The Good, the Bad, and the Scanned:

An Empirical Study of the Origins of Internet-wide Scanners

Georgios Koursiounis

Delft University of Technology

The Good, the Bad, and the Scanned:

An Empirical Study of the Origins of Internet-wide Scanners

by

Georgios Koursiounis

to obtain the degree of Master of Science
Software Technology Track
with a 4TU specialization in Cyber Security

at the Delft University of Technology,
to be defended publicly on Thursday June 13, 2024 at 14:00.

Student number:	5730511	
Project duration:	November 10, 2023 – June 13, 2024	
Thesis committee:	Prof.dr. G. Smaragdakis,	TU Delft, chair
	Dr. ir. H. J. Griffioen,	TU Delft, daily supervisor
	Dr. A. Katsifodimos,	TU Delft

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

The cover page was created with the help of Image Creator
by Microsoft Designer.

Preface

”The more I learn, the more I realize how much I don’t know.”
— Albert Einstein

This thesis project is carried out as part of the fulfillment obligations for the Master Computer Science withing the Cybersecurity group at the Faculty of Electrical Engineering, Mathematics, and Computer Science (EEMCS) of Delft University of Technology. Internet Measurements for Cybersecurity and Cyber Threat Intelligence (CTI) are two vivid and vast research topics with continuous evolvement and impact in understanding the cybersecurity landscape.

Throughout my Master’s, I put a high effort to follow diverse cybersecurity courses, score high and accumulate knowledge that would help me obtain a better understanding on the cybersecurity domain. Yet, during this thesis, I realized the truthfulness of an Albert Einstein’s quote (also discussed by Montesquieu): the more I learn, the more I realize how much I don’t know (for cybersecurity) and how much more I can/need to learn further. This thesis serves as the capstone of the Master programme.

For that reason, I would like to thank my advisors, George Smaragdakis and Harm Griffioen for offering me this opportunity, as well for their guidance and support throughout the project, which helped me grow personally and professionally. Also, I would like to thank my parents and all my friends for their support. Finally, I would like to thank the readers for taking the time to read this report.

Georgios Koursiounis
Delft, June 2024

Summary

Security researchers and industry firms employ Internet-wide scanning for information collection, vulnerability detection and security evaluation, while cybercriminals make use of it to find and attack unsecured devices. Internet scanning plays a considerable role in threat detection & response, and cyber threat intelligence. We adopt a data-driven approach, analyzing a large dataset of network traffic collected through a network telescope, to identify the origins of Internet scanners and their affiliations. We provide a traffic analysis of two monthly snapshots in two different years (2023 & 2024) of approximately 10 billion packets each. We also provide a methodology for data collection and aggregation of known/institutional scanners.

The study reveals that a small number of source IP addresses account for almost the entire portion of traffic volume, with 1% of total addresses contributing 97.38% of total traffic in June 2023 and 96.65% in February 2024. Traffic analysis identifies 40 to 44 known scanners, accounting for 0.36 to 0.62% of source IPs and 50.86 to 51.31% of total telescope traffic in each month. However, seven to ten organizations are responsible for around half of the total telescope traffic each month. The study also identifies 34 commercial bots, with a negligible footprint, accounting for up to 0.25% of total source IPs and less than 0.01% of total traffic per month. Mirai probes contribute 1 to 1.5% of monthly scanning traffic, with a burst in IP addresses in 2023. Similarly, traffic from Tor exit nodes appears small, constituting 0.01% of overall Darknet traffic and 0.04-0.06% of source IPs per month. The study also reports on the current usage of scanning software such as ZMap and Masscan, finding that around 40% of each monthly traffic volume contains the ZMap signature. Lastly, we highlight the further need for mutual exchange of threat intelligence among defenders, as well as the extension of data collection period and the establishment of a pipeline for continuous discovery and integration of known scanners from a research perspective, in order to efficiently differentiate institutional scanners and malicious actors, within an evolving cyber landscape.

Contents

Nomenclature

Abbreviations

Abbreviations	
AS	Autonomous System
ASN	Autonomous System Number
BGP	Border Gateway Protocol
DNS	Domain Name System
HTTP	Hypertext Transfer Protocol
IP	Internet Protocol
pps	Packets per second
TCP	Transmission Control Protocol
TTL	Time To Live
DHCP	Dynamic Host Configuration Protocol
CVE	Common Vulnerabilities and Exposures
SOC	Security Operations Center
FTP	File Transfer Protocol
UDP	User Datagram Protocol
ISP	Internet Service Provider
CIDR	Classless Inter-Domain Routing
SSH	Secure Shell Protocol
JARM	JWT-Secured Authorization Response Mode
IANA	Internet Assigned Numbers Authority

1 Introduction

1.1. Motivation

Over the past years, there has been a considerable increase in intensive Internet-wide scanning operations, mostly as a result of two developments: the creation of research scanning tools and the proliferation of botnets and malware targeting Internet hosts [3]. Internet scanning is a long-established technique serving both benevolent and malevolent purposes. On the one hand, it is used by security researchers to gather important information on the structure and behavior of the Internet and identify new types of vulnerabilities. It is also employed by security professionals and the industry to discover publicly exposed infrastructure and evaluate the security posture of services, and networks. On the other hand, it is used by cyber criminals to identify and exploit vulnerable services. Since the scanning technique is dependent on actions taken beyond the purview of security defenses and cannot be effectively mitigated by preventive controls, therefore significant threats to the security of online entities arise.

