomy from Liu and Fukuda (2018), backscatter is considered as mostly response packets to (D)DoS attacks carried out elsewhere on the internet. Backscatter is found by Sources that send multiple packets of UDP, TCP, or ICMP. For each category in the anomaly group, there are different rules. TCP is pin-pointed by an array of flags, UDP is categorized by port source from DNS, NTP, NetBIOS, and SNMP, and ICMP by ICMP type responding to ICMP echo.

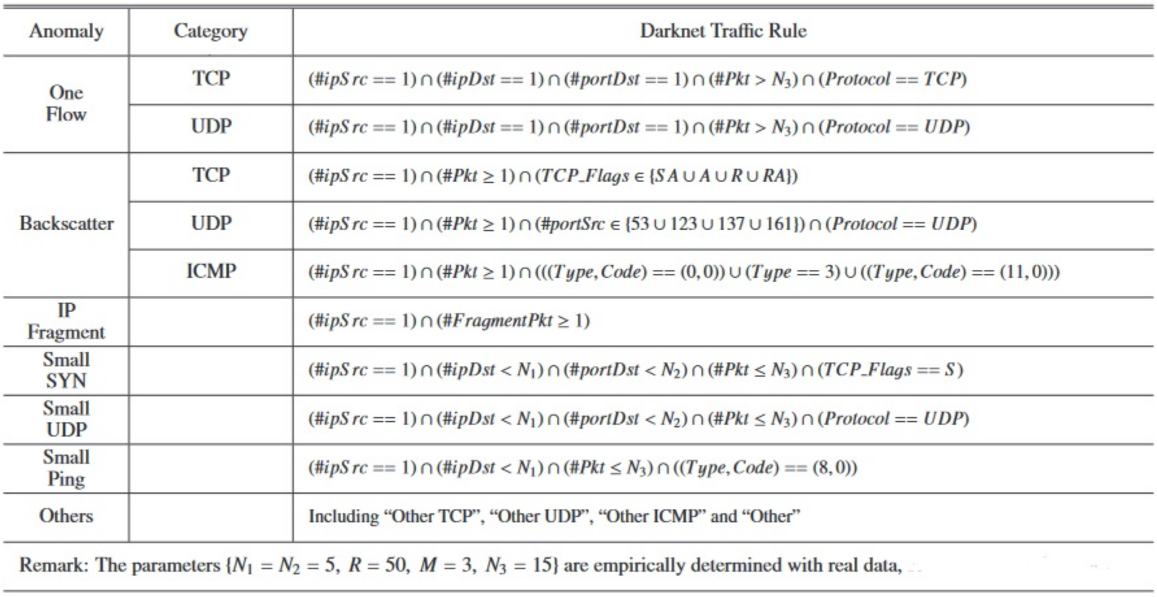
2.6.6 IP fragment

Traffic placed in this anomaly group contains exactly one source and one or multiple fragmented packets. This group represents DoS attacks or attempts to defeat packet filter policies (Liu and Fukuda 2018). TCP fragmentations scans are explained in *Stealth Scans* at section 2.3.5, as not a scanning technique, but rather a scan process. Fragmented packets are also presented in Mazel *et al.* (2014), explained as attacks to waste the resources of the host, forcing it to reassemble the packets.

2.6.7 Small activities

Small activities include “Small SYN”, “Small UDP”, and “Small Ping”. These anomalies originate from a single source based on a specific packet, unlike port scans, network scans, and one flow. “Small SYN” and “Small UDP” are packets from one source, to a small number of destinations, with less than 5 ports used, and a total of packets less or equal to *N3*=15. The difference is that Small SYN is TCP packets with SYN flags and Small UDP uses UDP protocols. These categories are included on the basis that Liu and Fukuda noticed that, what they categorized as, “Small SYN” and “Small UDP” events occurred often within certain time periods. Small Ping is also similar to the two other mentioned small events, except the packets are ICMP echo requests. The anomaly includes ICMP type == 8. It should be noted that Small events was often miscategorised in smaller datset (Liu and Fukuda 2018)

Table 2.4: One flow, Backscatter, IP fragment, and Small events (Liu and Fukuda 2018)



Anomaly	Category	Darknet Traffic Rule
One Flow	TCP	$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst == 1) \cap (\#Pkt > N_3) \cap (Protocol == TCP)$
	UDP	$(\#ipSrc == 1) \cap (\#ipDst == 1) \cap (\#portDst == 1) \cap (\#Pkt > N_3) \cap (Protocol == UDP)$
Backscatter	TCP	$(\#ipSrc == 1) \cap (\#Pkt \geq 1) \cap (TCP\_Flags \in \{SA \cup A \cup R \cup RA\})$
	UDP	$(\#ipSrc == 1) \cap (\#Pkt \geq 1) \cap (\#portSrc \in \{53 \cup 123 \cup 137 \cup 161\}) \cap (Protocol == UDP)$
	ICMP	$(\#ipSrc == 1) \cap (\#Pkt \geq 1) \cap (((Type, Code) == (0,0)) \cup (Type == 3) \cup ((Type, Code) == (11,0)))$
IP Fragment		$(\#ipSrc == 1) \cap (\#FragmentPkt \geq 1)$
Small SYN		$(\#ipSrc == 1) \cap (\#ipDst < N_1) \cap (\#portDst < N_2) \cap (\#Pkt \leq N_3) \cap (TCP\_Flags == S)$
Small UDP		$(\#ipSrc == 1) \cap (\#ipDst < N_1) \cap (\#portDst < N_2) \cap (\#Pkt \leq N_3) \cap (Protocol == UDP)$
Small Ping		$(\#ipSrc == 1) \cap (\#ipDst < N_1) \cap (\#Pkt \leq N_3) \cap ((Type, Code) == (8,0))$
Others		Including "Other TCP", "Other UDP", "Other ICMP" and "Other"

Remark: The parameters $\{N_1 = N_2 = 5,\ R = 50,\ M = 3,\ N_3 = 15\}$ are empirically determined with real data,

2.7 Initial comparison of Taxonomies

From 2008 to 2018 a lot has happened in the research field. We can not determine that these taxonomies are representative of the 10 years between. However, they are a good representation of conflicting approaches to research presented in chapter 2. The two taxonomies have the same goal of classifying traffic that is known to be *unproductive* traffic based on the traffic stemming from Network Telescopes as explained in section 2.2.

From the tables 2.8 from 2008, and 2.2, 2.3, and 2.4 from 2018, there is an immediate difference looking at the taxonomies. As stated in Barnett and Irwin (2008), their taxonomy has seven unique *scans*. As mentioned in section 2.5, these are Layer 2, ICMP, UDP, and TCP with flag options. In the 2018 taxonomy from Liu and Fukuda (2018), we find 23 different classifications of scanning traffic. That's more than 3 times the amount proposed by Barnett and Irwin (2008), which underlines the earlier statement of the 2008 taxonomy being a more straightforward approach to classifying scans. The taxonomy from 2018 is heavily based on classifying scans stemming from a single source, it does not offer the ability to find many-to-many and many-to-one which is often seen in scanning traffic as part of distributed/coordinated attacks explained in section 2.3.4, but with their approach to a specific classification of traffic stemming from one source, it would be simple to further analyze and compare traffic from multiple sources to find the attributes offered in Barnett and Irwin (2008). While the taxonomy proposed by Liu and Fukuda (2018) has a more scanning technique focus in their approach, Barnett and Irwin (2008) presents a taxonomy of overall scanning events (Mazel *et al.* 2014), based on level of Interaction of each source with a host (Pittman 2023)

The difference in using seven classifications vs 23 can be translated to the world of colour schemes. On one hand, we have the seven rainbow colours; Red, Orange, Yellow, Green, Blue, Indigo, and Violet. On the other hand, we have an unlimited array of colours that lies in between the seven rainbow colours. In the next section, we will further see if it is possible to compare the 23 anomalies from Liu and Fukuda (2018) with the much simpler classifications from Barnett and Irwin (2008).

2.7.1 Attribute Options

Although it is more obvious in the taxonomy from Barnett and Irwin (2008), both taxonomies offer an array of attributes complementing every classification in the taxonomy. In Barnett and Irwin (2008) we have *Scanning speed*, *Scan distribution*, *Destination protocols* and *Destination port*. *Scanning speed* is the only attribute of these that is not represented in the taxonomy from 2018. The three others are represented in the 2018 taxonomy, although with small differences. The exclusion of *scanning speed* is notable, as different scanning speeds often can highlight what kind of traffic is incoming, faster speeds are often linked with D(D)oS attacks, slower speeds are often linked with stealth scans as mentioned in 2.4.4.

Destination protocols are included in the 2008 taxonomy as a complement to the options of *Layer 2 scans*, which are not included in 2018, where the protocols are sub-categories of scan anomalies, and could never be categorised as *many*, as for layer 2 in the 2008 taxonomy. As mentioned in section 2.7, the exclusion of the possibility of many source IPs to a host IP removes the possibility of catching many-to-many, and many-to-one without further analysing, but one-to-one and many-to-many could be found with the *N1* parameter, used for a number of hosts. 2018 also offers the *N2*, *N3*, *R%* and *M* options, these are set limits for categorising traffic. They are explained in section 2.6.1.

2.7.2 TCP scanning traffic

TCP offers the most complexity between the two taxonomies. The addition of TCP flags makes the traffic more susceptible to getting analysed further, which makes TCP the protocol with the most options. They share similarities in their focus on identifying different types of scans using various criteria. Both Barnett and Irwin (2008) and Liu and Fukuda (2018) categorise scanning methods, with the latter further categorizing these scans as heavy or light based on the number of packets and the average number of packets per destination. In comparison, Barnett categorizes scans based on port, scanning speed, and distribution.

The two taxonomies differ in their criteria for further analysis of TCP traffic. In 2018 the set of flags is specifically said to be scan flags. In anomalies in the classifications, the ratio of these flags per packet is a part of that specific classification. Barnett and Irwin (2008) classify all TCP traffic in categories independent of flags, but three flags have their own category. The flags included in the taxonomy are TCP with SYN, TCP with ACK, and TCP with FIN. Although there are 6 flag options for TCP communications, these three flags were identified in Barnett and Irwin (2008), as being set specifically. The inclusion of the TCP flag's importance is described in section 2.4.

- 2008: SYN, ACK, FIN
- 2018 Scans: SYN, FIN, FIN/ACK, NULL

- 2018 Backscatter: SYN/ACK, ACK, RST, RST/ACK
- 2018 Small SYN: SYN

Liu and Fukuda (2018) includes a category for backscatter, which identifies TCP traffic with specific TCP flags. As earlier mentioned, backscatter traffic was found in Moore, Shannon, *et al.* (2004) to be a major part of data gathered on network telescopes. One flow is also added as an option to identify possible misconfiguration traffic identified in IBR from section 2.2.

Overall, both research offers different approaches for categorizing TCP scans. Liu and Fukuda (2018) taxonomy focuses on categorizing specific scanning techniques, while the taxonomy from Barnett and Irwin (2008) emphasizes identifying different types of TCP scans based scanning speed, and distribution.

2.7.3 UDP scanning traffic

The study conducted by Liu and Fukuda (2018) employs the same anomaly detection techniques for UDP traffic as those used for TCP traffic, with the notable exception of flag analysis. Unlike TCP traffic, UDP does not use flags in its communication protocol. By examining the usage of a specific range of ports, Liu and Fukuda (2018) is able to identify UDP backscatter traffic. Barnett and Irwin (2008) also utilizes the same network scanning taxonomy for UDP traffic as for TCP traffic, without the options for flags.

2.7.4 ICMP scanning traffic

ICMP traffic in the taxonomies proposed by Liu and Fukuda (2018) and Barnett and Irwin (2008) differs from that of TCP and UDP traffic. The lack of port use results in not including ICMP traffic in port scans and one flow in the 2018 taxonomy. Barnett and Irwin (2008) restricts ICMP scans based on ICMP types as opposed to ports. In contrast to the 2018 taxonomy, no other ICMP type than echo is considered in the taxonomy from Barnett and Irwin (2008), categorizing ICMP anomalies with a minimum limit of 1 ping packet.

It is worth noting that ICMP is primarily used for carrying application data. As a result, ICMP traffic is less common than TCP and UDP traffic. It is still essential to consider ICMP traffic in network scanning as it can be used in malicious activities, such as DDoS attacks and sweep scans

2.8 Summary

The research by Allman *et al.* (2007) is a great indicator of how scanning activity went from almost non to significant in the period between 1994-2007, as we can see in figure 2.2. Also Wustrow *et al.* (2010) shows us in table 2.1, an increase in scanning activity from 2006-2010. The spike shown in figure 2.2 is related to *Code Red* and *Nimda* Worms, who initialized the increased scanning activity. As the use of the Internet spiked in the 2000s, there was shown to be a connection between an increase in general network traffic and scanning traffic.

