Noroff University College

Bachelor in
Cyber Security

IPv4 Scanning: Differentiating Benign from Malicious Patterns

RUBEN SKJELSTAD

SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF BACHELOR IN **CYBER SECURITY**

SUPERVISOR
Professor Barry Irwin

Noroff University College, Norway
May, 2024

Mandatory Declaration

Declarations

The individual student is responsible for familiarising themselves with the rules and regulations regarding the use of sources, generated text and academic misconduct. Failure to declare does not release the student from their responsibility.

1.	I hereby declare that the submission answer is my own work, and that I have not used other sources other than as is referenced and cited correctly, or received help other than what is specifically acknowledged.	Yes
2.	**I further declare that this submission:** • Has not been used for another exam in another course at Noroff University College, at another department/university/college at home or abroad. • Does not refer to or make use of the work of others without acknowledgement. • Does not refer to my own previous work unless stated. • Has all the references given in the bibliography. • Is not a copy, duplicate or copy of someone else's work or answer. • Is not generated using AI generation tools.	Yes
3.	I am aware that a breach of any of the above is to be regarded as cheating and may result in cancellation of the exam and exclusion from universities and colleges in Norway, cf. University and College Act §§4-7 and 4-8 and Regulations on examinations §§ 31.	Yes
4.	I am aware that all components of this assignments may be checked for plagiarism and other forms of academic misconduct.	Yes
5.	I hereby acknowledge that I have been taught the appropriate ways to use the work of other researchers. I undertake to paraphrase, cite, and reference according to the acceptable academic practices, in accordance with the rules and guidelines, as taught.	Yes
6.	I am aware that Noroff University College will process all cases where cheating is suspected in accordance with the college's guidelines.	Yes

Publication Agreement

Authorisation for electronic publication of the thesis: Through submission you are accepting that Noroff University College has a perpetual, and royalty free right to retain a copy of work for its own internal use, and has the right to make work publicly available - considering any restrictions to publication.

Acknowledgements

I would start by thanking Professor Barry Irwin for his exceptional help in this field of study, his knowledge has been invaluable in guiding the direction and focus of this research. His expertise in network scanning and his willingness to share his insights have significantly enriched my understanding. I am deeply grateful for his support and guidance throughout this journey.

I would like to express my gratitude to Greynoise.io for granting me VIP access to their exceptional services. This access has significantly enhanced my research capabilities, allowing me to dive deeper into analysis with unparalleled insight. Furthermore, I extend my thanks to the Computer Security Incident Response Team of the South African Research and Education Network. Their trust in sharing valuable network data with me has been instrumental in advancing our mutual goals of network research. The collaboration and trust of both entities have not only enriched my work, but also underscored the importance of partnership in research. Thank you for your invaluable support and confidence in my endeavours.

Abstract

This research explores the distinction between benign and malicious IPv4 scanning activities. By examining scanning definition from previous studies using data gathered by the South African Research and Education Network provided by Professor Barry Irwin, this research aims to accurately identify and differentiate malicious intents from routine network maintenance. The study uses the Greynoise.io API to validate scanning activities and employs case studies to refine detection techniques and definition. The findings highlight the distinct observed behavioural patterns between benign and malicious scans. Benign scanning hosts prefer longer sporadic scans while malicious hosts prefer shorter network-wide scans. This work contributes to academic discourse by filling a notable gap in the literature on classification for network scanning.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Introduction

In today's digital world, IPv4 networks still remain the backbone of much of our global internet infrastructure. These networks are a target for continuous scanning activities. Although network scanning is a routine and often benign activity, it is also a tool used by malicious actors to look for vulnerabilities. This difference raises a significant question: How can one differentiate between benign and malicious scanning patterns?

Network scanning can be considered as a double-edged sword. On one side, network administrators, researchers, and security vendors use scanning to ensure that the health, performance, and security of their systems is where it needs to be. Benign scanning should be part of maintaining the robustness of the digital world and in organisations. On the other side, threat actors scan networks with the intent to identify vulnerabilities that can be exploited or a malware attack, leading to breaches in confidentiality, integrity, and availability.

Identifying the intent behind a scan is not always easy. Many characteristics of benign and malicious scans might be the same, making it a challenge to differentiate the two at first (Mazel et al., 2017). The concern of this differentiation gets worse when considering the potential consequences of mistaking one type of scan for another. Identifying a benign scan as a malicious scan can lead to unnecessary measures taken, while overlooking a malicious scan can result in significant vulnerabilities being left unnoticed (Marín et al., 2021).

This research explores the patterns, frequencies, and similarities of IPv4 scanning activities prevalent on the Internet. By understanding the difference between benign and malicious scans, the artifact can provide valuable information to network administrators

and future research, helping them develop better and more accurate response strategies.

1.2 Problem Statement

In the digital world of IPv4 networks, numerous scanning activities emerge daily (Richter & Berger, 2019). Although many of these scans have malicious intentions, seeking vulnerabilities or malware to exploit, others serve as routine for administrative, research, or maintenance objectives. Differentiating these scans according to their identity from patterns, frequencies, sources, and methods presents a difficult challenge (Barnett & Irwin, 2008). Is there a reliable way to detect benign routine scans from potentially harmful targeted ones by interpreting their characteristics? Accurately identifying these scans is crucial not only for understanding the digital landscape but also for maintaining robust cyber security defences. Thus, the research problem is the following.

- **How can we accurately differentiate between benign and malicious scanning in IPv4 networks based on specific patterns, behaviours, and characteristics?**

Secondary research question:

- **Can Threat Intelligence services be utilised to accurately distinguish benign scanning from malicious scanning.**
- **Do they exhibit any differences in observed behaviour**

1.3 Research Objectives

This research aims primarily to create a *research thesis to distinguish benign from malicious scanning, combined with two finalised scripts to identify scanning.*

In order to achieve this primary objective, the following secondary objectives have been identified:

- Find a source for network data.
- Create scripts to identify scanning activity.
- Pair the output from the scripts created, with Greynoise.
- Define the identity of benign and malicious scanning.
- Create case studies to further enhance the findings.
- Compare main studies with case studies.

1.4 Primary Data Set

Professor Barry Irwin from Noroff University provided the primary dataset for this study. The dataset being used for this research has been collected, and is owned by the research

supervisor. These data were collected using infrastructure hosted by the CSIRT for the South African Research and Education Network (SANREN) as part of an ongoing project on IBR collection and analysis. Portions of the data were collected prior to 2020 under the auspices of the Security and Networks Research Group at Rhodes University.

1.5 Artefact and Evaluation

The primary deliverable of this research will be an analytical artifact with the characteristics and identifiers of benign and malicious network scanning. Additionally, the research will produce a script to help identify scanning and distinguish benign from malicious intentions, with the aim of helping them understand benign and malicious scanning activities and their significance.

1.6 Ethical considerations

In conducting this research, ethical considerations around data privacy and handling sensitive information are paramount. The dataset used is in this research is owned by South African Research and Education Network, and any personal information will be anonymized to protect the privacy of individuals or organisations involved in the scans.

1.6.1 Scopes and limits

The research specifically targets network scans directed at network telescopes within South Africa, with a precise focus on differentiating between benign and malicious network scanning activities based on predefined criteria such as scan rates and destinations. It deliberately avoids the complexities of identifying specific scanning tools, which would require extensive resources, ensuring that the study remains focused on broader network behaviours.

1.7 Layout of Thesis

Chapter 2: Literature Review - Discussion on existing literature of network fundamentals, scanning methods, research organisations, and analysis tools.

Chapter 3: Methodology - Detailed methodological approaches, including data source acquisition, characterisation of scanning activities, correlation with external events.

Chapter 4: Analysis - Detailed analysis of network scans, including protocol distribution, geographical sources of scans, source IP addresses, port analysis, and behavioural analysis.

Chapter 5: Discussion - Discusses case studies, compares various methods, and summarising findings from different angles of analysis.

Chapter 6: Conclusion - Summaries the entire research, highlights the objectives met, contributions of the research, and suggestions for future work.

CHAPTER 2

Literature Review

2.1 Background

The following section is essential to introduce the reader to the fundamental concepts and existing research on the subject. 'Network Fundamentals' begins by introducing the fundamental concepts and terminologies associated with networking. This is critical to establish a basic understanding. 'Ports' then explore the specific gateways data travels in networks, explaining their significance and role. The following section, 'Network Scanning,' introduces the methodology of probing networks to find active devices, vulnerabilities, or open ports, laying the preparation for a more in-depth discussion. 'Different Types of Network Scans' then provides an in-depth exploration, detailing different methodologies and the specific purposes each of the scans serves in network analysis. 'Network Telescopes' takes a slight important detour to discuss modern systems that monitor enormous collections of Internet traffic. Finally, the 'Related Work' offers a review of the existing literature and research, establishing the current work in the field of the research. This structured progression ensures a logical flow of information and constructs a coherent understanding for the reader.

In Sections 2.2 and 2.2.1, the focus is Network Fundamentals and Internet Protocol version 4 (IPv4), where the basics of networking and the specifics of the IPv4 protocol are discussed. Section 2.2.2 delves into ports, an essential aspect of network communication and security.

Section 2.3 explores Network Scanning, a critical area for network security and management. This section is further divided into discussions on popular scanning tools such as Nmap, Zmap, and Masscan in Sections 2.3.1, 2.3.2 and 2.3.3, providing information on their functionalities and usage.

Section 2.4 elaborates on different types of network scan, breaking down various scan-

ning techniques such as TCP SYN scan, TCP Connect Scan, ACK Scan, UDP Scan, IP protocol scanning, and other scanning options. This part offers a detailed schematic of these scanning methods, their applications, and effectiveness.

In Section 2.5, the discussion expands to Vertical and Horizontal internet scans, explaining these strategies in the context of large-scale network analysis.

Section 2.6 and its subsections focus on research organisations such as Shodan and Censys, and other organisations, highlighting their role and contributions to network research.

Section 2.7 introduces the concept of Network Telescopes and explores specific types such as the cloud network telescope in 2.7.1.

Section 2.8 is dedicated to Analysis Tools, discussing various tools such as Wireshark, Tshark, and Greynoise.io API. This section examines the features of these tools, their usage in network analysis, and their roles in interpreting network data.

Finally, Section 2.9 discusses Related Work, reviewing previous studies and developments in areas such as identifying scanners, large-scale scanning, classifying scanning intentions, and visualising scanning activity.

2.2 Network Fundamentals

2.2.1 Internet Protocol version 4

Internet Protocol version 4, or IPv4, is the most used protocol for addressing Internet addresses in today's data (Jenani, 2017) Since its inception, IPv4 has been the cornerstone of the Internet's network protocol. The protocol was designed in 1980 and was modified to take the place of the Network Control Protocol (NCP) on ARPANET (Postel, 1981). The IPv4 protocol can host $2^{32} = 4,294,967,296$ unique Internet Protocol addresses, with its 32-bit architecture (Jenani, 2017). The internet packet is made up of 14 fields, whereas 13 of them are required for it to come through. The only field in the IPv4 header that is optional is the OPTIONS field (Postel, 1981). This is not necessary for the IPv4 packet to be sent and received. The internet and the use of devices have nearly outgrown the potential host addresses, and there was a need for another protocol to support the new devices. Fast forward to 2011, and IPv6 is constructed (Wu et al., 2013).

2.2.2 Ports

A port number is a unique number that specifies the connection endpoint to which a service should connect (Postel & Reynolds, 1992). If the connection endpoint is 80, it (in most cases) wants to connect to the server through the HTTP protocol (Postel & Reynolds, 1992). Another case is where the endpoint is port 22/tcp a connection is wanted on the SSH protocol, and then a TCP transport is needed (Irwin, 2011). The Internet Assigned Numbers Authority (IANA) is the organization responsible for coordinating Internet

numbers (IANA, 2023). In particular, IANA oversees the allocation and management of service names and port numbers [1], as detailed in their comprehensive directory. This is important as it is crucial when we are talking about network scanning.

2.3 Network Scanning

Network scanning is a fundamental aspect of both cyber security and network management (Marksteiner et al., 2020). It involves sending specific network data to devices on a particular network to gather information about active ports, operating systems, services, and vulnerabilities (Barnett & Irwin, 2008). This information is valuable not only for administrators who want to ensure that their networks are secure and for penetration testers looking to find weaknesses in a system, but also for malicious users who want to penetrate a device or network(Marksteiner et al., 2020). In the evolving landscape of cyber security, where new vulnerabilities and threats emerge daily, regular network scans can be the difference between a secure network and a compromised one.

Network scanning was first seen in 1992 (Boulanger, 1998). One of the first commercial scanning application was *Internet Security Scanner (ISS)* created by Christopher Klaus. However, network scanning was first popularised by Dan Farmer and Wietse Venema with the program *SATAN (System Administrator's Tool for Analyzing Networks)* in 1995 (Farmer & Venema, 2000). The ISS scanner and SATAN was quite similar although the latter one had advancements as SATAN had added some more network tests, and was later updated to work as a vulnerability scanner. Another network scanner released at the time was 'strobe - Super optimised TCP port surveyor' in 1995 by Julian Assange (Higgins, 2011). Strobe[2] was the first open source internet scanner to be published (Higgins, 2011).

TIME Magazine published "The Devil In The Network", and as the SATAN scanner gained popularity, concern about hackers also emerged. SATAN's fame is the root of other network scanners, like NMap, ZMap and Masscan.

Hendricks (2019) states that network and port scanning is essentially the same, the difference being the process in which each scan is performed. Hendricks (2019) stated that scanning can be the works of a worm or a virus which is self propagating and spreading, or a human wanting to target a specific network or organisation. Such a worm can be SSH worms and other generic OS worms that has infected a a computer and scans the internet for a vulnerable device (Wang et al., 2013).

2.3.1 Nmap

One of the most well-known tools for network scanning is Nmap[3] (Network Mapper). It is an open-source tool that allows for detailed network enumeration. Nmap enables users to identify active devices in a network and discover open ports, in addition to

[1] https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml
[2] https://github.com/whackashoe/strobe/blob/master/strobe.man
[3] https://nmap.org.

providing information about various characteristics of the network (Orebaugh & Pinkard, 2011). It can also help determine the operating system of the devices and the version of the services they are running, which is crucial for vulnerability assessment. In addition to Nmap, there are other tools like Masscan, which is known for its speed, and Zenmap, a graphical front-end for Nmap (Jafarian et al., 2023). These programs, when used appropriately and ethically, provide deep insight into network configurations, helping professionals improve defences and maintain robust network health (Wan et al., 2020);(Orebaugh & Pinkard, 2011). A discussion of the findings related to Nmap has been provided in Section 4.13.

2.3.2 Zmap

ZMap[4] is another open source network scanner specifically designed for global Internet research (Durumeric et al., 2013). Unlike traditional network scanners such as Nmap, ZMap was designed to scan the entire Internet in a rapid and complete manner (Durumeric et al., 2013). ZMap was developed by *Zakir Durumeric, Eric Wustrow, and J. Alex Halderman* (Durumeric et al., 2013) when they were looking for a more efficient way to analyze the global Internet. Traditional scanning tools, like Nmap, were robust and had many features, but were not designed for large Internet-scale scanning due to time constraints and the harsh results they produced (Orebaugh & Pinkard, 2011). Knowing this, a team from the University of Michigan set out to create a tool that would become ZMap.

Durumeric et al. (2013) states that a normal computer with no special hardware or kernel access could scan the entire IPv4 network for a specific open network-port and in a speed as fast as 97% of the theoretical fastest speed in a gigabit Ethernet line. In comparison to Nmap, ZMap could also send single-packet probes, like ICMP ping request scans, UDP Scans and TCP SYN scan. This opened up new possibilities for researchers, security professionals, and even attackers to gain insights into the global landscape of devices, services, and vulnerabilities present on the Internet.

ZMap is regularly used for academic research scans and malicious actors trying to find vulnerable sources (Durumeric et al., 2014). Censys is one of the organisations that uses ZMap for research purposes (Bennett et al., 2021). Indications for the use of ZMap were found and have been discussed in further detail in Section 5.8.2 and Section 5.8.1, and illustrated in Figure 5.8.1, Figure 5.6 and Figure 5.8.

2.3.3 Masscan

Another internetscale scanner is Masscan[5], developed by Robert David Graham (Trapkickin, 2015). Masscan was first published in October 2013 (Graham, 2013). Graham boldly claimed that Masscan could complete a scan of the entire internet in just 6 minutes, provided there was a 10 GbE connection available (Graham, 2013). A tug of war started

[4] https://zmap.io.
[5] https://github.com/robertdavidgraham/masscan.

between Robert Grayham with his Masscan against the developers on ZMap. Graham stated that Masscan was the scanner on top because of its advancement by using the Linux kernel, however, in Durumeric et al. (2013) they weaken Grayham's claims by experimentally outperforming them.

The key characteristics of Masscan include an asynchronous framework for thread-based tasks, comparable to ZMap (Durumeric et al., 2013). The primary benefit of this method is probe randomisation, which randomly distributes IP addresses as it scans the IPv4 address space. Graham's solution makes use of the BlackRock algorithm, which is based on a modified DES encryption method, and trades off some statistical randomness for faster packet processing (Trapkickin, 2015). This enables fast processing of high packet rates, such as 10 million packets per second. The goal of managing 10 million connections at 10 Gbps/10 Mpps with 10 microseconds latency is based on Masscan's C10M architecture. Trapkickin (2015) advises doing this by eliminating network operations from the kernel and, instead, using the *PF_RING ZC* driver to work outside the kernel and make a TCP/IP stack for itself.

By creating its own TCP/IP stack, Masscan operates in a way that might not be easily traceable or comparable to regular network traffic. This aspect is vital in understanding how malicious scanners could disguise their traffic. Furthermore, the technique of probe randomisation, which distributes IP addresses randomly during scanning, is significant. Malicious scanners might use this approach to avoid detection and appear less systematic.

2.4 Different Types of Network Scans

There are multiple types of network scans unique for its purpose. Each type of scan revolves around a type of network protocol. Researchers and administrators, as well as threat actors, use these types of scans. Whether it is transmission control protocol (TCP) scan or user diagram protocol (UDP) scan, their task is to enumerate the network. Nmap offers various scan types, each made for specific situations and providing unique information about the target network.

2.4.1 TCP SYN scan

The TCP SYN, also known as Half-Open Scan scan, is one of the most popular scanning techniques used with Nmap. It is sometimes referred to as a 'stealth' scan because it does not complete the traditional three-way handshake process (Orebaugh & Pinkard, 2011). Instead, it sends an SYN (synchronisation request) packet to the target address, and if the targeted address responds with a SYN-ACK (synchronisation acknowledgement), it marks the port open (Orebaugh & Pinkard, 2011). Nmap then sends an RST (reset) packet to unexpectedly close the connection before the handshake completes. This type of scan is relatively fast and evasive for intrusion detection systems (IDS), which makes it harder for intrusion detection systems to detect (Upadhya, 2020).

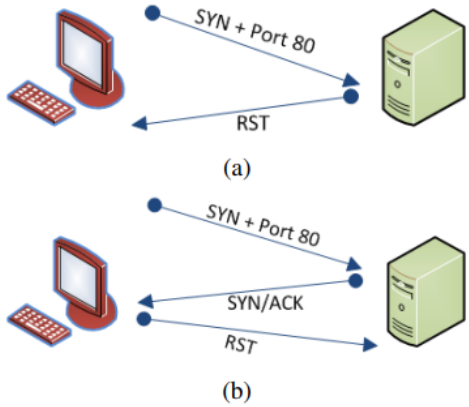

Figure 2.1: (a) TCP SYN scan on a closed port 80, (b) TCP SYN scan on a open port 80

2.4.2 TCP Connect Scan

The TCP Connect Scan type gets its name because it uses the connect system call to open a connection to every target port. Unlike the SYN scan, the connect scan completes the three-way handshake (Orebaugh & Pinkard, 2011) . This means it's easier to detect but does not require raw packet privileges, making it useful in environments where the SYN scan is not an option. Raw packet privileges, which means that this scan does not need elevated privileges to run a scan (Orebaugh & Pinkard, 2011).

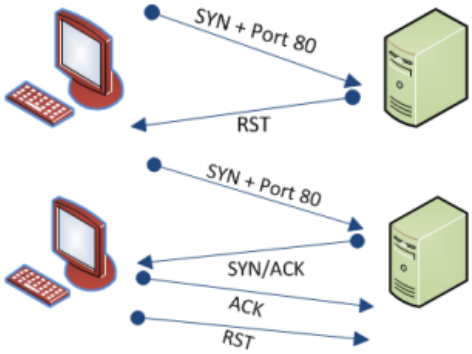

Figure 2.2: (a) TCP Connect scan on a closed port 80, (b) TCP Connect scan on a open port 80

2.4.3 ACK Scan

The primary objective of the TCP ACK scan is to map the firewall rules (Orebaugh & Pinkard, 2011). This scan type is not used to determine whether a port is open or closed but rather to see if it is filtered by a firewall. By sending an ACK packet, Nmap tries to identify ports that are filtered and those that are not filtered, and this helps to enumerate the involvement of firewalls in the network (Orebaugh & Pinkard, 2011).

2.4.4 UDP Scan

In a User Datagram Protocol (UDP) scan, the scanner typically sends an empty UDP header to every target port (Orebaugh & Pinkard, 2011). For many closed ports, the target device will respond with an Internet Control Message Protocol (ICMP) *Port Unreachable*

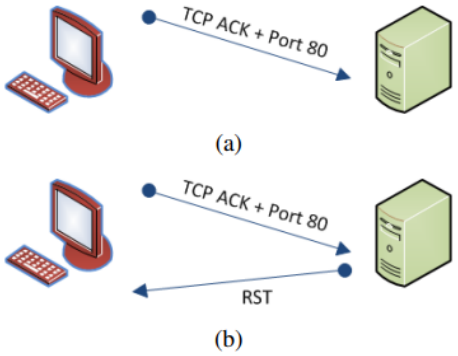

Figure 2.3: (a) TCP ACK scan on a closed port 80, (b) TCP ACK scan on a open port 80

message, indicating that there is no active service listening on that port. However, if a UDP port is open and listening for incoming data, the behaviour varies (Upadhya, 2020). Some services may respond with a protocol-specific response, while others might offer no response at all. This lack of consistent feedback makes UDP scans less definitive than their TCP counterparts (Orebaugh & Pinkard, 2011). In addition, some organisational networks might limit ICMP error messages, which will disrupt scan results. Despite this, UDP scans are still usable, especially when assessing network security, as many critical services and vulnerabilities rely on UDP.

Scanning UDP is challenging due to the fact that many services do not respond to empty probes, rendering the differentiation between open and filtered ports impractical (Orebaugh & Pinkard, 2011). However, for certain ports, Nmap is equipped with specific payloads that are harmless to dispatch and are likely to provoke a response.

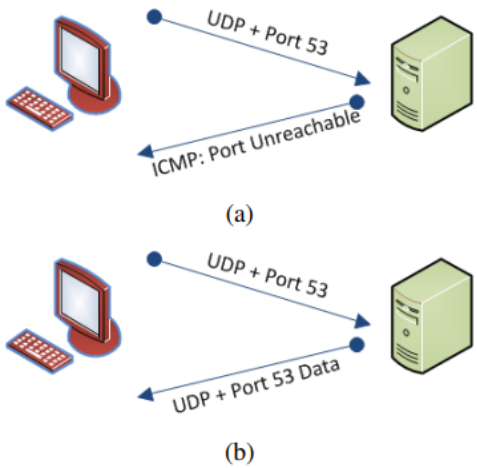

Figure 2.4: (a) UDP scan on a closed port 53, (b) UDP scan on a open port 53

2.5 Vertical and Horizontal internet scans

There is a significant difference between horizontal Internet scans and vertical internet scans. Aniello et al. (2011) defines a horizontal scan as a scan of a specific port on multiple IPv4 hosts, and a vertical scan as a scan that targets a single host on different ports. This understanding and definition are important when talking about port scanning, and could be used later in research (van Riel & Irwin, 2006).

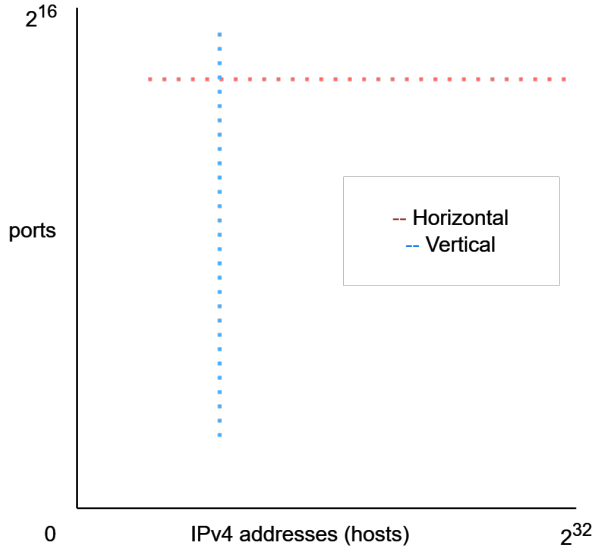

Figure 2.5: Illustration of the difference between vertical and horizontal Internet scanning

2.6 Research Organisations

Censys[6], Shodan[7], Project Sonar[8] by Rapid7 and the Shadowserver Foundation[9] are considered non-malicious organisations for several key reasons.

1. They operate with a high degree of transparency, often publicly documenting their methodologies, the data they collect, and the purposes for which it is used. This openness is in contrast to malicious actors that operate behind closed doors (University of Michigan, 2018).

2. These organisations involve themselves with the cyber security community and comply with ethical standards and legal frameworks. They provide valuable resources and services that contribute to the security and resilience of the Internet, such as identifying vulnerable systems so they can be secured and sharing threat intelligence that can prevent cyber attacks (Kaspersky, 2016).

3. Their research activities often include working with academic institutions, industry, and sometimes law enforcement, further legitimising their operations.

[6] https://censys.com
[7] https://www.shodan.io
[8] https://www.rapid7.com/research/project-sonar/
[9] https://www.shadowserver.org

4. Rather than exploiting vulnerabilities, they aim to expose potential threats and help remediate them, establishing a proactive approach to improving cyber health (Rapid7, 2019).
5. Their collective mission is to make the digital world safer for everyone, a goal inherently aligned with the principles of benignity, which are the opposite of malicious intentions (Kaspersky, 2016).

2.6.1 Shodan

Shodan.io[10] is a specialised search engine for the Internet of Things (IoT), offering a platform for users to find information about devices connected to the Internet rather than websites (Shodan, 2022). Shodan provides data on devices ranging from household electronics to industrial systems by scanning the Internet and collecting publicly available information from device banners, metadata that reveals details about the device's software and options (Shodan, 2022).

Launched in 2009, Shodan has become a tool for IT professionals and cyber security researchers to identify and secure network vulnerabilities and for businesses to conduct market research by examining the distribution of devices or software over the Internet (Shodan, 2022). Shodan also highlights the scale of publicly accessible information, highlighting the importance of securing devices against unauthorised access (Shodan, 2022). There is evidence of the appearance of shodan in case studies in Section 5.6.2

2.6.2 Censys

Censys[11] is a tool for searching, monitoring, and detecting devices that are on the Internet. Its technique includes regular scans of domain names and public IP addresses, creates an extensive dataset that can be used to identify emerging cyber threats and vulnerabilities in network devices (Censys, 2023a). It was developed by academic research of computer scientists at the University of Michigan (Censys, 2023a).

As mentioned in 2.3.2, Censys uses ZMap for its search on the Internet for connected devices on the Internet (Bennett et al., 2021). Censys can be used without paying and therefore can be used to search for probes used in this research (Censys, 2023b). Censys was found in a case study and its behaviour is discussed in Section 5.6.2.

2.6.3 Other Research Organisations

Rapid7 Project Sonar

Project Sonar[12] owned by Rapid7 is another organisation that conducts mass scans on the Internet to gain insight into global exposure of common vulnerabilities (Rapid7, 2017). Much like Censys and Shodan, the data collected by Project Sonar are used to improve

[10] https://www.shodan.io
[11] https://censys.com
[12] https://www.rapid7.com/research/project-sonar/

security decisions. They upload datasets so that researchers and network administrators can understand the changing landscape of the network (Rapid7, 2023).

Shadowserver Foundation

Shadowserver Foundation[13] is a non-profit security organisation, Shadowserver conducts mass scans of the Internet to identify malicious activity and vulnerable services, much like Project Sonar. They provide free daily network reports to network administrators, helping them mitigate potential threats and secure their systems (Shadowserver Foundation, 2024).

Greynoise.io

GreyNoise.io is a great contributor to the cyber security research domain, specialising in the analysis of internet scanning traffic. They work with differentiating cyber threats with background noise and benign actors (Greynoise, 2024). GreyNoise Labs, their research division, is a showcase of pioneering experiments and progress in cyber security, enhancing understanding of the wide range of internet behaviours, encompassing benign and malicious activities (Greynoise, 2024).

Furthermore, GreyNoise has contributed significantly to the cyber security space through numerous reports and studies. Their 2022 Mass Exploitation Report[14], offers detailed information on the major threat detection incidents of the previous year, shedding light on common vulnerabilities and exploitation patterns (Greynoise, 2022). This report shows their dedication to unraveling and sharing the difficult aspects of cyber security threats and protective measures. Greynoise will be discussed in more detail in Section 2.8.1 and Section 2.8.2.

Internet Systems Consortium, Inc.

Internet Systems Consortium, Inc. is a research organisation that specialises in the release of open source network packages (Internet Systems Consortium, Inc., 2024). ISC believes strongly in protecting the Internet with open source software from potential control by businesses or governments (Internet Systems Consortium, Inc., 2024). Isc.org[15] provides an API[16] to several TI feeds, providing important information about other research organisations, such as 2.6.1 Shodan, 2.6.2 Censys and the mentioned organisations in section 2.6.3.

[13] https://www.shadowserver.org
[14] https://www.greynoise.io/resources/greynoise-2022-mass-exploitation-report
[15] Isc.org
[16] https://isc.sans.edu/api/threatcategory/research?csv

2.7 Network Telescopes

Monitoring network threats is crucial in the digital world today. Network telescopes have been used since the early 2000s to deal with this monitoring of networks (Pearson & Irwin, 2018). To define a network telescope, a type of network monitoring system or sensor designed to passively and non-intrusively observe and collect data on Internet traffic (Moore et al., 2004). It is often used for security research, network analysis, and monitoring purposes. Network telescopes are typically deployed as network segments or 'sections of the IP address space that see minimal or no authentic traffic' (Pearson & Irwin, 2018). Security and Networks Research Group (SNRG) at Rhodes University has maintained some network telescopes since 2005 (Irwin, 2011).

Irwin (2011) study relating to the use of network telescopes conveyed significant insights into malicious activities, and how network telescopes can be used to monitor network traffic. Furthermore Irwin (2011) researches into mechanisms to aid in the differentiation of background network traffic and malicious. Additionally the study provides an insight into what the purpose of the malicious network data, if it is intended for a specific organisational network or just widespread hit or miss.

2.7.1 Cloud Network Telescopes

As an evolution of the network telescope, the cloud network telescopes have emerged. Cloud network telescopes serve the same purpose as a regular network telescope, only to be deployed in a cloud service provider (Bortoluzzi, Irwin, & Westphall, 2023). Sensors can be installed in different regions of the world to better understand the pattern differences that separate regions (Bortoluzzi, Irwin, Beiler, & Westphall, 2023). When deployed on a cloud platform, these telescopes leverage the elasticity and scalability of the cloud, allowing rapid adjustments to capture and analyse varying traffic patterns. Cloud network telescopes also serve as a more cost-effective way of collecting packets because of the cloud option to run sensors as instances instead of an onsite computer.

2.8 Analysis Tools

A thorough examination of network traffic data is essential for revealing significant insights into scanning patterns. The volume of data collected has increased tremendously as a result of the spread of digital communication networks, which requires the use of advanced tools and procedures for in-depth analysis. Greynoise.io's service will be integrated within a Python script to iterate over source addresses found to be scanning the network telescopes.

2.8.1 Greynoise.io API

The GreyNoise API is an integral part of the GreyNoise intelligence platform, designed to analyse and label Internet scan and attack traffic (Greynoise, 2023a). This API allows users to differentiate between benign and malicious activities in cyberspace, aiming to help security teams increase efficiency by focusing on genuine threats, thus reducing false positives (Greynoise, 2023a). By categorising IP intent, providing detailed context, and enabling automation of workflows, the API facilitates the reduction of noisy alerts and improves response to mass exploitation (Greynoise, 2023a).