The motivation for studying Internet-wide scanning is enhanced by its critical role in Threat Detection and Response (TDR) and Cyber Threat Intelligence (CTI). Early stages of cyber attacks often involve scanning. Active reconnaissance scans can be carried out by adversaries to obtain data for potential targets. The collected data can provide opportunities for additional reconnaissance (e.g., searching open domains or websites), the establishment of operational resources (e.g., developing or obtaining capabilities), and/or initial access (e.g., using external remote services or exploiting public-facing applications). Internet-wide scanning provides an overview of global network activity and identifies trends, abnormalities and potential threats. Detecting and analysing scanning traffic early enables the security industry and researchers to detect and address vulnerabilities before they are exploited. Understanding these activities can lead to the early identification of threats, allowing for proactive measures to prevent or mitigate potential harm. Furthermore, analyzing scanning data enhances Threat Intelligence, enabling security experts to stay informed about the tactics employed by cyber criminals and facilitating the development of effective defense strategies. Furthermore, Blue teams - which are in charge of protecting an organization's information systems and preserving its security posture - can gain insights into adversary tactics, techniques and procedures (TTPs) [7, 5, 6]. Study of Internet scanning helps organizations comprehend their shadow IT by exposing shadow devices, services and equipment on the network. This allows organizations to perform risk assessment by identifying publicly exposed services and weak points, enforce more effective security policies and reduce their attack surface.

The emerging challenge of this study lies in understanding the origins of Internet scanners and distinguishing them based on affiliations; whether they be research organizations, companies, or cyber criminals. The proposed research aims to investigate a sizable dataset of Internet scanning probes collected via a network telescope. Network telescopes observe and collect Internet background radiation (IBR) i.e. Internet traffic sent to a routed but unused address space [53]. This is commonly referred to as the *Darknet* address space. Darknets are an important source of information for security communities and can offer a worldwide view on Internet behavior.

1.2. Research Questions

This study adopts a data-driven approach to analyze the Internet scanning probes collected by a network telescope. We aim to answer the following research question:

Research Question

Who scans the Internet, how can we classify Internet scanners based on their origin and what are the discernible differences between scanning activities conducted by malicious actors and those associated with research organizations?

We divide the main research question into four sub-questions:

Q1. ***Can scanning probes be classified into specific categories based on their originating IP addresses, and if so, what are these categories?*** To classify the scanning probes into groups we take into consideration multiple parameters: we fingerprint traffic for known scanning tools such as Masscan and ZMap, identify scanning traffic from botnets such as Mirai, collect and compile a list of known/recurring scanners and map IP addresses to other academic establishments/research institutes, non-profit organizations and the security industry.

Q2. ***Are there identifiable differences in the scanning methodologies and techniques used by malicious actors compared to those used by legitimate research entities?*** We extract general statistics based on TCP and IP packet headers, analyse traffic per day, amount of traffic per source port and destination port, most popular destination ports (in whole and per day), number of distinct destination ports scanned within a day, number of distinct destination addresses scanned within a day, number of distinct destination IPs scanned in total, correlate destination ports with known services and Common Vulnerabilities and Exposures (CVEs), perform reverse DNS and identify aggressive hitters in terms of address dispersion, packet volume and number of distinct ports.

Q3. ***What are the geographic distributions of scanners, and can geographic patterns help distinguish between malicious and non-malicious scanning activities?*** In order to answer this research question, we investigate Autonomous Systems (ASes), analyse Internet Bad Neighborhoods (*BadHoods*), compute density and volume ratio of network prefixes, map the scanners to geolocation data to city and country level in terms of amount of scanners and total traffic volume.

Q4. ***To what extent can historical data and datasets from disperse source nodes be used to predict the intent behind a scanning operation, whether research-driven or malicious?*** We make use of external open-source, historical and proprietary datasets to analyse the scanning traffic. Datasets include known (acknowledged) scanners, IP blacklists, labelled datasets in terms of identified actors etc.

The research encompasses a multifaceted approach; *out-in* and in-out analysis. Commencing with the *out-in* analysis, this phase entails several procedures such as geolocation lookups, analysis of network prefixes and Autonomous Systems (AS), DNS reverse lookups, identification of aggressive hitters and known scanners. Subsequently, we transition into an in-out investigative phase where we reverse the focus by actively collecting recent and historical data about the scanners from external sources and archived datasets.

1.3. Contribution

The study aims to offer valuable insights into early detection and mitigation of cyber threats, supporting cybersecurity efforts by highlighting variations between scanning patterns of malicious actors and those of research institutions and industry firms. We make the following contributions:

- We provide a traffic analysis of two monthly snapshots in two different years (2023 & 2024) of approximately 10 billion packets each.
- We provide an up-to-date analysis for the modus operandi of scanning actors in terms of port scan strategy, address dispersion, ports targeted etc.