As highlighted by Anbar *et al.* (2013), scanning is the first step taken by attackers when launching an attack. Figure 2.1 shows how scanning practices can lead to the deployment of malware. In section

2.4, different attacks are explained to greater increase the knowledge of how packets can be manipulated to avoid detection and gather information to eventually perform an attack. We see how different combinations of flags for TCP can potentially avoid detection, while still getting a wanted result, for UDP and ICMP, finding hosts is the main concern. These techniques highlight the importance of knowing how scans behave compared to other network traffic.

When looking at network traffic, finding scans can be tricky. In a web of network traffic, finding anomalies is the key to withstand malicious activities for network administrators and security professionals (Anbar *et al.* 2013). In Pang *et al.* (2004) they found that analyzing data should be with a dataset consisting of mostly *nonproductive* would be beneficial to understand what this traffic is, as anomalies traffic is harder to find in a dataset with *normal* network traffic. Moore, Shannon, *et al.* (2004) introduced the network telescope where little to no legitimate traffic exists.

Network Telescopes are, as mentioned in section 2.2, a method to gather mostly illegitimate traffic. In the research by Borgnat *et al.* (2009), they explain that the dataset for analyzing anomalies in a network should not consist of *normal* data as this could raise questions of what *normal* and what *illegitimate* traffic is. We learn in Yegneswaran, Barford, and Plonka (2004) that the traffic present from a Network Telescope dataset is optimized for analyzing anomalies as *productive* connections would not normally access what Moore, Shannon, *et al.* (2004) calls an unused IP space. At the same time, the data would present a globally occurring event. Although these datasets could present problems with *nonproductive* data that is not scanning traffic, as shown in Pang *et al.* (2004), Network Telescope data consists of IBR that could be other than scanning traffic. Wustrow *et al.* (2010) later classified IBR into three groups as shown in table 2.1, but most of the data should be of a scanning nature.

With the knowledge acquired from Chapter 2, A test needs to be done. The research set out to test two taxonomies with the same datasets, to see how they compare. As established in section 2.7, the taxonomies from Barnett and Irwin (2008) and Liu and Fukuda (2018) demonstrate contrasts and commonalities regarding the research from the literature review. The goal of the next section is to see if they offer some of the same results despite their distinctions.

3

Methodology

As discussed in the preceding chapter, Network Telescopes captures IBR traffic. According to Wustrow *et al.* (2010), includes scanning traffic, backscatter, and misconfiguration. However, effective analysis of the massive volume of data collected by Network Telescopes necessitates advanced analytical techniques. In this regard, Wireshark has been found to be inadequate for analyzing data, as mentioned in Barnett and Irwin (2008) research. To address this challenge, this study will use a set of Python scripts that can analyze the data collected.

The primary objective of the study is to explore whether a set of taxonomies can be utilized to identify the types of scanning traffic identified in the literature review. To achieve this objective, the study will utilize two taxonomies previously introduced in research by Barnett and Irwin (2008) and Liu and Fukuda (2018). The study's approach will be quantitative, and it will test whether the observed anomalies in the dataset align with the respective research's claims. The study aims to compare the traffic classifications obtained from the two taxonomies using the same dataset, providing insights into the evolution of the field of categorizing network traffic from the 2000s to the 2010s

In this chapter, the dataset used for the research is introduced in section 3.1. The scripts for analyzing the data are covered in section 3.2. Further, what we are looking for in chapter 4 is introduced in section 3.3.

3.1 Dataset

The data was obtained through Security and Network Research Group (SNRG) based in Makhanda, South Africa. The dataset used for this research was collected over a 4-month period between

December 2020 and April 2021. Traffic was captured using a small /24 network telescope. Packets were captured directly with TCPdump and written to a file. Filters were used to ensure that only traffic directed at the monitored IP ranges was captured. Firewall rules on the capturing host and router ensured that no return traffic was possible. The dataset consists of four files, parted on the first day of a new month. The dataset is, as collected by a Network Telescope, not edited and consists of scanning traffic and other traffic. Table 3.1 provides a high-level summary of the data used for the research [1].

Start date	End date	Packets	Unique Src IPs
2020-12-01	2021-01-01	36M	1 380 774
2021-01-01	2021-02-01	34M	1 466 824
2021-02-01	2021-03-01	36M	1 171 514
2021-03-01	2021-04-01	37M	1 355 500

Table 3.1: Monthly Summary of dataset

Figure 3.1 show an overview of the traffic in December 2020. An even distribution of packets throughout the month as well as unique IP addresses are plotted. There is a notable spike in packets 19th of December. The spikes show an increase of 29% of packets compared to the rest of the month. Unique IP addresses are evenly distributed at around 4500 each day. The graph is also representative of the overall data from January, February, and March, although as we see in figure 4.1, there are spikes and drops throughout the four months. The traffic consists of 80% TCP data, UDP is around 18% and ICMP is around 2%.

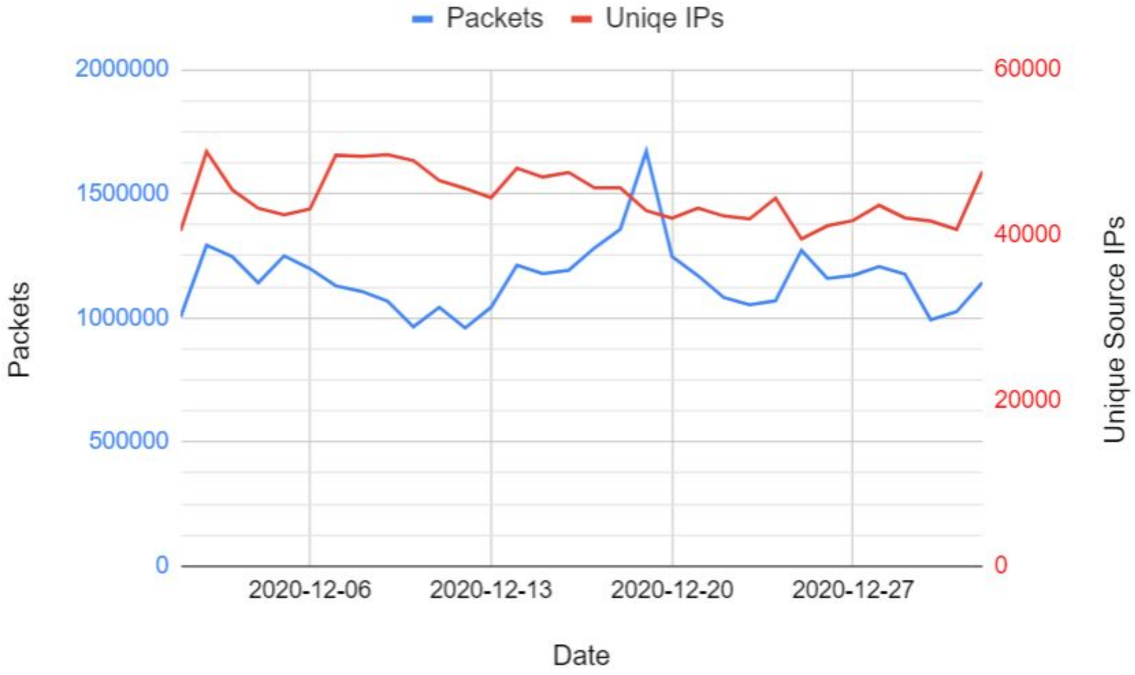


Figure 3.1: Packet and IP Address overview: December 2020

It must be mentioned that even in a real-life dataset, like the one utilized in this research, labeling traffic with 100% security requires an enormous amount of human expertise and time (Bhuyan *et al.* 2011). This research is based on the expertise from the research of Barnett and Irwin (2008) and Liu

[1] Counts of unique source IPs are reset each day

and Fukuda (2018). The researcher's job is to evaluate the differences between these approaches. Bhuyan *et al.* (2011) further mentions that the lack of available datasets results in many false alarms. Liu and Fukuda (2018) used two different datasets, the results are similar but not identical. This research will add a third dataset to the analysis of that particular taxonomy, exploring the variations of traffic seen in Network Telescopes.

3.2 Creation of Scripts

Testing of the two taxonomies will, as mentioned in the introduction chapter at 1, be done by creating two separate python scripts. The dataset is saved as packet capture files, and the python module *dpkt*[2] is used to parse the files. Dpkt offers simple packet parsing with definitions for the basic TCP/IP protocols. Working in dpkt and python will allow for effective analysis of large amounts of packets at the same time as constructing the two taxonomies for result-based research.

The taxonomies are, to the best of the researcher's ability, recreated. The recreations are created with the option of changing parameters and saving information for closer analysis of particular data in the next section. The taxonomy-scripts are based on tables 2.2, 2.3, and 2.4 for Liu and Fukuda (2018) and figure 2.8 for Barnett and Irwin (2008). The knowledge from the literature review factored in for the scripts to print valuable information in the next chapter. The scripts are found in Appendix-B.

3.2.1 Limitations

The research aims to recreate the scripts in a way that the output result will be as similar as possible to the original taxonomies. Although the approach of this goal was taken, there are some limitations to the script that affects some points, but not the overall analysis. The analysis will exclude Layer 2 scans which is a prevalent factor in Barnett and Irwin (2008), due to limitations in packet capture information.

For analyzing overview results, the research from Liu and Fukuda (2018) offers an overview of the traffic table. For Barnett and Irwin (2008) there are direct results to compare a successful adoption of the taxonomy against. Here the research will rely on the text from the study.

3.3 What are we looking for

In the research from Liu and Fukuda (2018), the study offers an evaluation of the classification to see how their two datasets compare to each other. A similar table of results is constructed for the dataset in the first part of the analysis, to see if the results are comparable. For the taxonomy proposed by Barnett and Irwin (2008), a similar table will be constructed for a better overview of their classification, of course, substituted anomaly categorizations to fit the taxonomy. From these tables, we will be able to see a greater overview of where the packets and IP sources are placed. Looking at these tables will aid us in better understanding how the taxonomies perform, as well as pointing toward the search of validating the recreations of both scripts.

For each taxonomy, an analysis of traffic placed in each category is undertaken, before comparing the two taxonomies. To compare the two taxonomies, percentages of unique IP sources that are placed in

[2] https://dpkt.readthedocs.io/en/latest/

each category will be checked against each other, to better understand how the categories correlate across the two taxonomies. These statistics will help to gain an initial understanding of differences in the taxonomies.

The scripts for the analysis are made to be flexible for broader analyzing capabilities. It will be of great use when looking closer at the traffic in each anomaly category to find certain events to compare across the taxonomies. By looking closer at certain traffic flows and source IP addresses, we will uncover where, or if, the traffic is picked up by the taxonomy/taxonomies. The analysis is evaluated using Wireshakr with smaller packet capture files for checking these events closer, and finding out what triggered or did not trigger the taxonomies to classify the event as anomalies.

3.4 Summary

The research set out to use Network Telescope data to verify the value of the taxonomies. In section 3.1, the dataset is introduced. The dataset covers a 4-month period, consisting of around 36 million packets each month. The composition of the data is 80% TCP traffic, 18% UDP traffic, and 2% ICMP traffic. In section 3.2, the creation and limitations of the scripts, striving to replicate the two taxonomies was introduced. For section 3.3, a methodology for the next chapter provides an understanding of how the analysis is chronologically built to reach the aim of the research. The aim of the analysis is to compare the results of the two taxonomies against each other. A methodology for the research has been covered in this chapter, and in the next chapter, a detailed look at the analysis is shown.

4

Analysis

The analysis chapter serves as the point of merging for this research. Following an extensive literature review of the data expected from a Network Telescope, this chapter presents an empirical examination of the dataset. Our primary objective is to identify network events, as discussed in section 2.2, and assess their placement in the two taxonomies, as well as their likelihood of being overlooked. Additionally, we will inspect the traffic distribution depicted in figure 4.1, in order to determine if any differences exist regarding the impact of data spikes on the distribution of traffic across anomaly categories.

Initially, aiming to examine the taxonomy presented in Barnett and Irwin (2008) to place the categorization of the data. Although no data statistics are used for comparison with our results, some of the research findings have been indicated in the published article. Our objective is to determine the primary categorization of the data and provide explanations for the distribution of each anomaly group. Secondly, evaluate how the taxonomy proposed in Liu and Fukuda (2018) classifies the traffic from our dataset. Their research includes two tables that we will use to compare our results with theirs.

In section 4.3, we will investigate if any correlations exist between the anomaly categories in the two taxonomies by conducting a comprehensive examination of the anomaly categories and exploring potential scans suggested in sections 2.3.1 and 2.4. Our aim is to establish if any associations exist between observed scanning activity and the placement of traffic in the categories of the two taxonomies. In the end, We will investigate the *Other* categories from the taxonomy proposed by Liu and Fukuda (2018) to determine its location in the taxonomy proposed by Barnett and Irwin (2008).