GreyNoise's API offers various outputs for different use cases, such as quick IP lookup, and detailed context queries for single or multiple IP addresses. The API endpoints, such as the multi-IP context endpoint, are designed to help determine if a given IP address is known for internet scanning activity (Greynoise, 2023a). The greynoise.io API will be used within the script, the flow chart can be found in Section 3.6.

Furthermore, GreyNoise ensures that its databases are updated in real time, collecting data from a series of sensors that monitor Internet-wide exploitation (Greynoise, 2023a). This results in a dynamic and timely intelligence service that is trusted by a diverse range of enterprises and security professionals (Greynoise, 2023a).

2.8.2 Greynoise.io Visualiser

The GreyNoise Visualiser is a unique and powerful tool developed by GreyNoise Intelligence, designed to provide a comprehensive view of Internet-wide scan and attack traffic. This visualiser is specifically designed to help security analysts and IT professionals understand and differentiate between benign and malicious Internet noise (Greynoise, 2023b).

Displays real-time data on IP addresses that saturate security tools with irrelevant traffic, allowing users to focus more effectively on targeted and emerging threats. The visualiser stands out for its ability to make internet noise comprehensible and actionable, thereby maximising the efficiency of security operations centres (SOCs) and aiding in defence against mass exploitation (Greynoise, 2023b).

By providing insights into common and unusual patterns of internet activity, it helps organisations make informed decisions about their cyber security posture. The tool is part of GreyNoise's larger suite of solutions and is trusted by enterprises, government organisations, and security researchers worldwide for its precision and utility in the realm of cyber security (Greynoise, 2023b).

The greynoise visualiser is used in Section 4.11 and regularly throughout Chapter 5 and has proven to be a superior service in mapping benign actors. The greynoise visualiser will be used in Chapter 4 and Chapter 5.

2.8.3 InetVis

InetVis[17], is a 3-D scatter plot visualisation tool that effectively visualises network events to better understand network traffic (Irwin & van Riel, 2007). InetVis takes a packet capture file as input and plots the resulting connections on a 3D plot (Irwin & van Riel, 2007).

The destination address within the packet capture file is represented on the blue X axis, which runs horizontally. This axis visualises the IP addresses that receive incoming network traffic (van Riel & Irwin, 2006).

The source address, indicating external Internet IP addresses that are initiating contact, is displayed on the red Z-axis, adding depth to the visualisation. This helps in identifying the external origins of network traffic (van Riel & Irwin, 2006).

Finally, the network ports, covering both the TCP and UDP ports, are mapped along the green Y-axis. This vertical axis helps in understanding which network ports are being used for incoming and outgoing traffic (van Riel & Irwin, 2006).

Below the main 3D plot is a 2D ICMP plane, marked in grey (van Riel & Irwin, 2006).

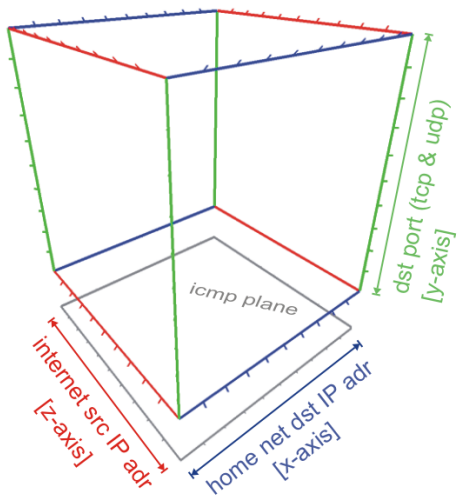

Figure 2.6: InetViz's 3-D Network Traffic Visualisation: Mapping Destination IP, Source IP, and Port Activity (van Riel & Irwin, 2006).

2.9 Related Work

This section explores some of the related work and highlights important studies and conclusions that closely relate to the topics of this research. These researchers have laid significant groundwork and provided insight within the themes and objectives of this study. By diving into their findings and methodologies, this thesis aims to connect their existing knowledge with fresh perspectives and further develop the academic discourse on this subject.

[17] https://github.com/yestinj/inetvis

Identifying Scanners: Section 2.9.1 examines the various strategies and tools that prior research has done to detect IPv4 scanning activities. It underscores the technological advancements and the analytical techniques that form the foundation for identifying potential network probes and sweeps.

Identifying Large Scanning: Section 2.10 presents relevant research used to detect large-scale network scanning activities. This analysis is crucial for understanding the patterns and intentions behind large network scans, providing insights into both benign and malicious intent.

Geographical Cybercrime Index: Section 2.12 mentions research done to index countries based on their competence and skill in cybercrime.

Classifying Scanning Intentions: Section 2.11 is dedicated to investigative the methods used to differentiate between benign and malicious IPv4 scans. It considers the frameworks and methods that researchers have developed to distinguish benign scanning activities from malicious.

Graphical view of Scanning Activity: Section 2.13 provides delving into methods other research has done to make a historical perspective on the trends and patterns of IPv4 scanning.

2.9.1 Identifying Scanners

Ghiëtte (2016) refers to IPv4 identification (IP-ID) as numbers in the IP header that can be used for fingerprint scanners. Some scanners prefer to have a hard coded IP-identification number in the source code. Ghiëtte (2016) research also found that there were 11 IP identification numbers that had a higher appearance in the dataset. This could suggest that some of these numbers were used by scanners.

Ghiëtte (2016) research made it possible to identify which scanner was being used in a port scan/network scan by identifying their fingerprints. ZMap was by far the easiest to identify due to is fingerprint, while Masscan and NMap being relatively harder due to it encoding of the fingerprints.

Identifying ZMap

ZMap can be identified by looking at the IPv4 identification (IP-ID) field. Looking at the source code of ZMap 5, the IPv4 identification field is a fixed value set to *54,321* (Ghiëtte, 2016). That means that a packet capture of a ZMap scan will most likely be indicated by the IP-ID field with a value of *54,321*, unless it has been changed to avoid detection. The chance that a packet with an IP-ID of *54,321* not being ZMap, is $\frac{1}{216}$ (Ghiëtte, 2016).

Identifying Masscan

For Masscan the process is a bit harder, because it uses XOR to find the IP-ID. The values used are the TCP sequence number and destination port, and IPv4 destination IP. The

formula for the Masscan IP-ID:

$$ip\_id = dst\_ip \oplus dst\_port \oplus tcp\_seq$$

All values are known when capturing a packet, therefore, the XOR can be reversed to get the IP-ID (Ghiëtte, 2016).

Identifying NMap

Finding the IP-ID of NMap is similar to Masscan but with more values in the XOR operation, making it harder to calculate. In the XOR operation for NMap the session key (seqmask), a number of attempts, a random port, and the ping sequence are encoded into a TCP sequence number (Ghiëtte, 2016). A simplified version of this encoding can be seen in Figure 2.7.

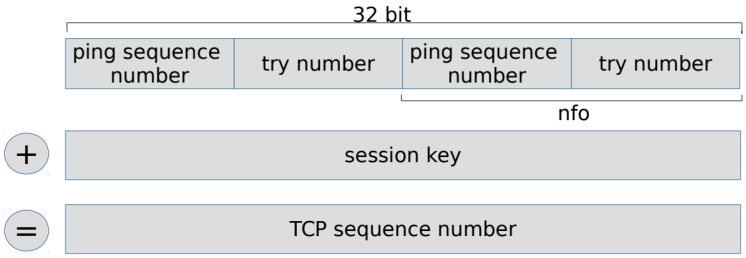

Figure 2.7: Simplified illustration of NMap's TCP sequence number generation process using XOR operation with a session key (Ghiëtte, 2016)

.

If enough packets from Nmap are captured, it is possible to reverse the encoding. When packets from the same source address are captured, they can be reversed using the XOR operation. If the same key *K* (session key) has been used in the decoding of the information in both source addresses, it can be reversed. XOR can be used on the session key (the result of Figure 2.7), to cancel the session key *K* and the result will be the encoded ping scanning try number and the sequence number of each packet (Ghiëtte, 2016). Despite all this effort, the try number and the sequence number are still encoded in the nfo (Figure 2.7), which is mirrored before the XOR (Ghiëtte, 2016). Due to this nfo mirroring, the result of the XOR of the two sequence numbers will also be mirrored (Ghiëtte, 2016). The interesting part is that there are only 65536 possible mirrored numbers for every result of the XOR operation of two sequence numbers.

Therefore, a likelihood of an XOR operation on two sequence numbers resulting in a reversed number is 1 in 65,536. This likelihood is the same as the chance that two packets share the same IP identification number. Hence, finding such a mirrored number as a result of the XOR operation strongly indicates NMap activity (Ghiëtte, 2016).

2.10 Identifying Large Scanning

Durumeric et al. (2014) defines a large scan as follows when the source IP address wants to establish a SYN connection to more than 100 of its own destination addresses, which

are hidden in their darknet, to the same port and at a frequency of 10 packets per second. This means that one has to look at the source address, and the number of destination addresses it has probed to, and from there look at the ports that were probed from the source address. If this is a match, one can go forward with looking at the packet per second.

One can not be certain if the definition as stated above concludes for a Internet-wide scan. But since the network telescopes are not responding to port scans, and the subnet is 'empty', one can draw conclusion of an internet wide scan Durumeric et al. (2014).

2.10.1 Censys

Bennett et al. (2021) have done incredible work in identifying scans from research organisations. Their research made it easier to identify scans from Shodan.io and Censys. By looking at source IP, they were able to identify benign scanning from Censys. The benign scanning from Censys came from the 198.108.66.0/23 subnet, and their experiment noted 189 unique IP addresses from this subnet (Bennett et al., 2021). This will be discussed in Section 5.6.2.

2.10.2 Shodan

As noted in 2.6.1 Shodan uses its own search engine to scan the Internet. This makes it harder to identify the patterns for the benign scan. Bennett et al. (2021) did it by using the tool *SecurityTrails* [18] to check the DNS history of $*.shodan.io$. In this way, Bennett and his group could check when the domain was first seen by the SecurityTrails API and when it was last changed (last seen).
In Bennett et al. (2021) research, if the first packet in a TCP session had a source IP that matched in the SecurityTrails API and the timestamp was between the first seen and the last seen, they wrote it down as a packet originating from a Shodan.io scan. Using this methodology, the group was able to pair 16 IP addresses with $*.shodan.io$. To further confirm their match, they compared the SHA1 hash of the 16 IP addresses they found with the Shodan *crawler* field, this is a hexadecimal string containing 40 characters that appears when querying Shodan.io. This method resulted in each IP being its own Shodan scanner.

2.11 Classifying scanning Intentions

In Irwin (2013) the purpose of the research was to analyse five network telescopes to potentially find malicious network activity. The research focused on ICMP, UDP, and TCP packets. The research with TCP packets was only considered if the SYN flag was set, as this could give a better indication of malicious network activity (Irwin, 2013). In Yamada and Goto (2013) an alternative technique is proposed to identify malicious packets. The

[18] https://securitytrails.com

study shows valuable results that could be used in this study. Observing only the Time-to-Live values, it is possible to distinguish between malicious and legitimate packets (Yamada & Goto, 2013).

Nkhumeleni (2014) showed that by looking at different variables in packet size and packet count it was possible to trace the packets to a specific malicious presence of a viral Worm, Virus or Denial of Service. Some of them being, W32.Rinbot4 that was indicated by variation of packet size, packet count rations of top ports, which could indicate the Conficker Worm, and big hoops of packets indicating a Denial of service. Since W32.Rinbot4 and Conficker Worm is older worms, it is not expected to find evidence of it, however Mirai evidence has been found and mentioned in Section 5.5, Section 5.6.1 and Section 5.10.

2.12 Geographical Cybercrime Index

Bruce et al. (2024) has created the 'World Cybercrime Index' (WCI) to get a better understanding of how the geographical cybercrime scene is today. The index is calculated and based on 5 categories, 'Technical products/services', 'Attacks and extortion', 'Data/identity theft', 'Scams', and 'Cashing out/money laundering' (Bruce et al., 2024).

The three categories relevant for this research are the 'World Cybercrime Index' it self, and the categories 'Technical products/services' and 'Attacks and extortion', which would use malicious network data. The 'World Cybercrime Index' scores from 0-100, where 100 is the most criminal. 'Technical products/services' and 'Attacks and extortion' also spans from 0-100.

Table 2.1: Top 5 Countries in relevant WCI categories

	WCI Overall	Technical products/services	Attacks and extortion
1	Russia	Russia	Russia
2	Ukraine	Ukraine	Ukraine
3	China	China	North Korea
4	USA	USA	China
5	Nigeria	Romania	USA

This index is relevant and a suitable comparison to the geographical heat maps is Section 4.5 and Section 5.7. The mentioned sections also expands past top 5.

2.13 Graphical view of Scanning Activity

The research done by Irwin (2013) has developed timelines for network data, with notation for spikes and decreases. This research presents remarkable information about the correlation between spikes and malware's being present in the cyber landscape. Anand et al. (2023) is another research that has informative timelines of network telescope scanning data. Bou-Harb et al. (2013) made some heat charts of the source country of scanning, this could be useful to see what countries to be aware of. Useful graphical

illustrations have been created for a high-level analysis, and are present in Section 4.5, Section 5.7 and Section E.

2.14 Summary

This chapter delves into various crucial aspects of network fundamentals and their security implications, categorically discussing the Internet Protocol, ports, network scanning, types of network scans, network telescopes, and analysis tools, in addition to reviewing related works in the field regarding network scanning. In addition, it explores the Internet-wide scanning tools that shed light on their capabilities and roles in identifying network vulnerabilities. The chapter extends to the conceptualisation and utilisation of network telescopes and cloud network telescopes, illustrating how these tools aid in monitoring and analysing internet traffic to detect and study malicious and benign activities. The section on analysis tools elaborates on the use of advanced software like Greynoise.io API and visualiser to distinguish between benign and malicious network activities effectively. The chapter concludes with a synthesis of related research work, demonstrating ongoing academic efforts to enhance the understanding of network scanning activities, classify their intentions, and visualise scanning patterns for better cyber threat assessment. This detailed review builds a comprehensive understanding of the current and evolving network security challenges and methodologies.

There is a notable gap in research around the classification of benign and malicious scanning. This research aims to address this gap by developing a clear and methodical approach to differentiate between these types of network scans and to look for distinguishable patterns. Using a detailed analysis of scanning patterns, rates, and geographic origins, this study improves understanding and allows for more accurate threat detection within IPv4 networks.

CHAPTER 3

Research Approach/Methodology

This chapter presents a comprehensive methodology that serves as the basis for research to distinguish between benign and malicious network scanning activities within IPv4 networks. The research begins with a rigorous definition of a network scan. This definition is pivotal to the study, focusing on instances where a source address probes 40 destination addresses at a rate exceeding 10 packets per second.

The scope of the investigation is precisely defined, centring on network scans towards network telescopes. Such a focus is crucial as it excludes localised, targeted scanning methods, aligning more closely with the study objectives. Furthermore, the research deliberately steers clear of the complexities involved in pinpointing specific scanning tools such as Masscan and NMap, acknowledging the substantial resources required for such analysis.

The methodology unfolds systematically, starting with the acquisition of an extensive dataset from Noroff University and SANREN. This dataset forms the backbone of the study and facilitates a thorough examination of scanning activities. The authors' approach encompasses a detailed analysis of scanning patterns, frequencies, and their correlations with external security events. This multifaceted and nuanced methodology is designed to clearly delineate between benign and malicious scanning activities.

The goal of this research is to clearly distinguish benign from malicious network scanning and to carry out more research with critical insights, enhancing digital defence strategies. By methodically dissecting network scanning activities, the research aims to contribute significantly to the field of cyber security, providing tools and knowledge for stronger, more effective digital security measures in a constantly evolving technological landscape.

3.1 Definition

The research requires a strong and solid definition for this research. This definition of a network scan is based on the definition Durumeric et al. (2014) presented and Irwin (2013) presented in their research. This is the definition of a scan in this research, called the Durumeric method:

A internet scan is a occurrence where a source address has probed 40 destination addresses connected to the network telescopes, on the same port on a rate of more than 10 packets per second.

In terms of malicious data packets, the packets that appear in the network telescope sensor are unsolicited (Irwin, 2011) and should be considered potentially malicious, until further processing has been done.

3.2 Scope of research

The primary goal of the research is to look at the network data provided by network telescopes placed in South Africa. The goal is to classify the benign from the potential malicious network data.

To achieve this goal, the research needs to look at these points:

1. Find IPv4 network addresses that have scanned the network telescope sensors, as per the definition in Section 3.1.
2. Analyse Source IP, Port, Amount of packets, packet rate of data these sources has probed to the sensors
3. Investigate the source addresses, by fetching useful data as country of origin, tags, has the source address been caught by greynoise.io[1] sensors.
4. Create geographical heat maps using python to investigate origin and geographical behaviour changes of benign and malicious actors
5. Create case studies to compare the Durumeric method output to
6. Receive more network data from SANREN

Section 2.7 on network telescopes mentions that the telescopes are placed in 'sections of the IP address space that see minimal or no authentic traffic' (Irwin, 2011). Resulting in the data to the network telescope sensor being unsolicited. Therefore, the network telescope sensors are not susceptible for reconnaissance scans from a single entity. For this reason, research will focus on Internet wide scans.

The research does not focus on scan methods, as the ones mentioned in Section 2.4 considering that this requires more time to process the data.

[1] https://viz.greynoise.io

Although identifying which scanner being used in the scanning could be beneficial, research does not pursue this task using the methods mentioned in Section 2.9.1. Consequently, including this process in the analysis of the dataset is outside of the scope of the research. This is to ensure that research remains concentrated on its primary objectives, allowing for a more efficient and effective exploration of the primary research question. The research focuses on the Greynoise tags for identifying scanners in Section 5.8.1 and Section 5.8.2.

3.3 Research Methodology

This section details the methodology used in conducting research aimed at differentiating between benign and malicious scanning activities on IPv4 networks. The methodology covers data sourcing, identification of scanning characteristics, analysis of patterns, correlation with greynoise.io, and case studies.

3.3.1 Data Source Acquisition

The first step in the methodology involves the acquisition of a comprehensive dataset that captures network scanning activities. The primary dataset for this study was provided by Professor Barry Irwin from Noroff University, collected through the infrastructure hosted by the Computer Security Incident Response Team (CSIRT) for the South African Research and Education Network (SANREN). The data is part of a broader project on Internet Background Radiation (IBR) collection and analysis, with portions accumulated under the aegis of the Security and Networks Research Group at Rhodes University.

The data consists of 649 062 444 packets of data divided into 12 files, one for each month, packets per month can be seen in Table 3.1. For more details of the data, see Section 3.7.

Data Source administration

It is crucial to address the handling of destination IP addresses to maintain the privacy and security of the sensors. Destination addresses will be compiled using the Classless Inter-Domain Routing (CIDR) notation, specifically at the /24 level or higher. In instances where individual IP addresses will be used, the final octet of the address will be obscured with a placeholder to distinguish between multiple addresses within the same subnet. For example, ’10.10.10.a’ could represent one address, while ’10.10.10.b’ represents another within the same /24 subnet.

For destination IP ranges, it is important to avoid disclosing specific addresses due to security. Only a high-level indication of the /8 CIDR block in which they operate will be provided, such as ’10.x.x.x,’ ’10/8,’ or ’10.0.0.0/8.’ If disclosure of a specific destination address or range is necessary, the second and third octets will be masked to prevent the identification of the exact location. An example format for such reporting is be ’10.x.x.128,’

where the 'x' serves as a placeholder. This is for safeguarding the monitored IP ranges from unauthorised disclosure.

3.4 Characterisation of Scanning Activities

To accurately distinguish between benign and malicious scans, it is crucial to define the identity of each type. This involves a thorough examination of the collected data to determine typical patterns, frequencies, sources, and methods associated with both types of scanning activities.

The research will focus specifically on the network packets that might indicate scanning activity, which means following a modified version of (Durumeric et al., 2013) definition of a scan, mentioned in Section 2.10.

To check for a scan according to the definition in Section 3.1. The research has to first check if the source addresses have tried to make a connection to 40 destinations of the network sensors. Furthermore, check for the packet per second rate of 10 to further confirm that it is a scan.

The method for identifying Zmap, detailed in section 2.9.1, as proposed by (Ghiëtte, 2016) could be beneficial. ZMap is a free tool that comes with Kali Linux[2] and can be used with malicious intent. Greynoise.io provides great indicators for the scanners used, mentioned in Section 5.8.1 and Section 5.8.2.

3.4.1 Pattern Recognition

The key focus is on the rate of packet transmission. Consistent with the definition of the research (Section 3.1), scans with a transmission rate exceeding 10 packets per second are flagged for further analysis. In addition, the research investigates the geographical spread of the target addresses. Benign scans often show a more random targeting pattern, whereas malicious scans may exhibit a concentrated focus on specific IP ranges.

The pattern recognition process also leads to cross-referencing the packet data with known benign sources, such as Censys or Shodan. The work in (Bennett et al., 2021), talked about in section 2.10, about identifying Censys and Shodan scans, will be used in cross-referencing for benign patterns. Section 5.6.2 will mention some of the most well-known research organisations.

3.4.2 Frequency Analysis

In the frequency analysis segment of the research, attention is focused on analysing the distribution and regularity of scanning activities. This involves analysing the chronological patterns of network scans, particularly how often and at what intervals these scans occur.

[2] https://www.kali.org

By examining the frequency and duration of scan events, the research aims to distinguish between systematic, persistent scanning activities often indicative of malicious intent and irregular, sporadic scans that might suggest benign purposes.

The analysis also involves assessing the persistence of the scanning activities. Continuous scanning over long periods, especially if targeting a specific range of ports or IP addresses, is marked as a potential indicator of malicious scanning, such as reconnaissance by cyber attackers. However, short-lived and non-repetitive scanning activities might align more closely with benign actions, Censys or Shodan.

3.5 Correlation with threat intelligence

Using the capabilities of the GreyNoise.io[3] API, as detailed in Section 2.8.1, the script is designed to retrieve information on external events that are currently within the GreyNoise.io databases. This process involves a comparison with the outputs generated by the script using the Durumeric method and case studies. Specifically, the script will execute a procedure in which it takes the IP addresses resulting from the Durumeric method and case study analysis and queries them against the GreyNoise.io databases. The outcome of this query is a detailed classification for each IP address regarding Source IP, Country of origin , Classification (benign, malicious, not seen), First Seen, Last Seen, and Tags. Such classifications are allowing for effectively discerning and eliminating data that is irrelevant or harmless internet noise.

3.5.1 Tools and Techniques for Scanning Activity Analysis

The methodology is achieved using tools and techniques designed to efficiently process and analyse the large dataset of network packets. Central to this approach is the use of a Python script leveraging different libraries as DPKT, Socket, CSV, os and datetime to look for scanning.

The script employs the DPKT library for parsing PCAP files, extracting relevant information from each packet, as source ips, and looking for destination addresses, how many packets it has sent, and for how long. These data form the basis for pattern recognition and frequency analysis, as outlined in section 3.4.2 and section 3.4.1.

A subsequent step in the script combines these details into a final, comprehensive dataset in a CSV. This CSV has the headers 'Date of Scan, Source IP, Network, Port, Distinct Destinations, Total Packets, Rate'. This dataset is then ready for further analysis, including pattern recognition and frequency analysis discussed previously in Section 3.4.2 and Section 3.4.1.

For the 'Correlation with External Events' in Section 3.5, Greynoise.io visualiser[4] is used. It is a tool for analysing IP addresses in quantity and scans their internal database for

[3] https://.greynoise.io
[4] https://viz.greynoise.io

matches. If there are matches, it classifies it as malicious, benign, or unknown. The visualiser delivers a report of country of origin, tags for external events, and name of the organisation that has done the scanning. All of this is necessary data in the research and will help with identifying benign and malicious activities.

3.6 Identifying Scanning

First objective of the analysis is to find scanning attempts in the provided dataset. There are nearly 650 million (Table 3.1) packets to be analysed to find out what is scanning and what not. The Python script noted in section 3.5.1 that is created to find network scanning amongst the huge amount of packets is made possible with the dpkt[5] library.

Processing PCAP Files:

The Python script opens the PCAP file and iterates over each packet, parsing the Ethernet frame to access the IP layer. Filters out packets that are not IP packets, as network scans are typically conducted at the IP level. Extracts source and destination IP addresses and normalises the destination IP to a network format (e.g., 192.168.1.x/24) for easier aggregation. Checks if the packet's IP data contain TCP or UDP payloads, extracts the destination port, and organises the packet information into the packet_data dictionary, indexed by source IP, destination network, destination port, and then destination IP, recording timestamps of packets.

Analysing Packet Data for Network Scans:

After processing all packets in a PCAP file, the script analyses the packet_data to identify potential network scanning activities. It looks for instances where a single source IP has sent packets to a more than 40 unique destination IP addresses within a network segment on a specific port, meeting the MIN_DESTINATIONS criterion. Then it calculates the total number of packets sent and the duration of the scan to calculate the packet rate per second. If this rate exceeds the RATE_THRESHOLD, the activity is considered a network scan. The results are aggregated into scan_results, with each entry containing the source IP, network, port, number of distinct destination IPs, total packets, and rate of the scan.

In Section 4.11.5, some considerations are noted as the script has a RATE_THRESHOLD of 10 packets per second, which is a fairly aggressive rate for scanning.

For the RATE_THRESHOLD, consider the following:

- A high threshold (e.g., 100 packets per second) might only catch very aggressive scans.
- A moderate threshold (e.g., 10-50 packets per second) could be suitable for detecting scans that try to be somewhat stealthy but are still relatively fast.

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

- A low threshold (e.g., 0-5 packets per second) would help in identifying very slow, stealthy scans aimed at flying under the radar.

Outputting Results:

For each PCAP file analysed, a unique CSV file is created in the specified output directory. The filename is derived from the original PCAP file's name with an added suffix to indicate that it contains scan results. The CSV file includes headers for Source IP, Network, Port, Distinct Destinations, Total Packets, and Rate. Each detected network scan is written as a row in the CSV file. A for loop is used to unify the search process across all months in 2023, whereas each of the generated CSV files is incorporated into the combined CSV file for convenience.

The script can be found and downloaded in Appendix F

3.6.1 Flow chart

The illustration shows the flow of the Durumeric method that is the basis of Chapter 4.

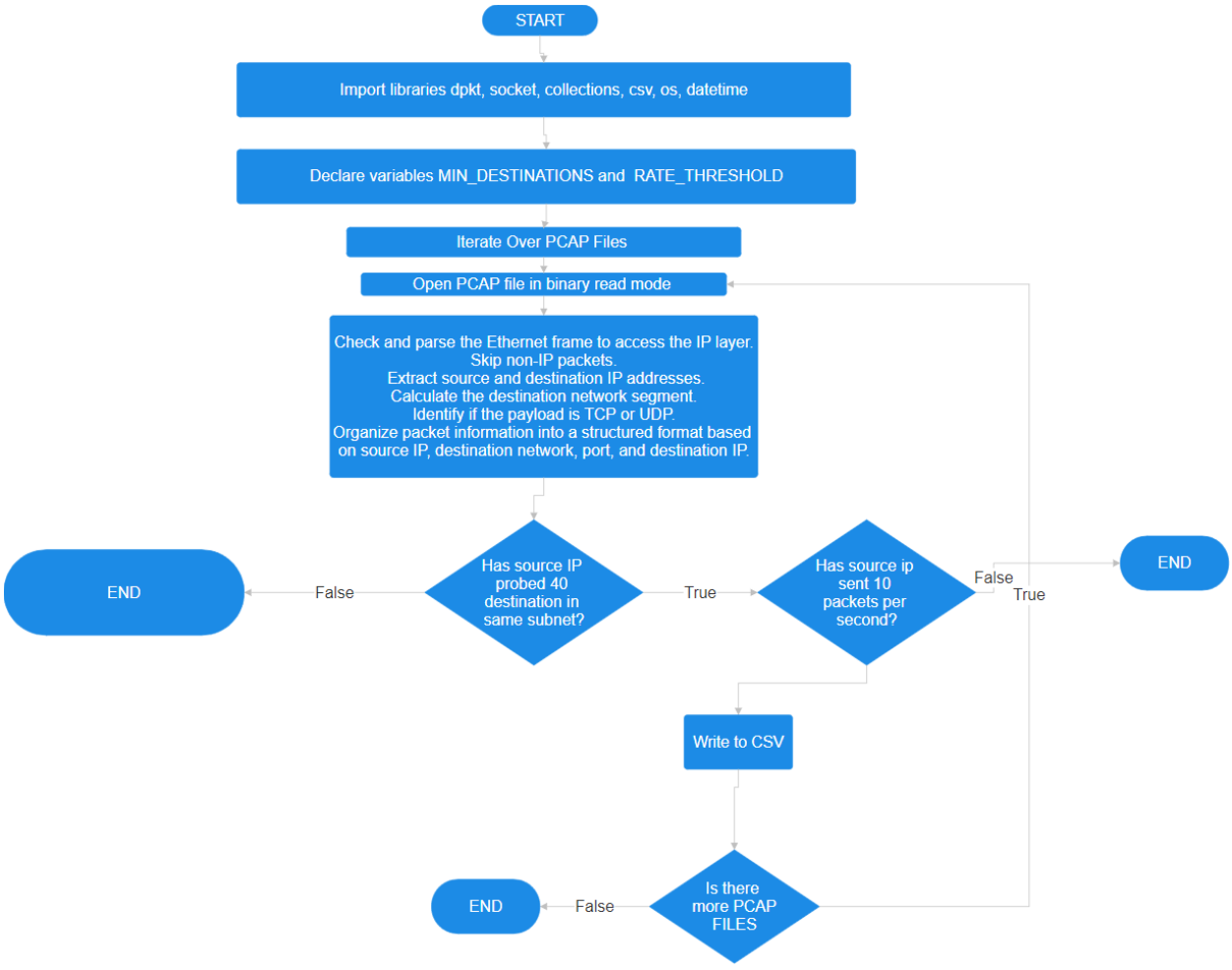

Figure 3.1: Flowchart of script

3.6.2 Case Study Script

Case Study script used in Chapter 5 enhances the detection mechanism by introducing additional parameters that account for the duration and intensity of scanning activities. These modifications can help identify not only the presence of network scans, but also their potential impact or threat level based on their characteristics.

The case Study Script considers the factors introduced in 3.6 but also includes the duration of the scan as a critical factor in its analysis, offering a more nuanced view of the scanning activities that could indicate different levels or intentions of threat. The case Study Script adds a layer of complexity and refinement to the scan detection logic in its analysis, aiming to provide a more detailed understanding of scanning behaviours by incorporating the duration of scans and adjusting the criteria for what constitutes a scan.

- MIN_DURATION: This parameter sets a minimum time frame for considering a sequence of packets as part of a network scan. By specifying a duration (in minutes), the script can differentiate between fleeting scanning activities and more sustained efforts. The inclusion of a minimum duration helps to identify scans that persist over a longer period, which might be more indicative of a methodical or systematic scanning approach by a benign actor, rather than a malicious one. The rationale here is that longer-duration scans could be more significant from a security perspective, possibly indicating a deliberate attempt to map out a network's structure or find vulnerabilities over time.

- MAX_RATE_THRESHOLD: While the the scripts mentioned i Section 3.6 sets a rate threshold at 0, essentially considering any rate of packet sending as potential scanning activity, the second script introduces a maximum rate threshold. This upper limit helps in identifying scans that fall within a specific rate range, excluding those that are too aggressive (which could be filtered out as noise or as a different type of attack). The idea is to focus on scans that are fast enough to be significant but not so fast that they're likely to trigger immediate alarms or be easily detected by rudimentary security measures. This nuanced approach allows for the identification of scans that are sophisticated enough to attempt to fly under the radar.

The case study script is tailored to identify scans that are not only widespread but also exhibit characteristics of being more calculated or stealthy. The adjustments made in the second script provide a richer context for each detected scan, offering insights into not just the breadth of the scan but also its persistence and intensity. The case study script is great for comparison between the script output of the methodology used in Section 3.6 and that of Section 3.6.2.

The script can be found and downloaded in Appendix F.