- We provide a methodology for data collection and aggregation of known scanner data, enriching the current datasets from literature and reporting up-to-date trends.
- We show that a small number of source IP addresses disproportionately accounts for almost the whole portion of traffic volume .i.e. 1% of total addresses accounts for 97.38% of the total traffic in June 2023 and 96.65% in February 2024, following the Pareto principle.
- We identify 44 (40 in Feb) known scanners that correspond to 0.36% (0.62% in Feb) of source IPs and account for 51.31% (50.86% in Feb) of the total telescope traffic, in June. However, we find that seven to ten organizations are essentially responsible for around half of the total telescope traffic in each month.
- We extend the notion of known scanners to commercial bots and identify a total of 34 bots. We find that bots have a negligible footprint, corresponding up to 0.25% of total source IPs and contributing less than 0.01% of the total traffic per month.
- We demonstrate that even though scanning traffic originates from over 12,000 Autonomous Systems, spread on more than 55,000 network prefixes and spanning over 220 countries, nonetheless over half of source IPs.
- We observe 60-65% common prefixes and 64-72% common ASes between the two months, suggesting a recurring behavior, especially from China, Egypt, Iran and India.
- We find that (original) Mirai probes account 1-1.5% of total scanning traffic per month. In 2023, we observe a burst in the number of IP addresses, with 48.20% of total source IPs sending at least one Mirai-fingerprinted packet. In February 2024, this trend falls to slightly below 20%.
- Similarly, we observe that traffic originating from Tor exit nodes remains negligibly small and accounts for merely 0.01% of the overall Darknet traffic, and corresponds to 0.04-0.06% of source IPs per month.
- We report on the current usage of scanning software such as ZMap and Masscan, and find around 40% of each monthly traffic volume has the ZMap signature. Conversely, Masscan traffic appears sparsely, less than 0.07% of each monthly traffic. ZMap is extensively used by the security industry and, notably, by many research institutes and universities. 40% to 50% of ZMap traffic is attributed to 20-25 known scanners per month.
- We showcase that grouping traffic into logical scans allows NAT remote detection in a global scale.
- Our findings stress out the need for further collaboration among defenders to exchange threat intelligence, as scanning poses a premature indication of potential upcoming cyber attacks.

1.4. Report Structure

The report is structured as follows. Chapter 2 sets the background on fundamental concepts such as scanning techniques and network telescopes. Then, we provide an overview of the current literature, in Chapter 3, regarding the study of Internet Background Radiation (IBR) and the applications of network telescopes. Chapter 4 details the research data and describes the data collection and refinement methodology. General characteristics of detected scanners are discussed in Chapter 5, followed by a reflection on the known scanners and known bots in Chapter 6. We group data into logical scans in Chapter 7 and discuss detected fingerprints. Lastly, Chapter 8 draws the conclusions of the analysis and proposes future research directions, and Chapter 9 summarizes and concludes our work.

2 Background

In the previous chapter, we defined the research questions and described the contributions of this study. We conduct measurements on **Internet-wide scanning** and study its origins, intention and overall behaviour by using a **Network telescope** that passively collects unsolicited traffic (**IBR**), including **TCP** packets. In this chapter, we set the background and elaborate on the concepts required to follow the rest of the study (highlighted keywords).

2.1. TCP/IP

The Internet Protocol (IP) is the primary protocol for routing packets across networks. It operates at the Network Layer of the Open Systems Interconnection (OSI) model and is responsible for sending packets from the source host to the destination host based on IP addresses. Transmission Control Protocol (TCP), on the other hand, functions at the Transport Layer and ensures that data is sent between applications in a timely, orderly and error-free manner. TCP creates connections using a three-way handshake, orders data segments with sequence numbers, and uses acknowledgment and re-transmission mechanisms to assure reliable delivery.

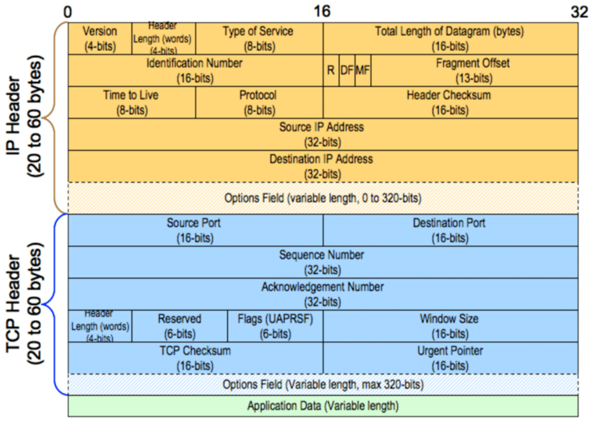

0
16
32
IP Header (20 to 60 bytes)
Version (4-bits)
Header Length (words) (4-bits)
Type of Service (8-bits)
Total Length of Datagram (bytes) (16-bits)
Identification Number (16-bits)
R
DF
MF
Fragment Offset (13-bits)
Time to Live (8-bits)
Protocol (8-bits)
Header Checksum (16-bits)
Source IP Address (32-bits)
Destination IP Address (32-bits)
Options Field (variable length, 0 to 320-bits)
TCP Header (20 to 60 bytes)
Source Port (16-bits)
Destination Port (16-bits)
Sequence Number (32-bits)
Acknowledgement Number (32-bits)
Header Length (words) (4-bits)
Reserved (6-bits)
Flags (UAPRSF) (6-bits)
Window Size (16-bits)
TCP Checksum (16-bits)
Urgent Pointer (16-bits)
Options Field (Variable length, max 320-bits)
Application Data (Variable length)

Figure 2.1: IP and TCP packet headers [32]

Figure 2.1 presents the IP and TCP packet headers. Only certain fields of interest are subsequently used throughout our study. Fields of interest in the IP header are the following: 1) IP address of the sender (32-bits) 2) IP address of the intended receiver (32-bits) 3) IP Identification (16-bits) which serves as an identifier generated by the sender to help in reassembling fragmented IP packets and 4) Time to Live or TTL (16-bits) that is used to time the datagram's lifespan, or number of network hops. Upon reaching zero, the datagram is discarded. Fields of interest in the TCP header are the following: 1) Source port (16-bits) 2) Destination port (16-bits) 3) Sequence number (32-bits) working as a counter that a host uses to record each packet sent out 4) Flags (URG, ACK, PSH, RST, SYN, FIN) or control bits which specify the function of the TCP segment and 5) Windows size (16-bits) that indicates how much data may be sent by the sender before receiving an acknowledgment response (ACK).

2.2. Internet-wide Scanners

Internet-wide scanning has evolved into a common measurement technique for observing online edge host activity. Internet scanning, made popular by programs like ZMap [30] and Masscan [36], has made it possible to study numerous topics such as like as censorship, operator behavior, botnets, outages, service rollout and security flaws.