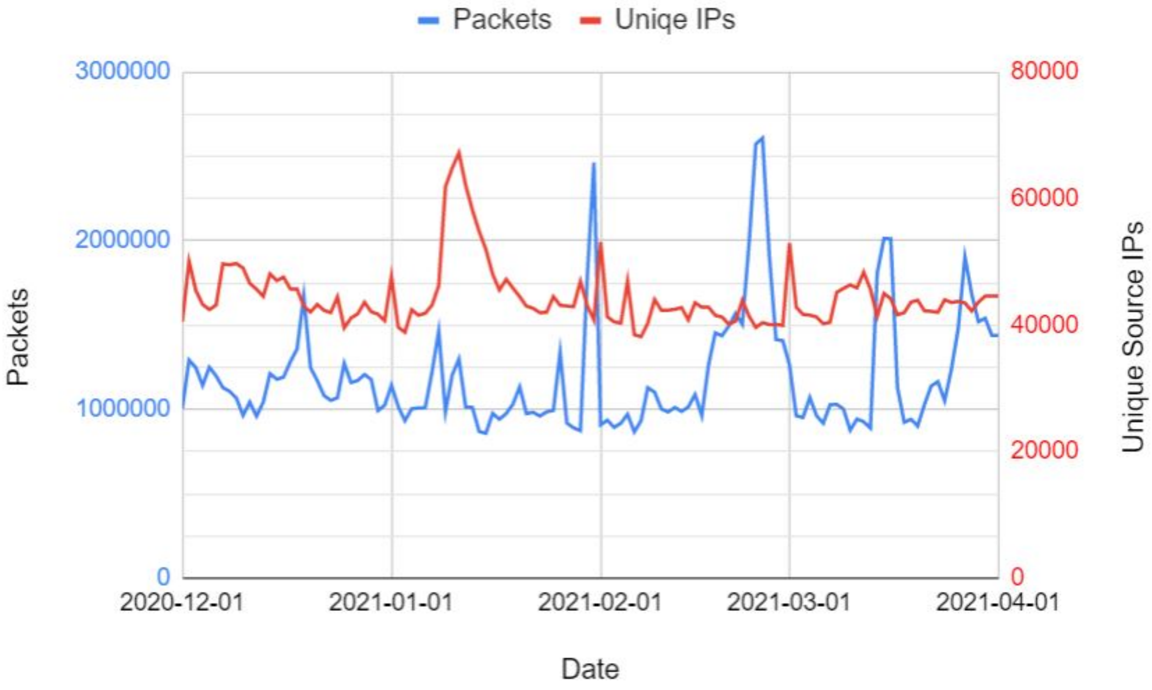


Figure 4.1: Packet and Source Address overview December 2020 to March 2021

4.1 Barnett & Irwin (2008) analysis

As mentioned in the section 3, the dataset consists of less ICMP and UDP traffic than TCP, which is normal. According to the research done by Pang *et al.* (2004), up to 95% of scanning traffic included SYN flags, which is a TCP indication for the state of connections. In a newer study conducted by Irwin (2022), TCP-SYN was said to be often used in spoofing attacks, only acquiring a single response packet to find active hosts as mentioned in section 2.4.5.

To get an overview of the data, table 4.1 shows the percentages of each distribution based on the total amount of traffic. In tables, 4.1 and 4.2, cells containing '-' represent less than 0.01%. The findings are consistent throughout the month of December. The table shows that most of the traffic was TCP, and a fairly big percentage was from one source to many hosts. The same distribution was most prevalent for ICMP and UDP. A closer look at the traffic in table 4.2 which is based on the number of packets for each protocol, shows UDP and ICMP show similar results as TCP, with slight variations. In the 2008 research, they found that One-to-One and Many-to-One were far more common than Many-to-One and Many-to-Many. In this chapter, the analysis confirms this statement.

Table 4.1: December 2020 barnett

Category	One-to-One	One-to-Many	Many-to-One	Many-to-Many
TCP	1.9%	86.08%	0.03%	0.01%
UDP	0.41%	10.72%	0.03%	-
ICMP	0.03%	0.72%	-	0.01%

Barnett and Irwin (2008) found that the most common TCP scan was One-to-Many with many ports. In this research, TCP One-to-Many accounted for 98% of TCP traffic and 86.08% of overall traffic. Most of this traffic was categorized as rapid or slow scans and primarily involved SYN flags. Further examination of the data reveals an average of 84% of traffic contacted many ports. Entries with many ports often used a combination of ports 80/tcp and 8080/tcp or ports 23/tcp, and 2323/tcp. These ports were also observed as the most active ports in Irwin (2022). Traffic targeting multiple ports was often classified as slow. For most of the slow traffic, multiple packets were used.

Those targeting a single destination port commonly utilized the same port as used the entries with many ports, in addition, port 22/tcp 445/tcp and 5555/tcp were used on a large number of hosts. 16% of One-to-many traffic contacted one port. Scans targeting a single port were often classified as rapid. Rapid scans mostly consisted of conversations with one or few packets, aligning with the findings in Irwin (2022) which showed port 80/tcp and 443/tcp were the only ports with significant traffic volumes. Based on the rapid nature of scans targeting a single destination port and multiple hosts, it is possible that this traffic could be attributed to backscatter. According to Wustrow *et al.* (2010), port 445/tcp is commonly used in backscatter traffic. The rapid nature of the traffic could also indicate that it is response packets to (D)DoS attacks, as they often consist of few packets.

In One-to-One scans, the source focuses on a single host and often tries multiple ports, these attributes are linked with “Half-open” scans, covered in section 2.4.3, where the source tries to find open ports. For the sources that tried one port, the majority of scans belonged to the “sweep-scans” category, as explained in section 2.4.5, where the source sends TCP-SYN packets to a host on a port to check if the host is active. The scans conducted on one port can be attributed to backscatter traffic.

Table 4.2: December 2020: Percentage of traffic from each protocol

Types		**Dataset (Day of month)**					
		1-7	*7-12*	*12-18*	*18-24*	*24-30*	*30-1*
One-to-One	TCP	2,29%	2,34%	1,97%	1,76%	1,76%	2,83%
	UDP	4,04%	3,93%	3,05%	2,86%	3,39%	5,20%
	ICMP	4,55%	4,30%	4,63%	3,61%	5,01%	5,20%
One-to-Many	TCP	97,67%	97,61%	97,99%	98,17%	98,19%	97,14%
	UDP	95,38%	95,92%	96,83%	96,98%	96,34%	94,30%
	ICMP	93,95%	93,88%	93,60%	94,72%	92,89%	92,12%
Many-to-One	TCP	0,03%	0,04%	0,02%	0,04%	0,03%	0,03%
	UDP	0,58%	0,15%	0,12%	0,16%	0,26%	0,50%
	ICMP	0,29%	0,19%	0,25%	0,41%	0,18%	0,05%
Many-to-Many	TCP	0,01%	0,01%	0,02%	0,03%	0,01%	-
	UDP	0,01%	-	-	-	0,01%	-
	ICMP	1,22%	1,63%	1,52%	1,25%	1,93%	2,63%

As with TCP, UDP traffic also exhibits a significant amount of One-to-Many scanning. However, the nature of this traffic is distinct from that of TCP. UDP scans often involve a source targeting one or two destination ports on multiple hosts, around 80% tried many ports. In contrast to TCP, the number of destination ports scanned in UDP is often smaller. The ports targeted by UDP scans are typically different from those of TCP. UDP often tried a combination of 8080/udp, 8081/udp, 8082/udp. These scans aim to detect vulnerable applications running on these ports and look for specific services associated with vulnerabilities. A notable percentage comes from source port 123/udp, which is linked with backscatter traffic, explained in section 2.6.5. One-to-Many UDP scans are generally categorized as medium or slow. Additionally, a significant portion of UDP traffic is categorized as One-to-One, constituting up to 5% of UDP traffic. All One-to-One UDP scans have a rapid nature and generally consist of one port in the 49152-65535 range and one or two packets. This type of scan could be attributed to backscatter since it does not possess the attributes regularly seen to identify active hosts or open ports because it only tries one or few ports and hosts.

Looking at table 4.2 in more detail, we can observe some variations between the protocols. ICMP traffic is generally categorized as One-to-Many, but we see more of the ICMP traffic classified as Many-to-Many, compared to TCP and UDP. The Many-to-Many conversations in ICMP are primarily in the slow-speed category and usually involve a large number of packets. These conversations often stem from hosts within the same /24 subnet. Although the number of unique source IPs for Many-to-Many is smaller than that of One-to-Many and One-to-One, the amount of packets for each entry is often higher than the other categories. As per the 2008 taxonomy, ICMP scan is based on the assumption that the initial flow had at least one echo ping packet. Therefore, we can infer that the echo (ping) packets are predominant in ICMP traffic. However, these conversations do not seem to be ICMP echo request scans as discussed in section 2.3. Rather, they appear to be coordinated scans where sources work together to target multiple hosts, often scanning an entire subnet at a time, as discussed in section 2.3.1.

The research by Barnett and Irwin (2008) found that ICMP traffic is less likely to use the “One to One” and “Many to One” classifications. In this research, we confirm that Many-to-One traffic is rare in our dataset, representing only 0.18% of the overall traffic. While the numbers are higher for UDP and ICMP, TCP traffic constitutes the largest proportion of traffic, with only 0.03% being classified as Many-to-One, lowering the average. One-to-One traffic is also seen as rare in ICMP in the research by Barnett and Irwin (2008), but in the December dataset, we found up to 5% of total ICMP traffic was classified as One-to-One scans. These conversations are all in the rapid category, and almost 100% of the traffic consists of a single packet. Since the 2008 taxonomy categorizes all ICMP conversations involving an echo (ping) packet, these one-packet flows must be Ping packets. The ICMP echo request technique discussed in section 2.3 involves sources sending ICMP echo packets to hosts to check their availability, usually for multiple hosts. Thus, we see such scans in the One-to-Many category. In contrast, One-to-One scans may represent attempts to find an active host in a subnet, check if the subnet is worth scanning thoroughly, most likley these flows are misconfiguration or backscatter traffic, probably a response to (D)DoS attacks. The sources of these ICMP packets are frequently observed sending only a single packet, indicating that this behavior is not indicative of scanning activity but may instead be attributed to misconfiguration.

4.2 Liu & Fukuda (2018) analysis

In this section, an overview of the results of the 2018 taxonomy is presented. A detailed table is presented with an explanation and analysis of the results. Table 4.3 provides a detailed breakdown of the overall traffic for the December 2020 dataset. It is important to emphasize that the main difference between the current research and the 2018 study lies in the duration of the analyzed data. Specifically, while Table 4.3 presents data from a single month, the 2018 study analyzed data spanning 12 years.

The percentages presented in table 4.3 are derived from the total amount of labeled sources for each protocol, with each category being represented by the number of entries from unique source IPs. Thus, each percentage represents the proportion of source IPs in each category compared to the total amount of labeled source IPs, which indicates the relative frequency of each anomaly in the dataset. Upon closer inspection of each anomaly, it is observed that several unique source IP addresses are often represented multiple times within the same anomaly. It is important to note that the taxonomy proposed by Liu and Fukuda (2018) also permits multiple labels for each source, and as such, it is not uncommon to see a considerable number of source IPs categorized in two anomalies, whereas it is less frequent for more than two categories. These instances will be examined in greater detail later.

In table 4.3, cells containing '-' represent less than 0.01%. The findings are consistent throughout the month of December. The *Other* anomaly in Table 4.3 ranges from a high of 2.3% to a low of 0.03%. This suggests that the proposed taxonomy effectively categorizes most of the traffic in the dataset, which aligns with the results presented in Liu and Fukuda (2018). The data categorized as "Other" consists of traffic without ports (for TCP and UDP) or ICMP type or code fields, which cannot be categorized. The "Other" anomaly reflects the findings of Liu and Fukuda (2018).

In figure 3.1, there are no significant data drops or spikes observed in the considered dataset. In the 2018 research, they used two datasets. One with a 7-day timespan, and one with a 1-day timespan. The dataset in this research encompasses longer time periods of 7 days, along with a 2-day period between December 30 and the first part of January 1. In the 2018 study, we explained that miscategorization of Small SYN could result from shorter time periods in section 2.6.7. It is noteworthy that approximately 72% of the source IPs in TCP traffic are categorized in the Small SYN category, while around 84% of the data in UDP traffic falls into the Small UDP category. Additionally, Small Pings at 79% also show a higher percentage than that reported in the 2018 research. This increase in percentages may be attributed to the shorter time period of one month analyzed in this study.