3.6.3 Data Visualisation

The data result of the research is valuable and should be used in visualisation. A visualisation is more straightforward for high-level analysis and easier for the public to comprehend. The CSV file is the result of the script in section 3.5.1 Tools and Techniques. This is imported to the greynoise script for querying and gives a country and tags as a response; these two are used in visualisation and figures.

Multiple visualisations will be made to strengthen the results:

1. **Graphical heat map Visualisation**

   - This is to show a graphical representation of where in the world the different malicious scans originate from

   Utilising the *geopandas* and *matplotlib* libraries, the research generated a static world map. This map visually represents the intensity of malicious scans across different countries, depicted through a gradient colour scheme. The intensity of the colour on the map correlates with the frequency of the scans, providing an intuitive and immediate understanding of the global distribution patterns of the scans.

2. **Graphs and Charts**

   - To give the reader useful information in illustrations.

   By using using *Google Sheets*, the research effortlessly creates graphs and charts, translating complex data into visually compelling representations. These visuals not only illuminate key findings, but also enhance understanding and communication of the research's insights with precision and clarity.

3.7 Data Overview

3.7.1 Packets received

To achieve the desired analysis on the provided dataset, it is necessary to understand the size and the sheer number of packets for each month in 2023.

Capinfos, a command line tool available in Linux, was used to retrieve the data presented in the table. By analysing packet capture files, capinfos provides essential information about packet capture file, including packet count, which were then used to populate the table's columns. This tool is particularly useful for fetching important data without having to spend several minutes running a script to count the number of packets.

Table 3.1 displays a complete breakdown of data relating to the number of packets sent or received during each month of the year 2023. This dataset is valuable for information on patterns and trends in data traffic over the course of the year.

Table 3.1: Number of packets for each month of 2023

Month	Number of packets	Percentage of total
January	44 855 656	6,91
February	48 194 871	7,43
March	63 323 583	9,76
April	53 412 852	8,23
May	53 048 977	8,17
June	49 365 513	7,61
July	65 254 169	10,05
August	67 926 541	10,47
September	57 651 847	8,88
October	61 420 726	9,46
November	48 468 301	7,47
December	36 139 408	5,57
Total	649 062 444	100,00

Looking at table 3.1, it is evident that the month with the highest number of packets is August, recording a significant 67,926,541 packets. This month is closely followed by July, with 65,254,169 packets, indicating a potential increase in network activity during the summer months, possibly due to vacations or seasonal events. In contrast, December has the lowest number of packets, totalling 36,139,408, which could be attributed to holiday periods and reduced business activity during that time.

In addition to the absolute numbers, the table also provides the percentage of each month's contribution to the total number of packets sent or received throughout the year. January, with 6.91% of the total, and February, with 7.43%, show lower percentages, possibly indicating a quieter start to the year. In contrast, August and July, with 10.47% and 10.05% respectively, represent the peak months of network activity, accounting for a significant portion of overall data traffic.

The graph in Figure 3.2 displays a bar graph that shows fluctuations in the number of packets received over a period of one year, from January to December of 2023. The y-axis represents the number of packets in millions, starting at 0 and increasing in increments, with the highest label showing 70,000,000. The x-axis represents the months of the year.

From January to March, there is a significant increase in the number of packets, reaching a peak in March. After March, there is a sharp decline in April, followed by a recovery with another peak in July, which is slightly higher than the March peak. The graph shows another higher peak in August, followed by a decrease in September, followed by a slight rise in October. After October, there is a marked decline, with the number of packets dropping significantly in November and December, ending the year at the lowest point on the graph.

In general, Table 3.1 and Figure 3.2 serve as a valuable resource for analysing data traffic patterns monthly in 2023, which can be useful to highlight which month could be more vulnerable to potential network scans.

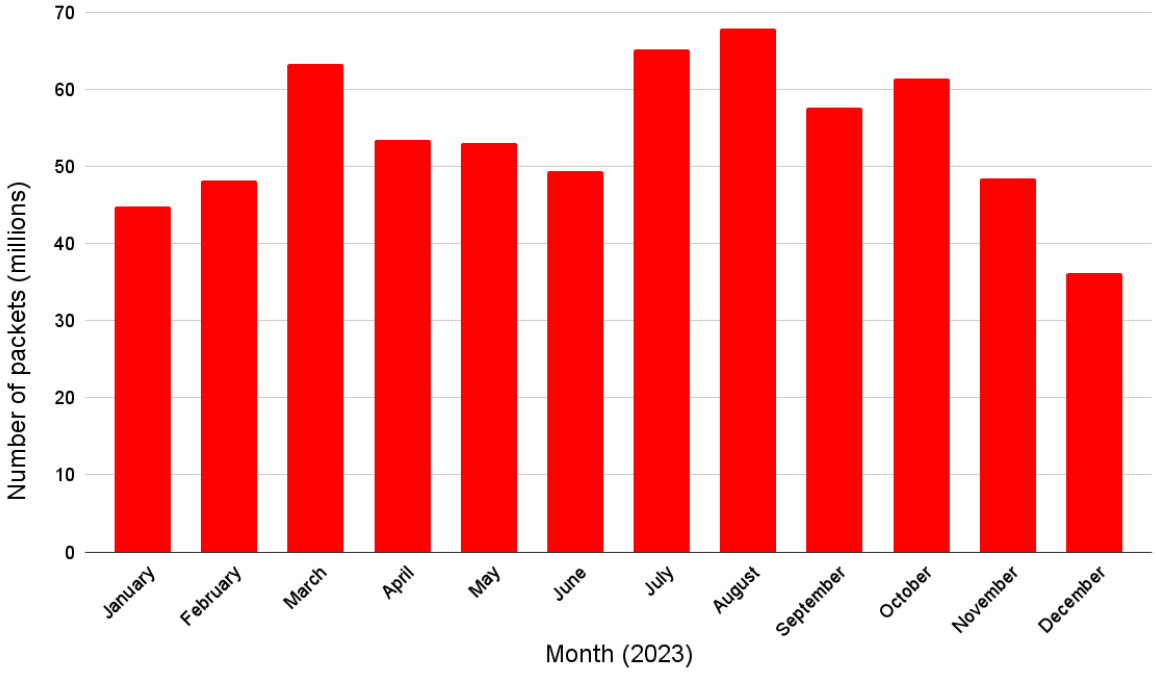

Figure 3.2: Visualisation of the number of packets for each month of 2023

3.7.2 Protocol Overview

This table presents a summary of network traffic data, focusing specifically on the distribution of packet counts and their respective percentages across three different network protocols: TCP, UDP, and ICMP. The table is structured with three columns: Protocol, Packet Count, and Percentage.

Table 3.2: Number of packets for top 3 protocols in 2023

Protocol	Packet Count	Percentage (%)
TCP	593 801 473	91.48
UDP	41 994 703	6.47
ICMP	12 972 768	1.99

Protocol: This column lists the network protocols observed in the dataset. The protocols shown are TCP (Transmission Control Protocol), UDP (User Datagram Protocol), and ICMP (Internet Control Message Protocol), which are fundamental protocols used in the Internet protocol suite for network communication.

Packet Count: This column shows the total number of packets observed for each protocol within the analysed dataset. There were 593,801,473 packets for TCP, 41,994,703 packets for UDP, and 12,972,768 packets for ICMP. These counts give a quantitative measure of how much data was transmitted using each protocol.

Percentage: This column represents the percentage of the total observed network traffic attributed to each protocol. The percentages show the relative volume of traffic for each protocol, indicating TCP's dominance in the dataset with 91.48% of the traffic, followed by

UDP with 6.47%, and ICMP with 1.99%.

3.7.3 Protocol Overview for each month of 2023

Appendix B provides 12 tables, each for one month of the year 2023. The tables contain values for protocol, packet count, and percentage. The protocol column lists the types of network protocols observed in the traffic data. The 'Packet Count' column indicates the total number of packets observed for each protocol during the month of 2023. Lastly, the percentage column shows the percentage of packets for each protocol relative to the total number of packets observed across all protocols during the month.

3.7.4 Port Overview

A full monthly overview is found in Appendix B

The **Port** column lists the network ports that have been observed in the traffic data. Network ports are communication endpoints and are integral to the process of directing data over the Internet and within local networks. Each port number identifies a specific application or service.

The **Traffic** row shows a percentage of total traffic attributed to each port. This value gives an idea of how much data is being directed to or from each port, relative to the total amount of traffic observed in the network analysis period. Highlights which ports are most active or highly used.

Unique Sources column shows the percentage of unique source addresses that have communicated with each port, relative to the total number of unique sources observed. A unique source is counted once, regardless of how many packets it has sent. This metric is useful for understanding how many distinct entities are interacting with each port.

A value of '0.00%' in table B.3 and table B.8 suggests that the number of unique sources, as a percentage of the total number of unique sources observed across all ports, is less than 0.00. This means that the percentage is so small that when rounded to two decimal places, it appears as '0.00%'.

3.8 Network telescope

The network telescope runs on a /24 network. The network telescope configuration results in 254 usable host addresses that can receive network packets. These network addresses are exposed and potentially vulnerable to network scan, including malicious scanning activities.

3.9 Summary

The methodology is designed to systematically find and differentiate between benign and malicious network scanning activities. By combining data analysis, pattern recognition, correlation with threat intelligence, and case studies, the research aims to provide a clear distinction between benign and malicious scanning intentions and a contribution towards research in network scanning.

CHAPTER 4

Analysis

4.1 Introduction

This chapter delves into the critical analysis of differentiating between benign and malicious scanning in IPv4 networks. The focal point of this research depends on identifying specific patterns, behaviours, and characteristics that distinctly categorise scanning activities. This analysis is essential, considering the escalating network security threats and the importance of maintaining robust cyber security measures.

The research question at the heart of this analysis is: '**How can we accurately differentiate between benign and malicious scanning in IPv4 networks based on specific patterns, behaviours, and characteristics?**'. The research also hopes to find answers to the secondary question: '**Can Threat Intelligence services be utilised to accurately distinguish benign scanning from malicious scanning.**'.These two questions underpins the entire research, guiding the analysis towards a nuanced understanding of network scanning activities.

4.1.1 Structure

(Section 4.4): Protocol Distribution - An examination of the distribution of network protocols observed during scanning activities, shedding light on the most frequently targeted protocols.

(Section 4.5): Geographical-location of Scans - This section explores the geographical origins of the scanning activity, mapping out the locations of scanners and potential hotspots.

(Section 4.6): Source IP Address - A detailed analysis of the source IP addresses involved in scanning, including their frequency, diversity, and potential patterns.

(Section 4.7): Port Analysis - Focusing on the ports targeted during scanning, this section provides insights into the most common ports scanned and their potential significance.

(Section 4.8): Autonomous System (AS) Analysis - Examining the Autonomous Systems associated with scanning activity to identify any trends or patterns at the network infrastructure level.

(Section 4.9): Destination Analysis - An analysis of the destinations or targets of the scanning activity, including their distribution and potential vulnerabilities.

(Section 4.10): Behavioural Analysis - This section investigates the behaviour of scanners, including scanning patterns, timing, and any anomalous behaviour that may be indicative of malicious or benign intent.

(Section 4.11): Correlation with Threat Intelligence - Exploring correlations between observed scanning activity and known threat intelligence sources to assess the potential risk and threat level.

(Section 4.12): Greynoise Tags - Discussion on the utilisation of Greynoise tags and data for enhancing the understanding of scanning activity and its context.

(Section 4.13): Summary - Summary of chapter with key findings.

4.2 Purpose of the Chapter

The purpose of this chapter is to systematically analyse the collected data set, applying the preceding methods in section 3.4 to uncover distinct patterns and behaviours. This analysis is instrumental not only in answering the research question, but also in contributing to the broader field of network security, particularly in understanding and identifying potential cyber security threats in IPv4 networks.

4.3 Overview

The comprehensive analysis conducted on network traffic data spanning the entire year of 2023 has developed a detailed report encapsulated within a combined CSV file. This document reveals a significant volume of scanning activity, totalling 23,415 attempts, directed towards the network segment 155.x.x.x/24. This activity has been distributed across 256 destination hosts within this subnet, averaging approximately 91 identified scans per host. This number not only underscores the persistent nature of scan attempts across the digital landscape but also serves as a critical indicator of potential security vulnerabilities and threats targeting this specific network range.

Table 4.1: Protocol distribution for attempted scans

Protocol	%
TCP	65.554
UDP	34.446

4.4 Protocol Distribution

Table 4.1 provides a concise overview of how different protocols were targeted in a set of attempted network scans, measured as percentages of the total attempts. Specifically, it depicts the distribution between the Transmission Control Protocol (TCP) and the User Datagram Protocol (UDP), two of the core protocols of the Internet protocol suite used for Internet scanning. According to the data presented, TCP scans constitute the majority, accounting for 65.554% of the attempts. This dominance of TCP could be attributed to its reliable, connection-orientated nature, which is often exploited for more sophisticated reconnaissance activities in cyber security threats.

However, UDP scans, making up 34.446% of the attempts, represent a significant portion of the scan attempts as well. UDP's connection-less design makes it a target for different types of scans, especially those aiming to discover services that respond predictably to unestablished requests, such as DNS servers or SNMP enabled devices. The distribution indicates that attackers or security organisations are using both protocols to explore network vulnerabilities, each chosen for its specific characteristics and the type of information or access it could generate. Understanding the geographical location of these scans could provide further insights into the origin of the attacks and potentially the targets most at risk, highlighting the importance of regional security measures and response strategies.

4.5 Geographical sources of scans

Table 4.2: Country of origin for attempted scans

	Country	Number of scans	%
1	United States (US)	6228	26.55
2	China (CN)	4687	19.98
3	Seychelles (SC)	1376	5.87
4	Netherlands (NL)	1319	5.62
5	Germany (DE)	1291	5.50
6	Hong Kong (HK)	1031	4.40
7	Canada (CA)	785	3.35
8	Russia (RU)	641	2.73
9	Belize (BZ)	517	2.20
10	Vietnam (VN)	463	1.97
	Σ_{10}	18338	78.17

Table 4.2 provides a detailed view of the global distribution of the source country behind these scans. Using data enriched through whois.cymru, a tool known for its effectiveness in finding IP addresses and network information, the table reveals that the highest number

of attempts originates from the United States (US), with a total of 26.55% of scans. This is closely followed by China (CN), which contributes 19.98% of the total. Such data is crucial to understanding the geographical distribution of potential cyber security threats and highlights the global nature of network scanning activities.

Significantly, the table also shows a diverse range of countries involved in the scanning activities, including smaller nations (in size and IP allocation), such as the Seychelles (SC), Belize (BZ), and Vietnam (VN). This diversity indicates that attempts to scan networks are not confined to any single region or limited to the most technologically advanced countries. Instead, it suggests a widespread interest or activity in exploring vulnerabilities in the digital landscape. The inclusion of countries such as the Netherlands (NL), Germany (DE), Hong Kong (HK), Canada (CA), Russia (RU), and others underscores the global spread of entities engaging in scanning activities, whether for benign research or potentially malicious intentions.

The use of whois.cymru[1] to enrich IP addresses in the analysis process emphasises the importance of reliable data sources in cyber security. This approach improves the accuracy of identifying the origin of network scans.

This data not only aids in immediate security posture improvements, but also contributes to a broader cyber security strategy and policy development, considering the international aspects of digital threats. Belize, Vietnam, and the like are common VPN hosting points (M. T. Khan et al., 2018), which is represented in this analysis. Therefore, one cannot know if the source is originating from these countries or not.

4.5.1 Heat map of scans

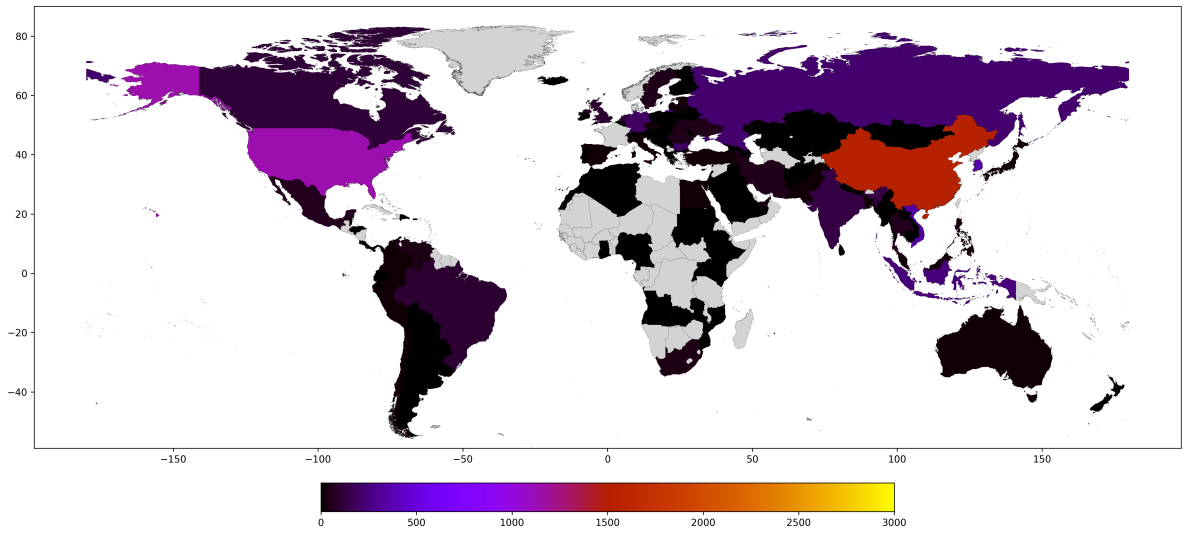

Figure 4.1: Geographical heat map of number of scans by country

[1] https://www.team-cymru.com/ip-asn-mapping

Figure 4.1 is a graphical illustration of total uncategorised port scanning towards the research network telescope, and is helpful when reading table 4.2. Countries with the darkest red hue are the largest sources of internet scanning, which could suggest a higher level of activity, whether it be for benign purposes like research or potentially for more malicious intentions. The darker shade of purple indicates moderate activity, while black and grey represent no significant activity detected from those regions. This map could be useful for understanding global patterns in internet scanning, as it depicts the origins of network scanning. The key take-away from Figure 4.1 is that scanning is widespread and needs to be categorised to get a grasp of where to look for incoming data.

Figures 4.2 and 4.3 are categorised geographical heat maps for the output of the Durumeric method. These figures are used in more comparable analyses in Section 5.8.

For benign actors a small purple hue is spotted in the Netherlands, and no other countries have reportedly scanned the network telescope for a benign cause.

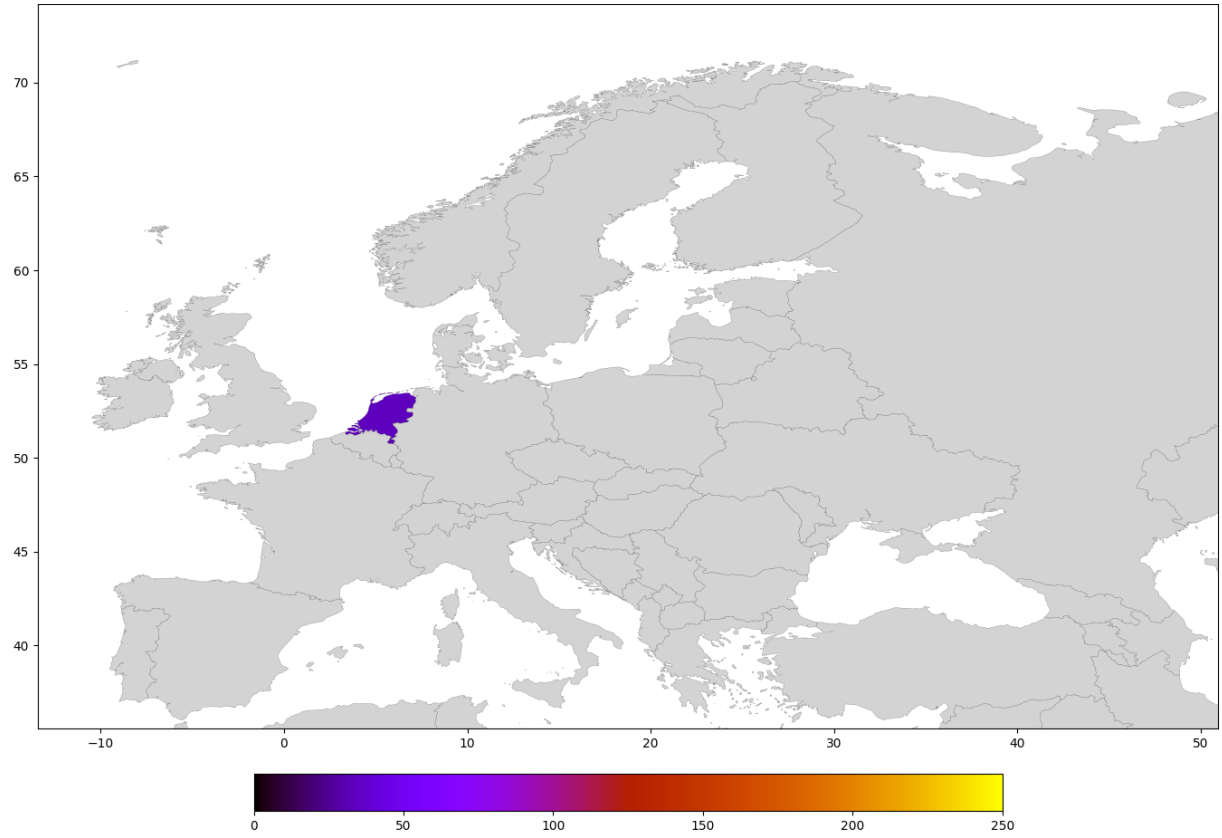

Figure 4.2: Geographical origin of benign scans from Durumeric method scans

In the malicious heatmap, China is seen as the most ”popular”. Indonesia is the second country to have been scanned for malicious purpose, closely followed by Russia and India.

When comparing the findings in Bruce et al. (2024) with Figure 4.3, there is a clear resemblance between the two. Bruce et al. (2024) ranks Russia, Ukraine and China in top 3, where as Figure 4.3 has China as a clear number one, Russia and Ukraine in the top 5.

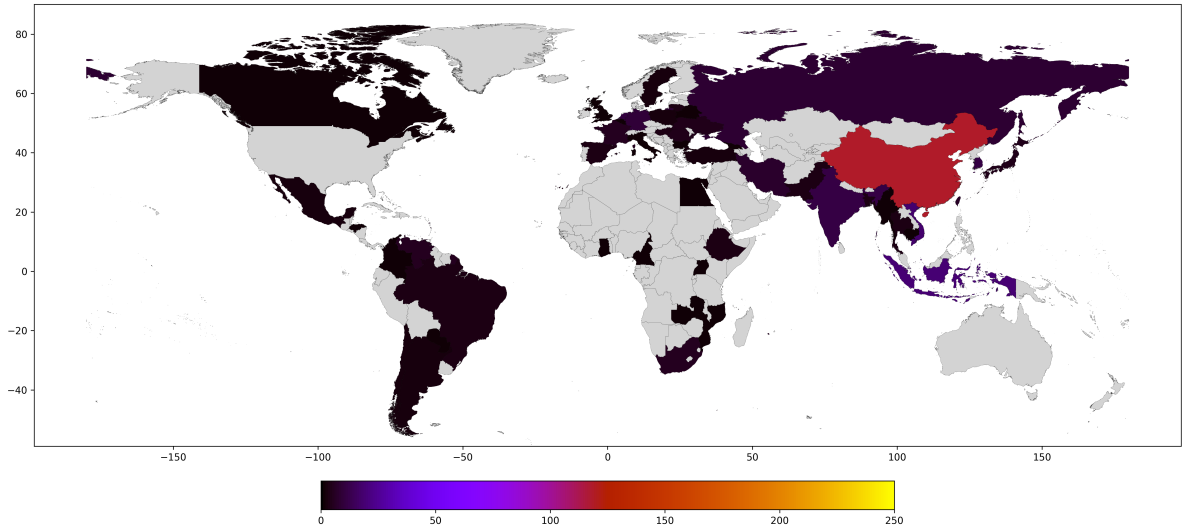

Figure 4.3: Origin of malicious scans from the Durumeric method scan output

The difference between the two is that the US is ranked number 4, where as in Figure 4.3 there is no counted malicious traffic from the US.

4.6 Source IP address

The detailed analysis of the network scanning activity presented in the tables provides a multifaceted view of the current scanning sources. The first table highlights the most active source IP addresses in terms of scanning attempts, offering a clear picture of where potential security threats are emanating. With IP addresses such as 151.106.35.1 and 168.80.174.2 leading in scanning attempts, there is an evident concentration of activity that needs deeper investigation. These addresses contribute significantly to the overall volume of network scans, with percentages indicating a substantial portion of total scanning activities. This suggests that certain nodes on the network are compromised or are specifically dedicated to probing and potentially exploiting network vulnerabilities.

Other addresses to keep an eye on are 118.123.105.85, 118.123.105.86 and 118.123.105.90 all come from the same netblock: 118.123.105.0/24. Combining the scanning attempts of all these three addresses would make the total of scans for 118.123.105.0/24 1081, which combines it is the netblock which has maliciously scanned the most. More analysis on netblocks follows in Section 4.6.1

The second table provides a layer of context to the raw numbers of scanning attempts, detailing the Autonomous System Numbers (ASN), Netblocks, countries, registering organisations, and dates of allocation associated with each IP address. This information is crucial for understanding the infrastructure from which the scans originate and can help in attributing the scans to particular networks or entities. For example, seeing a cluster

Table 4.3: Top 10 source IP addresses by scanning attempts and percentage of total

	Source IP	Scanning attempts	% of total
1	151.106.35.1	799	3.60
2	168.80.174.2	668	3.01
3	194.63.142.110	414	1.87
4	118.123.105.85	393	1.77
5	118.123.105.86	365	1.64
6	205.205.150.26	341	1.54
7	118.123.105.90	323	1.45
8	23.95.110.140	292	1.31
9	185.53.90.169	285	1.28
10	147.78.47.189	272	1.22
	Σ_{10}	4152	18.69

of IP addresses associated with a specific country or ASN can indicate a targeted campaign originating from that region or network. Additionally, understanding the registering organisation and allocation dates helps in tracing the administrative history of these IPs, providing clues about their intended use and the nature of their operators.

Table 4.4: Enriched information for Table 4.3

	ASN	IP-Address	Netblock	Country	Registry	Allocated
1	34088	151.106.35.1	151.106.32.0/20	DE	ripencc	1991-05-30
2	24567	168.80.174.2	168.80.174.0/24	SC	afrinic	1994-02-15
3	50113	194.63.142.110	194.63.142.0/24	SC	ripencc	2009-11-17
4	38283	118.123.105.85	118.123.105.0/24	CN	apnic	2007-09-12
5	38283	118.123.105.86	118.123.105.0/24	CN	apnic	2007-09-12
6	701	205.205.150.26	205.205.0.0/16	US	arin	1995-04-08
7	38283	118.123.105.90	118.123.105.0/24	CN	apnic	2007-09-12
8	36352	23.95.110.140	23.95.108.0/22	CA	arin	2013-08-16
9	215845	185.53.90.169	185.53.90.0/24	BZ	ripencc	2014-04-08
10	209588	147.78.47.189	147.78.47.0/24	LB	ripencc	2019-01-17

The historical data provided by the allocation dates reveal trends over time in the utilisation of IP addresses for scanning activities. Older allocations, such as the one from 1991 associated with the IP address 151.106.35.1, or 1995 and 1995 with 168.80.174.2 and 205.205.150.26, indicate that some networks have been operational for decades and may carry legacy vulnerabilities or policies that facilitate scanning activities. Conversely, more recent allocations may suggest the acquisition of IP blocks for the express purpose of conducting scans or could indicate of a malicious scanning intention.

4.6.1 Netblocks

The provided table offers a detailed look at the landscape of network scanning activities by highlighting the top ten netblocks based on the number of scanning attempts, alongside their respective percentages of the total scans. This breakdown is critical for identifying specific network segments that are particularly active in scanning, which can be indicative of either compromised systems within these blocks or organisations that are conducting

extensive network reconnaissance. Netblocks are gathered from whois.cymru[2]. The value labelled 'NA' in Tables 4.5 and 4.6, are values set to unidentified or unallocated ranges.

Table 4.5: Top 10 netblocks in order of scanning attempts from Durumeric method, with the percentage as of total scans

	Netblock	ASN	Country	Scanning attempts	% of total scans
1	104.152.52.0/24	14987	US	1457	6.21
2	118.123.105.0/24	38283	CN	1277	5.44
3	NA	NA	NA	1010	4.31
4	168.80.174.0/24	24567	SC	904	3.85
5	151.106.32.0/20	34088	DE	880	3.75
6	94.102.61.0/24	202425	NL	618	2.63
7	185.53.90.0/24	215845	BZ	515	2.20
8	185.224.128.0/24	49870	NL	434	1.85
9	194.63.142.0/24	50113	SC	414	1.76
10	60.176.0.0/12	4134	CN	352	1.50
	Σ_{10}			7861	33.5

The leading netblock, 104.152.52.0/24, stands out with a significant 6. 21% of total scanning attempts, suggesting a concentrated source of scanning activity within this range. Similarly, the 118.123.105.0/24 netblock follows closely, accounting for 5.44% of total scans, which underscores the presence of another major player in the scanning ecosystem. Interestingly, the value labelled 'NA', which contributes to 4.31% of the scans, suggests a large number of unidentified or unallocated ranges, highlighting the challenges of attributing all scanning activities to specific sources.

Table 4.6: Enriched information about the top 10 netblocks in Figure 4.5

	AS Name	Netblock	ASN	Country
1	RETHEMHOSTING, US	104.152.52.0/24	14987	US
2	CHINANET-SCIDC-AS-AP CHINANET,CN	118.123.105.0/24	38283	CN
3	NA	NA	NA	NA
4	QTINC-AS-AP QT Inc., JP	168.80.174.0/24	24567	SC
5	GDY-FRANCE, DE	151.106.32.0/20	34088	DE
6	IP Volume inc, SC	94.102.61.0/24	202425	NL
7	TECHOSERVERS, GB	185.53.90.0/24	215845	BZ
8	AS49870-BV, NL	185.224.128.0/24	49870	NL
9	SUPERSERVERSDATACENTER, CZ	194.63.142.0/24	50113	SC
10	CHINANET-BACKBONE, CN	60.176.0.0/12	4134	CN

The 2 China netblocks in Table 4.5 contribute to a total of 6.94% of total scans, indicating a large network operation or hosting services in these countries. The netblock sizes are mostly /24, which corresponds to 256 IP addresses per netblock. However, two entries (151.106.32.0/20 from Germany and 60.176.0.0/12 from China) cover a larger range, indicating a broader allocation of IP addresses, typical for larger hosting environments or ISPs. The China netblock, ranked 10th, covers an address range with 1,048,574 unique addresses.

There are several different continents represented in the table, the entries from the United

[2] https://www.team-cymru.com/ip-asn-mapping

States and China are expected, given these countries' significant online presence and infrastructure. However, the presence of netblocks registered in the Seychelles and Belize could reflect specific business practices, such as offshore hosting, for regulatory, privacy or economic reasons. Another explanation questions around data sovereignty, security, and compliance with local and international laws in these countries, making them a hotspot for cyber crime.

DE (Germany) and NL (Netherlands) are typical locations for European data centres and hosting services, known for their infrastructure and connectivity (Data Center Map, 2024).

The AS NAMES provide insights into the types of services these networks offer, but could also be forged to mislead. For example, 'RETHEMHOSTING' suggests a hosting service, while running it through a threat intelligence shows something different (more in Section 4.11), while 'CHINANET-BACKBONE' indicates a major infrastructure provider in China.

Some AS names like 'QTINC-AS-AP QT Inc., JP' and 'TECHOSERVERS, GB' might reflect specific tech companies or hosting services, though the country codes in AS names (like JP and GB) might not always match the country listed. This could be a case where 'QTINC-AS-AP QT Inc., JP' or 'TECHOSERVERS, GB' is located in multiple countries and AS name is a country tag for location or department. On the other hand, it could be a case where the AS is setting up a fake AS name to be viewed as a legitimate organisation. As 'TECHOSERVERS, GB' which would look like a Great Britain organisation which are set up in Germany, actually are set up in Belize.