2.2.1. Port Scanning Techniques based on Traffic Type

Port scans enable remote attackers to identify a computer's location, services, and operating system. A port scan is a connection attempt to each port on a target machine to detect which ports are open and consequently which services are active. A port scan can identify open ports on a target machine or network, indicating active services. Port scanning techniques may be classified according to the protocol and intention into the following [17, 59]:

- **TCP SYN Scan** The most used scan option is TCP SYN scan. In order to function, a SYN packet must be sent and either an RST for a closed port or a SYN+ACK response must be received. The port status is unknown if there is no response
- **TCP Connect Scan**: Rather than crafting its own packets, this scan makes advantage of the connect() operating system API. However, application layer services can detect and log the scan once a full connection has been completed
- **UDP Scan**: Depending on whether the port is open or closed, this scan sends UDP packets and waits for a UDP or ICMP response
- **SCTP INIT Scan**: This scan makes use of SCTP INIT chunks and anticipates receiving an ICMP error response if the port is closed or an INIT+ACK response if it is open
- **TCP NULL/FIN/Xmas Scans**: These scans attempt to bypass the firewall by using unusual or illegal TCP flag combinations in requests. NULL scan has all TCP flags set to false, whereas Xmas scan employs the FIN+PSH+URG combination of TCP flags
- **TCP Maimon Scan**: This scan transmits a FIN+ACK request like a FIN scan. The host returns an RST with details about the port's state
- **Custom TCP Scan**: With this scan, any flag combination may be specified
- **ACK Scan**: This scan sends ACK packets in order to map firewall rules. However, it does not offer much more than determine if a port is filtered
- **TCP Window Scan**: This scan uses the same protocol as the ACK scan to create ACK requests, but it additionally takes into account the size of the TCP window, allowing it to distinguish between an open and closed port
- **TCP Idle Scan**: This scan generates incremental IP Identification values by taking use of the so-called zombie host
- **IP Protocol Scan**: To identify supported protocols, this scan repeatedly loops over Internet protocol numbers
- **FTP Bounce Scan**: This scan uses an exploit of the FTP PORT command which, when supported, enables the use of a remote FTP server in passive mode to determine the target host's port state.

2.2.2. Port Scanning Techniques based on Destination Address

Staniford et al. [66] define three types of port scans based on destination address:

- **Vertical Scan**: A vertical scan involves scanning many or all ports of a single machine to identify the services it runs. This type of scan is employed by attackers with specified targets who try to fingerprint open and listening ports on the host and exploit potential vulnerabilities of running services. This scan is simple to detect as it just requires local (single-host) detection capabilities.
- **Horizontal Scan**: A horizontal scan involves pinging a port over a substantial percentage of a network or internet-wide systems. Attackers try to identify servers that expose a certain service since they possibly have an exploit for it. This type of scan can offer insights on frequently used Common Vulnerabilities and Exposures (CVE) by malicious actors or, in some cases, provide indications of zero-day attacks. Example of this strategy is the original Mirai Botnet which targeted Telnet ports TCP/23 and TCP/2323 [4].

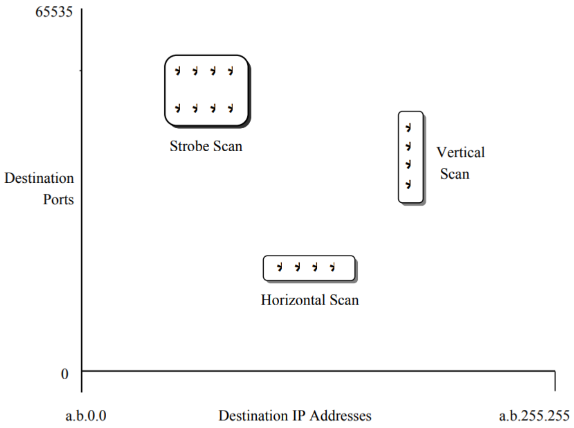


Figure 2.2: Scanning Strategies [33]

- **Block Scan**: A block scan is a combination of vertical and horizontal scans that target multiple ports across multiple hosts. This scanning strategy is also referred by some authors as strobe scan, whereas block scan targets all (65535) ports [33, 14]. For this study, we do not make a distinction between strobe and block definitions, since scanning on all ports is not observed frequently and, therefore, there is no need for further separation. An example of this strategy is Censys which scans every day the whole IPv4 space on 137 popular ports based on IANA-assigned services [31].

Visualizing network scanning strategies can contribute to the deeper comprehension of emerging trends and identification of common patterns in scanning campaigns among cybercriminals, researchers and the security industry, as shown in figure 2.2.

2.2.3. Identification of Scanning Tools

Tools for probing and analyzing networks for a variety of uses, such as vulnerability identification, network mapping and security evaluation, are known as network scanning tools. These tools collect data on devices and services on targeted hosts or networks employing a variety of methods such port scanning, host discovery and service enumeration. They help administrators and security researchers to identify active hosts and often contribute to vulnerability assessment and security auditing. However, they are also used by cybercriminals to identify publicly exposed vulnerable services. Ghiette et al. [34] have demonstrated that it is possible to remotely distinguish between the tools used for scanning by analyzing the inbound traffic. In this study, we examine the following tools:

Masscan. Masscan is a high-speed scanner developed for large-scale IP address and port scanning. It is noted for its remarkable speed, which allows it to scan the full IPv4 address space in five minutes [36]. It utilizes asynchronous transmission to increase scanning efficiency and can be tailored for specialized scanning tasks. It can be efficiently fingerprinted based on the Identification field of the TCP header as follows [34]:

$$IPId = g_{L16}(DstIP) \oplus DstPort \oplus g_{L16}(Seq)$$

Function g_{L16} denotes the extraction of the lower 16 bits of a bit-represented variable. We extract the lower 16 bits out of the 32-bit destination IP address of the IP packet header, the lower 16 bits out of the 32-bit Sequence number of the TCP header and the 16 bits of the destination port included in the TCP header. $\oplus$ denotes the XOR bit-wise operation.