N% Is based on unique source IPs for each protocol

The research by Liu and Fukuda (2018) highlights that Small events, including Small SYN, Small UDP, and Small Ping, make up a substantial portion of the traffic. In line with the 2018 research, our results also show a slight increase in Small SYN events for the smaller 2-day dataset, explained in section 2.6.7. The high percentages of Small events suggest that most sources are likely to send a smaller number of packets to evade intrusion detection systems (IDS) (Liu and Fukuda 2018). Upon analyzing the most commonly used ports for small events, it was observed that TCP traffic primarily targeted well-known ports such as 22/tcp, 23/tcp, 80/tcp, 81/tcp 8080/tcp, and 445/tcp. The primary objective of these connections is to determine if the host was running services linked to these ports. For UDP, scans targeted less commonly used ports. Small UDP events often attempted to reach dynamic/private ports in the range of 49152-65535. These findings are consistent with those of the 2018 research. These scans search for specific vulnerabilities linked with services that could be running on these ports.

For ICMP small Pings this research differs from the study by Liu and Fukuda (2018). We observed that the percentage of Small Ping traffic in our dataset was approximately 79%. Section 4.1 highlighted a higher number of One-to-One ICMP traffic, which consists of a lot of the same traffic we see in Small Pings. In the 2018 taxonomy, anomalies are based on one source IP to less than 6 hosts, which aligns with the results of the 2008 taxonomy and Wireshark, indicating that many sources send single ICMP echo packets to only one source. These events are identified by the Small Ping anomaly category.

Table 4.3: December 2020: Percentage of all labeled Inbound IP Sources. Fukuda

Types			**Dataset (Day of month)**					
			1-7	*7-12*	*12-18*	*18-24*	*24-30*	*30-1*
Port Scans	TCP	H	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
		L	1.02%	0.98%	1.12%	1.08%	1.05%	1.06%
	UDP	H	0.02%	0.01%	0.03%	0.03%	0.04%	0.06%
		L	0.29%	0.39%	0.41%	0.53%	0.47%	0.29%
Network scan	TCP	H	0.38%	0.39%	0.45%	0.43%	0.39%	0.4%
		L	10.61%	10.26%	11.48%	12.64%	11.74%	11.57%
	UDP	H	0.14%	0.18%	0.19%	0.18%	0.19%	0.09%
		L	3.59%	4.78%	4.92%	4.91%	4.96%	8.72%
	ICMP	H	0.32%	0.3%	0.29%	0.32%	0.32%	0.58%
		L	0.41%	0.38%	0.38%	0.41%	0.42%	0.88%
one flow	TCP		6.29%	4.73%	4.89%	4.29%	4.49%	2.63%
	UDP		3.43%	6.27%	5.53%	1.13%	5.53%	2.65%
backscatter	TCP		2.26%	6.78%	4.2%	2.88%	5.08%	3.02%
	UDP		7.36%	1.46%	0.73%	0.31%	0.19%	0.1%
	ICMP		17.69%	20.07%	22.73%	23.83%	19.83%	16.75%
IP Fragment			0.05%	-	-	-	-	0.01%
Small SYN			74.91%	68.35%	71.67%	73.56%	70.43%	76.42%
Small UDP			78.61%	82.74%	85.07%	89.43%	85.22%	85.38%
Small Ping			81.54%	79.25%	76.56%	75.4%	79.38%	81.69%
Other TCP			4.53%	8.49%	6.18%	5.12%	7.20%	4.9%
Other UDP			6.33%	4.17%	3.14%	3.48%	3.4%	2.68%
Other ICMP			0.05%	-	0.04%	-	0.05%	0.1%
Other			2.3%	1.65%	0.24%	0.06%	0.03%	0.04%

When examining the category of one flow, a decrease in TCP traffic can be observed, ranging from over 6.29% down to 2.63%. The fluctuations in results are similar to those found in the study conducted by Liu and Fukuda (2018). Interestingly, the first 7 days of the month and the last 2 days of the dataset stand out as different from the average. As we have previously explained the results of the shorter 2-day period could stem from miscategorization of Small SYN. It is worth noting that the 6.29% observed during the first 7 days is simply a coincidence, as closer inspection of the data reveals that it is merely a continuation of the same data seen in the other 7-day stretches, with more entries.

In Section 2.6.4, it was noted that *one flow* are often a result of misconfigurations. The most common destination ports used in the *one flow* category are 22/tcp and 23/tcp. These same ports were observed as the most popular in the research conducted by Liu and Fukuda (2018). Additionally, ports 81/tcp and 443/tcp were commonly used in TCP *one flow*. The same port was observed for One-to-Many in the 2008 taxonomy in the entries with one destination port. In the case of UDP, the most commonly used destination ports for one flows were 53/udp and 5060/udp. Since the one flow anomaly is based on categorizing miscofngigurations, these port destinations could be a result of misconfiguration or responses to (D)DoS attacks. Some of the same ports were commonly found in the research conducted on background radiation by Wustrow *et al.* (2010).

backscatter is another anomaly used to categorize background radiation. It is noted that destination ports 80/tcp and 445/tcp are commonly used in backscatter, which is consistent with the findings of the 2018 research and Wustrow *et al.* (2010). The port usage is explained by the emergence of the Conficker botnet in 2008 (Wustrow *et al.* 2010; Irwin 2013), but as seen in this research, these ports continue to be commonly used. In terms of UDP backscatter, the most commonly used destination ports are 123/udp, 137/udp, and 53/udp. This indicates that the majority of the traffic has the same source and destination port, as the anomaly is based on traffic originating from ports 53/udp, 123/udp, 137/udp, or 161/udp

ICMP backscatter is the second most popular anomaly in ICMP. backscatter traffic is mostly responses to (D)DoS attacks (Liu and Fukuda 2018). In the 2018 research, they also found that ICMP backscatter traffic is a high percentage of the ICMP traffic with fluctuating results for each year, this concurs with the results from table 4.3, which shows that around 20.15% of ICMP traffic was backscatter. The fact that the ICMP backscatter anomaly is categorized by response ICMP messages, as explained in section 2.6.5, makes the category a reliable indicator of the amount of background radiation on the internet, as explained in section 2.2.

4.2.1 Network & Port Scans

The network scans anomaly is one of the major anomalies we observe in table 4.4. The table shows the difference between the unique source IPs and the number of entries for each anomaly. As we can see, the Port Scans and network scan anomalies allow for multiple entries from the same source IP, resulting in a significant increase in the number of events for these anomalies. This is because network scans often involve scanning multiple IP addresses within a given network range, which can generate a large number of entries for a single source IP.

Table 4.4: December 2020: Unique sources vs. entries

Anomaly	**Port Scan**		**Network Scan**		**One flow**	**Backscatter**	**Small events**	**Other**
	H	*L*	*H*	*L*				
Unique Sources	0.01%	0.93%	0.37%	10.2%	4.4%	3.91%	74.52%	5.66%
Number of entries	0.5%	36.61%	0.2%	23.89%	1.91%	1.71%	32.7%	2.47%

Unique Source IPs: N% Is based on number of unique Source IPs in each entire

Number of entire: N% Is based on number of entries

In the 2018 study, a significant increase in light TCP *network scan*s from 2008 to 2009 was observed, indicating that attackers have increasingly preferred to use TCP network scans with light traffic since 2008 (Liu and Fukuda 2018). The research also found that the Mirai botnet outbreak in 2016 placed most of the IP sources in the TCP light network scan category. Mirai used TCP-SYN packets to exploit weak passwords in Telnet port 23/tcp (Liu and Fukuda 2018). The *network scan* anomaly entries are based on one IP source to one Port destination to multiple destination IPs. The Mirai attack was frequently seen using port 23/tcp, this is why an increase in network scans was observed in the 2018 research since the anomaly is based on connections to one port. In our dataset, we found that port 23/tcp was used in around 9% of the conversations. We also observed that traffic with the ports that are used multiple times in network scans also appears in the *Port Scan* anomaly.

The Port Scan anomaly is based on a single source IP to a single destination IP, but with a number of ports exceeding 4. Since flows are based on IP-to-IP interactions, there are often multiple entries from the same source IP. Only about 3% of the entries in the Port Scan anomaly are from unique source IPs, whereas in the network scan category, about 43% of the entries are from unique source IPs. Further analysis reveals that most of the source IPs categorized in Port Scan, around 90%, are also found in the network scan anomaly group. This suggests that many of the source IPs categorized as Port Scan are also placed in network scans. On the other hand, for traffic categorized as network scans, only 30% of the source IPs were also found in the Port Scan anomaly group. This indicates that network scans are more diverse in terms of the source IPs involved. The entries exclusively categorized as Port Scans tend to involve smaller amounts of data, with most of the traffic from a source IP consisting of fewer than 30 packets. In contrast, the entries categorized as network scans tend to involve larger amounts of data, indicating more comprehensive probing activity.

Light Port Scans appear to be less frequent for UDP than for TCP. In the case of UDP Port Scans, only about 19% of unique IP sources are also represented in UDP network scans. It is worth noting that the percentage of heavy Port Scans in UDP is higher compared to TCP. While heavy Port Scans in TCP constitute only 0.41% of the total TCP traffic, the same figure for UDP is 1.28%. This suggests that UDP sources are more likely to send multiple packets to a host. This behavior can be attributed to the UDP Scan scanning technique mentioned in section 2.4.6, which is commonly used to scan for active hosts with UDP. Unlike TCP, UDP does not require any flags, which makes UDP scan more effective, and scanners are more likely to attempt scanning many ports and sending multiple packets

When examining the overall traffic distribution, it becomes apparent that both Port Scans and network scans are less common in UDP than in TCP, this is also true for ICMP network scans. For both UDP and ICMP, there is a higher percentage of Small events, suggesting a correlation between a high percentage of Small Events equals fewer Port and network scans indicated by Liu and Fukuda (2018). As mentioned in section 4.2, we observe a decrease in network scans and an increase in Small Events during a two-day period in December. Additionally, Table 4.3 shows that UDP and ICMP, which have fewer network scans, have a higher percentage of Small events. This indicates that UDP is more frequently used to locate active hosts or known vulnerable open ports. This result aligns with the findings of Barnett and Irwin (2008), whose taxonomy revealed that UDP data was predominantly sent from one source to multiple hosts, targeting multiple ports. Similarly, ICMP Small Pings are primarily used to identify active hosts.

As we have observed the correlations between the two categories, the reason for the Port Scan category to have a larger number of entries in table 4.4 is due to the nature of the traffic. An example of this is shown in table 4.9. This phenomenon occurs because larger traffic flows from a single IP source often attempt to contact multiple hosts on many ports, resulting in a higher number of entries in the Port Scan category compared to the network scan category

The *Other* anomaly is observed to have relatively similar percentages of TCP and UDP traffic, while ICMP traffic is notably lower. This can be attributed to the differences in how ICMP traffic is analyzed in the script, as ICMP traffic typically uses echo(ping) packets and has fewer rules for categorisation. As a result, most of the ICMP traffic is categorized in either network scans or Small ping categories. Conversely, traffic from ICMP categorized in the *Other* anomaly primarily consists of corrupt IP addresses. In comparison, TCP and UDP traffic is more accurately categorized based on the specific ports and number of scan packets used, resulting in more accurate results. Further analysis of traffic placed in the *Other* anomaly will be conducted later on.

4.3 Comparison

In the following sections, comparisons to demonstrate the differences in how the two taxonomies categorize the traffic from the same dataset are provided. The tables presented in the following sections include parentheses that denote the number of times a particular source IP appears in a given category, as well as the number of each category in those entries. For example, in Table 4.9, the entry for Light network scans (200) implies that a specific source IP is present 200 times in that anomaly. Additionally, 1(1) indicates that all entries had only one port, while 256(256) denotes that all entries contacted 256 destination IPs. Finally, 256(256) shows that all entries sent 256 packets. For these specific traffic events, Wireshark has been used to verify the taxonomy scripts written in Python. As Wireshark handles a larger amount of data poorly, all events are found from a smaller dataset. The events covered are from 30.12.2020 from 08:54-11:40.

Before delving into the details of the traffic, it is important to examine the correlations between the categories in each taxonomy. Table 4.5 illustrates these correlations based on all unique source IPs from the 2018 taxonomy, and their placement in each category The table aims to show where the source IPs from the anomaly groups proposed in Liu and Fukuda (2018) are placed in the categories proposed in Barnett and Irwin (2008). The reason for the total never reaching 100% is small differences in the Python scripts, except for backscatter, which will be explained later.