4.7 Port Analysis

Ports are the gateway to different applications and software on the internet, and therefore it is important to analyse when looking at the scans. One can analyse the ports tried and maybe try to match ports with already existing vulnerabilities that are using that specific port. In this way, a network administrator can be observant on specific connections on specific ports.

Table 4.7: Top 10 scanned ports using Durumeric method

	Port	Number of scans	%
1	445/tcp	1017	4.59
2	5555/tcp	936	4.22
3	22/tcp	772	3.48
4	1433/tcp	769	3.47
5	5060/udp	581	2.62
6	139/tcp	471	2.12
7	5432/tcp	452	2.04
8	80/tcp	331	1.49
9	443/tcp	291	1.31
10	123/udp	277	1.25
	Σ_{10}	10007	26.59

Table 4.7 reflects a broad spectrum of cyber interests, from traditional web services (ports 80/tcp and 443/tcp) to specific database (ports 1433/tcp and 5432/tcp), device (port 5555/tcp) and communication services (port 5060/tcp). The high incidence of scans on ports associated with known vulnerabilities and essential services highlights the need for robust security measures. It also underscores the importance of monitoring and securing less common ports (like 5555/tcp or 5060/tcp) which might be overlooked but are clearly targeted by malicious actors.

Other interesting ports are docker ports, which are port that are opened to use applications like web server or database services. Table 4.8 illustrates popular docker ports with the number of scans and percentage.

Table 4.8: Docker ports, number of scans and percentage

	Port	Number of scans	%
7	5432/tcp	452	2.04
8	80/tcp	331	1.49
9	443/tcp	291	1.31
12	8080/tcp	190	0.81
15	3306/tcp	133	0.57
	Σ_N	1397	6.22

4.7.1 Vulnerabilities linked to top 10 scanned ports

Each of the ports listed in the table represents common services that may have associated vulnerabilities. In the following, a general overview of vulnerabilities typically associated with these ports is discussed. The actual vulnerabilities depend on the specific software version and configuration running on these ports, and there are many cases where the scan is without a payload, so the intention of the scan remains unclear.

445/tcp SMB - Server Message Block: Vulnerabilities in SMB can include remote code execution, denial of service, and unauthorised access. Notable vulnerabilities include those exploited by WannaCry ransomware (EternalBlue) and NotPetya. NotPetya spreads using SMB, same as the EternalBlue exploit (Ongun et al., 2021).

5555/tcp Multiple use port: The Fortinet FortiNAC is susceptible to a security breach that could allow an external hacker to infiltrate the system without authorisation. This risk comes from a vulnerability to command injection. An attacker, by sending a specifically designed request to the service running on 5555/tcp, could use this flaw to replicate files from one area of the device to another local area within the same device (NVD - CVE, 2023). Android also uses this port (Costantino & Matteucci, 2019).

Backdoor.Win32.FTP.Ics is involved in an Unauthenticated Remote Command Execution issue - the malware operates by monitoring TCP port 5555. External attackers with access to the system can execute commands through the backdoor, thereby gaining control over the compromised host (Malvuln, 2022).

22/tcp SSH - Secure Shell: Vulnerabilities may include brute-force attacks on weak passwords, exploitation of outdated SSH versions, or unauthorised access through misconfigured permissions (Raikar & Meena, 2021).

1433/tcp Microsoft SQL Server: Common vulnerabilities include SQL injection, unauthorised access with default credentials, or exploiting misconfigurations to execute remote commands (Hassan Kilavo & Dudu, 2023).

5060/udp SIP - Session Initiation Protocol: Vulnerabilities might involve SIP flooding (DoS), eavesdropping on VoIP calls, or unauthorised access to VoIP services (H. M. A. Khan et al., 2021).

139/tcp NetBIOS: This port can be exploited for information disclosure, such as revealing user names or browsing shared folders, and for man-in-the-middle attacks (Jayasuryapal et al., 2021).

5432/tcp PostgreSQL Database: Similarly to other database services, vulnerabilities include SQL injection, unauthorised database access, or remote code execution (Susanto et al., 2020).

80/tcp HTTP - Hypertext Transfer Protocol: Being the standard web service port, vulnerabilities include various web application attacks such as cross-site scripting (XSS), SQL injection, and insecure server configurations (Tanakas et al., 2021).

443/tcp HTTPS - HTTP Secure: Although HTTPS is more secure than HTTP, vulnerabilities may still include SSL/TLS vulnerabilities, misconfigured certificates, or man-in-the-middle attacks exploiting weak cipher suites (Jagamogan et al., 2022).

123/udp NTP - Network Time Protocol: Vulnerabilities include the exploitation of NTP servers to carry out reflection/amplification DDoS attacks or exploiting outdated NTP versions for unauthorised system access (Putu et al., 2020) and (Rudman & Irwin, 2015).

It is important to note that simply having these ports open does not inherently mean a system is vulnerable; rather, it is the combination of open ports with misconfigurations, outdated software, or unpatched vulnerabilities that poses a security risk. Regular security assessments, the application of security patches, and the following best practices in configuration and authentication are crucial steps to mitigating these vulnerabilities.

4.7.2 Mirai Ports

Mirai is a widespread botnet which is looking for open telnet servers and starts by scanning specific open ports, port 23 and port 2323. The port usage ratio is 1/10 (Snehi & Bhandari, 2021). Therefore, a port scan on 2323 will happen for every 10th scan on port 23.

The source addresses in Table 4.9 have only scanned these two ports, indicating a targeted MIRAI scan, instead of a widespread scan targeting multiple ports. Table 4.9 presents a concise list of source IP addresses that have been identified as exclusively

Table 4.9: Source Addresses that have only tried to connect to port 23/tcp or 2323/tcp

Source address	AS	Country
45.61.184.12	53667	US
217.146.82.142	25369	GB
217.146.82.141	25369	GB
180.189.95.16	9770	KR
42.57.188.229	4837	CN
113.53.133.86	23969	TH

attempting connections to port 23 or 2323. There is no other appearance of these source addresses within the total of 23415 scans and is therefore in the scope for a Mirai scan.

AS53667 The first source address in table 4.9, 45.61.184.12 (AS53667) is registered to *PONYNET-15* in United States. PONYNET-15 is owned by FranTech Solutions. PONYNET-15 positions itself as a ”bulletproof” hosting provider, thus claiming resilience against complaints regarding illicit activities on their platforms. This claim is due to their operations in countries with inadequate cyber legislation or limited resources dedicated to combating cybercrime. Given PONYNET's claims as a bulletproof host, it has been identified as a known malware distributor (HP Wolf Security, 2019).

Greynoise has not seen that 45.61.184.12 sends packets to any of their sensors. To verify the lack of findings from GreyNoise, an IP address abuse reporting service is used. Abuseipdb [3] receives reports from private persons and organisations if abuse of a IP-address has occurred. Abuseipdb has 189 reports on this source address, regarding port scanning on port 23/tcp, but has 0% confidence that it is being abused. Combining HP Wolf Security (2019) findings, PONYNET's bulletproof hosting and abuseipdb reports, 45.61.184.12 can presumably be linked to Mirai botnet.

AS25369 The interesting part of Table 4.9 is that there are two source addresses that belong to the same netblock, 217.146.82.142 (AS25369) and 217.146.82.141(AS25369). According to a whois service[4] the source addresses belong to 'BANDWIDTH-AS Hydra Communications Ltd'.

AS9770 180.189.95.16 (AS9770) originates from South Korea and is registered by *LG HelloVision Corp.*. LG HelloVision Corp is a large tech company from South Korea, and is a likely victim of the Mirai botnet, due to its activity. It varies between scanning port 23/tcp and port 37215/tcp, the latter being a port to achieve remote code execution on Huawei devices (NVD - CVE, 2017). The scanning of port 23 might be to spread Mirai to other devices. When checking 180.189.95.16 abuseipdb reports of being 88% confident of abuse, of 135 reports. Most of the reports are from destinations that received port scanning on port 23, and some even reported it trying a different username on open telnet ports. The reports from both greynoise and abuseipdb strongly suggest that one of the connected devices from *LG HelloVision Corp* has

[3] https://www.abuseipdb.com/
[4] https://whois.domaintools.com/

been infected with Mirai.

AS4837 42.57.188.229 (AS4837) registered to *CHINA UNICOM China169 Backbone* and originates from China. The classifications are unknown, as greynoise has not received any malicious nor benign activity from this source address. The search in abuseipdb returned 0% confident of being abused, but has 7 reports of it scanning port 23/tcp. There is no linking to Mirai, other than it only scans port 23.

4.8 Autonomous System (AS) Analysis

This section focus on the appearance and roles of Autonomous Systems (AS) in the context of Durumeric method scanning. This section assesses how AS numbers are represented in network scan data, identifying the frequency and distribution of ASes across various scan events.

Table 4.10: Top 10 AS numbers in number of scans with percentage of total scans

AS Number	Number of scans	%
AS14987	1457	6.21
AS4134	1363	5.81
AS38283	1279	5.45
AS14061	1104	4.71
NA	1010	4.31
AS24567	904	3.85
AS34088	880	3.75
AS202425	780	3.33
AS4837	700	2.98
AS14618	628	2.68
Σ_{10}	10105	43.08

Table 4.10 and figure 4.4, provide a analysis of the top 10 Autonomous System (AS) numbers based on their appearances in the main dataset obtained from network scanning against the network telescope. The pie chart visually represents the distribution of these AS numbers, highlighting their proportional contribution to the dataset, while the accompanying table quantifies their specific occurrences and percentages.

The bar graph in Figure 4.4 illustrates that of the top 10 the AS number 14987 is the most prevalent, accounting for 6. 21% of the dataset, closely followed by AS4134 with 5.81% and AS38283 with 5.45%. Other notable AS numbers include AS14061 (4.71%), AS24567 (3.8%), and AS34088 (3.7%). The unallocated AS number labelled “NA“ encompasses 4.31% of the dataset, indicating a diverse array of unallocated AS numbers in the top 10.

Table 4.10 complements the bar graph by providing exact figures: AS14987 appears 1457 times, AS4134 is present 1363 times, and AS38283 appears 1279 times, among others. These precise data underscore the dominance of the top 10 AS numbers, which together (Σ_{10}) account for 43.08% of the occurrences in the dataset, while the total count of appearances is 10105.

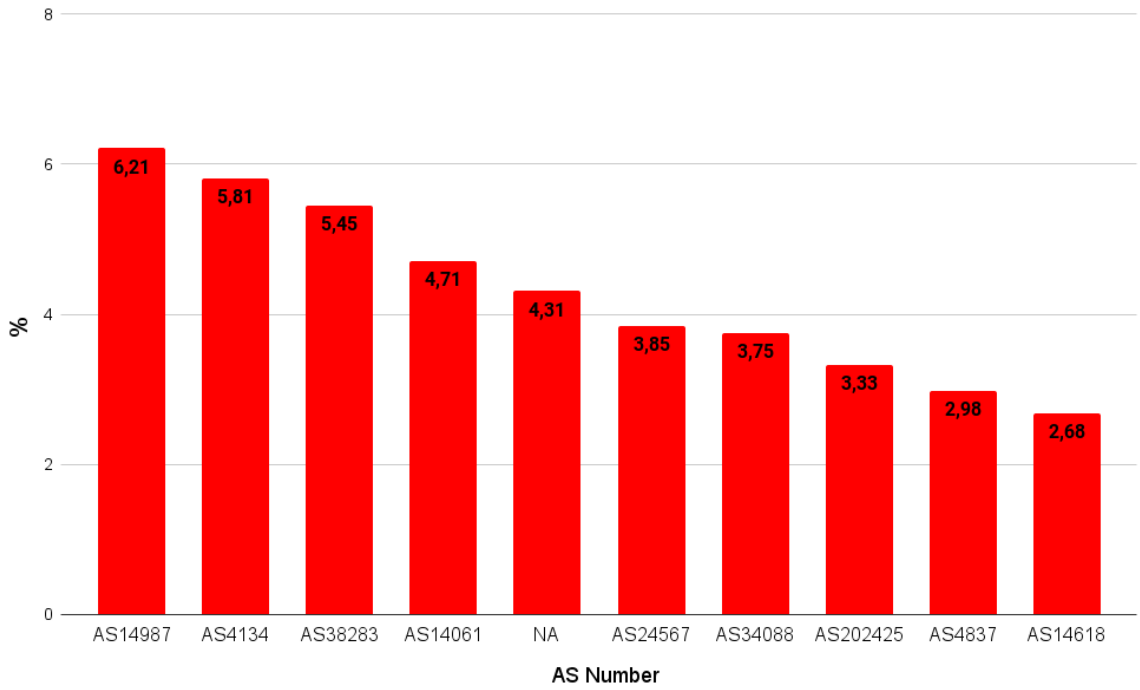

Figure 4.4: Bar chart to compliment table 4.10

4.8.1 AS numbers with the highest scanning activity

AS4134 China Telecom Backbone: This AS is associated with China Telecom, one of the largest providers of internet services in China. It is part of a backbone network that spans various provinces, including Zhejiang, Henan, Hainan, Ningxia, Tianjin, and many others, facilitating extensive internet connectivity across the region.

AS38283 CHINANET SiChuan Telecom Internet Data Center: Operated by China Telecom as well, this AS specifically serves the Sichuan province in China. It plays a crucial role in connecting China's operators to international IPv4 operators, highlighting its importance in the global landscape of the Internet.

AS4837 CHINA UNICOM China169 Backbone: This AS is part of China Unicom, another major telecommunications provider in China. It encompasses a wide range of networks in different regions, including Shanghai, Guangzhou, Shenzhen, and Guangdong province, among others. The AS connects with numerous other networks globally, underscoring its extensive reach and connectivity.

AS202425 IP Volume Inc: This AS is known for its extensive list of peering connections with numerous other ASNs around the world. It imports routes from a wide range of ASNs, demonstrating its interconnected nature and pivotal role in global Internet traffic routing. IP Volume is mentioned later on in the research in section 4.11.3 about benign activity, but IP Volume is an ISP that hosts many others and should not be considered benign.

AS14618 Amazon.com, Inc.: Associated with Amazon Technologies Inc., this AS primarily

operates in the United States, particularly in Ashburn, Virginia. It supports a vast array of IP addresses that are utilised for Amazon's cloud services and data centres, indicating its critical function in supporting Amazon's extensive online services and infrastructure.

4.9 Destination analysis

Destination analysis can reveal patterns, behaviours, and potential vulnerabilities within a network, making them a key component in identifying malicious activities. By examining these addresses, one can detect irregularities, predict possible attack vectors, and develop more effective defence mechanisms. Therefore, a thorough analysis of destination addresses in scanning activities provides essential insights into patterns and behaviour of the scans and can give valuable information about scanning types, regularities, and irregularities.

4.9.1 Distinct destinations

To understand the behaviour of scans, it is important to look at the distinct destinations each scan has probed to. The minimum number of probed destinations to define a scan using the Durumeric method is 40.

Table 4.11: Top 10 Distinct Destination addresses with counts

	Distinct destinations	Number of scans	%
1	256	13622	58.18
2	254	2269	9.69
3	255	970	4.14
4	253	249	1.06
5	252	159	0.68
6	251	144	0.61
7	250	138	0.59
8	248	124	0.53
9	249	123	0.53
10	247	102	0.44
	Σ_{10}	16564	76.45

Table 4.11 lists the top 10 distinct destination addresses (to which the scanner has probed) along with their occurrence counts. This information can be valuable for understanding scanner behaviour, identifying potential targets within a network, or discerning patterns of scanning. This analysis is relative to a /24 network; other subnet sizes would not be appropriate for accurate results within this context.

The highest count, 12,214 scans, probed all 256 possible addresses within the /24 network segment that the network telescope has. This indicates a comprehensive scanning approach that aims to discover and possibly exploit every target within the network range. This kind of scanning could be indicative of automated tools or bots methodically mapping the network. The scanning of 256 destination addresses in a /24 network is strange due

to x.x.x.0 and x.x.x.256 not being usable hosts and is likely due to automated machines just scanning the internet for hosts.

The remaining entries show varying levels of scanning activity, with some scans targeting nearly all addresses (e.g., 254 addresses, where every host is usable), while others are more selective, probing fewer addresses (as few as 105 for the least probed destination). The variation in the number of addresses probed may reflect different objectives or strategies of the scanners, such as targeting specific services or machines known to occupy certain IP addresses.

From an analytical perspective, the data in table 4.11 is valuable to understanding what types of scanner this Durumeric method is picking up. The Durumeric method detects large net scan, as 76.45% of the scans have targeted 247 destinations or more. It helps in understanding the scope and focus of scanning activities within a particular network segment. By analysing these patterns, one can identify trends and possibly anticipate future scanning based on the observed scanning behaviour. Recognising how scanners operate and which areas they target, enhanced security measures can harden network resilience against such scans.

4.9.2 Destination Ip-address

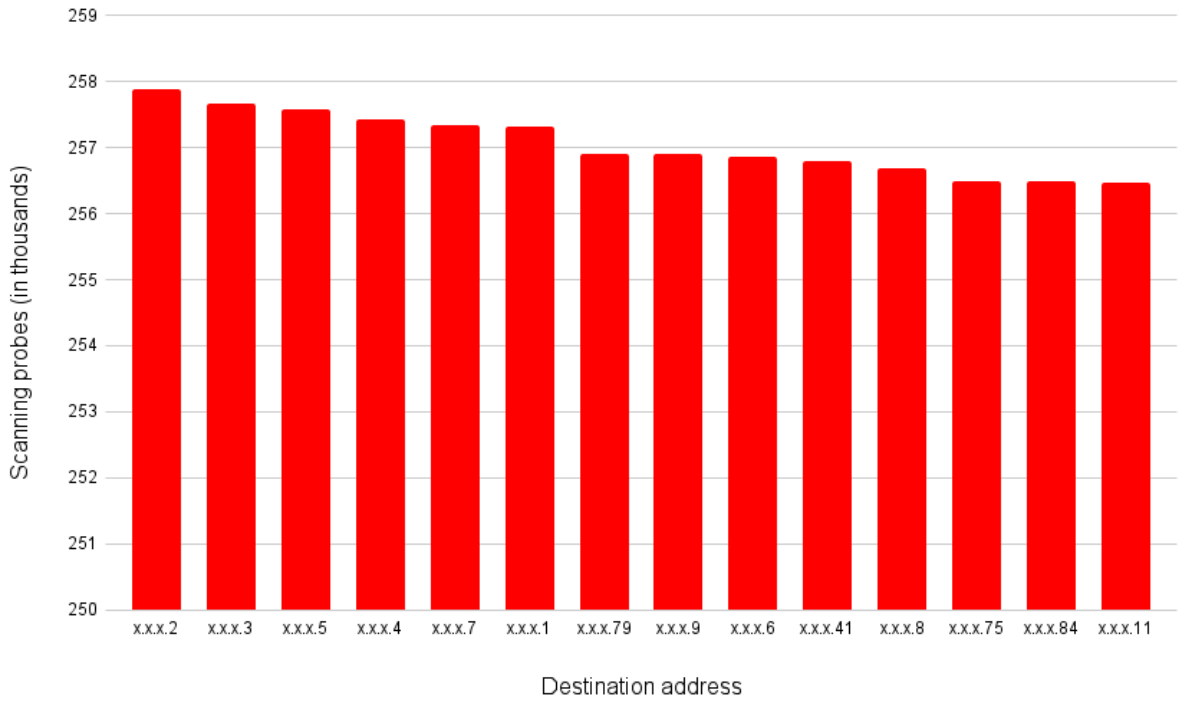

Figure 4.5: Top 10 destination IP address based on scanning probes

Figure 4.5 presents the number of scanning probes that target specific destination addresses, measured in thousands. The numbers are from January 2023 through December 2023. Destination addresses are anonymised, each address having a unique identifier at the end.

The first observation is the relatively uniform level of scanning activity across all destination addresses. The number of probes for each address ranges narrowly between approximately 256,000 and 258,000, suggesting a consistent level of interest or targeting by the entities conducting the scans.

Although the overall activity is uniform, there are slight variations in the number of probes to different addresses. This could indicate varying levels of vulnerability or interest from the scanners' perspective. Addresses with slightly higher probe counts might be considered more valuable or more likely to be vulnerable by the attackers.

There does not appear to be a clear pattern in the distribution of probes based on the anonymised address identifiers (e.g., x.xx.2, x.xx.3, etc.). The variations seem random rather than showing a trend, which might suggest that the scanning is broad and indiscriminate rather than targeted based on specific characteristics of the destination addresses.

The close grouping of probe counts suggests a widespread scanning campaign rather than targeted attacks against specific addresses. It could be part of a reconnaissance phase by attackers looking for vulnerable systems across a range of addresses, or an internet-wide scan from research organisations (covered in Section 2.6), or lastly Mirai scans (covered in section 4.7.2). The slight differences in probe counts could be due to minor variations in scanner configurations, network accessibility, or response rates from the targeted addresses.

For organisations or entities responsible for the destination addresses, this graph highlights the fact that potential attackers are always looking for vulnerabilities to exploit, and thus scanning large portions of networks at a time.

Table 4.12: Top 10 population of scanning destination for network scans

	Destinations IP	Number of scans	%
1	.256	12214	55.17
2	.254	2331	10.53
3	.255	973	4.39
4	.253	249	1.12
5	.252	161	0.73
6	.251	145	0.65
7	.250	138	0.62
8	.248	124	056
9	.249	124	0.56
10	.247	105	0.47
	Σ_{10}	16564	74.8

4.10 Behavioural Analysis

Figure 4.6 compares the number of scans captured each month (represented by the red line) with the connections those scans have made (blue line), over the course of 2023 for all scans. 'Connections for scans' is calculated as multiplying the total scans by the total

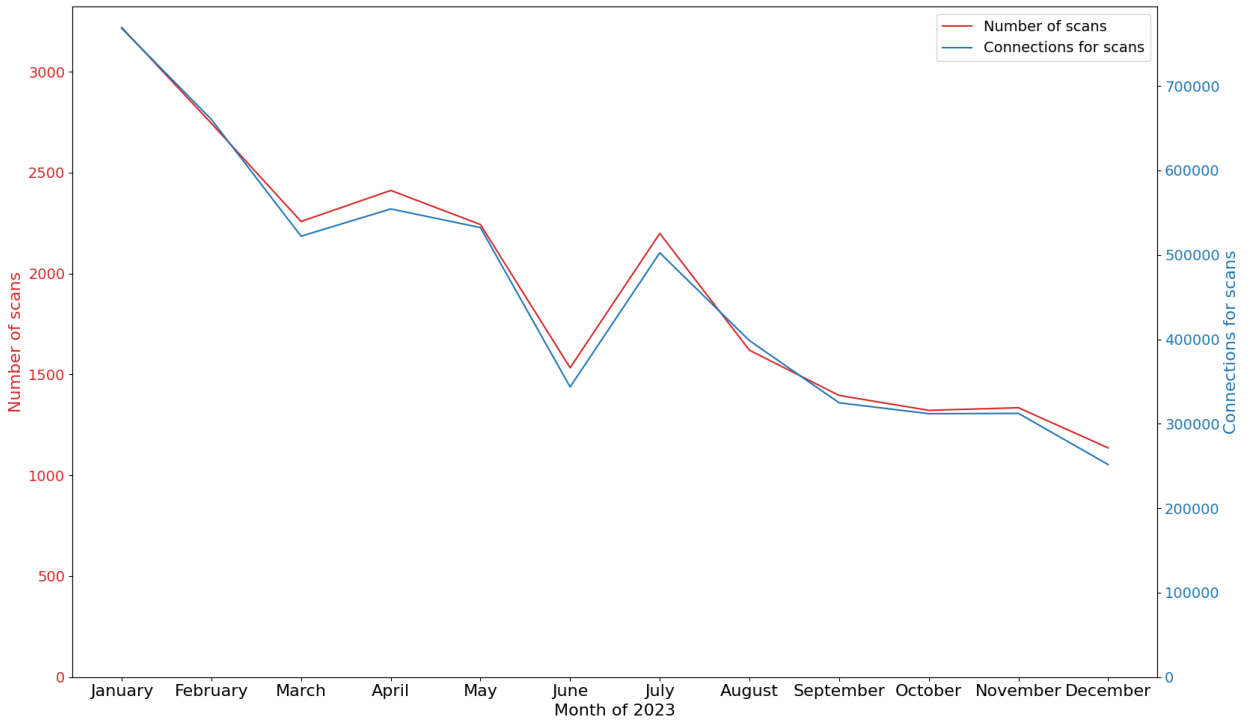

Figure 4.6: Number of scanning per month of 2023 versus number of scan connections

number of distinct destinations.

$$totalscans * distinctdestinations$$

Both the number of scans and the connections follow a somewhat parallel trend throughout the year, suggesting a correlation where an increase or decrease in scans tends to be accompanied by a similar change in connections. This parallel movement indicates that, in general, the number of scans reflects the connections made by these scans.

The year starts off with the highest volume of scans throughout 2023, followed by a decline that follows to March. April has a small peak that could be explained by Easter breaks. There is a decline in both scans and connections in May and June, followed by a sharp peak in July. The sharp decline could indicate the end of this event, a patch of a vulnerability, or a break after all the other scanning before the seasonal vacations. The July spike could be due to the summer vacation as people are less aware of security and staff are on vacation, thus ramping up the scanning.

Scans and connections continue to decline in August, although not to the level of June, suggesting a partial rebound or correction after the June dip. This could be due to efforts to increase scanning activity or the natural variance in user engagement.

The gradual decline in both scans and connections towards the end of the year could be attributed to several factors, including seasonal trends, such as holidays, changes in user behaviour, or external factors affecting the scanning activities.

The strong correlation between scans and connections throughout most of the year implies a stable relationship; however, sudden fluctuations (spikes and drops) suggest inter-

mittent periods of instability. These variations may be influenced by external elements that affect both the frequency of scans and their success in establishing connections. A month-by-month analysis could identify particular months in which there is a disparity between the increase in scans and the corresponding increase in connections. For example, a substantial increase in scans accompanied by only a slight increase in connections could suggest problems with the quality of scans or their efficiency in creating significant connections.

Overall, the number of scanning appears to reflect the connections made, as seen in the parallel trends. However, the effectiveness (connections per scan) might vary and specific spikes or drops could be crucial points for further investigation to understand what influences the scan-to-connection ratio.

4.10.1 InetVis

In Section 2.8.3 the blue line in figures 4.7 and 4.8 is the destination address ranging from .0 to.255 in the network telescope /24 network. The green line is port numbers, 1 being at the bottom and 65535 at the top. The red line is the source IP address and can be set to isolate netblocks. In figures C.1 and C.2 the red line can be clearly seen as a netblock is isolated.

Figure 4.8 shows the InetVis view plot of network scans from the netblock 104.152.52.0/24 to the network telescope. The snapshot was taken from 2023/01/11 and has isolated scans from the netblock with most total scans ranging from January through December 2023, from Table 4.5. The InetVis view plot shows several horizontal scans rather than vertical ones, as explained in 2.8.3. Scans on port numbers on the smaller range are preferred from scans from this netblock. Scans from 104.152.52.0/24 favours more scans within a smaller time-frame compared to Figure 4.8. The side angle of the snapshot can be viewed in Appendix C.

Figure 4.8 shows the InetVis view plot of network scans from netblock 118.123.105.0/24. The snapshot was taken from 2023/01/01 and has isolated scans from the second most active scanner netblock from table 4.5. The InetVis view plot shows fewer horizontal scans compared to 4.7. The netblock scans a small amount of ports numbers expanded over several days. Scans from 118.123.105.0/24 favour less scans within a longer time frame compared to Figure 4.8. Appendix C presents a side angle of the ’118.123.105.0/24’ scans.

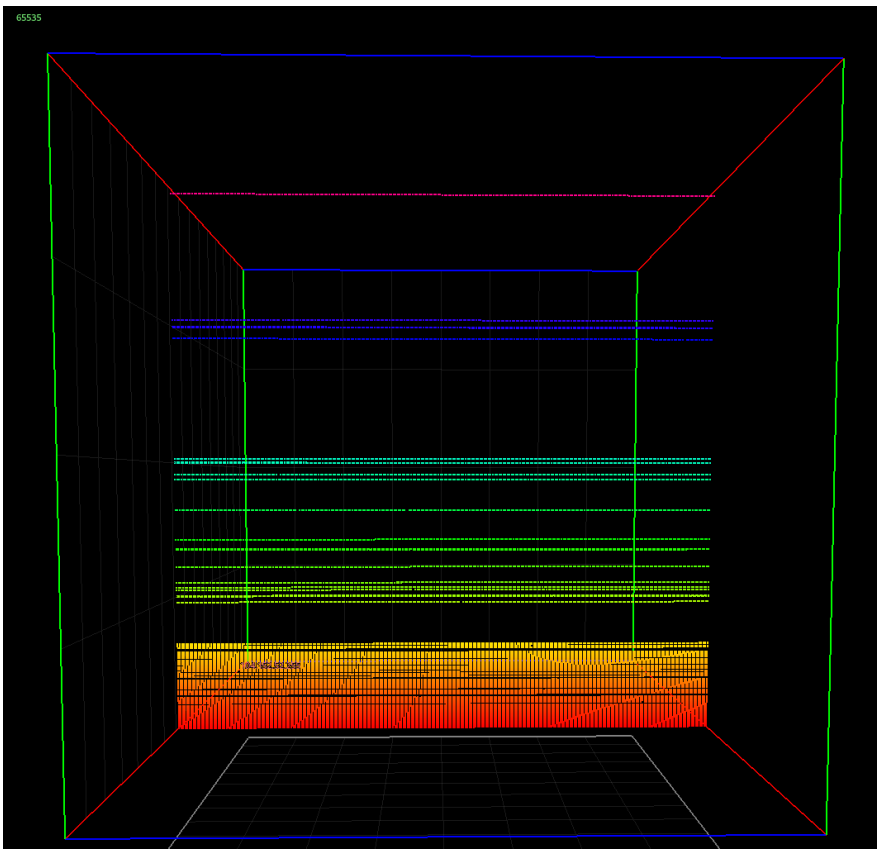

Figure 4.7: InetVis view plot of network scans from ’104.152.52.0/24’

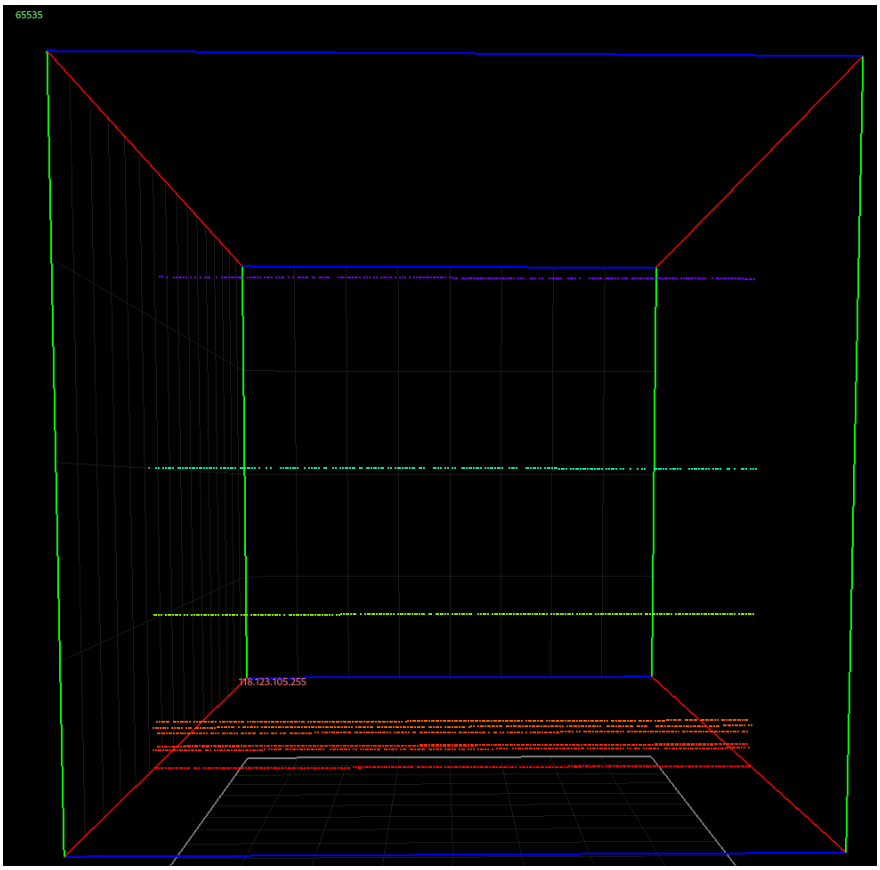

Figure 4.8: InetVis view plot of network scans from ’118.123.105.0/24’

4.11 Correlation with Threat Intelligence

GreyNoise.io is a powerful tool for distinguishing between benign and malicious activity online. One can utilise it by monitoring and analysing the vast amounts of internet background noise – which includes the benign yet ubiquitous pings, scans, and bot activity – versus the targeted, potentially harmful malicious actions. Greynoise is a reputable research organisation, so findings from their sensors are powerful findings towards this research.