ZMap. Zmap is a fast open-source network scanner for effective Internet-wide scanning. It is capable of scanning the entire public IPv4 address space on a single port within 45 minutes [30]. Furthermore, it is highly configurable and can be used for a variety of security research applications such as vulnerability assessment. It is effectively fingerprinted by the hard-coded $IPId = 54321$.

2.3. Network Telescopes

2.3.1. Definition

Moore et al. [53] define a network telescope as a globally routed - yet unused - IP address space that carries little or no legitimate traffic. More precisely, a network telescope comprises IP addresses that are not currently allocated to any active device or service. These addresses are usually chosen from a pool of IP addresses that have been assigned but not currently used by an organization. A network telescope - alternatively referred to as blackhole, Internet sink, darkspace or Darknet [19] - passively monitors all incoming unsolicited traffic, or Internet Background Radiation (IBR).

Therefore, a telescope is a useful instrument for network security and Internet measurement. It offers a distinct viewpoint on malevolent actions by enabling the observation of large-scale and small-scale remote security events. Furthermore, it facilitates threat landscape analysis and study of malicious activities such as denial-of-service attacks, scanning behaviour, malware propagation etc. We perform a thorough examination of the related work in Section 3. Lastly, it is beneficial to research and education because it provides a better understanding of network security issues and allows the development of more effective solutions. For example, the UCSD (University of California San Diego) Network Telescope has been employed to generate datasets for numerous publications [19].

2.3.2. Internet Background Radiation (IBR)

Security researchers can obtain insights about numerous large-scale events taking place on the Internet by analysing telescope collected data, or broadly put, any IBR traffic that is passively captured at specified network vantage points. Events include security incidents such as malware propagation [52], botnet outbursts [4, 10, 24], Denial-of-Service (DoS) attacks [51] and long-term cyber attacks [9]. Study of unsolicited (IBR) traffic is also proved to be useful when it comes to studying Internet scanning trends [29], misconfigurations [13], Internet connectivity and outages as a result natural disasters [26], country-level censorship [25] and network failures [12].

According to Wustrow et al. [77], IBR traffic may be classified into three categories based on the underlying root causes: 1) scanning that originates from scanning campaigns and infected hosts scanning for vulnerable targets 2) backscatter which is usually caused by Denial-of-Service (DoS) attacks and 3) misconfigurations that occur due to software or hardware issues. Based on the above definition, they consider TCP SYN packets as scanning traffic. Next, TCP SYN+ACK, RST, RST+ACK and ACK packets are considered backscatter traffic most likely produced as response by victims receiving spoofed traffic that is allegedly coming from the darknet. The remaining traffic is classified as misconfigurations.

On the other hand, Collins et al. [21] adopt a different diving IBR into four categories: 1) short IBR traffic 2) scanning 3) backscatter and 4) other. Short IBR traffic includes the total traffic received from IP addresses which send less than or equal to four packets of traffic a day. Similarly to Wustrow et al., TCP SYN packets are considered scanning, whereas backscatter is TCP SYN+ACK, FIN and RST. The remaining traffic is not considered.

2.3.3. Practical Considerations & Limitations of Network Telescopes

According to Moore et al. [53], a network telescope operate under two assumptions 1) uniform IP address selection and 2) accurate targeting rate observation. First, IP addresses of incoming traffic are selected unbiased and at random from the entire IPv4 address space. However, sometimes IP addresses are not chosen randomly because certain regions might be cut off (e.g., blocked address space of certain countries), more biased than others (e.g., certain countries are favoured) or due to buggy randomness generation. Second, we assume that the measured targeting rate appropriately reflects host targeting rates. In some cases, a network telescope may underestimate the targeting rate due to network overload, limitations in processing and storing traffic, internet routing instabilities and difficulty distinguishing concurrent events.

Darknets present certain limitations. In particular, they are segregated form the rest of the address space and frequently declared by research networks or institutions [61]. That renders them easily identifiable by external observers (e.g. cybercriminals, other researchers) who avoid sending traffic to the telescope address space. Furthermore, telescopes capture only traffic destined to their own address space, making it impossible to detect scanning campaigns against targets different than either the whole IPv4 space (or a considerable portion of it) or the telescope address spaces themselves [61]. Additionally, as passive instruments they do not interact and capture stateless TCP scanners, who return to their identified potential victims to complete the TCP handshake [39].

2.3.4. Single and Multiple Packet Detection

A geometric distribution is used to calculate the probability of detecting a single probe packet when a scanner selects target IP addresses uniformly randomly [53]. Given that a network telescope observes a portion of the whole IPv4 address space, we consider the probability p that a single packet transmitted by the scanning host reaches our telescope. The quantity of packets seen at the telescope in the event that the host delivers multiple packets is described by a binomial distribution with parameter p.

2.3.5. Flow Timeout Problem

The Flow Timeout Problem refers to the decision between grouping and splitting a sequence of packets. When a using network telescope to track Darknet events, three constraints should be considered [53]: 1) the probability of packets being observed by the telescope should be calculated, rather than splitting events based on a (reasonable) packet rate 2) timeout limit should be large enough to avoid erroneous splitting due to the nature of events and 3) the timeout limit should be kept as short as possible to prevent the grouping of unrelated events. We provide our scan definition and relevant timeout limits in methodology section 4.2.3.