As we see in the table Port Scans are predominantly placed in the One-to-Many category, which is unsurprising as source IPs often appear multiple times in this anomaly. In fact, only 3% of the entries were from unique source IPs. The high percentage of 86.99% in the One-to-Many category confirms this observation. As previously noted in section 4.1, One-to-Many scans usually involve many ports. 12% of Port Scans are placed in the One-to-One category, which often includes many ports, such as 23/tcp, 2323/tcp, 80/tcp, and 8080/tcp. All scans placed in the Port Scans anomaly must contact at least four ports, indicating that a significant portion of the One-to-One and One-to-Many scans involved multiple ports.

Table 4.5: December 2020: Taxonomy Correlation

Anomaly	Port Scans	network scans	one flow	backscatter	Small events	Other
One-to-One	11.8%	-	40.76%	19.28%	71%	7.47%
One-to-Many	86.99%	99.3%	52.28%	24.92%	24.67%	92.45%
Many-to-One	1.18%	-	3.37%	12.91%	1.87%	0.02%
Many-to-Many	-	0.59%	1.7%	0.01%	0.02%	-
Total	99.97%	99.89%	99.77%	53.57%	97.81%	99.94%

Upon examining the network scans category, it can be observed that a majority of the traffic is placed in the One-to-Many subcategory. These scans involve multiple hosts being contacted by a single source IP. It is worth noting that the 2018 taxonomy permits multiple entries from each source, which could result in an inflated number of entries for certain categories. As mentioned in section 4.2.1, network scans comprised 60% of duplicated source IPs, indicating that many of the One-to-Many scans used multiple ports. This finding is similar to that of Port Scans, where a significant portion of One-to-Many scans was found to involve multiple ports being contacted by a single source IP

The one flow anomaly is distributed more widely throughout the 2008 taxonomy. Notably, 3.37% of the IPs categorized in one flow also appeared in Many-to-One, which was the least common distribution in the 2008 taxonomy. One flow consists of Source IP to destination IP on one destination port with more than 15 packets. The results in table 4.5 show that one flow entry often contacted the same IP host on different ports. Most of the entries in one flow and Many-to-One were from UDP with a rapid speed, and the most popular port for these entries was 5060/udp. For the backscatter traffic, only 53.57% of the source IPs were categorized in the 2008 taxonomy presented by Barnett and Irwin (2008). This was mainly due to ICMP types other than echo, which were not categorized in the 2008 taxonomy. In fact, none of the ICMP traffic from the backscatter was placed in the 2008 taxonomy. Furthermore, a high percentage of each categorization was observed except Many-to-Many. An interesting observation is that the two categories created to categorize backscatter and misconfiguration traffic from Network Telescopes, namely backscatter, and One-Flow, are both distributed across all categorizations in the 2008 taxonomy.

For the Small events, the majority of the traffic is classified under the One-to-One category. Since Small events are defined by a limited number of destination IPs, destination ports, and packets, we can infer that One-to-One events mostly consist of Source IPs contacting few ports. These events were mostly identified as rapid scans.

Overall, table 4.5 shows that the Source IPs placed in the anomalies proposed in Liu and Fukuda (2018) are for the most part distributed across all categories proposed in Barnett and Irwin (2008). We see most of the source IPs are placed in One-to-Many, this is a result of most of the traffic from the 2008 taxonomy also being categorized under One-to-Many. This means that all source IPs from the 2018 taxonomy have a higher chance of being placed in One-to-Many. It does not mean that most of the traffic placed in Port scans is actually One-to-Many scans.

Table 4.6 presents data, where the unique IP sources placed in categories proposed in Barnett and Irwin (2008) are placed in the anomalies proposed in Liu and Fukuda (2018). The *Total* shows that IP Sources are duplicated in the 2018 taxonomy. Notably, the tables show that Many-to-One Source IPs were the most duplicated IPs in the 2018 taxonomy. Looking closer at the data. Most of the entries in Many-to-One stem from UDP Backscatter and Small UDP traffic. The rules for UDP Bacsckatter are based on the traffic from source ports 53/udp, 123/udp, 137/udp, or 161/udp. The rules for Small UDP are few destination IPs, destination ports and packets. It seems that a high percentage of this traffic is duplicated, based on UDP backscatter rules and Small UDP rules. As the traffic is from many source IPs to one destination IP, and few packets and ports, the traffic in UDP many-to-many aligns with information from section 2.2 about backscatter. This is also true for One Flow, which was mostly categorized in categories ending with Many, inferring, that most One Flow scan IPs are duplicates.

Table 4.6: December 2020: Taxonomy Correlation - mirrored

Anomaly	**One-to-One**	**One-to-Many**	**Many-to-One**	**Many-to-Many**
Port Scans	0.2%	1.78%	0.03%	-
Network scans	-	26.48%	-	16.67%
One flow	0.35%	0.96%	0.21%	8.33%
Backscatter	1.39%	1.3%	25.65%	5.56%
Small Events	98.78%	58.47%	90.98%	19.44%
Other	0.66%	21.06%	0.29%	-
Total	101.39%	110.05%	117.16%	50%

As in table 4.3, Small events cover most of the traffic. Small events are by far the most popular anomaly in the 2018 taxonomy, therefore the IP sources from the 2008 categories have a bigger chance to be placed there. Only One-to-Many and Many-to-Many had a below-the-average percentage in that anomaly. This shows that many of the entries in these categories contacted more than 4 Destination IPs and Destination ports. Notably, the source IPs in many-to-many was only covered 50% by the IPs categorized in the 2018 script. This is due to some corrupted IP source addresses not being categorized in the 2018 Python code.

The other traffic is mostly categorized under One-to-Many, but it is important to note that, as most of the IP source was of One-to-Many nature, IPs will always have a bigger chance of being placed here.

4.3.1 Scanning techniques refresh

Before diving deeper into specific traffic, a refresh of previously discussed scanning techniques from section 2.4 is provided. Table 4.7 presents a summary of all scans covered in the next sections. The table includes the Scanning technique Group (Group), The specific Technique name(Techniques) with the protocol used for the scan (PR). For each specific scanning technique, the entries have an initial message sent (SM), the answer for closed ports or unavailable hosts (RMC), and for open ports or active hosts (RMO). Far-right, a *checkmark*, shows if the technique is immune to firewall detection and an *x-mark* if not, this table is a simplified version of table 1 from Bou-Harb *et al.* (2014).

Table 4.7: Scanning Techniques

Group	Technique	PR	SM	RMC	RMO	IFD
Open scan	Open scan	TCP	SYN	RST	ACK	✖
Half-Open scan	Half-Open scan	TCP	SYN	RST	ACK	✖
Stealth Scan	SYN/ACK scan	TCP	SYN/ACK	RST	-	✔
	XMAS scan	TCP	URG,PUSH,FIN	RST	-	✔
	ACK scan	TCP	ACK	-	RST	✔
Sweep scan	ICMP Echo scan	ICMP	ICMP Echo	ICMP	ICMP Replay	✖
	TCP-SYN scan	TCP	SYN	RST	ACK	✖
Miscellaneous scan	UDP Scan	UDP	UDP pkt	ICMP[3]	UDP Data	✖

After assessing table 4.7, the next section provides the comparison of specific traffic from the two taxonomies reviewed in this research.

4.3.2 Port Scans

Port Scanning is a commonly used technique in network reconnaissance, where an attacker sends packets to a target host on a range of ports to determine which ports are open and what services are running on them. The purpose of Port Scanning is to gather information about a target system and identify potential vulnerabilities that can be exploited. Table 4.8 demonstrates that the number of ports contacted is a critical factor in identifying a Port Scan. The more ports contacted, the higher the likelihood that the traffic is a Port Scan. A typical Port Scan involves a single IP address sending multiple packets to many ports. According to the 2018 taxonomy, a Port Scan is identified when more than four ports are contacted, which is consistent with the findings in the research by (Yegneswaran, Barford, and Ullrich 2003). A vertical scan, as introduced in section 2.3, is a One-to-One scan that shares similar attributes with a Port Scan. To classify a scan as a vertical scan, it must contact five or more ports. In the case of Table 4.8, the host was contacted on 12 different ports, which satisfies the criteria for both vertical scans defined in Yegneswaran, Barford, and Ullrich (2003) and a Port Scan defined in Liu and Fukuda (2018).

In the 2008 taxonomy, the distribution of traffic for Port Scan is classified as One-to-One. This means that the source of the traffic only contacted one host. However, if there were multiple hosts contacted, it would be classified as One-to-Many. To determine the distribution in the 2018 script, we need to check if the traffic belongs to any other anomaly groups. In Table 4.8, the traffic is only categorized in the Light Port Scan group, which confirms that the traffic is placed correctly in one-to-one and Port Scan.

Table 4.8: Port Scans

Field	**2008**	**2018**
Protocol	TCP	
Source IP	193.122.96.x	
Anomaly	TCP One-to-One Rapid	TCP Light Port Scan
Number of Ports	12	12
Number of Destination IPs	1	1
Number of Packets	36	36

According to Bou-Harb *et al.* 2014, Port Scans are one of the most common scanning techniques, but our research indicates that the amount of traffic exclusively categorized in Port Scans is relatively minor compared to network scans according to the 2018 taxonomy. In Table 4.9, we provide an example of how traffic from a single source IP can be categorized in both Port Scans and network scans. This is a common occurrence in the 2018 taxonomy since it is based on one IP source with multiple rules, and the traffic from that source is categorized into smaller pieces or categorized multiple times with the same traffic. We see an example of this in table 4.9. As mentioned in section 4.2, we see a big amount of the traffic from the Port Scans anomaly also represented in the network scan category. In the 2008 taxonomy, the data shown in Table 4.9 have been categorized only as a One-to-Many scan. However, the 2018 taxonomy allows for a more specific classification of scanning techniques, which allows traffic flows to be split into smaller pieces. The source in tabel 4.9 in fact tried 200 ports on every host on the /24 netblock. Sending one packet to each hosts.

Table 4.9: Port Scan/Network scan

Field	**2008**	**2018**	
Protocol	TCP		
Source IP	89.248.165.x		
Anomaly	One-to-Many Rapid	Light Port Scan(256)	Light network scan(200)
Number of Ports	200	200(200)	1(1)
Number of Destination IPs	256	1(1)	256(256)
Number of Packets	51200	200(200)	256(256)

UDP scanning is a technique that attackers use to identify open ports on target systems (Bhuyan *et al.* 2011). As mentioned in section 2.4.6 and shown in table 4.7, the connection attempts in this technique receive ICMP type 3 replies for open ports or no response at all, and thus the ports are deemed closed by the attacker.

In table 4.10, we see an example of a UDP scan where the source IP is sending multiple UDP packets to 18 different destination ports. This traffic can be classified as an attempt to identify open ports on the target system. The 2008 taxonomy categorizes this traffic as Many-to-One based on the speed of the traffic (Rapid) and multiple source IPs trying to reach the same destination IP using the same protocol. In addition to the source IP seen in table 4.10, there are three other source IPs (96.127.175.x, 92.118.161.x, and 188.169.45.x) trying to reach the same destination IP. In the 2018 taxonomy, the traffic from the three additional source IPs is categorized in the Small UDP anomaly, based on packet count (less than 16), destination IP count (less than 5), and destination port count (less than 5). Both taxonomies categorize all these UDP connections, although in different ways. This demonstrates that different taxonomies can provide different views of the same data, which is a common theme looking through the data.

Table 4.10: UDP scan

Field	2008	2018
Protocol	UDP	
Source IP	158.69.23.x	
Anomaly	UDP Many-to-One Rapid	UDP Light Port Scan
Number of Ports	18	18
Number of Destination IPs	1	1
Number of Source IPs	3	1
Number of Packets	54	18

4.3.3 Sweep Scans/network scans

Section 2.4 discusses one of the most common scanning techniques, the Sweep Scan. The goal of a sweep scan is to identify hosts. Table 4.11 Shows a typical TCP scan with a SYN packet, as we see in table 4.7, the initial message for many TCP scans is a SYN packet. The scan could be attributable to Open scan, half-open scan or TCP-SYN scan. A SYN scan consists of one source sending TCP SYN packets to multiple hosts, trying to find active hosts. Scanners can determine this by checking if the answers from the TCP SYN packets for active hosts.

Table 4.11 presents a typical example of a SYN scan. The table shows that the IP address 51.81.255.x has sent multiple TCP SYN packets to numerous host IPs. The table also highlights the clear distinction in the categorization of traffic between the two taxonomies. While both taxonomies classify the traffic in only one anomaly, the 2018 taxonomy has multiple entries from the same IP source, which supports the notion that this is a TCP-SYN scan placed in the sweep scan group, seem in table 4.7, as the IP address is attempting to reach multiple hosts using the same TCP-SYN scanning technique.