4.11.1 Greynoise classification

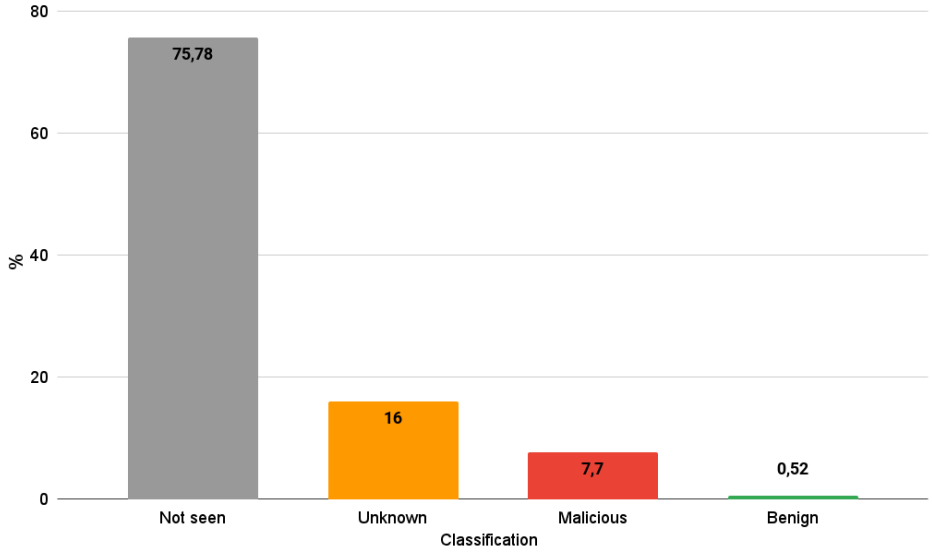

Figure 4.9: Distribution of scanning classifications based on this research's scanning data

Figure 4.9, based on enriched data from Greynoise.io and the scanning results of this study, represents internet traffic and threat intelligence through the monitoring of IP addresses by their sensors.

Not seen (75.78%) This is the largest category, indicating that the majority of IP addresses scanned or observed by GreyNoise have not triggered any of their sensors. This means that these IPs have not engaged in activity that GreyNoise tracks, which could include a range of potentially harmful behaviours. It suggests that a significant portion of Internet traffic is not hitting GreyNoise's specific sensors, or that it is benign and does not exhibit characteristics typical of known threats.

Unknown (16.00%) These are IPs that GreyNoise has observed but cannot classify as benign or malicious. This ambiguity could arise from insufficient data, IPs showing signs of both benign and suspicious behaviours, or engaging in activities not yet fully understood or categorised by GreyNoise. This represents a significant portion of the study and indicates the challenges in classifying internet behaviours where intent is unclear.

Malicious (7.70%) These IPs have been identified to be engaged in harmful activities. Although this is a smaller percentage compared to 'Not seen' and 'Unknown,' it still represents a substantial number of IPs considering the vastness of internet space. This indicates that a notable fraction of Internet traffic is engaged in malicious activities.

Benign (0.52%) This is the smallest category, showing that very few of the observed IP addresses are classified as harmless or safe. This low percentage could indicate that GreyNoise's sensors are primarily designed to detect and analyse potential threats, rather than to identify safe traffic, or that a bigger portion of scans can be categorised as not malicious.

Table 4.13: Greynoise classification of seen IPs

Classification	%
Unknown	67.7
Malicious	30
Benign	2.3

While Figure 4.9 shows the percentage of all IPs, Table 4.13 visualises the classification of IP addresses that have been detected by GreyNoise.io sensors. The chart is divided into three segments, each representing a different category of Greynoise IP classification. Malicious makes up 30% of the IPs, indicating that nearly a third of the IP addresses in the scans have been classified as malicious. A very small percentage of the IPs spotted by greynoise sensors, accounting for 2.3%, represents benign IPs. The largest portion of the greynoise classification, at 67.7%, is designated for unknown classifications.

Table 4.13 serves as a valuable tool in section 5.2 and section 5.3, particularly in examining additional scan behaviour to ascertain whether known benign or malicious actors use distinct scan configurations.

4.11.2 Benign activity

Table 4.13 shows that benign activity only accounts for 0.52 percent of total activity; it is easier to map out its behaviour than malicious and unknown. Total 34 source IP are deemed known benign actors throughout the 12 months of 2023. Research organisations from Section 2.6 has not been spotted using the Durumeric method of defining a scan; other benign actors have been found using this method.

Table 4.14: Netblocks with number of known benign source IPs

Netblocks	number
94.102.61.0/24	33
80.82.70.0/24	1

4.11.3 CriminalIP

Table 4.14 shows that 33 out of 34 source IP originates from the same /24 network, 94.102.61.0/24. This is registered on *IP Volume inc* (AS202425) in the Netherlands.

The benign actor behind this scan is *CriminalIP*[5]. *CriminalIP* uses Artificial Intelligence and machine learning to provide cyber threat intelligence to paying users (CriminalIP, 2024). Their service is similar to greynoise's, which is to provide threat intelligence around IP addresses. Greynoise has 93 entries from *CriminalIP*, where all 93 are marked as benign.

CriminalIP has scanned the network telescope 618 times during the months of 2023, where each case has been 247 destinations or more. There is no special scanning pattern by *CriminalIP*, each of the source addresses from the subnet is scanning different each time. Figure 4.10 is a output of a script made for searching the main scan result CSV, and is showing an example of a scanning pattern from *CriminalIP*. The IP address 94.192.61.43, from *CriminalIP* has consistently scanned identical ports throughout 2023, including ports 873, 83, and 636, as seen in Figure 4.10, occasionally targeting additional ports as well as seen in Table 4.15. Rate value in Figure 4.10 is the rate that the packets has been sent. A rate of 10 indicates that 10 packets has been sent per second, as per definition in Section 3.1.

```
Source IP: 94.102.61.43, Port: 873, Distinct Destinations: 256, Total Packets: 256, Rate: 10.392572522125192,
Source IP: 94.102.61.43, Port: 873, Distinct Destinations: 256, Total Packets: 256, Rate: 10.951026443062291,
Source IP: 94.102.61.43, Port: 83, Distinct Destinations: 256, Total Packets: 256, Rate: 10.543071581132747,
Source IP: 94.102.61.43, Port: 636, Distinct Destinations: 256, Total Packets: 256, Rate: 10.33018837192625,
Source IP: 94.102.61.43, Port: 83, Distinct Destinations: 256, Total Packets: 256, Rate: 10.55232134312207,
```

Figure 4.10: CriminalIP scanning pattern

Table 4.15: CriminalIP scanned ports

Ports	Number of scannings
83	5
548	2
636	5
873	7
888	1
999	1
1080	1

4.11.4 Open Port Statistics

The other known benign activity captured in 2023 was from 80.82.70.217, originating from 80.82.70.0/24. This also is registered on *IP Volume inc* (AS202425). The actor is *Open Port Statistics*, and there is little to no information about them on the internet. The only notation of *Open Port Statistics* is in a blog post[6] where their benign intentions are questioned. Greynoise has 3 entries from *Open Port Statistics* collected from their sensors, 2 marked as benign and 1 as unknown.

The IP address 80.82.70.217 was only active on March 7th, 2023, conducting all 133 scans that day. Their scanning behaviour lacks a discernible pattern, initially focusing on larger port numbers and ending with smaller numbers ranging from 6000 to 8000 as seen

[5] https://www.criminalip.io
[6] https://isc.sans.edu/diary/Whats+the+deal+with+openportstatscom/26912

in Figures 4.11 and 4.12. Notably, scans of the smaller port numbers do not cover all 256 destinations, instead targeting between 189 and 210 distinct destinations.

```
Date: 07.03, Source IP: 80.82.70.217, Port: 35634, Distinct Destinations: 255, Total Packets: 255, Rate: 10.804577446043105,
Date: 07.03, Source IP: 80.82.70.217, Port: 35180, Distinct Destinations: 256, Total Packets: 256, Rate: 11.358155403322733,
Date: 07.03, Source IP: 80.82.70.217, Port: 46332, Distinct Destinations: 255, Total Packets: 255, Rate: 11.142363650276174,
Date: 07.03, Source IP: 80.82.70.217, Port: 64739, Distinct Destinations: 256, Total Packets: 256, Rate: 10.783977523759958,
Date: 07.03, Source IP: 80.82.70.217, Port: 38011, Distinct Destinations: 256, Total Packets: 256, Rate: 14.465688486331704,
Date: 07.03, Source IP: 80.82.70.217, Port: 22945, Distinct Destinations: 256, Total Packets: 256, Rate: 11.907603699688202,
Date: 07.03, Source IP: 80.82.70.217, Port: 38696, Distinct Destinations: 256, Total Packets: 256, Rate: 14.064869255909503,
Date: 07.03, Source IP: 80.82.70.217, Port: 38465, Distinct Destinations: 256, Total Packets: 256, Rate: 14.433388427290266,
```

Figure 4.11: Open Port Statistics Scanning pattern of port numbers in the higher end

```
Date: 07.03, Source IP: 80.82.70.217, Port: 7293, Distinct Destinations: 212, Total Packets: 212, Rate: 12.813290982827825,
Date: 07.03, Source IP: 80.82.70.217, Port: 7269, Distinct Destinations: 191, Total Packets: 191, Rate: 11.340819010544793,
Date: 07.03, Source IP: 80.82.70.217, Port: 7747, Distinct Destinations: 218, Total Packets: 218, Rate: 12.959209946471796,
Date: 07.03, Source IP: 80.82.70.217, Port: 7755, Distinct Destinations: 202, Total Packets: 202, Rate: 11.96854986391654,
Date: 07.03, Source IP: 80.82.70.217, Port: 7532, Distinct Destinations: 227, Total Packets: 227, Rate: 13.697515000474036,
```

Figure 4.12: Open Port Statistics Scanning pattern of lower port numbers

4.11.5 Considerations

When considering the concept of benign versus malicious network scan, it is important to recognise that methodologies can vary significantly across different actors. This research's definition of a network scan prioritises quick, extensive sweeps. This approach, while effective for rapid detection, contrasts with the strategies employed by some benign research organisations, which may opt for prolonged, more discrete scanning processes. Organisations, like those mentioned in Section 2.6 often spread their scans over extended periods to reduce the likelihood of disrupting network performance or alerting detection systems. Such methods aim to blend into normal network traffic, thus avoiding the potential pitfalls of triggering security defences or hindering user activities. Although both approaches seek to scan the network, the difference in tactics underscores the varied priorities between immediate visibility and long-term, unobtrusive observation.

This consideration is explored further in the case studies in Chapter 5.

4.12 Greynoise tags

Section 4.12 explores the classification and analysis of IP addresses according to their observed behaviours as classified by Greynoise.io. Greynoise assigns tags to these IPs, categorising the behaviour. In Subsection 4.12.1, the discussion centres on the benign tags that represent activities considered harmless or legitimate. Section 4.12.2 delves into malicious behaviour based on greynoise tags.

4.12.1 Greynoise benign tags

Figure 4.13 illustrates output from Greynoise.io with information on the most common tags associated with a set of IP addresses that are considered benign. These tags categorise

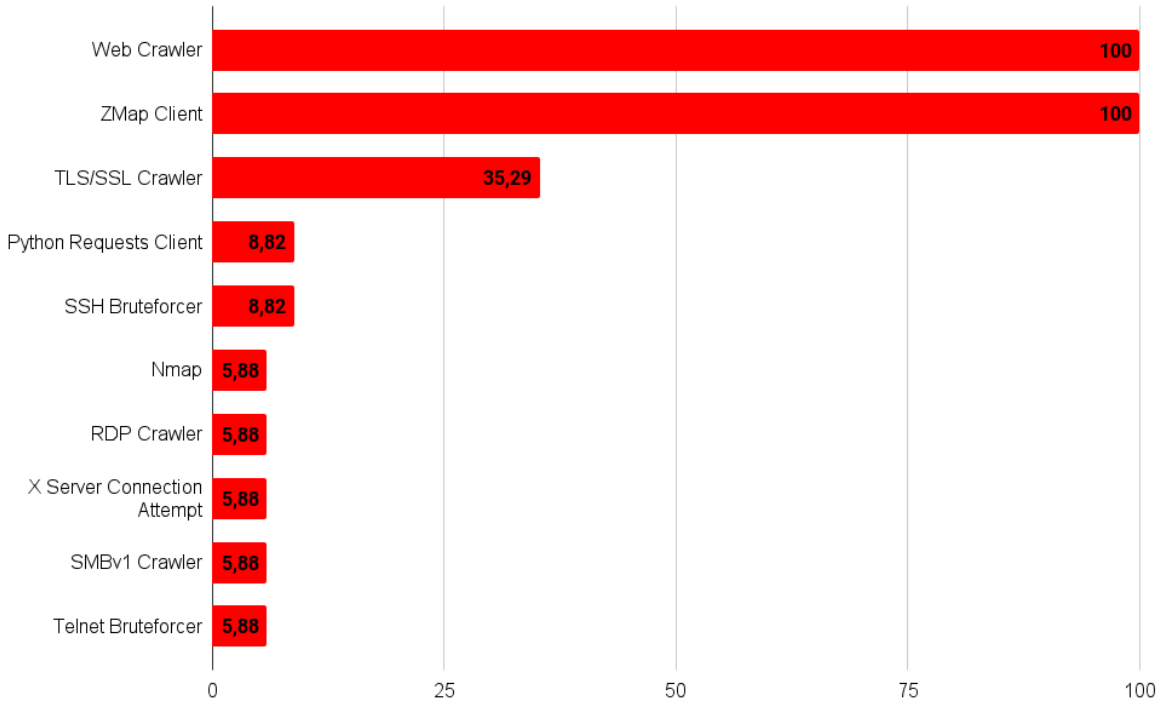

Figure 4.13: Benign tags associated with the observed IPs from scanning data

the types of activities or behaviours typically associated with these IP addresses. Here is a breakdown and analysis of the top 10 tags provided:

Web Crawler (100.00%): This tag indicates that all IPs are associated with web crawlers. Web crawlers are automated online software that indexes web page information and content; they are commonly used by search engines to update their content databases. This high percentage suggests that a significant proportion of benign IP activity is related to web indexing or data collection for legitimate purposes.

ZMap Client (100.00%): Similarly, all IPs are marked as ZMap clients. ZMap is an open source network scanner that is typically used for security auditing and network performance measurement. The fact that this tag is present in all IPs could indicate widespread use of ZMap for benign purposes like research or security assessments.

TLS/SSL Crawler (35.29%): More than a third of the IPs are associated with TLS/SSL Crawlers. These are specific types of crawlers that focus on collecting information related to transport layer security (TLS) and secure socket layer (SSL) protocols. This activity is often related to security research, such as identifying misconfigured certificates or understanding the adoption rate of security practices.

Python Requests Client (8.82%): A smaller portion of the IPs are marked as using Python Requests, a popular HTTP library in Python. This suggests that these IPs are part of systems or scripts that automate HTTP requests for various purposes, which could range from data scraping to automated testing of web services.

SSH Bruteforcer (8.82%): This tag indicates that a small percentage of the IPs have been

associated with attempts to forcefully access systems via SSH (Secure Shell). Although the context here labels these IPs as benign, this activity is generally considered malicious, suggesting that these might be controlled environments for testing or IPs incorrectly marked as benign.

Nmap (5.88%): This is a small fraction associated with Nmap, a network discovery and security auditing tool. Similar to ZMap, it is widely used for legitimate purposes like network mapping and security scanning.

RDP Crawler (5.88%): This indicates a minor percentage related to Remote Desktop Protocol (RDP) crawling. Crawling RDP instances can be for various purposes, including security assessments and research.

X Server Connection Attempt (5.88%): A small proportion of IPs are trying to connect to X servers, which are used in Unix systems for graphical interfaces. This could be for benign purposes like automated testing or research.

SMBv1 Crawler (5.88%): This tag indicates scanning or crawling for SMBv1 servers, an older version of the Server Message Block protocol. Although SMBv1 is known for vulnerabilities, crawling could be for legitimate reasons such as security auditing.

Telnet Bruteforcer (5.88%): Similarly to SSH Bruteforcer, this small percentage of IPs is associated with attempts to brute-force Telnet, a network protocol known for its lack of security. This activity typically raises security concerns, although in this benign context, it might be related to security testing or research.

In summary, Figure 4.13 suggests a mix of legitimate scanning, security research, and automated web activities. However, the presence of tags associated with bruteforcing activities, even in a benign context, highlights the importance of context and intent in cyber security assessments. It is also crucial to note that benign IPs can sometimes be involved in security testing, which may explain why activities typically considered malicious, like SSH or Telnet bruteforcing, are included in this list.

4.12.2 Greynoise malicious tags

The tags represented in Figure 4.14 show specific types of malicious or potentially unwanted activities associated with the observed IPs. These tags help categorise the types of malevolent activities these IPs are involved in.

SSH Bruteforcer (62.03%): A significant majority of malicious IPs are involved in SSH bruteforcing, where attackers attempt to gain unauthorised access to devices or servers by guessing the SSH credentials. This high percentage indicates a prevalent threat from actors trying to exploit weak or common passwords in Secure Shell (SSH) services.

Web Crawler (32.80%): Although web crawling is not inherently malicious, when associated with malicious IPs, it suggests that these crawlers are being used for malicious purposes, such as data theft, website scraping without consent, or reconnaissance for further attacks.

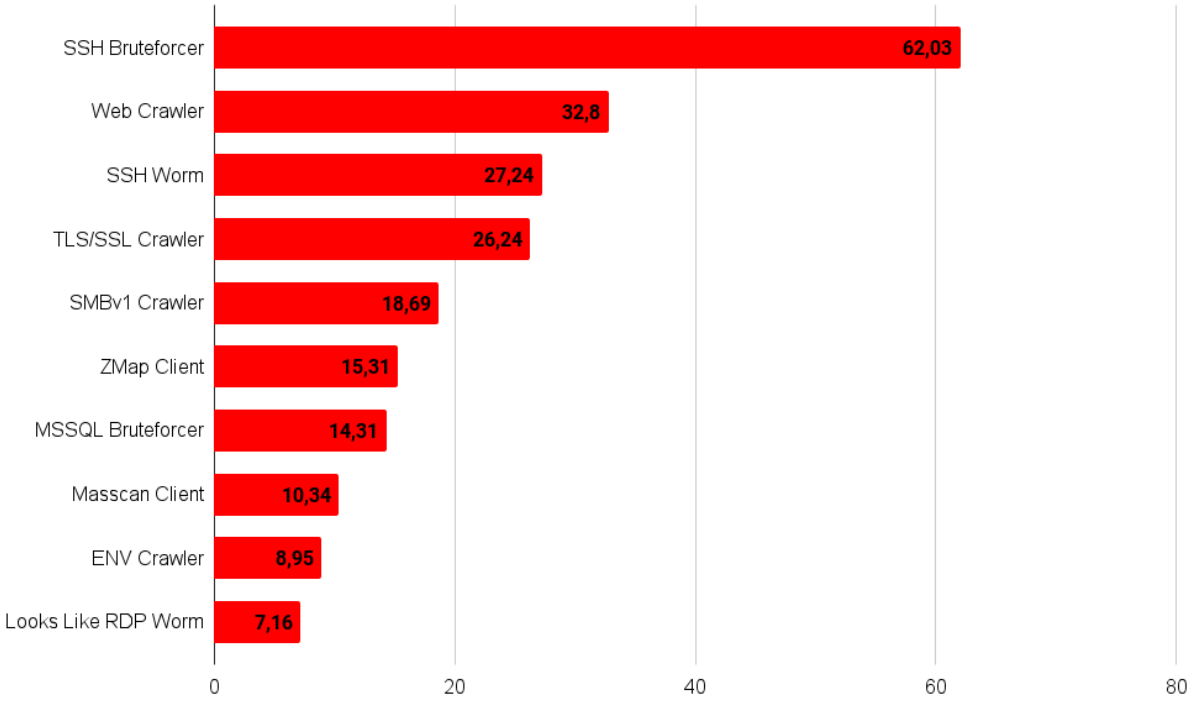

Figure 4.14: Malicious tags associated with the observed IPs from scanning data

SSH Worm (27.24%): More than a quarter of the IPs are tagged as SSH worms, indicating that these IPs are part of a network of infected devices that spread malware, particularly targeting SSH services. This type of malware can replicate itself across networks, leading to widespread security breaches.

TLS/SSL Crawler (26.24%): Similarly to web crawlers, TLS/SSL crawlers associated with malicious IPs are likely scanning for vulnerabilities in SSL and TLS configurations. They could be looking for expired certificates, misconfigurations, or other weaknesses to exploit in encrypted communications.

SMBv1 Crawler (18.69%): These IPs are scanning for devices using the outdated SMBv1 protocol, which is known for its vulnerabilities, most notably exploited by the WannaCry ransomware attack. This suggests targeted attempts to find and exploit systems that still run this insecure version.

ZMap Client (15.31%): A smaller percentage of IPs are using ZMap, a tool for network scanning. In a malicious context, this tool might be used to quickly identify vulnerable devices on the Internet for various illicit purposes, such as finding targets for further exploitation.

MSSQL Bruteforcer (14.31%): These IPs are attempting to brute-force Microsoft SQL databases. This activity indicates targeted attacks aiming to compromise databases, steal sensitive information, or gain unauthorised access to internal networks.

Masscan Client (10.34%): Masscan is another network scanning tool, similar to ZMap but capable of scanning the entire internet in under 6 minutes for open ports. In the hands of malicious actors, it is used to identify potential targets quickly.

ENV Crawler (8.95%): This tag indicates scanning for environment files (.env) which often contain sensitive information such as database passwords or API keys. Accessing these can provide attackers with critical information to further exploit web applications.

Looks Like RDP Worm (7.16%): A smaller portion of the IPs appear to be part of RDP worms, which are malware that specifically targets Remote Desktop Protocol connections. This suggests a focus on exploiting RDP to spread malware, ransomware, or conduct espionage.

4.13 Summary

In Chapter 4, the research delves into the critical task of distinguishing benign from malicious network scanning within IPv4 environments, based on the definition in Section 3.1. Anchored by the question of how to differentiate scanning activities based on specific patterns, behaviours, and characteristics, the analysis uses a dataset to explore network scanning based on the Durumeric definition. Finding and describing the behaviour of benign actors has been difficult due to the poor number of benign actors spotted using the Durumeric definition.

Geographical analysis pinpoints the main scanning origins from the US and China, with notable activities from nations like the Seychelles and Belize. Further analysis is applied to source IP addresses, revealing several IPs frequently involved in scanning, suggesting that a limited number of sources may be responsible for a significant volume of scans. The examination of targeted ports highlights potential vulnerabilities malicious scans may target, while Autonomous System analysis identifies specific networks' involvement in scanning activities. Behavioural analysis throughout the year reveals variations in scanning intensity, possibly reflecting tactical shifts among scanners or external influences.

CHAPTER 5

Discussion

The decision to conduct four case studies to examine the network scanning patterns of benign versus malicious network data come from the need for comprehensive and comparative analysis, and poor benign activity in Durumeric method. By investigating four different setups in the scanning script separately, this research can determine distinctive patterns, behaviours, and signatures unique to benign and malicious activity. This approach allows for a deeper understanding of the intricacies involved in distinguishing between normal network activities and potentially harmful ones. Moreover, by comparing the findings from the scanning datasets, this research tries to identify anomalies and deviations more effectively. In general, the four-case study approach provides a nuanced perspective, facilitating more accurate identification and mitigation of benign and malicious actors.

As described in section 3.6, RATE_THRESHOLD variable is a minimum number for packet per second. In two instances, Sections 5.2.1 and 5.3.1, this threshold is set to 1. Forcing the script to look for scans that have sent at least 1 packet per second.

A new variable was added to the script, as described in Section 3.6.2, to only look for scans that had run for more than the set number of seconds and then write to the output CSV. This new variable was used in Sections 5.5, 5.5, and 5.6.2.

5.1 Structure

(Section 5.2): Case Study 1 - Focuses on scanning behaviours over a duration of 10 minutes, analysing slow and fast scans and their implications in network security.

(Section 5.3): Case Study 2 - Examines scanning behaviours over a duration of 5 minutes to detect differences in scan intensity and to validate the findings from Case Study 1.

(Section 5.5): Case Study 3 - Extended scanning over 24 hours to observe the persistence and distribution of scan patterns and their geographical implications.

(Section 5.6): Case Study 4 - Investigates the behaviours and characteristics of scans lasting 10 and 30 days to distinguish between benign and malicious actors.

(Section 5.8): Comparison - Analyses the variance between the Durumeric method outcomes and the case studies to contextualise the findings within broader research questions.

(Section 5.9): Earlier years - Analysis of 2019 and 2020 to compare actor behaviour against recent years.

(Section 5.10): Findings - Summary of findings from case studies in Chapter 5.

5.2 Case Study 1 - 10 minute scans

This case study is divided into two parts, Sections 5.2.2 and 5.2.1, where RATE_THRESHOLD is set to 0 and 1, where the intention is to catch slow scans.

5.2.1 Scanning longer than 10 minutes with packet rate 1

The output of the script that checked for scans that ran for more than 10 minutes with at least 1 packet per second revealed 18 scans, of which 14 were from unique sources.

From these 14, 5 (35.71%) of the sources have been identified, while 9 (64. 29%) are unidentified, according to Greynoise.io's visualiser[1].

Table 5.1: Greynoise classifications from case study 1

Classification	%
Unknown	60
Malicious	40
Benign	0

Table 5.1 displays the classification of IP addresses based on GreyNoise.io analysis, with an additional category that is not present in the earlier charts:

40% of the IPs is classified as malicious. 60% of the IPs is has a behaviour not fully characterised, making their intentions or nature unclear. Notably, there is no benign activity, suggesting scanning behaviour that benign actors do not use.

However, this scanning configuration is not surprisingly returning any findings, as this would mean that the actor was sending 1 packet per second, in 10 minutes resulting in 600 packets divided over 256 hosts. The fact that there are identified source IPs using this behaviour could be a topic of research in the future.

[1] https://viz.greynoise.io/analysis

5.2.2 Scanning longer than 10 minutes with packet rate 0

The script output that checked for scans that ran for more than 10 minutes with at least 0 packets per second, excluding the RATE_THRESHOLD variable in the script. This variation of the script checks for scans that have run for more than 10 minutes and does not have a set packet rate, thus looking for extremely slow scans.

The output was 1,313,110 scans through 2023. This is an increase of 5507%, over the output of the Durumeric method in Section 4.3, and substantially larger than the scans found in Section 5.2.1. From these *1,313,110* scans 151,580 unique source IPs were found, where 26,263 (17.33%) of them were identified by Greynoise and 125,295 (82.66%) have not been seen by the Greynoise sensors yet.

Table 5.2: Classification of scans longer than 10 minutes with packet rate of 0

Classification	%
Unknown	42.7
Malicious	50.9
Benign	6.3
Riot	0.1

Table 5.2 presents another set of classifications for IP addresses, based on data from GreyNoise.io.

RIOT is a feature that helps identify IP addresses used by common business services, in this case *CDN77*, *Cloudfare CDN* and *Amazon Global Accelerator*, which are highly unlikely to be a source of attack. RIOT serves as a counterbalance, allowing security administrators to filter out benign activity and reduce false positives in their network traffic analysis (GreyNoise, 2023). Table 5.2 shows a worrying prevalence of IPs that are classified as malicious (50. 9%), which could indicate increased security risks within the analysed dataset. It also shows that a smaller but noteworthy portion of the data remains unclassified (42.7%), and a minor portion is benign or associated with the 'Riot' category (0.1%). The benign activity is up to 6.3%, an increase of 4% over the Durumeric method in Section 4.11.1. The average rate of packets these *1,313,110* scans have is *0.00968*. That would be approximately 34.85 packets per hour, which would complete all 256 hosts in 7 hours and 20 minutes, which is considered as a slow scan.

5.3 Case Study 2 - 5 minute scans

In this section, the research explores the behaviour of scans longer than 5 minutes with different packet rate. This gives valuable insights towards the main output in Chapter4.

5.3.1 Scanning longer than 5 minutes with RATE_THRESHOLD 1

This case study behaviour found 73 scanning longer than 5 minutes, with a packet rate of more than 1. This would imply that the scanners sent more than one packet to more than

44 hosts on the \24 network.

Table 5.3: Greynoise classifications for scans longer than 5 minutes and a packet rate of 1

Classification	%
Unknown	38.9
Malicious	33.3
Benign	22.2
Riot	5.6

The source IPs of this behaviour was given to Greynoise, which provided insights into the nature and classification of internet traffic as analysed by GreyNoise. From the data, 52 unique IP addresses were discovered using a scanning behaviour longer than 5 minutes and a packet rate of 1 per second. This would accumulate to over 300 packets for the whole scan, divided into 256 hosts. This would imply that there was an uneven distribution of packets sent to the hosts. Of these, 17 IP addresses (32.69%) were identified by GreyNoise, which means that sufficient information was available to classify them according to their behaviour or characteristics. 34 IP addresses (65.38%) remained unidentified, indicating a lack of sufficient data to classify their activity or perhaps that they represent newer or less common sources of traffic.

7 (41%) IPs were classified as malicious, 6 (35%) IPs were marked as unknown. Such IPs might be under investigation, or they might be involved in new or atypical activities that do not clearly fall into established categories. 4 (24%) IPs were found to be benign, their activities are considered safe, authorised, or otherwise not harmful. 1 (6%) IP was classified in the category of 'Riot'. This classification refers to IPs involved in large-scale coordinated activities that are not necessarily malicious, but are of mass activity.

The benign actor caught by this behaviour is registered to IP Volume (Section 4.8.1), and the actor is CriminalIP, which is mentioned in Section 4.11.3.

This scanning behaviour caught 55 more scans than the behaviour in Section 5.2.1. Furthermore, it caught benign activity and 1 RIOT actor. Which the behaviour in Section 5.2.1 did not.

5.3.2 Scanning longer than 5 minutes with RATE_THRESHOLD 0

This behaviour is similar to the one in Section 5.2.2, as both utilise a packet rate of more than 0, meaning that the script does not check for the packet rate but the length of the scan.

There were 1320638 scans that used this behaviour throughout 2023, with the average packet rate being *0.0124*. Assuming they scan all 256 hosts, the average scan would complete in 5 hours and 42 minutes.

Table 5.4 shows values almost exactly the same as in Table 5.2 except from the 0,1% of RIOT scans. The analysis suggests that there are no noticeable distinctions between scans exceeding 10 minutes in duration and those extending 5 minutes, particularly with

Table 5.4: Greynoise classifications for scans longer than 5 minutes and a packet rate of 0

Classification	%
Unknown	42.8
Malicious	50.9
Benign	6.3

regard to a packet rate of 0.

5.4 Difference between case study 1 and case study 2

In the longer scan duration with a packet rate of 1 in Section 5.2.1 and Section 5.3.1, there was a notable decrease in the quality of the returned data numbers. Therefore, this section does not use this in its comparison between the two.

There are no big differences in scans that were longer than 10 minutes versus 5 minutes. There were 7528 more scans in a shorter period of time, indicating that the scans span a longer period of time than 10 minutes. Furthermore, there is a noticeable difference in the average rate, again suggesting that there is scanning far past 10 minutes. The packet rate in table 5.5 is packet per second, a lower packet rate means a slower scan. Since there are extremely small packet rates, the difference is also small. The 0.00281 packet rate difference is a difference of 10.116 packets per hour. The research continues with a case study of 10-hour long scans to see if there is a change in the number of benign versus malicious scanning. Figure 5.5 illustrates the differences in a table.