3 Related Work

3.1. Study of Internet Background Radiation (IBR)

Several studies have been carried out on this topic with the use of network telescopes. Specifically, they have been used extensively to study various security incidents such as malware and botnet outbursts [10], network outages [25], Internet scanning trends [29], DDoS attacks [51] etc. Nonetheless, the rising scarcity and the subsequently elevated economic worth of the IPv4 address space has led many network telescopes to steadily shrink over time (e.g. CAIDA, Merit) and researchers to adopt a plethora of ways to collect and study IBR. A summary is presented in table 3.1.

Several concepts have been adopted to overcome the emerging challenges and limitations of conventional network telescopes. Firstly, Richter et al. [61] examined scanning behavior through the analysis of unsolicited traffic that is intercepted at the firewalls of around 89,000 hosts in a Content Distribution Network (CDN). Authors argue that this approach has an advantage over typical darknets since it is dispersed among 1,300 networks with traffic coming from live servers. It comprises 178,000 addresses in 2,800 prefixes and collect 19.4 billion packets (1.2 Terabytes) within a one-month period (Nov 2018). According to the research, localized scans account for around 30% of scan activity. Compared to more extensive Internet-wide scans, these localized scanning efforts show unique characteristics in target selection and scanned services, indicating that traditional darknets could be underestimating focused scanning activity inside certain network prefixes or areas. Another novel concept called *meta-telescope* was presented by Wagner et al. [74]. A meta-telescope offers broad coverage of the dark space regarding the size and topological placement comparing to the common telescopes. The authors are able to capture traffic towards over 350,000 /24 blocks across over 7,000 autonomous systems, resulting in a wide coverage of dark address space. They partner with 14 IXPs and three network telescopes collecting 86,667 billion flows (880 Petabytes) from IXPs and an average of 1.99 million packets daily from each telescope, for a certain time period in April 2023. The examination of IBR directed towards these networks discloses differences according to the kind of network, geographical location and destination network, offering important information on patterns of Internet usage. The study also addresses the difficulties and lessons learned from using a meta-telescope in real-world scenarios. Lastly, Hiesgen et al. [39] examined Internet-wide scan traffic with Spoki, a real-time network telescope for asynchronous TCP SYN packet capture. After examining data for three months, the research finds that a large percentage of TCP SYNs have irregular features, suggesting highly focused scanning operations coming from different regional perspectives, including a considerable amount from malicious sources.

A number of researchers, such as Tao et al. [11] and Iglesias et al. [41], deployed clustering and deep learning techniques to analyze Darknet data and identify novel attack patterns, long term scanning activities and victims. One more example of this research approach is DarkVec by Gioacchini et al. [35]. DarkVec identifies groups of senders on darknets that present comparable behaviors by using word embedding techniques like Word2Vec. Thorough testing shows that this approach offers the potential for automated pattern learning in traffic analysis since it can accurately correlate unidentified sender IP addresses with existing labels and discover new scan groups. Their dataset is generated by an academic /24 darknet which collects approximately 67 million packets (63.5 million for training and 3.5 million for testing) in a 30-day period. Other researchers, like Brownlee [18], focused on measuring the IBR inter-arrival time to detect new activities. Benson et al. [13] examined the frequency of networks sending IBR and proposed a new qualitative framework for assessing unsolicited traffic as an appropriate data source for extracting Internet-wide properties.

Table 3.1: Summary of datasets, methods and results from previous work (ordered by publication year)

Author	Topic	Data Source	Collection Period
Wagner et al. [74] (2023)	Meta-telescope	14 IXPs & 3 network telescopes & ISP NetFlow records	24-30 Apr 23
Anand et al. [3] (2023)	Aggressive Hitters	Merit ORION NT (approx. 500,000 IPs)	Jan 21 - Oct 22
Collins et al. [21] (2023)	Acknowledged Scanners	USC-ISI (768 IPs)	Sep 22
Kallitsis et al. [45] (2022)	Changes of scanning behavior	Merit (/10 & /13 darknets)	Sep 16 & 20 Feb 22
Hiesgen et al. [39] (2022)	Spoki	4 /24 IP prefixes across the US and EU	Apr - Jun 20
Gioacchini et al. [35] (2021)	(DarkVec) Word Embeddings	Academic /24 darknet	Mar 21
Griffioen et al. [38] (2020)	Quantifying AS IP Churn	Network telescope (approx. 65,000 IPs)	9 months (Mar - Dec 18)
Griffioen et al. [37] (2020)	Slow, distributed scanners	3 partially populated /16 networks (65,000 IPs)	2 months in 2018
Richter et al. [61] (2019)	Distributed Network Telescope	178,000 addresses in 2,800 prefixes	Nov 2018
Antonakakis et al. [4] (2017)	Mirai Botnet	Merit (4.7 million IPs)	Jul 16 - Feb 17
Ghiette et al. [34] (2016)	Port scan toolchain identification	2 /16 telescope (128,000 IPs)	18 month (focus on Apr 15)
Irwin et al. [44] (2013)	Study of malicious activity	5 /24 network telescopes in TENET South Africa	5 month period
Benson et al. [13] (2015)	IBR opportunistic analysis	UC San Diego (UCSD-NT) & Merit	Jul - Sep 12 & Jul - Aug 13
Durumeric et al. [29] (2014)	Internet-wide scanning	Merit (5.5 million IPs)	Jan 13 - May 14
Wustrow et al. [77] (2010)	IBR revisited	Merit	2010 & 2006-2010 for each dataset