The 2018 taxonomy categorizes all the traffic from the IP address as Light network scan. The nature of the network scan anomaly, which is based on anomalies of one source and one port, causes the traffic to be split into multiple entries, unlike the 2008 taxonomy, which gathers all the traffic from the source IP and categorizes it as a “One-to-Many” scan. Table 4.11 indicates that the total number of ports in the 2008 anomaly is 19, which does not match the 13 from the 2018 taxonomy. A confirmation using Wireshark, confirms that 19 ports were contacted, but six of these were to a small number of hosts. As a result, that traffic was not categorized as network scan in the 2018 taxonomy, because there must be five or more IP destinations. This traffic also avoided being placed in potential categories such as one flow and small SYN’s, with too many ports and IP destinations, respectively, for small SYN and too few packets per flow for one flow. This occurs because every anomaly is based on one source IP. For example, for “Small SYN”, the source IP would need to have 196 destinations, 19 ports, and 330 packets, which would not follow the Darknet traffic rules outlined in Table 2.4. The same is true for one flow, which should have more than 15 packets per IP source-IP destination-Port destination flow, based on probable misconfigurations established in Wustrow *et al.* (2010). Instead, the remaining traffic is placed in TCP other.

Table 4.11: SYN scan

Field	**2008**	**2018**
Protocol	TCP	
Source IP	51.81.255.x	
Anomaly	TCP One-to-Many Rapid	Light network scan (x13)
Number of Ports	19	1(13)
Number of Destination IPs	196	16(5-47)
Number of Packets	330	16(5-60)
SYN packets	100%	100%(100%)

Sweep scans, a type of network scanning technique, are frequently conducted using the ICMP, which is a portless protocol. Although an example of a sweep scan can be observed in table 4.11 using the TCP, a different approach is taken when utilizing the ICMP protocol. In table 4.11 and 4.12, it can be seen that both ICMP and TCP sweeps scans are categorized as Light network scan in the 2018 taxonomy. The traffic in this category comprises less than 4 packets per destination. However, as ports are not used in ICMP traffic, the classification of network scan for ICMP categorizes sources that exceed 4 destination IPs contacted.

Table 4.12: ICMP sweep scan

Field	**2008**	**2018**
Protocol	ICMP	
Source IP	3.112.218.x	
Anomaly	ICMP One-to-Many Rapid	ICMP Light network scans
Number of Destination IPs	11	11
Number of Source IPs	1	1
Number of Packets	11	11

Figure 4.2 depicts a sweep scan graphed in InetVis, a traffic analysis tool (van Riel and Irwin 2006; Irwin and van Riel 2007; Irwin and van Riel 2008). The scan’s source IP is linked to significant portions of the destination network, represented by the blue horizontal axes. The green horizontal line shows the destination ports, which in this case, are high port numbers. High port numbers are often utilized to avoid restrictions on more commonly used ports, such as port 80/tcp for HTTP and 21/tcp for FTP. In section 2.3.1, horizontal scans are identified as a scanning technique. Figure 4.2 provides an explanation for why these scans are referred to as horizontal scans. The depicted scan originates from a single source IP and targets multiple destination IPs. The scan observed in Figure 4.2 is classified as both a One-to-Many and a network scan.

4.3.4 backscatter traffic

Backscatter traffic has been extensively mentioned in both the taxonomies developed in this research and in section 2.2 in this research. The 2018 taxonomy specifically includes a separate category for backscatter traffic due to its unique characteristics. This type of traffic mainly consists of response

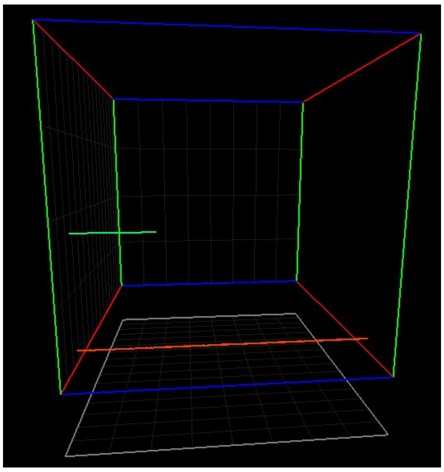


Figure 4.2: Horizontal scan shown using Inetvis

packets to Distributed Denial of Service (DDoS) attacks and is identified by an array of TCP flags, UDP ports, or ICMP response messages. According to Wustrow *et al.* (2010), ICMP packets constitute a significant portion of packets in this category. as observed in 4.3.

Table 4.13 provides an example of traffic that has been classified in the backscatter anomaly group in the 2018 taxonomy. Interestingly, the same traffic is categorized as One-to-Many in the 2008 taxonomy. The literature review section reveals that stealth scans exhibit similar characteristics to backscatter traffic, as Stealth-scans and backscatter are attributable to TCP SYN-ACK packets. Table 4.13 presents traffic that could represent a stealth attack. The traffic is slow but still aligns with the maximum threshold of one hour between each connection, as stated by Yegneswaran, Barford, and Ullrich (2003). Additionally, table 4.7 shows that SYN/ACK scans are included in the stealth scans category. The traffic in table 4.13 only consists of TCP SYN-ACK packets. If this were a stealth scan, a more precise classification would be network scan or Port Scans in the 2018 taxonomy. However, traffic identified by TCP SYN-ACK packets is in the ruleset proposed by Liu and Fukuda (2018) classified as backscatter table, indicating that SYN-ACK packets could be associated with backscatter traffic.

Table 4.13: Stealth scan or backscatter

Field	**2008**	**2018**
Protocol	TCP	
Source IP	51.255.81.x	
TCP Flags	SYN-ACK	
Anomaly	One-to-Many Slow	backscatter
Number of Ports	250	250
Number of Destination IPs	168	168
Number of Packets	278	278

4.3.5 Uncategorized/Other

An addition to the 2018 taxonomy is the *Other* categories. These are not specific anomalies, but a gathering of unplaced traffic. This category provides a way to capture the traffic that was not placed in any other specific categories. In contrast, the 2008 taxonomy categorizes all traffic as long as it has an IP source and IP destination, along with ports for UDP and TCP. Table 4.14 and 4.15 provide examples of the type of traffic that can be found in the “Other” category in the 2018 taxonomy. Table 4.14 shows multiple TCP connection attempts from a single source IP to 17 different destination addresses, all of which are unique but not necessarily for all destination ports. All the packets sent from this source are TCP SYN packets. Additionally, in Wireshark, we observe a UDP connection to from the same IP to a single source and a NAT-PMP external address request connection at the start of the activity.

NAT-PMP (Network Address Translation Port Mapping Protocol) is a protocol used to manage the automatic forwarding of network traffic through a router or gateway (Boucadair *et al.* 2011). It is commonly used in small networks where multiple devices share a single public IP address. NAT-PMP allows devices on the local network to request the opening of a specific port on the router for incoming traffic. In this case, the external address request connection is made by the source IP to map the NAT port for incoming traffic. The traffic in table 4.14 is also placed in the slow category in the 2008 taxonomy. The properties of this traffic suggest that it could be a Slow SYN attack. However, SYN attacks are not typical of stealth attacks.

Table 4.14: Stealth or Other

Field	**2008**	**2018**
Protocol	TCP	
Source IP	102.165.30.x	
Anomaly	TCP One-to-Many	Other TCP
Destination IPs	17	17
Destination Ports	12	12
Packets	17	17
SYN Packets	17	17

If we look at table 4.15, the use of various TCP flags combinations is provided. Unconventional flags are a common technique used by attackers to evade detection and classification by intrusion detection systems. The XMAS scan, which uses the FIN, PSH, and URG flags in combination, is one such technique that is commonly employed to scan for open ports. The goal of the XMAS scan is to identify open ports without arousing the suspicion of the target host. The SYN flag, which is typically used to initiate a TCP connection, is not typically used in an XMAS scan as it may trigger the target host’s intrusion detection system.

The combination of flags used in the traffic observed in table 4.15 includes FIN, SYN, RST, PSH, ACK, and URG. This combination avoids detection by intrusion detection systems that rely on a specific combination of flags to identify malicious traffic. In this case, the source IP was considered but not categorized as a network scan, one flow, backscatter, or Port Scan in the 2018 taxonomy. Interestingly, the combination of flags used in this traffic was not deemed to be an indication of scan packets in the 2018 taxonomy. For network scan, there were too few destination IP addresses for the traffic to be categorized, and for backscatter, the traffic was filtered out for having too few packets linked with one destination IP. The 2008 taxonomy placed this traffic in the One-to-Many category as it was a connection from one source IP to many destination IPs. It was considered to be a connection with a FIN flag. It is placed in the slow speed category. As seen in section 2.4, the stealth scan category includes scans with different combinations of flags from the usual. The combination of the flags, and the low-speed connections, may point to that this scan could be a stealth attack. Although, it could also be deemed as other traffic

Table 4.15: XMAS scan or Other

Field	2008	2018
Protocol	TCP	
Source IP	119.45.157.x	
TCP Flags	FIN, SYN, RST, PSH, ACK, URG	
Anomaly	TCP One-to-Many Slow With FIN	Other TCP
Number of Ports	3	3
Not categorised in	N/A	network scan, one flow backscatter, Port Scan
Number of Destination IPs	3	3
Number of Packets	3	3

4.4 Performance of each taxonomy

After examening the dataset provided using a Network Telescope, individually for each taxonomy, comparing categorizations, and at specific traffic, an understanding of how the taxonomies categorize data is acquired. The paragraphs that follow aim to discuss the strength and weaknesses and establish an area of use for each taxonomy. Finding these requires the discussion of how well these taxonomies handled classifying the traffic.

As mentioned in the methodology chapter at 3, the dataset utilized in this study spans from December 2020 until March 2021. When examining traffic categorized as Other, our findings are similar to those of Liu and Fukuda (2018). The Other category is relatively low, averaging at 0.62% as shown in Table A.2 in Appendix-A. However, for the other traffic specific to each protocol, the percentages are higher. In the research conducted in 2018, the authors found similar results for UDP and ICMP, as found in this research. With regards to TCP, the Other category is higher than expected, averaging at 5.45% compared to an average of 0.31% in 2018. This implies that the TCP traffic has seen some changes in the traffic.

In section 4.3.5, we provide examples of Other traffic that often exhibit attributes similar to stealth scans. Considering the *strict* rule set in the taxonomy from 2018, an average of 3.33% other traffic, implies that the ruleset established is still relevant for categorizing most traffic. For the taxonomy in Barnett and Irwin (2008), traffic that was not categorized was only 0.02%, this is mostly ICMP traffic that did not send ping packets, which was not categorized as scans (Barnett and Irwin 2008). Otherwise, the taxonomy does not set any rules for the traffic to be categorized, and will always have low other percentages.

In a recent study conducted by Irwin (2022), the most commonly targeted TCP services in a Network Telescope dataset, were identified. The author implies that traffic targeting 23/tcp is attributed to traffic stemming from IoT botnets conducting active scanning. If we look closer at the dataset, the most targeted port in the network scan anomaly was 23/tcp. In table 4.16, showing the top 8 ports from network scans across the first week of each month, port 5555/tcp was also in the top 3 most targeted ports. Port 5555/tcp is linked with active scanning toward the Android Debug Bridge (CVE-2014-1909). Port 6379/tcp, a default port used by REDIS database (Irwin 2022), was also in the top 8 ports. Port 22/tcp, 80/tcp, and 8080/tcp. is also regularly found in the top ports scanned. As noted by Irwin (2022), targeting these ports shows clear signs of scans of vulnerable services.

We have also observed traffic that aligned with the properties of IBR traffic in for example in sections

Table 4.16: Top 8 destination ports in TCP network scan

Rank	TCP Port	%
1	23/tcp	8.8%
2	80/tcp	1.24%
3	5555/tcp	1.21%
4	8080/tcp	1.18%
5	22/tcp	1.16%
6	445/tcp	0.74%
7	443/tcp	0.73%
8	6379/tcp	0.6%

4.3, 4.3.4. Notably, backscatter and one flow which was found to be distributed across all four categories in the 2008 taxonomy as seen in table 4.6. We also found many of the source IPs in the Small events to align with the properties for backscatter and misconfiguration traffic in section 4.2, especially Small pings were, in some instances, seen sending only one ICMP message to one source.

From the observations done in the above sections, we can say that most of the traffic is similar to what Wustrow *et al.* (2010) classified as Internet background radiation traffic, and as seen in table 2.1, most of the traffic should consist of scanning traffic, while a smaller part is attributable to backscatter and misconfigurations. Although, the last-mentioned categories were observed with fluctuating numbers (Wustrow *et al.* 2010)

It is worth noting that while the taxonomy proposed by Barnett and Irwin (2008), provides a more simplistic taxonomy for traffic, the taxonomy categorizes all traffic, except some ICMP traffic, and provides a classification of distributions of possible scanning. The authors aimed to extend the definitions of scanning traffic previous to 2008, to provide a better understanding of its complexity. Looking at the taxonomy from Liu and Fukuda (2018), the authors propose multiple anomalies, while that taxonomy offers stricter rules for each categorization, while at the same time categorizing almost all traffic. We saw in table 4.6, that these anomalies were, for the most part, distributed in all categorizations proposed by Barnett and Irwin (2008), which confirms their statement of classification of scanning is a complex activity, as scanning techniques and scanning distribution is not easily comparable.