Table 5.5: Differences in scanning longer than 10 minutes versus longer than 5 minutes with packet rate 0

	10 minute rate 0	5 minute rate 0	Difference
Scans	1313110	1320638	7 528
Average packets per second	0.00968	0.01249	0.00281
Benign	1,649	1650	1
Malicious	13241	13277	36

5.5 Case study 3 - 24 hour scans

As Section 5.3.2 notes, to see a difference in the scanning behaviour, the research must look for scans longer than 10 minutes. This section looks at scans that have taken place for more than 24 hours. looking at Greynoise data, geographical location and behaviour.

There were 848734 scans longer than 24 hours; this is 464376 scans shorter than case study 1. 126,800 unique IPs were discovered using this behaviour. Of these 126800 IPs, 22772 (17.96%) were identified, while 104028 (82.04%) were unidentified.

1There were only *79* benign actors in difference between 10 minute and 24 hour scans, implying that most benign actors are scanning for more than 24 hours. There was a

reduction of 1065 in the number of malicious IPs detected, and this reduction suggests possibly a variation in behaviour patterns.

Table 5.6: Greynoise classifications for scans longer than 24 hours

Classification	%
Unknown	39.0
Malicious	54.1
Benign	6.9

The classification of 24 hours scans in table 5.6 is almost the same as the 5 minutes scan in table 5.2. Malicious intentions are the majority of scans with 54.1%, unknown intentions are 39%, and benign scanning intentions are 6.9%.

By comparing tables 5.6 and 5.2, the percentage of malicious data has increased by 3.1 percentage points over 10 minute scans, while the unknown category increased by 3.8 percentage points. This could imply a difference in behaviour. The benign category has seen a negligible increase of 0.6 percentage points, maintaining its status as a minor component of the data. In general, while the proportion of known threats has increased, the uncertainty has decreased.

An interesting finding is that 79657 unique source IPs have only targeted port 23, port 2323 or both. This is a notable increase from the 5 source IPs in Table 4.9.

5.5.1 Destination address analysis

Table 5.7 presents data on the top 10 distinct destinations according to the number of scans and their corresponding percentages of the total. The most scanned destination number (all 256 hosts) had 191458 scans, accounting for 22,56% of the total scans. The rest of the scanned destination numbers, ranging from 40 to 57, show a descending order of scans from 17779 to 12384, with percentages gradually decreasing from 2.09% to 1.46%. This distribution of targeted hosts highlights varied interest in scanning behaviour, indicating sporadic scans that could be trying to distinguish it self for cyber defence sensors, or because of more targeted sporadic scans.

Table 5.7: Top 10 Distinct Destinations with counts and percentages for 24 hour scans

	Distinct destinations	Number of scans	% of total scans
1	256	191458	22.56
2	40	17779	2.09
3	41	16867	1.99
4	42	16154	1.90
5	43	15058	1.77
6	44	14353	1.69
7	45	13692	1.61
8	46	13086	1.54
9	47	12619	1.49
10	57	12384	1.46
	Σ_{10}	323450	38.1

5.6 Case study 4 - 10 and 30 day scans

5.6.1 10 day scans

There were 563575 scans that were more than 10 days long and 77733 unique IPs. Greynoise has identified 17596 (22.64%) of the unique IPs. 9884 (56%) of the identified IPs are malicious, 1561 (9%) benign, and there are 6,151 (35%) not yet classified, thus set as unknown. Mirai-based malicious scans were 37% (6572) of the 10-day-long malicious scans, which was a decrease of just 1000 from 24-hour-long scans, implying that the Mirai based scans prefer scans longer than 10 days to look for Mirai victims. More tags will be discussed in Section 5.8.2.

5.6.2 30 days scans

There were a total of 105786 scans that lasted for 30 days, originating from 12981 unique IP addresses. Among these, Greynoise identified 5175 IPs, accounting for 39.87% of the unique addresses. Of the identified IPs, 2539 (49%) were flagged as malicious, 1178 (23%) as benign, while 1458 (28%) remained unclassified and were labelled unknown. This big leap in benign percentage to 23% is an interesting find and would suggest that benign actors prefer extremely long scans. Additionally, among the 30 day long malicious scans, 36% (1871) were associated with the Mirai malware strain, while 28% (1437) were attributed to 'SSH BruteForcer' activities.

In the output of 30 day scans, a number of popular research organisations mentioned in Section 2.6 was found. Shodan.io, Censys, ShadowServer.Org, Bitsight and Cortex Xpanse to name a few.

Table 5.8: IPs associated with research organisations

Benign actor	Number of IPs
AdScore	2
Bitsight	105
Censys	147
CriminalIP	16
CrowdStrike Falcon Surface	10
Cortex Xpanse	448
CyberGreen	1
CyberResilience	12
Intrinsec	9
ipip.net	11
ONYPHE	1
Shodan	36
ShadowServer.Org	359
University of California San Diego	1
Σ_{14}	1158

This finding disproves the claim that Censys uses the netblock 198.108.66.0/23, which was noted in Section 2.10.1 as proposed by Bennett et al. (2021). A full list of IPs belonging to benign actors can be viewed in the Appendix D.

```
Found 306 matches:
Source IP: 162.142.125.85, Port: 9999, Distinct Destinations: 67, Total Packets: 80, Rate: 3.0502290997123754e-05,
Source IP: 162.142.125.85, Port: 2380, Distinct Destinations: 82, Total Packets: 97, Rate: 3.6811769356180506e-05,
Source IP: 162.142.125.85, Port: 49152, Distinct Destinations: 96, Total Packets: 121, Rate: 4.5588611443928805e-05,
Source IP: 162.142.125.85, Port: 53, Distinct Destinations: 74, Total Packets: 85, Rate: 3.247977177037854e-05,
Source IP: 162.142.125.85, Port: 18082, Distinct Destinations: 82, Total Packets: 102, Rate: 3.818047409017122e-05,
Source IP: 162.142.125.85, Port: 623, Distinct Destinations: 82, Total Packets: 101, Rate: 3.8610759654897524e-05,
Source IP: 162.142.125.85, Port: 8888, Distinct Destinations: 141, Total Packets: 199, Rate: 7.468012923264471e-05,
```

Figure 5.1: Figure showing scanning behaviour of a source IP attributable to Censys

Censys's scanning behaviour is shown in Figure 5.1, with the IP 162.142.125.85. The methodology used by Censys is shown to emphasise targeted exploration over broad and indiscriminate sweeps. These IP activities across multiple ports, ranging from 53 to 18082, highlight a precise engagement strategy with a calculated number of distinct destinations, varying from 67 to 141, and total packet counts that gently exceed these figures. Such data suggests a nuanced scanning operation, potentially focusing on specific vulnerabilities, services, or responses, rather than casting a wide net across the Internet. The slowness of the scans (from roughly 3.05 to 7.47 packets per ten thousand seconds across the noted ports) is evidence of carefully chosen methodology to not disrupt normal network traffic.

```
Found 651 matches:
Source IP: 93.174.95.106, Port: 8334, Distinct Destinations: 54, Total Packets: 57, Rate: 2.134088799352263e-05,
Source IP: 93.174.95.106, Port: 5008, Distinct Destinations: 43, Total Packets: 49, Rate: 1.880855719351747e-05,
Source IP: 93.174.95.106, Port: 21379, Distinct Destinations: 48, Total Packets: 51, Rate: 1.9130008418232184e-05,
Source IP: 93.174.95.106, Port: 6002, Distinct Destinations: 54, Total Packets: 61, Rate: 2.291819771476073e-05,
Source IP: 93.174.95.106, Port: 25105, Distinct Destinations: 46, Total Packets: 54, Rate: 2.049281612730605e-05,
Source IP: 93.174.95.106, Port: 2628, Distinct Destinations: 46, Total Packets: 51, Rate: 1.9124937799037e-05,
```

Figure 5.2: Figure showing scanning behaviour of a source IP attributable to Shodan.io, figure shows port that has been scanned, how many destinations were scanned, how many packets were sent in that scan and in what packet per second rate

Figure 5.2 shows the scanning behaviour attributed to the source IP 93.174.95.106, linked to Shodan, which exhibits an unconventional and methodical approach that deviates from the more commonly observed broad-spectrum scans seen in Section 4.11.2. This IP's scanning behaviour is characterised by its sporadic targeting and engagement, as suggested in 4.11.5 with a relatively small number of distinct destinations, as opposed to conducting simultaneous scans of large blocks of addresses (e.g., 256 hosts at once). For example, across various ports such as 8334/tcp, 5008/tcp, 21379/tcp, 6002/tcp, 25105/tcp, and 2628/tcp, the number of distinct destinations ranged from 43 to 54, with the total packets sent being slightly higher than the destinations, indicating a highly targeted and restrained scanning strategy.

Moreover, the total number of packets sent does not directly correlate with the number of distinct destinations, suggesting a nuanced approach where the number of packets is not merely a function of the number of targets, but likely tailored to the specific characteristics or responses of each destination. The scan rates, which are remarkably low (ranging from approximately 1.88 to 2.29 packets per 10000 seconds), further underscore the cautious and extremely slow pace of these scans. Such an approach is suggestive of a strategic and considerate operation, aiming to minimise any potential impact on the

scanned systems.

5.7 Comparing geographical heatmaps of benign and malicious IPs

Some of the figures in this section can be found in Appendix E. This section shows a scanning length ranging from 24 hours to 30 days. These figures are not for accurate interpretation, but rather for higher-level interpretation.

Section E.1 presents a sequence of figures that depict the geographical distribution of network activities identified as non-malicious.

Section E.2 contrasts Section E.1 by mapping out the origins of network traffic deemed malicious. These two sections are critical for understanding abnormal traffic patterns and establishing baselines for abnormal network behaviour.

5.7.1 Benign

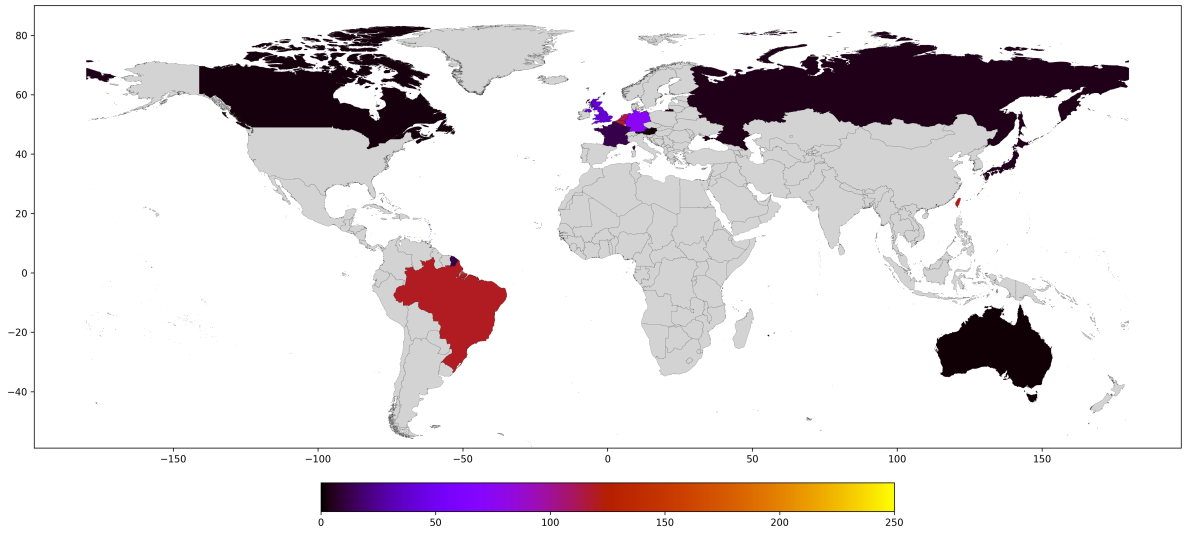

Figure 5.3: Geographical origin of benign scans from 24 hour long scans

When looking at the benign actors utilizing 24-hour scans in Figure 5.3 and 10-day scans in Figure E.3 in Appendix E, one might deduce that scanning activity is somewhat constant during these two scanning length, meaning that the benign actors scanning for 24 hours was the same that scanned for over 10 days. In Appendix E Figure E.4 shows the origin of benign actors using a 30 day long scans, showing that there was no benign actor from Canada (*BinaryEdge.Io*) or Russia (*Academy For Internet Research*) scanning for over 30 days, there was less benign actors from Brazil (*Cortex Xpanse*), also the benign actors from United Kingdom (*AdScore*) used less 30 day scans.

The scans of the Durumeric method in Figure 4.2 which only had benign actors from

the Netherlands (*CriminalIP* and *Open Port Statistics*) are quite different from the others, suggesting that the methodology or the definition of scanning is different from those used in time-bound scans.

The 30 day scans also had French Guiana, Taiwan and Japan, and Switzerland as origin country for the 30 day scans. This can be explained by being close to the more colourised countries and therefore could be a VPN access point for scanning from the more colourised countries. For example, French Guiana could be VPN for Brazil, Japan for Taiwan, Switzerland for The Netherlands, or Germany.

5.7.2 Malicious

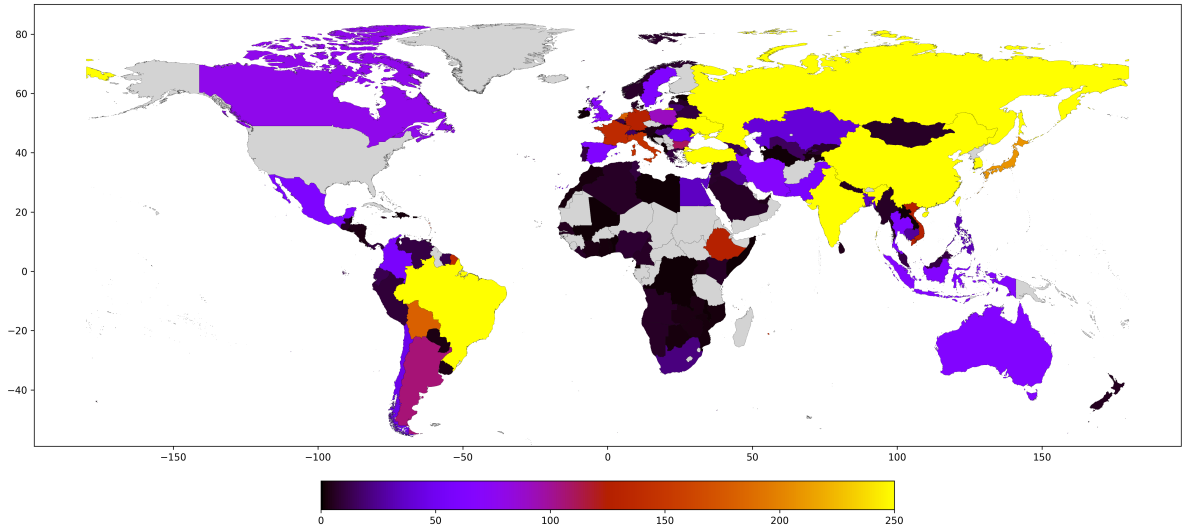

Figure 5.4: Origin of malicious scans longer than 24 hours

From the malicious actors that were responsible for 40% to more than 50% in some cases has more coverage in the geographical heat maps, a difference in colour intensity in Figure 4.3 which is IP data from the Durumeric method, from the set length scans in Figure 5.4, Figure E.7 and Figure E.8, the latter found in Appendix E. The Durumeric method uses a modified version of Durumeric et al. (2013) definition of a scan and therefore different from the others, as mentioned in Section 3.1.

There is a lack of US based malicious IPs targeting the network telescope in both the set length scans and the Durumeric method. Malicious actors from Norway, France, Italy and most of Africa are not present in Figure 4.3. China is the most intense colour country in both Figure 4.3 and Figure 5.4. Between Figure 5.4 and Figure E.7 (Appendix E) showing 24 hour long scans and 10 hour long scans there are no differences. However, the colour intensity in Africa and Ukraine has degraded. The 30 day long scans (Figure E.8, Appendix E) have less colour intensity than the 10 day scans (Figure E.7, Appendix E), except for China, which still has over 250 malicious actors. Russia, Brazil, and India

are around 100 malicious actors on 30 day scans.

5.8 Comparing Durumeric method output against Case studies

It is difficult to compare the Durumeric method with the output of the longer scan period, since the Durumeric method uses a modified version of Durumeric et al. (2013) definition of a scan. However, there are behaviour to compare, like greynoise tags, actor number

Table 5.9: Differences between the Durumeric method outcome and the case studies

	Durumeric method	10 minute	24 Hour	10 days	30 days
Scans	23416	1313110	848734	563575	105786
Malicious	424	13021	12205	9884	2539
Benign	34	1649	1578	1561	1178
Malicious:Benign Ratio	12.47	7.89	7.73	6.33	2.15

The lower Malicious:Benign ratio means that the distance between the number of malicious versus benign scans has decreased. Figure 5.5 is complementary to Table 5.9, showing the appearance of malicious and benign actors in each of the scan behaviours and the Durumeric method.

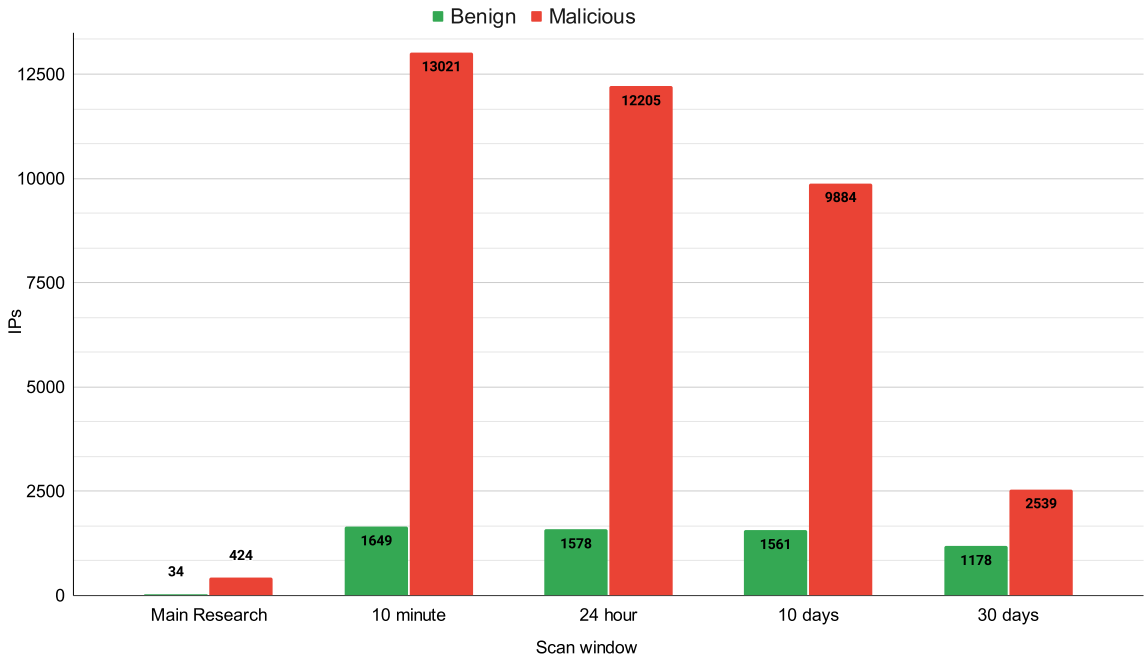

Figure 5.5: Benign (green) and malicious (red) actors across different scan window

Figure 5.5 suggests that benign scanning favourites longer scans, 30 days, was the behaviour of the research organisations mentioned in 2.6. For malicious scanning, a shorter period of scans was preferred, as the number decreased exponentially as the scan length went up. This decrease in malicious scans explains the decrease in the Malicious:

benign ratio in Table 5.9, suggesting that benign scans prefer longer scans.

5.8.1 Benign tags

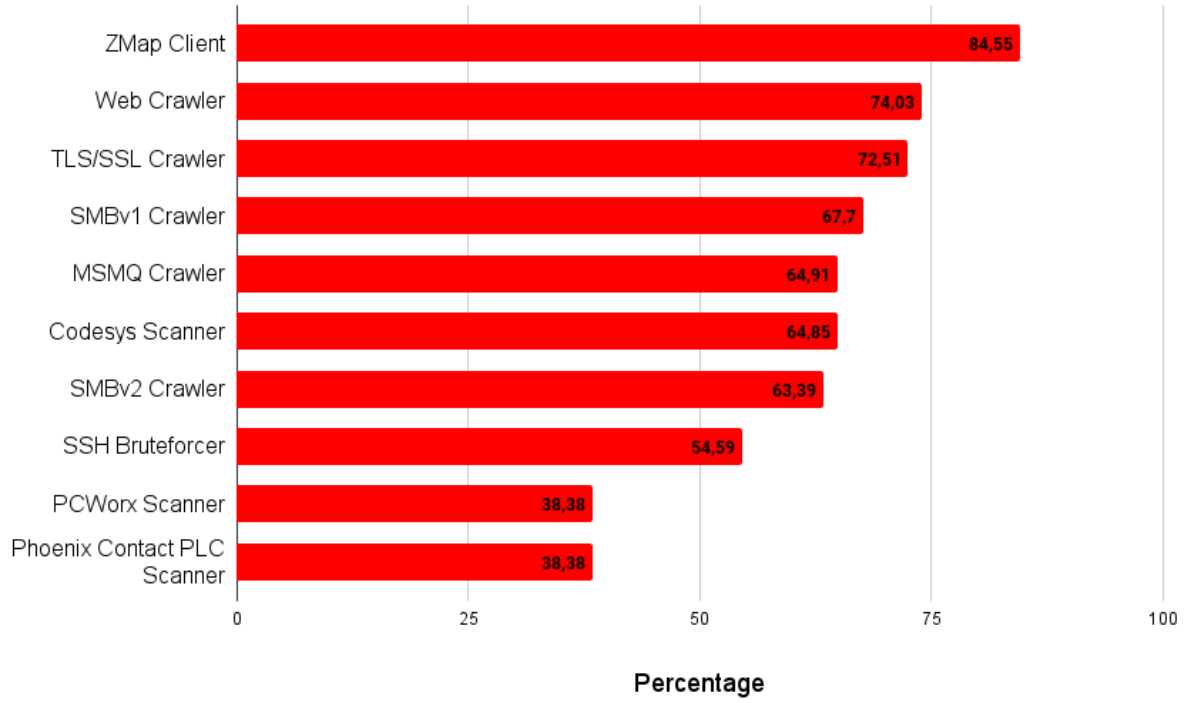

Figure 5.6: Greynoise benign tags for 24 hour long scans

Similarly to the output of the benign Durumeric method in Table 4.13, the benign 24 hours, 10 days, and 30 days also have ZMap Client and Web Crawler as the top 2 tags. While Table 4.13 has them at 100%, Table 5.6 has them at 84,55% and 74,03%, indicating that some of the benign actors are using scanners other than ZMap. As suggested in Section 2.6, Shodan.io uses its own scanner. The TLS/SSL Crawler is also the 3rd most popular in both Figure 4.13 and Figure 5.6

Due to the lack of benign actors in the Durumeric method, it was found to be difficult to compare more tags.

5.8.2 Malicious tags

The difference between Figure 4.12 and Figure 5.7 is a change in the distribution of activity types. In the 24 hour scan, Mirai is the dominant activity, while in the Durumeric method scan, 'SSH Bruteforcer' takes the lead. This might indicate a change in the threat landscape or a shift in attacker focus over the specified period. Mirai scans are not present in the Durumeric method scan, since they do not appear in Figure 4.12. This could imply that Mirai-associated activities were more sporadic in case studies than in the Durumeric method scan. The presence of different types of activity in Figure 4.12, such as 'ENV Crawler' and 'Looks Like RDP Worm', suggests that new threats emerged

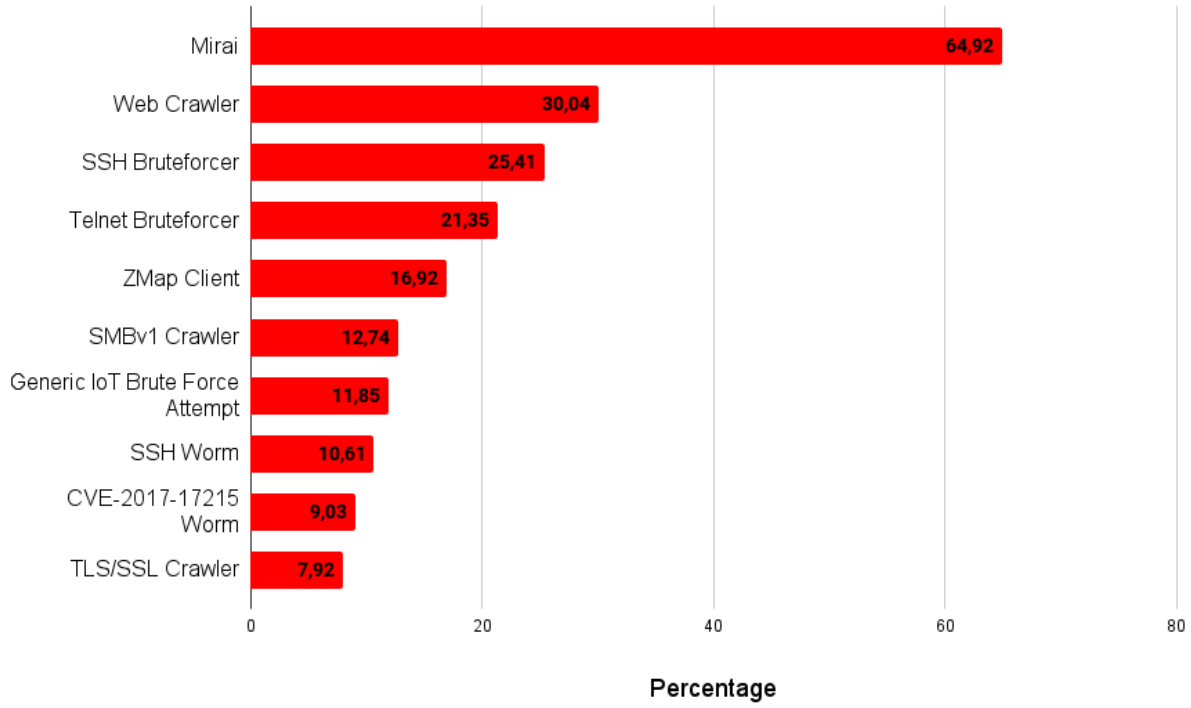

Figure 5.7: Illustrates malicious tags in 24 hours long scans

or were detected during the Durumeric method scan that were not present in the case studies.

A comparative analysis of the two charts Figure 5.7 and Figure 5.8, it is evident that there has been a significant increase in activity associated with Mirai and web crawlers for longer scan periods. Mirai, which is particularly notorious for its involvement in large-scale network attacks, has seen an almost 10% jump in its tag percentage. Specifically, it went from 64. 92% to 73. 82%, while the activity tagged as web crawler has decreased by approximately 5%, moving from 30.04% to 23.9%.

5.8.3 Distinct destination

In Section 4.9.1 it was discussed how the Durumeric method script showed how over 50% of the scan had tried scanning all 256 hosts. In Table 4.11 it was also shown that 76.45% of the scans had scanned 247 or more distinct destinations in the scan.

In Section 5.5 Table 5.7 showed that in the 24 hour scans 22.56% of the scans targeted all 256 hosts, while the rest of the top 10 was of the lower end near the minimum limit for distinct destination. This could be due to the fact that the actors are trying to be invisible to the sensors or trying not to disrupt normal traffic, as discussed in Sections 4.11.5 and 5.6.2.

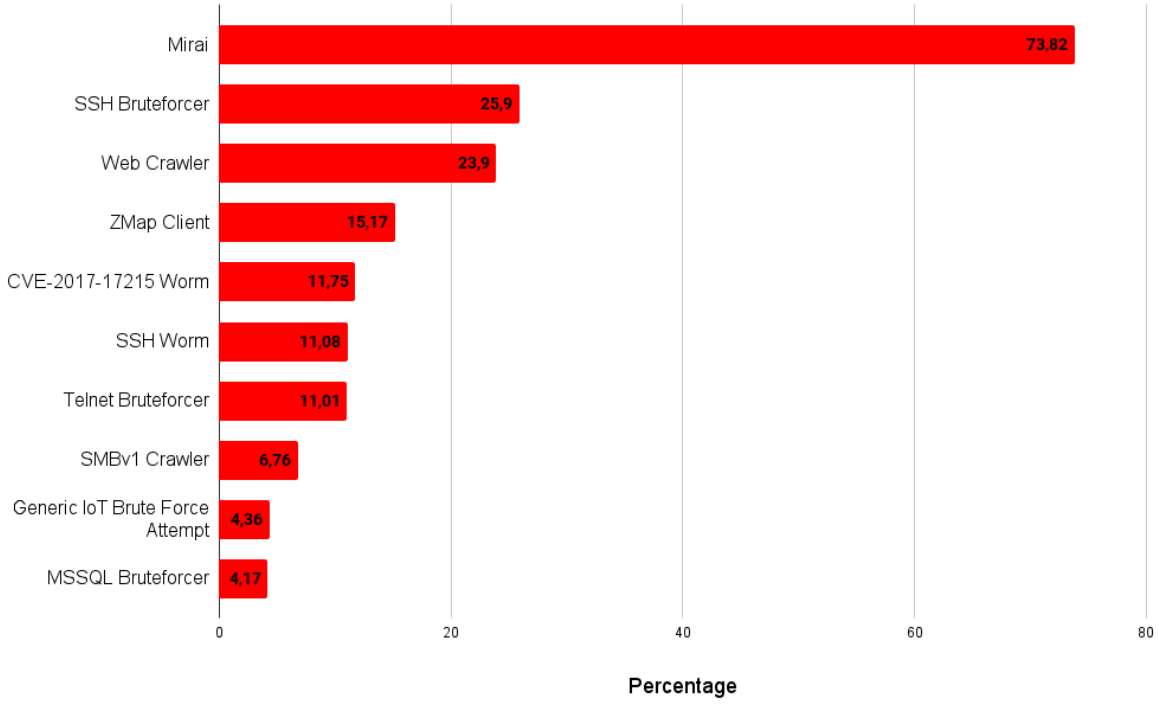

Figure 5.8: Portrays malicious tags in 30 day long scans

Table 5.10: Average days seen for source IPs for benign and malicious actors

Scan window	Benign	Malicious
Durumeric method	703	726
24 hours	1114	543
10 days	1121	550
30 days	1013	556

5.8.4 Average Days of Appearance

Table 5.10 presents an analysis of the Greynoise sensor data on the average days of appearance for source IPs. The average day off appearance for each classification and for each scan window is calculated by

$$\text{Average days of appearance} = \frac{\sum_{i=1}^{n}(\text{First Seen}_i - \text{Last Seen}_i)}{n}$$

n represents the total number of classified IPs. First Seen_i and Last Seen_i denote the first and last day when the i-th IP was observed. The numerator sums the days that each IP was seen, calculated by the difference between the first and last days of observation for each IP. The denominator n is the total number of IPs, thus calculating the average days of appearance.

Analysing table 5.10, a distinction is made between benign and malicious Internet activities. Benign IPs demonstrate a long-term presence, averaging 703 days in Durumeric method scans, indicative of stable, legitimate operations that form the backbone of internet services. Interestingly, malicious IPs maintain a slightly longer presence than benign

ones in Durumeric method scans, averaging 726 days. This suggests that malicious activities can persist through sophisticated evasion techniques or by leveraging compromised yet legitimate services.

The trend changed drastically in the case studies, with benign IPs showing an increase in average days of presence, such as an average of 1121 days in the 10-day scan window. On the other hand, malicious IPs have a reduced presence in these shorter spans, down to 543 days for 24 hour long scans, indicating potentially sporadic activities or more effective detection and mitigation efforts. Table 5.10 shows that benign actors tend to live longer than malicious source IPs. The table also shows how vastly different the Durumeric method scans are compared to the case studies.

5.9 2019 and 2020

Table 5.11: Total scan for each scan length in 2019, 2020 and 2023

	24 Hour	10 days	30 days
2019	638190	250572	31219
2020	770584	357133	40855
2023	848734	563575	105786

Table 5.11 lists the total number of scans performed on three different scan lengths (24 hours, 10 days, and 30 days) for the years 2019, 2020, and 2023. It is a straightforward representation of data, where each row corresponds to a year and each column corresponds to a total count of scans for a specified duration. This is data gathered on scans on the same netblock within 155/8.