In an effort to understanding the motives and tools employed by the various Internet-wide scanning actors, several researchers performed fingerprint analysis on IBR traffic. Ghiette et al. [34] investigated the identification of recurring patterns in port scan packets to determine the adversary's toolchain. Empirical examination of scan traffic from two /16 networks (128,000 IPs collecting IBR of 8 TB in total) reveals patterns in the use of open-source port scan tools (Zmap, Masscan etc.) and geographical regions. They also demonstrate the feasibility of remote identification of scanning software based on traffic analysis. Antonanakis et al. [4] presented a comprehensive analysis on Mirai botnet. They examined its influence on various types of devices, the evolution of Mirai variations and the botnet's quick expansion to 600,000 infections. The study underscores the vulnerability of IoT ecosystems and suggests that Mirai represents a fundamental change in botnet development, demonstrating how easily beginner approaches may largely attack devices. The report provides technical and non-technical solutions to reduce this risk. For this study, researchers obtain traffic probes from a telescope at Merit Network composing 4.7 million IPs over a seven month period (Jul 16 - Feb 17). By taking advantage of flaws in the random number generation of the Mirai botnet, Griffioen et al. [38] extended the research scope presenting a novel method for detecting and measuring IP churn within ASes. Through the utilization of these vulnerabilities, the technique is able to reliably re-identify IoT devices that have already been infected, even if they reappear with new IP addresses. Due in part to Mirai's extensive dissemination and proactive scanning behavior, this permits extensive and passive monitoring of IP churn across provider netblocks. For this study, authors employed a network telescope of approximately 65,000 IP addresses collecting 6.5 billion packets (864 GB of traffic).

It is known that a subset of Internet scanners interact with the public via websites making no intention to conceal their scanning goals. Collins et al. [21] explored acknowledged (ACKed) scanner who offer a range of services including corporate engagement, non-profit work and education. The study highlights the importance of distinguishing between various scanner types in order to prevent operational difficulties and false research findings. Additionally, they provide quantitative analysis based on a 30-day sample of darkspace data from the USC Information Sciences Institute (USC-ISI) network and maintain a list of more than 40 ACKed scanners. Similarly, Trapickin [73] indexed several known scanner entities and performed entity classification per organization type, intention, spatial distribution, publication of results and used scanning software. However, this analysis includes only a number of known scanners and does not search thoroughly their scanning behavior.

Lastly, Anand et al. [3] employed a network telescope from Merit Network to conduct an empirical analysis of aggressive IPv4 scanners within a two-year period and assessed the network impact using supporting data from two academic Internet Service Providers (ISPs), packet streams from a third network and honeypot data. The two types of scanners that are commonly used are criminal actors looking to exploit susceptible targets and benign research-oriented organizations. The study provides insight into the actions of aggressive IPv4 scanners through a longitudinal empirical investigation, discovering that these scanners take up a sizable amount of processing power on ISP routers and underscoring the possibility of interference with vital network functions.

4

Data Collection & Methodology

In this chapter, we introduce the research data and describe the data collection and refinement methodology. We perform a macroscopic breakdown of the IBR traffic and, finally, we discuss ethical considerations.

4.1. Data Collection

Our initial dataset consists of all the TCP traffic received by an organization's network telescope comprising three IP ranges. Traffic routed to the unused space of the network ranges is captured and stored in .pcap format. Ingress traffic to ports TCP/23 (Telnet) and TCP/445 (SMB) is dropped by policy. Due to the dynamic nature of these networks, the size of the network telescope is dynamic and fluctuates between the equivalent sizes of 1 /16 and 3 /16 networks. An allocated address returns to the pool of unused IP addresses after i) a client device shuts down ii) releases its DHCP lease explicitly (i.e. disconnects from the network) or iii) its DHCP lease duration has expired. Any traffic routed to the unroutable addresses is captured by the Darknet. Moore et al. [52] claim that address ranges used by a network telescope should ideally be routed but not utilized, which filters out all regular activity. However, these address ranges may include active machines, provided that the normal traffic can be eliminated from the captured data upon inspection. The noise introduced due to the current Darknet structure is refined during data selection, described in section 4.3.

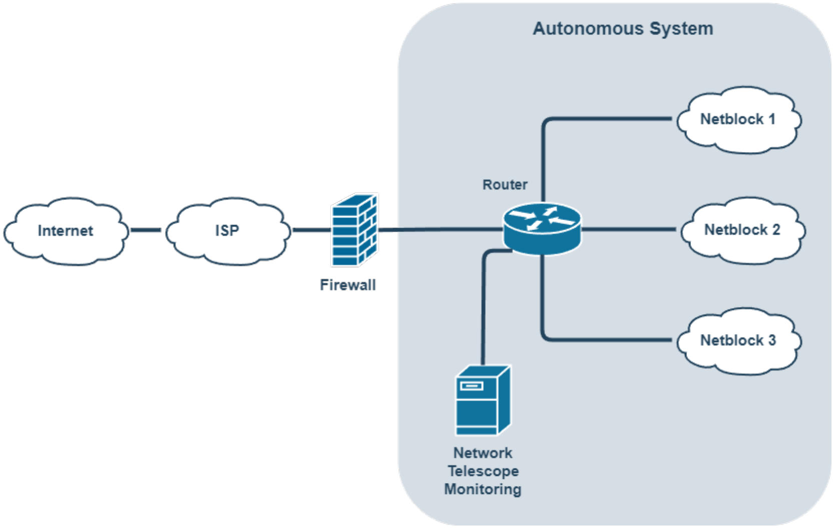


Figure 4.1: Network topology of our Darknet

Our initial dataset includes two one-month snapshots and corresponds to periods 01 - 30 June 2023 and 01 - 29 February 2024. We note a gap in the dataset in 04 - 08 June 2023 possibly due to a network failure (outage). The darknet receives approximately 10.18 billion packets from over 1.57 million source IP addresses in June 2023 and around 12.64 billion packets from over 2.17 million source IP addresses in February 2024.