The 2018 taxonomy brings value to the confirmation of scanning techniques. The validity of the rule set for each anomaly proposed by Liu and Fukuda (2018) is confirmed in section 4.3, as most traffic classified in the anomalies was found to be fitted correctly. However, as we saw in section 4.3.5, stealth scans could evade being categorized in the taxonomy as well as evade Firewalls seen in table 4.7. The iteration of the taxonomy used in this research also showed that most of the traffic placed in port scan anomaly was also found in network scan anomaly. The duplication of IP sources was avoided in the taxonomy proposed by Barnett and Irwin (2008).

The 2008 taxonomy, compared to the 2018 taxonomy, mainly shows the distribution of the scanning activity. Accounting for Speed and Distribution, the taxonomy provides an overview of traffic events. As the categorizations are more open than the 2018 taxonomy, the data must be closely analyzed to find popular scanning techniques in a dataset, but it provides an overview of activity. The 2018 taxonomy is mostly based on categorizing scanning techniques and findings from analyzing IBR traffic.

4.5 Summary

For the analysis chapter, the primary objective was to identify scanning events and assess their placement in the two taxonomies. The above sections provide the information we set out to identify. For the first section, found at 4.1, we confirmed the statements made by Barnett and Irwin (2008) of One-to-One and Many-to-Many being the most popular scan distribution options, we also found that most scans targeted many ports. In contrast to the research, we found that IMCP was more prevalent in One-to-one, than the authors claimed in the research. For UDP traffic, most of the traffic was one-to-many, with an average of 80% scanning many ports. We observed a number of coversations from source ports 53/udp, 123/udp, 137/udp, and 161/udp, which are commonly related to backscatter.

In the analysis of the taxonomy proposed by Liu and Fukuda (2018), found in section 4.2, most findings from 2018 were confirmed, as well as the validity of the script, which provided similar percentages in table 4.3 as seen in the research by Liu and Fukuda (2018). Showing that the taxonomy, 5 years after introduction, still provides similar results. Small events were indeed the most populated anomaly, containing 74.52% of unique IP sources. We also confirmed that the ICMP backscatter rules, still is valid, as it was the second most popular anomaly for ICMP traffic. This result correlated with the findings in the research from 2018. For Port and Network Scans, there were multiple occurrences of duplicates of traffic, as expected. In these categories, we saw similarities with the traffic from One-to-Many, in the 2008 taxonomy, as most IP sources contacted multiple hosts on multiple ports.

Section 4.3, Comparison showed the complexity of traffic stemming from a Network telescope. Both tables provided in that section, 4.5 and 4.6, showed that there is no real correlation between the two taxonomies. The results rather showed, that the most populated categories will show the highest percentages of correlations, which implies that categorizing based on distribution and scanning techniques is different. The only certainty is that Network scans will always be placed in categorizations ending with Many, as the rules for that anomaly are based on one source with multiple destination IPs.

Looking closer at specific Source IPs, we once again confirmed the validity of both scripts used in this research, as the same traffic was checked in both taxonomies and confirmed in Wireshark. The most notable takeaway was the evasion of stealth scans from the 2018 taxonomy, and also we got a hands-on look at the duplications of some of the traffic placed in the anomalies proposed by Liu and Fukuda (2018).

5
Conclusion

This research presents an analysis of Network Telescope data classifications using the taxonomies proposed by Barnett and Irwin (2008) and Liu and Fukuda (2018). The dataset used was collected on a /24 IP v4 netblock, covering a 4-month period between December 2020 and April 2021. The relevance of the taxonomies was compared with research on both Network Telescope data and scanning techniques from the last three decades, spanning from the 1990s to the 2020s. Findings proved the continued validity and use of both these taxonomies.

5.1 Summary of Research

The aim of the project was to identify potentially outdated research in the classification of Network Telescope data, which could pose as a weakening factor in the cyber security field. To achieve this, two taxonomies were evaluated, with a 10-year difference, using the same dataset. This allowed us to observe trends that either supported or contradicted their continued relevance.

The literature review, chapter 2, provides an overview of related works on Network Telescope data (section 2.2) and scanning techniques (section 2.3). It also includes an examination of the two taxonomies proposed by Barnett and Irwin (2008) and Liu and Fukuda (2018). They are discussed in sections 2.8 and 2.6. In light of the research conducted in the first half of the Literature review chapter, an initial comparison between the two taxonomies is found at 4.3.

In chapter 3, the methodology for further analysis is explained. The chapter begins with an introduction to the data gathered through a Network Telescope, which shows an even distribution of unique IP sources and the number of packets for each month. In the methodology chapter, the creation of the two taxonomies was introduced. Both scripts utilize Python with the dpkt module for parsing packets from packet capture files. These programs were used in the analysis chapter. Section 3.3 explains that the analysis aims to create two tables showing an initial overview of how the two taxonomies categorize data.

In chapter 4, an initial overview of how the two taxonomies categorize data is presented through two tables: table 4.1 for (Barnett and Irwin 2008) and table 4.3 for (Liu and Fukuda 2018). The analysis then compared the categorizations of source IPs across the two taxonomies, searching for notable correlations in section 4.3. Further analysis delved deeper into specific traffic in section 4.3.2, leading to section 4.4 where the performance of both taxonomies was discussed

5.2 Research Objectives

Looking back at the introduction to the project in chapter 1, a problem statement was introduced regarding the possibility of outdated network scanning taxonomies being a potential weakening factor in the field of cyber security. To address this issue, a primary objective was established in section 1.2

- Use Network Telescope data to confirm the relevance of two established taxonomies

In order to achieve this objective, the following secondary objectives were identified:

- Investigate suitable Python development approaches
- Recreate the two taxonomies in Python for analyzing packet capture files
- Identify data from Network Telescopes
- Use Taxonomies to classify Network Telescope data

In Appendix-B, the artifacts used in the analysis are discussed. However, there are some limitations to the creation of the scripts, as explained in section 3.2.1. One of the main limitations is that Layer 2 scans, which are prevalent in the taxonomy from Barnett and Irwin (2008), were not considered. Additionally, no previous research to compare the results shown in table 4.2 against was found, other than some findings from the original study. There were some differences in the data that did not mirror the results from the original study, mainly regarding ICMP traffic. The validity of the 2008 taxonomy script was discussed in sections 3.2, 4.1, and 4.4. In these sections, it was found that the artifact delivered results similar to what was expected, despite the limitations.

In the creation of the taxonomy from Liu and Fukuda (2018), one main difference was observed in the TCP OTHER anomaly, which showed a higher average percentage of categorization. This difference was attributed, in section 4.4, to a possible indication of a changing landscape in Network Telescope data. However, for the OTHER anomalies, similar results to the original research were obtained, and for this reason, the recreation of the script was deemed valid in sections 4.2 and 4.4.

After confirming the validity of the artifacts for both taxonomies, the research proceeded to compare them in section 4.3. The results showed that, with a few exceptions, there were no correlations between the categories in the two taxonomies. One exception was the NETWORK SCAN anomaly from the 2018 taxonomy, which was only placed in distributions with many destination IPs from the 2008 taxonomy. For the remaining categories, most of the source IPs from the 2018 taxonomy were classified as ONE-TO-MANY scans, which had the most entries. It was concluded that there was a higher chance of finding a correlation in that category (section 4.3). However, the majority of the traffic was distributed among all categories from the 2008 taxonomy, indicating that there were no significant correlations between the two taxonomies in terms of how they categorized source IPs. These findings were discussed in sections 4.3 and 4.5. Additionally, when mirroring the analysis, similar results were obtained, with most IP sources found in Small events, which had the highest percentage of unique IP source addresses in table 4.3. It was concluded that there were no notable correlations between the two taxonomies in terms of how they categorized source IPs.

In chapter 2, previous studies on Network Telescope data and scanning techniques were discussed in sections 2.2 and 2.4. These findings were then used in chapter 4 to evaluate the validity of both taxonomies. The taxonomy from Liu and Fukuda (2018) was found to classify most traffic into the fitting anomaly in section 4.3, but it was also noted that stealth scans could evade categorization in section 4.3.5. The different results from ICMP traffic are discussed in sections 4.2 and 4.4. In section 4.3, the 2018 taxonomy showed its value by offering stricter rules for categorization, which confirmed scanning techniques introduced in section 2.4 and table 4.7, as noted in section 4.4.

The taxonomy proposed by Barnett (Barnett and Irwin 2008) was found to classify all traffic except the ICMP traffic which was placed in the ICMP BACKSCATTER category in the 2018 taxonomy. As discussed in section 4.3.4, this was due to the 2008 script only classifying ICMP flows with ICMP echo packets. Wireshark was used to confirm the distribution and events of TCP, UDP, and ICMP traffic categories in the taxonomy, found in sections 2.8 and 4.3. The 2008 taxonomy was based on the distribution of sources, as explained in section 2.3, rather than specific categorizations of scanning techniques. This approach gives a great overview of the traffic, for further analysis in the comparison section at 4.3, the researcher had to search for specific IPs in each category to identify potential scanning techniques, discussed in section 2.4. The value of the 2008 taxonomy lies in providing an overview of the traffic, including speed, distribution, and port distribution, which is important information for closer analysis of the traffic, as explained in section 4.4.

The above paragraphs show that both taxonomies have use areas when analyzing Network Telescope data from the 2020s. The statements made in the original research by Barnett and Irwin (2008) and Liu and Fukuda (2018) were still found to be mostly correct in chapter 4, thus we can conclude this research by confirming the validity and use of both taxonomies.

5.3 Research Contribution

By comparing the two taxonomies proposed from Barnett and Irwin (2008) and Liu and Fukuda (2018) with the same dataset, this research provides a new look at different approaches to classifying data from Network Telescopes. As mentioned in the introduction, there has been no recent work in this area. As far as the author is aware, these two taxonomies have not been compared using the same dataset before, bringing a new understanding to different approaches. To accomplish this research, two Python scripts have been developed to replicate the proposed taxonomies, which can be used for future research. These scripts are provided in Appendix-B.

5.4 Future Work

From what we learned in this research, the lack of correlation between the anomalies in the 2018 taxonomy and the categorizations in the 2008 taxonomy, stands out as notable. As both taxonomies were concluded to offer value to the research field. While the taxonomies are compared and analyzed with the same data, a logical step for future work is to curate a *zoo* of traffic capture data to closer compare these taxonomies, or for comparing other network scanning taxonomies, to confirm or affirm their validity and contribution to the research field. These traffic capture zoo's could be found with the use of the artifacts from this research, as they are both validated as relevant. The *Zoo* of scanning traffic would contribute to a standard of Network Telescope data for other researchers to use. Future work should commence with:

- Explore differences and overlap between the two taxonomies in greater detail.

- Merge the two taxonomies with the view to developing an improved unified taxonomy.

- Curate a *Zoo* of Packet Captures with relevant data. This would be a to provide a library for other researchers to observe scan traffic in isolation in order to aid with future research.

References

Allman, M., Paxson, V., and Terrell, J. (2007). “A brief history of scanning”. In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement - IMC ’07*. San Diego, California, USA: ACM Press, page 77. ISBN: 978-1-59593-908-1. DOI: 10.1145/1298306.1298316. (Visited on 11/09/2022).

Anbar, M., Ramadass, S., Taha, A. A., Manasrah, A., Altaher, A., Aljmmal, A., and Almomani, A. (2013). “Investigating Study on Network Scanning Techniques”. In: DOI: 10.4156/jdcta.vol7.issue9.37.

Arkin, O. (1999). “Introduction to intelligence gathering techniques”. In.

Barnett, R. J. and Irwin, B. (2008). “Towards a taxonomy of network scanning techniques”. In: *Saicsit ’08*. Saicsit ’08. Wilderness, South Africa: ACM, pages 1–7. ISBN: 978-1-60558-286-3. DOI: 10.1145/1456659.1456660.

Bhuyan, M. H., Bhattacharyya, D. K., and Kalita, J. K. (2011). “Surveying Port Scans and Their Detection Methodologies”. In: *The Computer Journal* 54.10, pages 1565–1581.

Borgnat, P., Dewaele, G., Fukuda, K., Abry, P., and Cho, K. (Apr. 2009). “Seven Years and One Day: Sketching the Evolution of Internet Traffic”. In: *IEEE INFOCOM 2009 - The 28th Conference on Computer Communications*. Rio De Janeiro, Brazil: IEEE, pages 711–719. ISBN: 978-1-4244-3512-8. DOI: 10.1109/INFCOM.2009.5061979. (Visited on 02/01/2023).