For each scan length, the number of scans increased every year from 2019 to 2023. The most substantial growth over the years appears in the 30-day scan category, which grew 238% from 2019 to 2023, and 159% from 2020 to 2023. The 10 day scan category also shows a significant increase in 2023, more than doubling the number from 2019. The 24 Hour scan numbers have consistently grown but at a more gradual pace compared to the other two categories.

This data could suggest that there is an increasing trend in the number of scans being performed as the years progress. The 30 day scans show a significantly increased focus, which could imply a growing interest or requirement for longer-term scanning operations.

5.9.1 Benign and malicious actors

For benign actors, the percentage of total scans saw a minor decrease from 0.000597% in 2019 to 0.000555% in 2020, suggesting a stable number of benign actors if Table 5.11 is taken into context. However, this trend reversed dramatically by 2023, with the percentage jumping to 0.001859%. This indicates a significant increase in benign scanning activities, as the total number of scans also has increased significantly.

Table 5.12: Benign and malicious actors as a % of total scans for years 2019, 2020 and 2023

	Benign actors as a % of total scans	Malicious actors as a % of total scans
2019	0.000597	0.000915
2020	0.000555	0.001377
2023	0.001859	0.014380

On the other hand, the data on malicious actors reveal a more alarming trend. Starting at 0. 000915% of the total scans in 2019, there was a noticeable increase to 0.001377% in 2020, followed by a substantial increase to 0. 014380% in 2023. If Table 5.11 is taken into context, the increase is more alarming. This sharp increase suggests that malicious scanning activities have grown exponentially, reflecting a heightened threat landscape. The increase in malicious scans can be associated with a variety of factors, including the proliferation of cybercriminal groups and possibly a higher number of vulnerabilities being targeted as digital transformation accelerates, and large-scale botnets such as Mirai.

Table 5.12 shows that most of the actors from 2019 through 2023 are considered unknown. However, Section 3.1 defines every unknown probe from the network telescope as unsolicited (Irwin, 2011) and should be considered potentially malicious, until it proves benign.

5.10 Findings

Big benign actors, such as Shodan, Censys, Shadowserver, and BitSight, play a crucial role in mapping and understanding the vast expanse of the Internet. A key aspect of their operational methodology is the use of a scan duration strategy, often extending up to 30 days, as proven in Section 5.6.2. This is a finding to the consideration suggested in Section 4.11.5, and is a methodology to prevent hindering normal network traffic. This find is also supported by Section 5.6.2 where Figure 5.1 and Figure 5.2 illustrate the slow and inconsistency of the total packets from the number of distinct destinations.

In total opposite, the research has found evidence of malicious scans that prefer faster scans. In Section 5.8, Figure 5.5 shows that there is a downward trend in malicious actors as the duration of the scan increases. Table 5.9 also indicates the preference for faster scans in malicious actors. For faster malicious scans ”Bruteforcers” and ”Web Crawlers” are more commonly used than in slower scans, as seen in comparing Figure 5.7 with Figure 5.8.

Slow malicious scans can be attributed to Mirai scanning, as indicated by the high number of malicious scans. Faster scans can be used to find Mirai victims, as SSH and Telnet brute forcers are in the 4th and 5th place in malicious tags (Section 5.7). Section 5.5 shows that through the entire 24 hour scan CSV *79657* unique source IPs were found to be targeting port 23, port 2323 or both, which could be attributable to Mirai Zombies trying telnet.

In both Section 4.5 and Section 5.7.2 China can be placed as the number one origin country in malicious scans. This can be attributed to the geopolitical situation the world is in now and would place China as one of the top malicious hacking superpowers. For the case studies, both Brazil, China, Russia, India, and Turkey can be set as the top malicious scanners, this is due to the sheer number of malicious scans compared to the findings in Chapter 4. This is not surprising, as Bruce et al. (2024) reported these as the top 10 (with the exception of Turkey) in 'World Cybercrime Index - Overall' and top 14 (again with the exception of Turkey) in 'World Cybercrime Index - Technical products/services' and top 16 with Turkey in 'World Cybercrime Index - Attacks and extortion'.

As for the origin of benign actors, in Section 4.5.1 The Netherlands is the only country with colour. This is due to the poor data output from the Durumeric-method scans compared to the case studies. For case studies Brazil is the number one in benign scans, for longer scans Brazil was the most colorised country followed by Germany and The Netherlands. For the shorter scans in Figure E.2 and Figure E.3, other countries such as the UK, France, Canada, and surprisingly Russia are to be found.

In Section 4.9.1, Table 4.11, it was found that half of the source IPs were scanning all 256 hosts in the network telescope. This is not the case for the case studies as Figure 5.1 and Figure 5.2 indicate in Section 5.6.2. In Section 5.5 the discussion of how the scans had distributed its distinct destinations through Table 5.7. It showed a complete opposite of Table 4.11

A note of the discussion is the poorly data number that the Durumeric method scanning script has given. This has resulted in a poor number of scans, thus resulting in poor number of benign actors and a poor number of malicious actors. The Durumeric method was targeting 4 - 25 second scans, due to the 10 packets per second, and a minimum limit of 40 distinct destinations.

5.11 Summary

Chapter 5 integrates the findings of Chapter 4, focusing on the distinct patterns, behaviours, and signatures of benign and malicious network activities through a structured analysis of four case studies. This chapter addresses the research questions posed in Section 1.2 by analysing the data gathered from various scanning scripts and settings.

The data collected and analysed in the case studies provide a nuanced understanding of Chapter 4, contributing to a more effective identification of benign activities. The research gap identified in Section 2.14 is being addressed through the investigations presented in Chapters 4 and 5.

For further details and the final conclusions drawn from the comprehensive analysis presented in this chapter, refer to the concluding Chapter 6 of the thesis. This structure ensures a thorough examination of network scanning behaviours and provides a solid

foundation to address the research questions set forth in the study.

CHAPTER 6

Conclusion

6.1 Introduction

This chapter summarises and reflects on the research undertaken into differentiating benign from malicious IPv4 scanning patterns. It incorporates the findings and insights obtained in Chapter 4 and Chapter 5. In addition, it serves as a structured overview of the discussions that follow, which include a summary of the research in Section 6.2, an assessment of the research objectives in Section 6.3, the contributions of this work to the field in Section 6.4, and potential future research in Section 6.5.

6.2 Summary of Research

6.2.1 Chapter 2

The literature review in the thesis provided a foundational understanding of IPv4 network scanning, essential for identifying the key differences between benign and malicious scans. It contextualised the research within existing studies, highlighting gaps and aligning the research objectives with the needs for advanced diagnostic tools in cyber security.

6.2.2 Chapter 3

This chapter described the methodology for processing and analysing large quantities of network packet data, using specialised scripts and algorithms to uncover patterns of scanning activities. Through the development of a Python script for parsing PCAP files and employing the Greynoise API, for further enrichment. The study distinguished between benign and malicious scans with notable precision.

6.2.3 Chapter 4 and 5

These chapters processed and analysed the PCAP files for network data, using scripts and algorithms to discover patterns of scanning activities from the methodology. The integration of geographical heat maps and the study of case studies further enriched the findings, providing a complex view of the network scanning dynamics.

6.3 Research Objectives

This section outlines the key questions that guided the focus of this study. The research was designed to fill a notable void in the existing literature.

- **How can we accurately differentiate between benign and malicious scanning in IPv4 networks based on specific patterns, behaviours, and characteristics?** This research developed and validated several scripts that analyse network traffic data to distinguish between benign and malicious scanning. By using packet data, the scripts were made to identify scanning from source IP addresses from packet number and probed destination addresses within a subnet.

  These scripts produced a CSV that was queried to Greynoise.io. The greynoise.io output helped identify and separate the differences between malicious and benign scanning.

  To distinguish between harmless and harmful network scanning, one should examine the duration of the scan, as well as the variability in packet sizes and the range of targeted addresses. Scans that are longer and show differences in packet size and destination tend to indicate benign activity, shown in sections 5.6.2 and 5.8. However, shorter scans that have uniform packet sizes and focus on specific addresses could indicate malicious intent, as discussed in Sections 5.5 and 5.8. Recognising these patterns is essential to accurately classify the nature of the scan and implement effective cyber security strategies.

The primary research goal was supported by the secondary research question:

- **Can Threat Intelligence services be utilised to accurately distinguish benign scanning from malicious scanning.** The research incorporated Threat Intelligence (TI) services (Greynoise.io), as a critical component in enhancing the detection capabilities of the developed models and scripts. Greynoise provided additional layers of information regarding IP, known malicious entities, known benign entities, exploit types for malicious entities, scanning tags for benign entities, and historical security incidents associated with certain network behaviours. By integrating Greynoise into the analysis framework, the research demonstrated an increase in the predictive accuracy of the models. This integration allowed for a more dynamic response, demonstrating that TI services are not only useful but necessary to maintain up-to-date and effective network security measures.

- **Do benign and malicious scanning exhibit any differences in observed behaviour** The research distinctly highlights the divergent behaviours of benign and malicious network actors through comprehensive case studies. Benign actors predominantly engage in extended, methodical scans that without disrupting network functionality. These activities are often attributed to reputable research institutions and utilise specific, noninvasive methodologies. In stark contrast, malicious actors prefer short, intense scanning activities aimed at quickly identifying and exploiting vulnerabilities or malware, characterised by higher packet rates and a lack of regard for network disruption. This fundamental difference in approach underscores the contrasting objectives between these groups; security and exploration versus exploitation and disruption.

This research aimed to primarily create a *research into validate and distinguish benign from malicious scanning. The results were used to develop scripts for others to use*.

To achieve this primary objective, the following secondary objectives have been identified.

- Find a source for network data.
  - The network source data came from Professor Barry Irwin from Noroff University by the CSIRT for the South African Research and Education Network (SANREN).
- Create scripts to identify scanning.
  - Two scripts were developed: one using the Durumeric method for detecting a range of scanning activities with custom algorithms, and another tailored for case studies to address specific research questions and provide focused analysis. Details of how these scripts operate are illustrated in a flowchart in section 3.6.
- Incorporate data enrichment using Greynoise.io.
  - Script outputs were analysed using Greynoise.io API to differentiate between benign and malicious scanning, leveraging its comprehensive database and sophisticated algorithms to enhance the understanding and effectiveness of scanning behavior analysis. Approximately 20% of the total IP addresses were found in the greynoise.io database, 80% is considered unknown scanning intentions.
- Determine the identity of benign and malicious scanning.
  - With help from greynoise it was easier to look at behaviour differences between benign and malicious scanning. Benign scanning was observed to prefer longer scans, more sporadic scans, has a longer appearance time, meaning that they do not appear and disappear as fast as malicious ones.
- Create case studies to further enhance the findings.
  - Case studies were developed to offer detailed insights into scanning behavior across different scenarios, enriching the overall findings of the research. These case studies were accurately designed to examine specific instances of scan-

ning activity, shedding light on factors such as duration, intensity, and potential implications.

  - Four distinct case studies were conducted, each exploring the unique aspects of scanning behaviour. These case studies can be found in Section 5.2, Section 5.3, Section 5.5, and Section 5.6, offering detailed analyses and insight into the observed scanning phenomena. Through the integration of case study analyses, the research not only validated the findings of the primary scanning detection methods but also offered nuanced perspectives on the dynamics and implications of scanning behaviour in practical contexts.

- Compare main studies with case studies.
  - A thorough comparison between the conventional definitions of scans, as proposed by Durumeric et al. (2014), and the findings from case studies was performed to explore discrepancies and nuances, challenging existing assumptions and expanding the conceptual framework of scanning behavior. The synthesis of this comparison was detailed in Sections 5.4 and 5.8, offering a comprehensive review and fostering a deeper understanding of scanning activities.

6.4 Research Contribution

This research has successfully developed an analytical script capable of distinguishing between benign and malicious IPv4 scanning patterns with the help of Greynoise.io. This script, which constitutes the main artefact of the study, integrates complex algorithms and takes advantage of network packet data to differentiate benign from malicious network data. Throughout the research process, it was observed that benign activity spanned longer than malicious activities and that benign activities scanned more sporadic and at a random packet number than malicious ones. Malicious traffic favourites short scans, and average days of appearance on malicious actors was below 50% of benign ones. Although some IPv4 scanning patterns clearly aligned with established malicious activities, others subtly mirrored benign intents, with malicious scans lasting longer than 30 days.

The findings in Chapter 4 challenge the existing practice and methods in the literature as defined in Durumeric et al. (2014). Their definition of a scan gave poor results for telescope data because it is unidirectional. This thesis is a good research contribution toward an approach to classification of network scans, especially with unidirectional traffic.

The significance of this research lies in its practical application to real-world network security management. By accurately classifying scanning activities, network administrators can prioritise security measures more effectively, ensuring resources are allocated to mitigate truly malicious threats rather than benign probes. This not only optimises network security investments, but also improves the overall resilience of network infrastructures against potential cyber attacks. The development of these scripts and algorithms ad-

dresses a crucial need in cyber security, as identified in the introductory chapter, providing a robust tool that aids in the subtle but critical task of distinguishing harmful network activities from harmless ones. This research contributes a nuanced perspective to the ongoing dialogue in network security, emphasising the complexity of network behaviours and the need for advanced analytical tools to navigate this terrain.

In addition to its practical applications, this research contributes to academic discourse by filling a notable literature gap. It offers a comprehensive overview of network scanning classification, enriching the body of knowledge with empirical data and theoretical insights. This makes it a cornerstone reference for current and future studies in the field, inspiring further research and development in the quest to secure digital infrastructures.

6.5 Future Work

Building on this foundational work, future researchers can explore several avenues:

- **Enhancing Scan Detection Algorithms**: Future work could refine the developed scripts and algorithms to enhance accuracy in distinguishing between benign and malicious scans. This refinement would involve incorporating analysis of bi-directional traffic, examining both incoming and outgoing network interactions. By analysing how data is sent and received, algorithms can better understand the context of traffic patterns and identify coordinated or suspicious activities that may not be apparent from unidirectional data alone. Additionally, integrating advanced machine learning techniques that adapt to changing behaviours of network threats will be critical, as these techniques can learn from a broader set of data inputs, including the temporal and spatial characteristics of bi-directional traffic, thereby improving the detection and classification of network scans as attack strategies continue to evolve.
- **Longitudinal Studies**: There is a significant opportunity to conduct further longitudinal studies to monitor the evolution of scanning tactics over time. Such studies would help in understanding how scanning strategies develop and change in response to network defences, which can be crucial for developing more dynamic and responsive security measures.
- **Geographical and Political Context**: Expanding the analysis to consider how geopolitical contexts influence scanning activities could provide insights into targeted cyber threats associated with specific regions. This extension would be particularly beneficial for global networks operating in politically sensitive environments, where cyber activities are often influenced by international relations.

This thesis, grounded in accurate data analysis and a comprehensive methodological framework, not only advances the understanding of network scanning activities, but also lays a robust foundation for future explorations in cyber security. This approach ensures that the work remains relevant and that future research can build directly on the estab-

lished findings and methodologies.

References

Anand, A., Kallitsis, M., Sippe, J., & Dainotti, A. (2023, May 11). *Aggressive Internet-Wide Scanners: Network Impact and Longitudinal Characterization*. arXiv: 2305.07193 [cs].

Aniello, L., Lodi, G., & Baldoni, R. (2011). Inter-domain stealthy port scan detection through complex event processing. *Proceedings of the 13th European Workshop on Dependable Computing*, 67–72. https://doi.org/10.1145/1978582.1978597

Barnett, R. J., & Irwin, B. (2008). Towards a taxonomy of network scanning techniques. *Proceedings of the 2008 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries: Riding the Wave of Technology*, 1–7. https://doi.org/10.1145/1456659.1456660

Bennett, C., Abdou, A., & Van Oorschot, P. C. (2021). Empirical Scanning Analysis of Censys and Shodan [Accessed: 2023-10-18]. *Workshop on Measurements, Attacks, and Defenses for the Web*. https://doi.org/10.14722/madweb.2021.23009

Bortoluzzi, F., Irwin, B., Beiler, L. S., & Westphall, C. M. (2023). Cloud telescope: A distributed architecture for capturing internet background radiation. *2023 IEEE 12th International Conference on Cloud Networking (CloudNet)*, 77–85. https://doi.org/10.1109/CloudNet59005.2023.10490018

Bortoluzzi, F., Irwin, B., & Westphall, C. M. (2023). A cloud-native framework for globally distributed capture and analysis of internet background radiation, 1–4. https://doi.org/10.23919/CISTI58278.2023.10211290

Bou-Harb, E., Debbabi, M., & Assi, C. (2013). Cyber scanning: A comprehensive survey. *IEEE communications Surveys & Tutorials*, *16*(3), 1496–1519. https://doi.org/10.1109/SURV.2013.102913.00020

Boulanger, A. (1998). Catapults and grappling hooks: The tools and techniques of information warfare [Accessed: 2023-10-16]. *IBM Systems Journal*, *37*(1), 106–114. https://doi.org/10.1147/sj.371.0106

Bruce, M., Lusthaus, J., Kashyap, R., Phair, N., & Varese, F. (2024). Mapping the global geography of cybercrime with the world cybercrime index. *PLOS ONE*, *19*(4), 1–16. https://doi.org/10.1371/journal.pone.0297312

Censys. (2023a). *About Censys - Censys* [Accessed: 2024-01-16]. https://about.censys.io/

Censys. (2023b, December). *Search pricing - Censys* [Accessed: 2024-01-16]. https://censys.com/search-pricing/

Costantino, G., & Matteucci, I. (2019). Candy cream - hacking infotainment android systems to command instrument cluster via can data frame. *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, (19), 476–481. https://doi.org/10.1109/CSE/EUC.2019.00094

CriminalIP. (2024, January). *Ai spera | criminal ip* [Accessed: 2024-03-18]. https://www.criminalip.io/about/aispera

Data Center Map. (2024). *Western europe data centers* [Accessed: 2024-04-20]. https://www.datacentermap.com/western-europe/

Durumeric, Z., Bailey, M., & Halderman, J. A. (2014). An Internet-Wide view of Internet-Wide scanning. *23rd USENIX Security Symposium (USENIX Security 14)*, 65–78. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/durumeric

Durumeric, Z., Wustrow, E., & Halderman, J. A. (2013). Zmap: Fast internet-wide scanning and its security applications, 605–620. https://www.usenix.org/system/files/conference/usenixsecurity13/sec13-paper_durumeric.pdf

Farmer, D., & Venema, W. Z. (2000). Improving the security of your site by breaking into it. https://www.dcs.ed.ac.uk/home/rah/Resources/Security/admin_guide_to_cracking.pdf

Ghiëtte, V. (2016). *Classifying scanners: Mapping their behaviour* [Master's thesis, Delft University of Technology]. http://resolver.tudelft.nl/uuid:8b842a9e-563e-4a44-b46f-becfb6e8af18

Graham, R. (2013, September). *Masscan: the entire Internet in 3 minutes* [Accessed: 2024-01-16]. https://blog.erratasec.com/2013/09/masscan-entire-internet-in-3-minutes.html

GreyNoise. (2023). *Understanding riot* [Accessed: 2024-04-22]. https://docs.greynoise.io/docs/riot-data

Greynoise. (2022). *Greynoise intelligence dives deep into the cybersecurity landscape with its 2022 mass exploitation report* [Accessed: 2024-01-23]. https://www.greynoise.io/press/greynoise-intelligence-cybersecurity-landscape-2022-mass-exploitation-report

Greynoise. (2023a). *Using the greynoise enterprise api* [Accessed: 2024-01-25]. https://docs.greynoise.io/docs/using-the-greynoise-api

Greynoise. (2023b). *Using the greynoise visualizer* [Accessed: 2024-01-25]. https://docs.greynoise.io/docs/using-the-greynoise-visualizer

Greynoise. (2024). *GreyNoise | Product and features* [Accessed: 2024-01-23]. https://www.greynoise.io/greynoise-product

Hassan Kilavo, S. I. M., & Dudu, R. G. (2023). Securing relational databases against security vulnerabilities: A case of microsoft sql server and postgresql. *Journal of Applied Security Research*, *18*(3), 421–435. https://doi.org/10.1080/19361610.2021.2006032

Hendricks, W. (2019). *An Analysis of Internet Background Radiation within an African IPv4 Netblock* [Master's thesis, Rhodes University]. http://hdl.handle.net/10962/103791

Higgins, M. (2011). *Julian assange: Wikileaks founder: Wiki leaks founder*. ABDO.

HP Wolf Security. (2019, June 11). *Mapping out a malware distribution network* [Accessed: 2024-03-04]. https://threatresearch.ext.hp.com/mapping-malware-distribution-network/

IANA. (2023). Service names and port numbers. *Internet Assigned Numbers Authority*. https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml

Internet Systems Consortium, Inc. (2024). *About us* [Accessed: 2024-02-01]. https://www.isc.org/about/

Irwin, B. (2013). A baseline study of potentially malicious activity across five network telescopes. *2013 Cyber Conflict (CyCon)*, *5*, 18 pages. https://doi.org/https://catalog.caida.org/paper/2013_1_irwin_b_researchgate_6547386

Irwin, B., & van Riel, J.-P. (2007). Inetvis: A graphical aid for the detection and visualisation of network scans. *Conference on Vizualisation Security (VizSec2007)*. https://doi.org/https://www.cs.ru.ac.za/research/g02v2468/publications/Irwin-VizSEC2007_draft.pdf

Irwin, B. (2011, January). *A framework for the application of network telescope sensors in a global IP network* [Doctoral dissertation, Rhodes University]. http://hdl.handle.net/10962/d1004835

Jafarian, J. H., Abolfathi, M., & Rahimian, M. (2023). Detecting Network Scanning Through Monitoring and Manipulation of DNS Traffic [Accessed: 2023-09-18]. *IEEE Access*, *11*, 20267–20283. https://doi.org/10.1109/ACCESS.2023.3250106

Jagamogan, R. S., Ismail, S. A., Hassan, N. H., & Abas, H. (2022). Penetration testing procedure using machine learning. *2022 4th International Conference on Smart Sensors and Application (ICSSA)*, 58–63. https://doi.org/10.1109/ICSSA54161.2022.9870951

Jayasuryapal, G., Pranay, P. M., Kaur, H., et al. (2021). A survey on network penetration testing. *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*, 373–378. https://doi.org/10.1109/ICIEM51511.2021.9445321

Jenani, M. (2017). Network security, a challenge [Special Issue: TECHSA-17, Government Arts College for Women, Ramanathapuram]. *International Journal of Advanced Networking & Applications*, *08*(05), 120–123. https://www.ijana.in/Special%20Issue/TPID28.pdf

Kaspersky. (2016). *What are IOT search engines shodan and censys and what are they capable of* [Accessed: 2024-01-16]. https://www.kaspersky.com/blog/shodan-censys/11430/

Khan, H. M. A., Inayat, U., Zia, M. F., Ali, F., Jabeen, T., & Ali, S. M. (2021). Voice over internet protocol: Vulnerabilities and assessments. *2021 International Conference on Innovative Computing (ICIC)*, 1–6.

Khan, M. T., DeBlasio, J., Voelker, G. M., Snoeren, A. C., Kanich, C., & Vallina-Rodriguez, N. (2018). An empirical analysis of the commercial vpn ecosystem. *Proceedings of the Internet Measurement Conference 2018*, 443–456. https://doi.org/10.1145/3278532.3278570

Malvuln. (2022, February). *MVID-2022-0499* [Accessed: 2024-03-01]. https://malvuln.com/advisory/584bc06128469423f9e50e8a359d18ac_B.txt

Marín, G., Caasas, P., & Capdehourat, G. (2021). Deepmal - deep learning models for malware traffic detection and classification. In P. Haber, T. Lampoltshammer, M. Mayr, & K. Plankensteiner (Eds.), *Data science – analytics and applications* (pp. 105–112). Springer Fachmedien Wiesbaden. https://doi.org/https://doi.org/10.1007/978-3-658-32182-6_16

Marksteiner, S., Jandl-Scherf, B., & Lernbeiß, H. (2020). Automatically Determining a Network Reconnaissance Scope Using Passive Scanning Techniques [Accessed: 2023-09-18]. In X.-S. Yang, S. Sherratt, N. Dey, & A. Joshi (Eds.), *Fourth International Congress on Information and Communication Technology* (pp. 117–127, Vol. 1027). Springer Singapore. https://doi.org/10.1007/978-981-32-9343-4_11

Mazel, J., Fontugne, R., & Fukuda, K. (2017). Profiling internet scanners: Spatiotemporal structures and measurement ethics. *2017 Network Traffic Measurement and Analysis Conference (TMA)*, 1–9. https://doi.org/10.23919/TMA.2017.8002909

Moore, D., Shannon, C., Voelker, G., & Savage, S. (2004, June). *Network Telescopes: Technical Report* (tech. rep.). Cooperative Association for Internet Data Analysis (CAIDA). https://catalog.caida.org/paper/2004_tr_2004_04

Nkhumeleni, T. M. (2014). *Correlation and comparative analysis of traffic across five network telescopes* [Master's thesis, Rhodes University]. http://hdl.handle.net/10962/d1011668

NVD - CVE. (2017, March). *Nvd - cve-2017-17215* [Accessed: 2024-03-04]. https://nvd.nist.gov/vuln/detail/CVE-2017-17215

NVD - CVE. (2023, June). *Cve-2023-27997* [Accessed: 2024-03-01]. https://nvd.nist.gov/vuln/detail/CVE-2023-27997

Ongun, T., Spohngellert, O., Miller, B., Boboila, S., Oprea, A., Eliassi-Rad, T., Hiser, J., Nottingham, A., Davidson, J., & Veeraraghavan, M. (2021). Portfiler: Port-level network profiling for self-propagating malware detection. *2021 IEEE Conference on Communications and Network Security (CNS)*, 182–190.

Orebaugh, A., & Pinkard, B. (2011). *Nmap in the enterprise: Your guide to network scanning*. Elsevier. https://doi.org/10.1016/B978-1-59749-241-6.X0001-5

Pearson, D. T., & Irwin, B. (2018, August). *OpenDTI: A threat intelligence collation and dissemination framework* (tech. rep.). https://doi.org/10.13140/RG.2.2.30907.18729

Postel, J. (1981, September). *Internet Protocol* (tech. rep. No. 791). https://doi.org/10.17487/RFC0791

Postel, J., & Reynolds, J. (1992, July). *Assigned Numbers* (tech. rep. No. 1340). https://doi.org/10.17487/RFC1340

Putu, I., Pratama, A. E., Agus, P., & Pratama, E. (2020). Tcp syn flood (dos) attack prevention using spi method on csf: A poc. *Bulletin of Computer Science and Electrical Engineering*, *1*(2), 63–72. https://doi.org/https://doi.org/10.25008/bcsee.v1i2.7

Quittner, J. (1995, April). *The Devil In The Network* [Accessed: 2023-11-02]. https://content.time.com/time/subscriber/article/0,33009,982835,00.html

Raikar, M. M., & Meena, S. (2021). Ssh brute force attack mitigation in internet of things (iot) network: An edge device security measure. *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)*, 72–77. https://doi.org/doi={10.1109/ICSCCC51823.2021.9478131}

Rapid7. (2017, August). *Scanning All The Things* [Accessed: 2024-01-16]. https://www.rapid7.com/blog/post/2013/09/26/internet-wide-probing-rapid7-sonar/

Rapid7. (2019). *Attack surface security monitoring with insightvm and project sonar* [Accessed: 2024-01-16]. https://www.rapid7.com

Rapid7. (2023). *About open data | Rapid7* [Accessed: 2023-10-23]. https://opendata.rapid7.com/about/

Richter, P., & Berger, A. (2019). Scanning the Scanners: Sensing the Internet from a Massively Distributed Network Telescope [Accessed: 2023-09-14]. *Proceedings of the Internet Measurement Conference*, 144–157. https://doi.org/10.1145/3355369.3355595

Rudman, L., & Irwin, B. (2015). Characterization and analysis of ntp amplification based ddos attacks. *2015 Information Security for South Africa (ISSA)*, (14), 1–5. https://doi.org/10.1109/ISSA.2015.7335069

Shadowserver Foundation. (2024). *Who we are | The Shadowserver Foundation* [Accessed: 2024-01-16]. https://www.shadowserver.org/who-we-are/

Shodan. (2022). *What is Shodan? - Shodan Help Center* [Accessed: 2024-01-16]. https://help.shodan.io/the-basics/what-is-shodan

Snehi, M., & Bhandari, A. (2021). Apprehending mirai botnet philosophy and smart learning models for iot-ddos detection. *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, 501–505.

Susanto, C. O. N., Rizko, K. N. F., & Purbohadi, D. (2020). Security assessment using nessus tool to determine security gaps on the repository web application in educational institutions. *Emerging Information Science and Technology*, *1*(2), 58–62.

Tanakas, P., Ilias, A., & Polemi, N. (2021). A novel system for detecting and preventing sql injection and cross-site-script. *2021 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, 1–6.

Trapkickin, R. (2015). Who is scanning the internet. *FI & IITM SS 15*, 81–88. https://doi.org/10.2313/NET-2015-09-1_11

University of Michigan. (2018). *Censys, mapping the internet one device at a time* [Accessed: 2024-01-16]. https://innovationpartnerships.umich.edu/stories/censys-mapping-the-internet-one-device-at-a-time/

Upadhya, A. (2020). A survey on different port scanning methods and the tools used to perform them. *IJRASET*, *8*, 3018–3024. https://doi.org/10.22214/ijraset.2020.5505

van Riel, J.-P., & Irwin, B. (2006). Inetvis, a visual tool for network telescope traffic analysis. *4th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa*, 85–89. https://doi.org/10.1145/1108590.1108604

Wan, G., Izhikevich, L., Adrian, D., Yoshioka, K., Holz, R., Rossow, C., & Durumeric, Z. (2020). On the Origin of Scanning: The Impact of Location on Internet-Wide Scans [Accessed: 2023-09-14]. *Proceedings of the ACM Internet Measurement Conference*, 662–679. https://doi.org/10.1145/3419394.3424214

Wang, Y., Wen, S., Xiang, Y., & Zhou, W. (2013). Modeling the propagation of worms in networks: A survey. *IEEE Communications Surveys & Tutorials*, *16*(2), 942–960.

Wu, P., Cui, Y., Wu, J., Liu, J., & Metz, C. (2013, Autumn). Transition from IPv4 to IPv6: A State-of-the-Art Survey [Accessed: 2023-09-18]. *IEEE Communications Surveys & Tutorials*, *15*(3), 1407–1424. https://doi.org/10.1109/SURV.2012.110112.00200

Yamada, R., & Goto, S. (2013). Using abnormal TTL values to detect malicious IP packets [Accessed: 2023-09-21]. *Proceedings of the Asia-Pacific Advanced Network*, *34*, 27. https://doi.org/10.7125/APAN.34.4

APPENDIX A

Monthly Protocol overview

Appendix A provides a comprehensive monthly breakdown of protocol usage throughout the year 2023, detailing the distribution of network packets by type across three primary protocols. TCP (Transmission Control Protocol), UDP (User Datagram Protocol), and ICMP (Internet Control Message Protocol). This data is summarised in a series of tables, from Table A.1 to Table A.12, for each month from January to December, respectively.

Each table presents the total count of packets for each protocol and their respective percentages of the overall network traffic for that month. This information is critical for understanding the composition of the traffic and can help in analysing trends, planning network capacity, or assessing security protocols.

This appendix serves as a resource for 3.7.3.

Table A.1: Protocol Overview for January 2023

Protocol	Packet Count	Percentage (%)
TCP	40,346,775	89.99
UDP	3,293,002	7.34
ICMP	1,196,247	2.67

Table A.2: Protocol Overview for February 2023

Protocol	Packet Count	Percentage (%)
TCP	43,644,786	90.58
UDP	3,455,070	7.17
ICMP	1,084,032	2.25

Table A.3: Protocol Overview for March 2023

Protocol	Packet Count	Percentage (%)
TCP	58,568,842	92.51
UDP	3,527,967	5.57
ICMP	1,215,482	1.92

Table A.4: Protocol Overview for April 2023

Protocol	Packet Count	Percentage (%)
TCP	48,708,500	91.25
UDP	3,502,966	6.56
ICMP	1,166,087	2.18

Table A.5: Protocol Overview for May 2023

Protocol	Packet Count	Percentage (%)
TCP	48,402,391	91.26
UDP	3,550,373	6.69
ICMP	1,084,613	2.04

Table A.6: Protocol Overview for June 2023

Protocol	Packet Count	Percentage (%)
TCP	44,641,083	90.44
UDP	3,604,743	7.30
ICMP	1,114,225	2.26

Table A.7: Protocol Overview for July 2023

Protocol	Packet Count	Percentage (%)
TCP	60,704,993	93.04
UDP	3,479,186	5.33
ICMP	1,063,455	1.63

Table A.8: Protocol Overview for August 2023

Protocol	Packet Count	Percentage (%)
TCP	62,743,313	92.38
UDP	4,039,140	5.95
ICMP	1,133,912	1.67

Table A.9: Protocol Overview for September 2023

Protocol	Packet Count	Percentage (%)
TCP	52,217,869	90.68
UDP	4,242,655	7.37
ICMP	1,121,287	1.95

Table A.10: Protocol Overview for October 2023

Protocol	Packet Count	Percentage (%)
TCP	56,745,906	92.49
UDP	3,517,252	5.73
ICMP	1,088,453	1.77

Table A.11: Protocol Overview for November 2023

Protocol	Packet Count	Percentage (%)
TCP	44,018,239	90.89
UDP	3,413,006	7.05
ICMP	1,000,854	2.07

Table A.12: Protocol Overview for December 2023

Protocol	Packet Count	Percentage (%)
TCP	33,057,083	91.51
UDP	2,364,660	6.55
ICMP	704,121	1.95

APPENDIX B

Monthly Port Data Overview

Appendix B presents a detailed overview of the activity of the network port for each month of a specific year, providing a deep dive into the percentage of traffic and unique source interactions associated with various ports. This appendix is structured as a series of tables, each corresponding to a different month, from January to December 2023. Each table is formatted to show the percentage of traffic and the percentage of unique sources communicating through the most significant ports.

The data in these tables are essential for understanding the distribution of network traffic within the network telescope across different ports over time. It helps to deepen our understanding of the size and structure of the network data with which this research works.

Table B.1: Month: January

Port	Traffic (%)	Unique Sources (%)
23	14.73	3.69
22	2.69	3.24
80	2.16	5.71
8080	1.63	5.19
61953	1.30	0.41
3389	1.23	2.86
443	1.17	2.48
445	1.17	7.66
5555	1.16	2.45
81	0.90	4.52

Table B.2: Month: February

Port	Traffic (%)	Unique Sources (%)
23	11.17	3.92
57454	5.31	0.02
56246	5.31	0.02
80	2.00	4.93
22	1.84	3.50
5555	1.50	1.68
8080	1.27	5.83
3389	1.19	1.25
443	1.15	2.09
445	1.08	7.31

Table B.3: Month: March

Port	Traffic (%)	Unique Sources (%)
53994	22.65	0.00
23	8.59	4.37
46870	1.82	0.03
80	1.65	5.14
22	1.42	3.25
443	1.11	1.99
8080	0.86	7.02
445	0.82	7.89
45568	0.81	0.08
47513	0.81	0.09

Table B.4: Month: April

Port	Traffic (%)	Unique Sources (%)
23	12.80	3.54
46344	5.31	0.01
46334	4.77	0.02
46870	2.37	0.03
80	1.93	4.93
22	1.77	2.66
60702	1.34	0.05
443	0.96	2.82
445	0.95	7.51
1024	0.80	1.63

Table B.5: Month: May

Port	Traffic (%)	Unique Sources (%)
23	10.19	4.77
40045	7.23	0.02
40069	4.04	0.02
22	1.79	2.63
80	1.46	6.02
443	0.91	3.01
60000	0.90	0.50
445	0.87	8.23
5555	0.87	3.05
3389	0.76	1.64

Table B.6: Month: June

Port	Traffic (%)	Unique Sources (%)
23	8.88	6.38
51382	4.94	0.04
22	1.57	2.16
80	1.55	5.71
8080	1.11	6.79
443	1.01	3.72
445	0.89	8.44
60000	0.88	0.64
5555	0.88	3.67
3389	0.86	1.33

Table B.7: Month: July

Port	Traffic (%)	Unique Sources (%)
23	7.02	6.96
41443	3.33	0.02
41168	2.04	0.03
54436	1.96	0.07
53633	1.95	0.10
52874	1.95	0.07
54752	1.95	0.07
54929	1.95	0.07
54393	1.95	0.07
55664	1.95	0.07

Table B.8: Month: August

Port	Traffic (%)	Unique Sources (%)
58870	16.23	0.00
23	7.66	5.42
40474	1.89	0.04
80	1.31	5.13
22	1.00	3.15
58914	0.96	4.46
8080	0.88	5.28
443	0.85	3.37
3389	0.84	1.18
45845	0.76	0.19

Table B.9: Month: September

Port	Traffic (%)	Unique Sources (%)
23	10.15	4.41
80	1.71	4.78
58914	1.65	4.55
22	1.24	3.38
443	1.08	4.04
8080	1.02	4.53
3389	0.96	0.98
445	0.71	8.61
43051	0.65	0.15
46031	0.65	0.14

Table B.10: Month: October

Port	Traffic (%)	Unique Sources (%)
23	13.47	3.30
50203	4.17	0.05
443	1.93	1.69
42665	1.83	0.04
22	1.62	3.14
80	1.38	4.43
58914	1.37	1.58
20824	1.00	0.02
8080	0.94	5.76
3389	0.79	1.00

Table B.11: Month: November

Port	Traffic (%)	Unique Sources (%)
23	11.75	5.19
22	1.93	1.85
58914	1.64	3.28
80	1.39	5.40
443	1.11	3.05
8080	1.06	4.78
8728	0.99	0.07
3389	0.96	1.19
445	0.77	9.18
50897	0.72	0.33

Table B.12: Month: December

Port	Traffic (%)	Unique Sources (%)
23	8.22	7.09
50289	3.77	0.05
22	1.58	2.51
80	1.28	6.05
8080	1.02	5.32
3389	1.00	1.42
52838	0.96	0.22
443	0.96	3.97
445	0.72	9.68
58914	0.67	13.42

APPENDIX C

InetVis

This appendix serves as a complimentary to figures 4.7 and 4.8 in Section 2.8.3. Figures C.1 and C.2 are different angles of the same view plot of isolated netblocks.

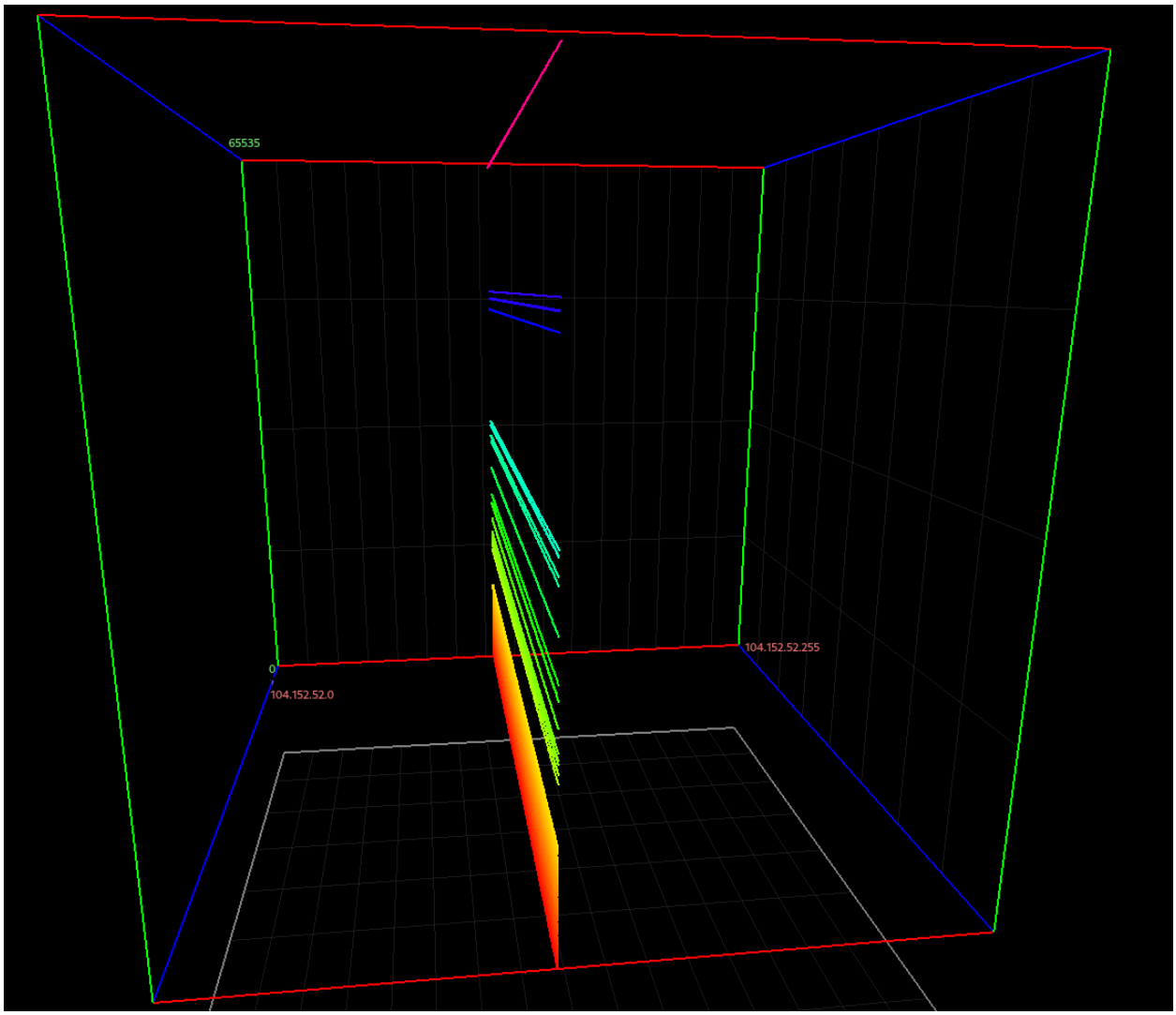

Figure C.1: InetVis view plot of network scans from ’104.152.52.0/24’

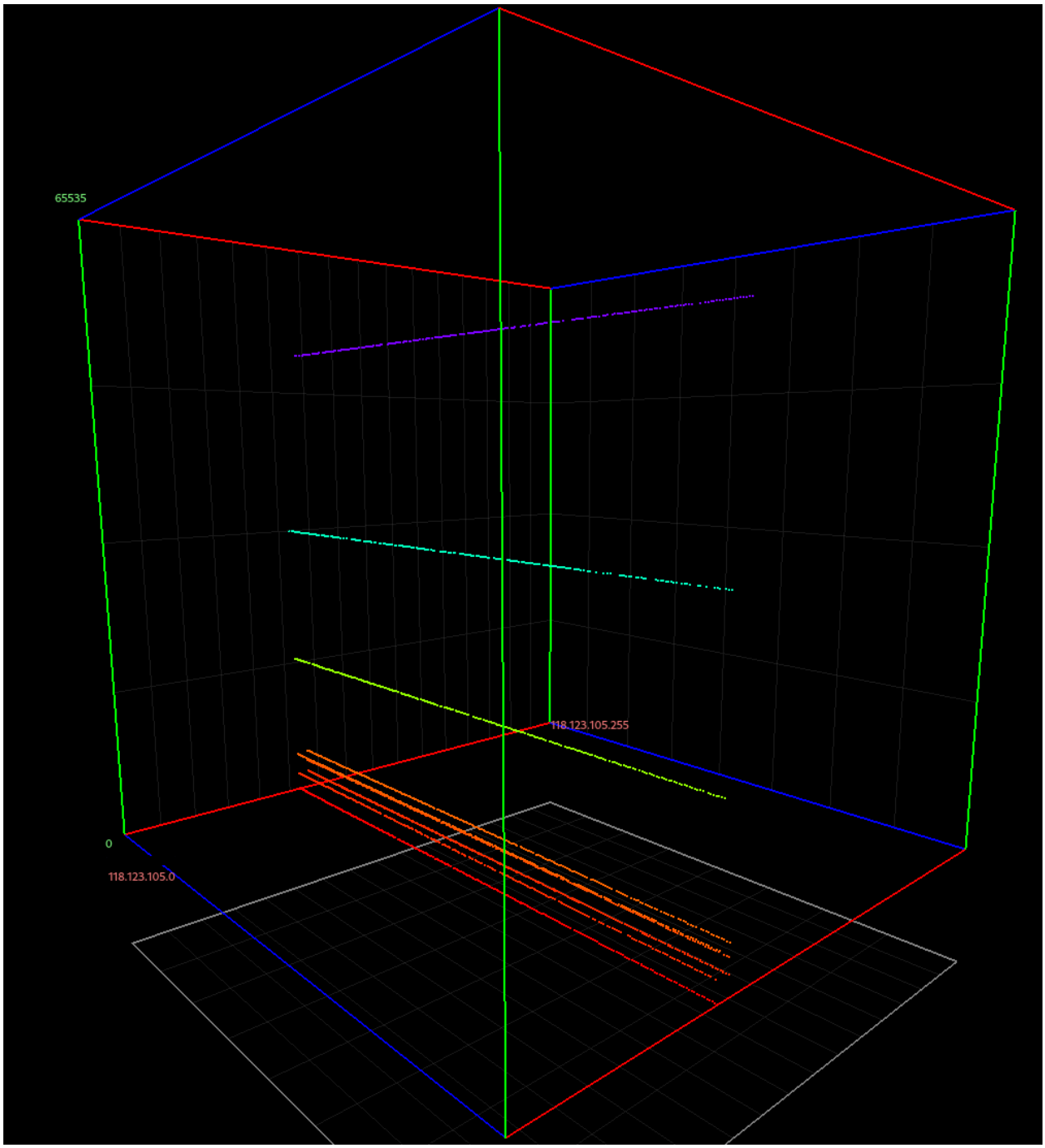

Figure C.2: InetVis view plot of network scans from ’118.123.105.0/24’

APPENDIX D

30 days scan benign IPs

Attached a link to the GitHub project. 'ip,actor.txt' contains IP addresses associated with various research organisations. The list includes 1158 IPs that are potentially linked to benign organisations. This comprehensive list serves as an addition to Table 5.8 in the research. It is important to note that these IPs are not fully verified to belong to the specified entities; they were sourced from greynoise.io.

https://github.com/neburdotcom/skjelstad_bachelor_2024NUC

APPENDIX E

Geographical heatmaps from benign and malicious actors

Appendix E offers a visual exploration of geographical data, presenting heat maps that illustrate the origin of network traffic classified as benign or malicious. This appendix is divided into two subsections, each containing a series of figures that represent different time frames, ranging from 24-hour periods to 30-day intervals.

This appendix serves as a valuable additional in-depth resource for Section 5.7.

E.1 Benign actors heatmaps

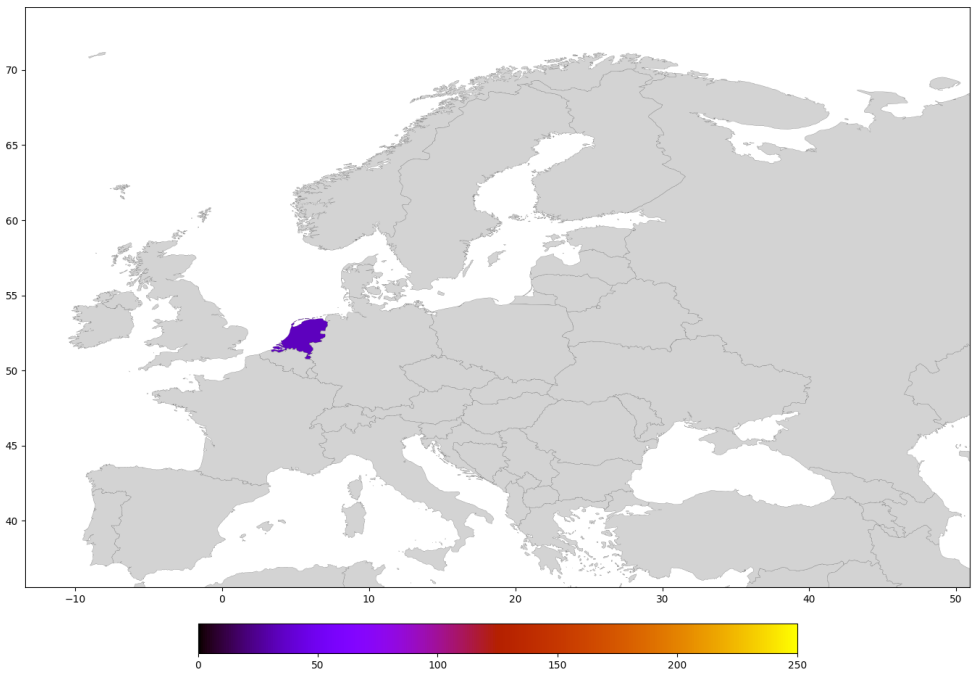

Figure E.1: Origin countries: Durumeric method Scan benign actors from January 2023 through December 2023

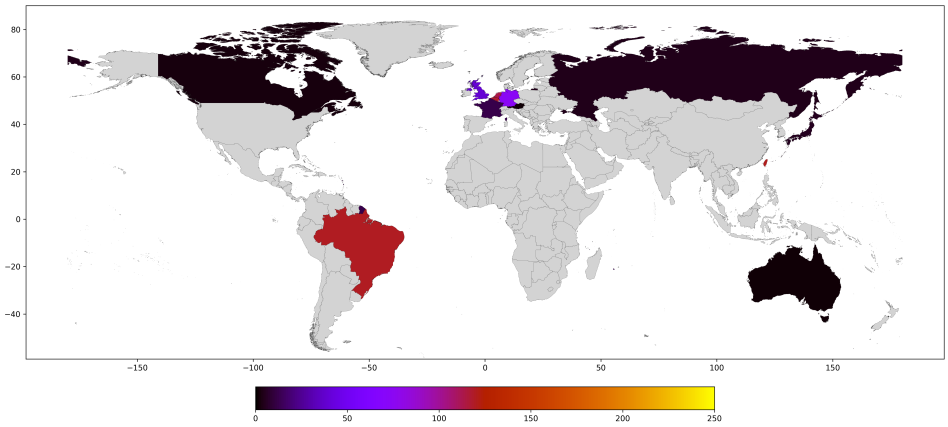

Figure E.2: Origin countries: 24 hours benign actors from January 2023 through December 2023

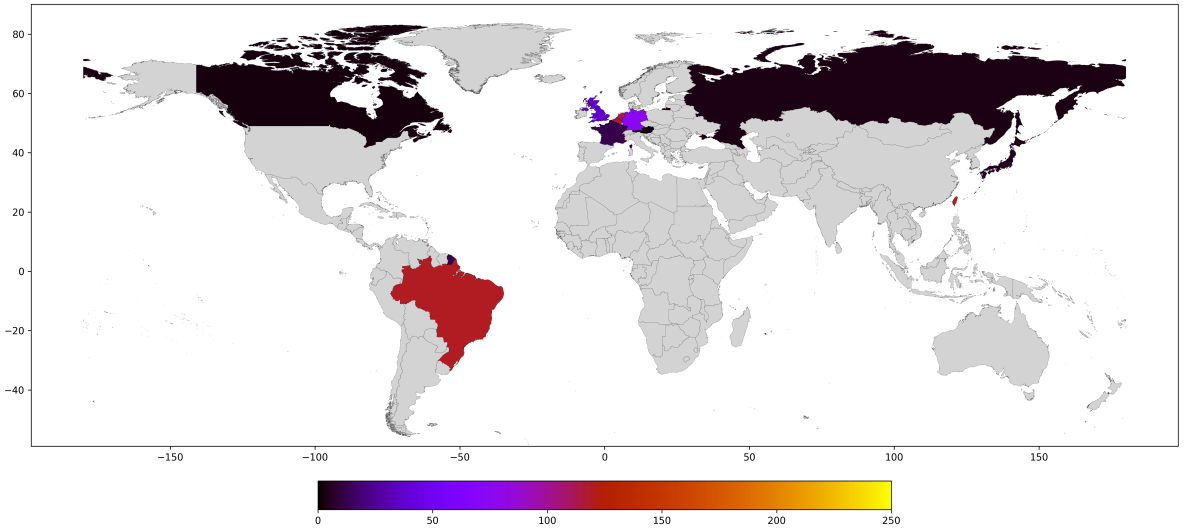

Figure E.3: 10 day scan benign actors origin countries from January 2023 through December 2023

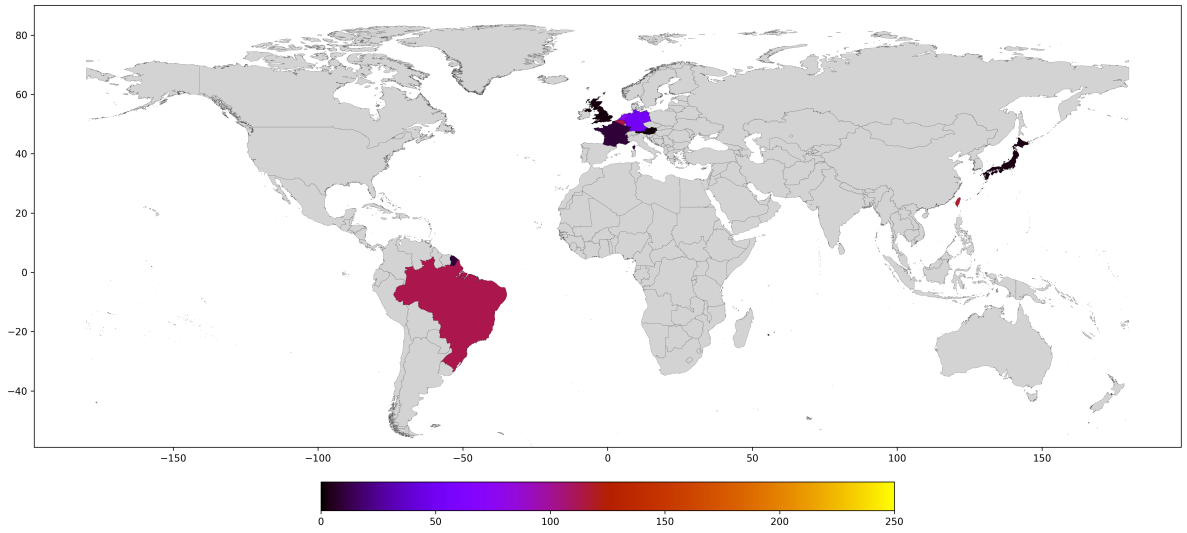

Figure E.4: 30 day scans benign actors origin countries from January 2023 through December 2023

E.2 Malicious actors heatmaps

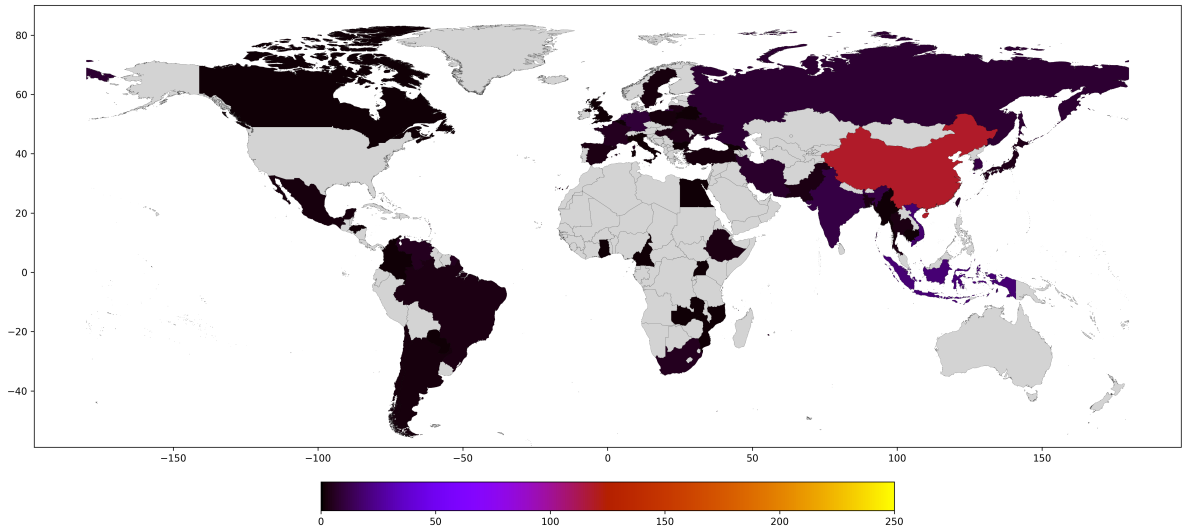

Figure E.5: Main scan malicious actors origin countries from January 2023 through December 2023

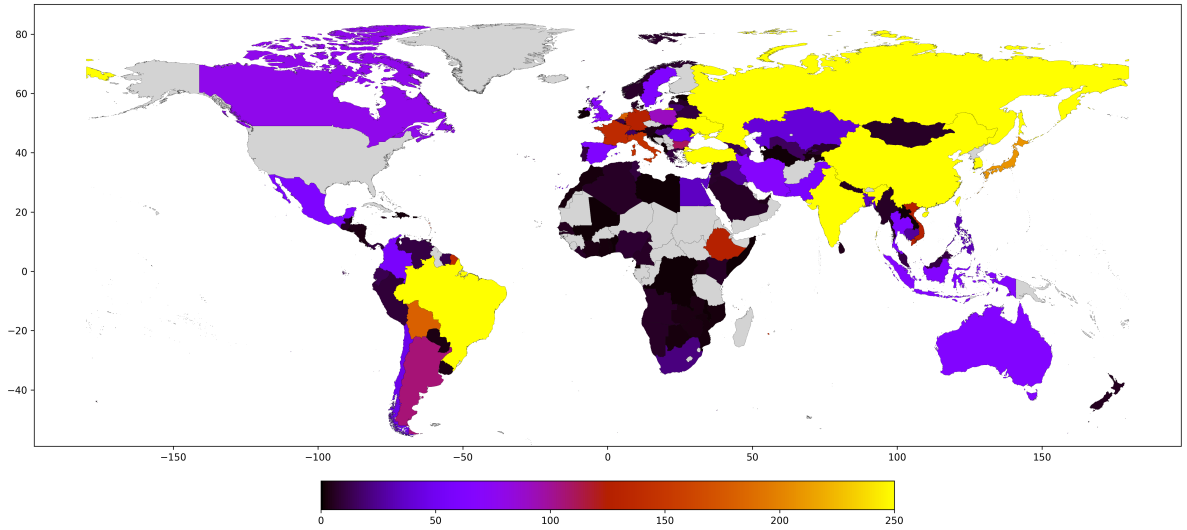

Figure E.6: 24 hours malicious actors origin countries from January 2023 through December 2023

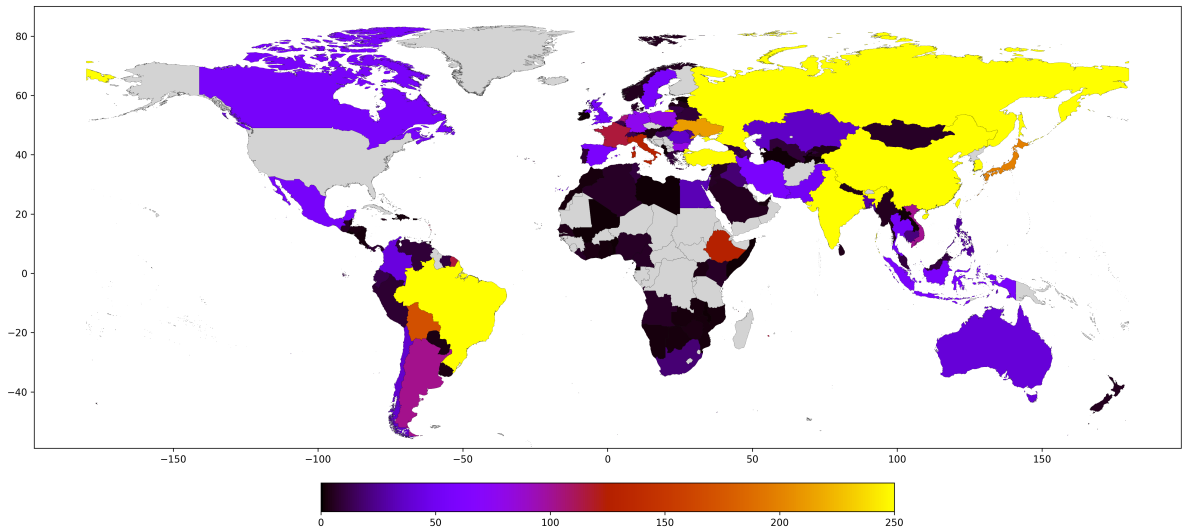

Figure E.7: 10 days scan malicious actors origin countries from January 2023 through December 2023

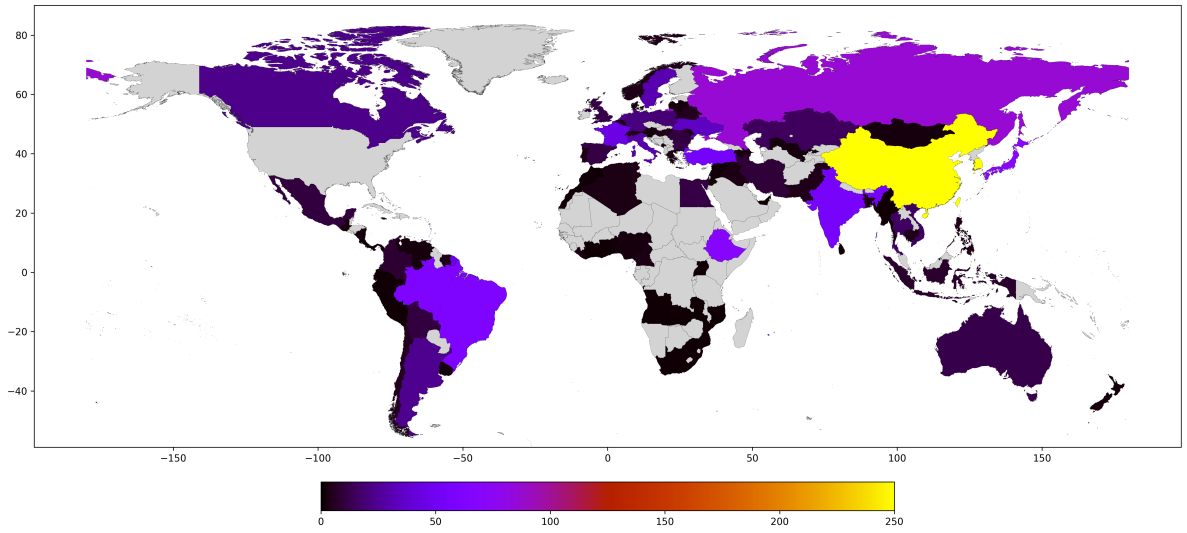

Figure E.8: 30 day scan malicious actors origin countries from January 2023 through December 2023

APPENDIX F

Scripts

Appendix F provides access to practical resources, specifically the scripts mentioned in Section 3.6, and applied in Chapter 4 and Chapter 5. This appendix contains a direct link to a GitHub repository where the script can be downloaded and used.

The inclusion of this script is intended to supplement the theoretical and analytical discussions presented in Chapters 4 and 5 with a practical tool that can be directly applied by researchers. By providing this script, the document facilitates the implementation of the concepts and techniques discussed in Chapter 3, allowing the user to engage with the material in a hands-on manner.

The repository contains two Python scripts: one implementing the Durumeric method and another designed for the four case studies, showcasing practical applications of the methodologies discussed. The CSV output of the scripts can also be found in the GitHub.

https://github.com/neburdotcom/skjelstad_bachelor_2024NUC

Word count metrics

NUC Bachelor Project Word Count:
Total Sum count: 26257 Words in text: 25115 Words in headers: 395 Words outside text (captions, etc.): 734 Number of headers: 142 Number of floats/tables/figures: 61 Number of math inlines: 10 Number of math displayed: 3
NOTE: References are excluded.