Bou-Harb, E., Debbabi, M., and Assi, C. (2014). “Cyber scanning: A comprehensive survey”. In: *IEEE Communications Surveys and Tutorials* 16.3, pages 1496–1519. ISSN: 1553877X. DOI: 10.1109/SURV.2013.102913.00020.

Boucadair, M., Ford, M., Roberts, P., Durand, A., and Levis, P. (June 2011). *Issues with IP Address Sharing*. Request for Comments RFC 6269. Num Pages: 29. Internet Engineering Task Force. DOI: 10.17487/RFC6269. (Visited on 04/04/2023).

Cornish, P., Livingstone, D., Clemente, D., and Yorke, C. (Sept. 2011). *Cyber Security and the UK’s Critical National Infrastructure*. Technical report. Chatham House. https://www.chathamhouse.org/sites/default/files/public/Research/International%20Security/r0911cyber.pdf.

Cowie, J., Ogielski, A., Premore, B., and Yuan, Y. (Jan. 2001). *Global routing instabilities during Code Red II and nNimda worm propagation*. Technical report. RENESYS.

de Vivo, M., Carrasco, E., Isern, G., and de Vivo, G. O. (Apr. 1999). “A review of port scanning techniques”. In: *ACM SIGCOMM Computer Communication Review* 29.2, pages 41–48. ISSN: 0146-4833. DOI: 10.1145/505733.505737.

Fontugne, R., Borgnat, P., Abry, P., and Fukuda, K. (Nov. 2010). “MAWILab: combining diverse anomaly detectors for automated anomaly labeling and performance benchmarking”. In: *Proceedings of the 6th International COnference*. Philadelphia Pennsylvania: ACM, pages 1–12. ISBN: 978-1-4503-0448-1. DOI: 10.1145/1921168.1921179. (Visited on 02/01/2023).

Fukushima, M. and Goto, S. (1999). “Analysis of TCP flags in congested network”. In: *1999 Internet Workshop. IWS99. (Cat. No.99EX385)*. Osaka, Japan: IEEE, pages 151–156. ISBN: 978-0-7803-5925-3. DOI: 10.1109/IWS.1999.811007. (Visited on 02/02/2023).

Green, J., Marchette, D., Northcutt, S., and Ralph, B. (1999). “Analysis Techniques for Detecting Coordinated Attacks and Probes”. In.

Irwin, B. (2013). “A Source Analysis of the Conficker Outbreak from a Network Telescope”. In: *SAIEE Africa Research Journal* 104.2, pages 38–53. DOI: 10.23919/SAIEE.2013.8531865.

Irwin, B. and van Riel, J.-P. (2008). “Using InetVis to Evaluate Snort and Bro Scan Detection on a Network Telescope”. In: *VizSEC 2007: Proceedings of the Workshop on Visualization for Computer Security*. Edited by J. R. Goodall, G. Conti, and K.-L. Ma. Berlin, Heidelberg: Springer Berlin Heidelberg, pages 255–273. ISBN: 978-3-540-78243-8. DOI: 10.1007/978-3-540-78243-8_17.

Irwin, B. (Sept. 2012). “Network Telescope Metrics”. In: Southern African Telecommunications and Applications Conference (SATNAC).

– (Dec. 2022). “Evaluation of Mauritian IPv4 address space within Internet Background Radiation data”. In: *International Conference on Intelligent and Innovative Computing Applications* 2022, pages 103–111. ISSN: 16944607, 1694464X. DOI: 10.59200/ICONIC.2022.011. (Visited on 04/17/2023).

Irwin, B. and van Riel, J.-P. (2007). “InetVis: a Graphical aid for the Detection and Visualisation of Network Scans”. In: *In Conference on Vizualisation Security (VizSec2007)*.

Kuhrer, M., Hupperich, T., Rossow, C., and Holz, T. (2014). “Hell of a Handshake: Abusing TCP for Reflective Amplification DDoS Attacks”. In.

Labovitz, C., Iekel-Johnson, S., McPherson, D., Oberheide, J., and Jahanian, F. (2010). “Internet inter-domain traffic”. In.

Liu, J. and Fukuda, K. (Jan. 2018). “An evaluation of darknet traffic taxonomy”. In: *Journal of Information Processing* 26. Publisher: Information Processing Society of Japan, pages 148–157. ISSN: 18826652. DOI: 10.2197/ipsjjip.26.148.

Mazel, J., Fontugne, R., and Fukuda, K. (Aug. 2014). “A taxonomy of anomalies in backbone network traffic”. In: *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*. Nicosia, Cyprus: IEEE, pages 30–36. ISBN: 978-1-4799-0959-9. DOI: 10.1109/IWCMC.2014.6906328. (Visited on 01/31/2023).

Mirkovic, J. and Reiher, P. (Apr. 2004). “A taxonomy of DDoS attack and DDoS defense mechanisms”. In: *ACM SIGCOMM Computer Communication Review* 34.2, pages 39–53. ISSN: 0146-4833. DOI: 10.1145/997150.997156. (Visited on 02/02/2023).

Moore, D., Shannon, C., Voelker, G. M., and Savage, S. (2004). *Network Telescopes: Technical Report*. Technical report. CAIDA, San Diego Supercomputer Center, University of California, San Diego.

Moore, D., Voelker, G. M., and Savage, S. (May 2001). “Inferring Internet Denial-of-Service Activity”. In: *ACM Transactions on Computer Systems.* 24.2, pages 115–139. ISSN: 0734-2071. DOI: 10.1145/1132026.1132027.

Oberheide, J. and Karir, M. (2006). *Honeyd Detection via Packet Fragmentation*. Technical report. MERIT. https://jon.oberheide.org/files/merit06-honeyd.pdf.

Orebaugh, A. and Pinkard, B. (2008). *Nmap in the enterprise: your guide to network scanning*. OCLC: ocn180880716. Burlington, MA: Syngress Publishing. ISBN: 978-1-59749-241-6.

Pang, R., Yegneswaran, V., Barford, P., Paxson, V., and Peterson, L. (2004). “Characteristics of internet background radiation”. In: *Proceedings of the 4th ACM SIGCOMM conference on Internet* *measurement - IMC '04*. Taormina, Sicily, Italy: ACM Press, page 27. ISBN: 978-1-58113-821-4. DOI: 10.1145/1028788.1028794. (Visited on 11/09/2022).

Pittman, J. M. (Jan. 2023). “Machine Learning and Port Scans: A Systematic Review”. In: arXiv:2301.13581 [cs] version: 1. http://arxiv.org/abs/2301.13581 (visited on 02/07/2023).

Shaikh, S. A., Chivers, H., Nobles, P., Clark, J. A., and Chen, H. (Nov. 2008). “Network reconnaissance”. In: *Network Security* 2008.11, pages 12–16. ISSN: 13534858. DOI: 10.1016/S1353-4858(08)70129-6. (Visited on 02/03/2023).

Staniford, S., Hoagland, J. A., and McAlerney, J. M. (Jan. 2002). “Practical automated detection of stealthy portscans”. In: *Journal of Computer Security* 10.1-2, pages 105–136. ISSN: 18758924, 0926227X. DOI: 10.3233/JCS-2002-101-205. (Visited on 09/09/2022).

Taylor, D. E. (Sept. 2005). “Survey and taxonomy of packet classification techniques”. In: *ACM Computing Surveys* 37.3, pages 238–275. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/1108956.1108958. (Visited on 02/02/2023).

Van Riel, J.-P. and Irwin, B. (2006). “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. Cape Town, South Africa: Association for Computing Machinery, pages 85–89. ISBN: 1595932887. DOI: 10.1145/1108590.1108604.

Wustrow, E., Karir, M., Bailey, M., Jahanian, F., and Huston, G. (Nov. 2010). “Internet background radiation revisited”. In: *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. Melbourne Australia: ACM, pages 62–74. ISBN: 978-1-4503-0483-2. DOI: 10.1145/1879141.1879149. (Visited on 02/01/2023).

Yegneswaran, V., Barford, P., and Plonka, D. (2004). “On the Design and Use of Internet Sinks for Network Abuse Monitoring”. In: *Recent Advances in Intrusion Detection*. Edited by D. Hutchison *et al.* Volume 3224. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, pages 146–165. ISBN: 978-3-540-23123-3. DOI: 10.1007/978-3-540-30143-1_8. (Visited on 02/01/2023).

Yegneswaran, V., Barford, P., and Ullrich, J. (June 2003). “Internet Intrusions: Global Characteristics and Prevalence”. In: *SIGMETRICS Performance Evaluation Review* 31.1, pages 138–147. ISSN: 0163-5999. DOI: 10.1145/885651.781045.

A

Appendix-A: Data

For the tables below, cells containing '-' represent values of less than 0.01%.

Table A.1: December-March: Percentage of traffic from each protocol from the 2008 taxonomy

Types		Dataset (1-7 Day of Month)				Average
		Dec	Jan	Feb	Mar	
One-to-One	TCP	2,29%	2,07%	2,29%	2,29%	2,24%
	UDP	4,04%	3,63%	4,6%	6,21%	4,61%
	ICMP	4,55%	3,63%	4,81%	3,98%	4,24%
One-to-Many	TCP	97,67%	97,88%	97,66%	97,64%	97,71%
	UDP	95,38%	96,27%	95,26%	93,51%	95,1%
	ICMP	93,95%	93,95%	94,1%	94,93%	94,23%
Many-to-One	TCP	0,03%	0,04%	0,05%	0,02%	0,04%
	UDP	0,58%	0,1%	0,14%	0,27%	0,27%
	ICMP	0,29%	0,13%	0,58%	0,01%	0,25%
Many-to-Many	TCP	0,01%	0,01%	-	0,04%	0,02%
	UDP	0,01%	-	-	0,01%	0,01%
	ICMP	1,22%	2,29%	0,51%	1,08%	1,27%

Table A.2: December-March: Percentage traffic from each protocol from 2018 taxonomy

Types			**Dataset (Day of month)**				**Average**
			Dec	**Jan**	**Feb**	**Mar**	
Port Scans	TCP	H	0.01%	0.01%	0.01%	0.02%	0.01%
		L	1.02%	1.01%	0.96%	1.07%	1.01%
	UDP	H	0.02%	0.01%	0.03%	0.05%	0.03%
		L	0.29%	0.47%	0.34%	0.29%	0.35%
Network scan	TCP	H	0.38%	0.38%	0.33%	0.4%	0.37%
		L	10.61%	11.2%	12.59%	10.32%	11.18%
	UDP	H	0.14%	0.21%	0.29%	0.27%	0.23%
		L	3.57%	5.29%	4.15%	4.01%	4.26%
	ICMP	H	0.32%	0.33%	0.27%	0.32%	0.31%
		L	0.41%	0.43%	0.41%	0.41%	0.41%
one flow	TCP		6.29%	5.20%	4.37%	6.21%	5.51%
	UDP		3.43%	3.49%	5.56%	7.87%	5.09%
backscatter	TCP		2.26%	3.47%	3.22%	3.18%	3.03%
	UDP		7.36%	1.4%	1.67%	2.24%	3.17%
	ICMP		17.69%	22.98%	24.31%	30.25%	23.81%
IP Fragment			0.05%	-	-	-	0.01%
Small SYN			74.91%	73.04%	73.03%	72.74%	73.43%
Small UDP			78.61%	84.38%	84.38%	81.72%	82.27%
Small Ping			81.54%	76.26%	74.97%	68.97%	75.43%
Other TCP			4.53%	5.7%	5.5%	6.06%	5.45%
Other UDP			6.33%	4.73%	3.58%	3.54%	4.55%
Other ICMP			0.05%	-	0.05%	0.05%	0.03%
Other			2.3%	0.02%	0.01%	0.16%	0.62%

N% Is based on unique source IPs for each protocol

B

Appendix-B: Scripts

Scripts developed for, and used in the processing of data for this research can be found on GitHub at https://github.com/miadog007/Evaluation-of-Scanning-Taxonomies

These scripts are written in python and have a dependency on the dpkt library for packet processing. Scripts are organised into two directories based on the taxonomic analysis they implement:

- Barnett, R. J. and Irwin, B. (2008)
- Liu, J. and Fukuda, K. (2018)

In each of the above directories, there is a `main.py` which can be run against the required Packet Capture(PCAP) datafile.

Word count metrics

NUC Bachelor Project Word Count:
Total Sum count: 20609 Words in text: 20254 Words in headers: 161 Words outside text (captions, etc.): 193 Number of headers: 73 Number of floats/tables/figures: 34 Number of math inlines: 1 Number of math displayed: 0
NOTE: References are excluded.