4.2. Analysis of IBR

4.2.1. Scanning, Backscatter and Misconfigurations

In this section, we analyse the composition of telescope collected data - or IBR traffic. More precisely, we follow the definition by Wustrow et al. [77], discussed in section 2.3.2. IBR traffic is classified into three categories:

- Scanning (TCP SYN)
- Backscatter (TCP SYN+ACK, RST, RST+ACK and ACK)
- Misconfigurations (other)

Traffic decomposition of our IBR dataset for June 2023 shows that scanning comprises 98.61% of the total telescope traffic, corresponding to an average of around 401.67 million packets per day. Backscatter and misconfigurations comprise 1.22% and 0.17% of the total telescope traffic respectively. Notably, backscatter traffic reaches at most approximately 13.19 million packets per day (and on average 4.98 million packets/day), whereas misconfiguration packets reach at most 2.86 million daily (and on average 699,224 packets/day).

IBR composition of the February 2024 dataset demonstrates comparable results. Scanning comprises 98.50% of the total telescope traffic, corresponding to approximately 344.81 million packets per day. Backscatter and misconfigurations comprise 1.40% and 0.1% of the total telescope traffic, respectively. Additionally, backscatter traffic reaches at most approximately 12.47 million packets per day (and on average 4.92 million packets/day), whereas misconfiguration traffic reach at most 564,677 packets per day (and on average 342,638 packets/day).

4.2.2. Traffic Type Analysis

Table 4.1: IBR Composition (Jun 23)

TCP Flags	# of Packets (approx.)	% of Total Traffic
SYN	10.04 billion	98.61%
SYN+ACK	99.03 million	0.97%
RST	3.85 million	0.04%
RST+ACK	2.42 million	0.02%
ACK	19.26 million	0.19%
Other	17.48 million	0.17%

Table 4.2: IBR Composition (Feb 24)

TCP Flags	# of Packets (approx.)	% of Total Traffic
SYN	9.99 billion	98.50%
SYN+ACK	90.15 million	0.89%
RST	23.83 million	0.23%
RST+ACK	2.32 million	0.02%
ACK	26.37 million	0.26%
Other	9.93 million	0.10%

Tables 4.1 and 4.2 provide a breakdown of IBR traffic volume in terms of TCP flags, for June and February. To begin with, the TCP SYN flag - used to initiate a TCP connection - dominates the traffic with approximately 10 billion packets each month, accounting for around 98.50-98.60% of the total traffic. On the other hand, backscatter accounts for roughly 1.22% in June and 1.40% in February. SYN+ACK - typically sent in response to a TCP SYN packet to acknowledge connection requests - comprises around 99 million packets, or 0.90-1% of the total traffic. RST and RST+ACK flags - used to reset connections - together account for 0.06-0.26% of the total traffic. Notably, RST flag presents an approximately 519% increase between the two months, from 3.85 million packets in June to 23.83 packets in February. Additionally, The ACK flag - used to acknowledge the receipt of a packet - accounts for 0.19-0.26% of the total traffic. The rest flag combinations represent the remaining 0.10% to 0.17% of the total traffic.

The definition by Wustrow et al. [77] provides a comprehensive tool for classifying IBR. Nonetheless, there are also other TCP flag combinations employed by various software scanning tools like Nmap [69], contributing potential research value and thus should be examined. We detect 3,046 packets in June (510 in February) that contain the Nmap XMAS scan signature (PSH+URG+FIN). Similarly, there are only 584 packets in June (124 in Feb) exhibiting the FIN scan fingerprint and 103 packets in June (14 in Feb) that carry the SYN+FIN scan fingerprint. Conversely, the number of packets for NULL scan (all flag bits set to zero) is around 11.80 million in June (5.75 million in Feb), whereas for XMAX scan (all bits set to one) [46] is 225,651 in June (978,550 in Feb). Consequently, TCP flag combinations other than the TCP SYN flag do not appear to demonstrate any significant research value in our dataset and, therefore, can be safely ignored.

4.2.3. Scan Definition

Essential to our study of Darknet data is the notion of a scanning event, or scan. A scan is essentially a grouping of packets pertaining to the same activity under specific parameters. Grouping of traffic into logical scans facilitates the analysis of fingerprints and, further, the identification of scanning actors, their intentions (e.g. malicious botnets) and scanning tools and software (e.g. Zmap, Masscan). We adopt the methodology by Durumeric et al. [29], assuming the probability distributions described in section 2.3.4.

We define a scan by a given source address as a sequence of probes that hit at least 100 distinct Destination IP addresses in our darknet at a minimum estimated Internet-wide scan rate of 100 pps (packets per second). We consider only TCP SYN traffic. We do not require a scan to be on a fixed port number. Given the dynamic nature of our Darknet, we consider the Darknet size as the number of destination IP addresses that received unsolicited traffic (IBR). Therefore, the darknet size is 65,321 addresses in June (equivalent to 1 /16 network and 0.0017% of the public IPv4 space) and 177,027 addresses in February (0.0047% of the public IPv4 space).

The inter-arrival times of the probes to any address of the telescope are less than a given timeout interval. Based on the work by Moore et al. [53], a scanner probing random IPv4 addresses at the rate of 100 pps will appear in our darknet with 99% confidence within 2,614 seconds (approximately 43 minutes) and with 99.9% confidence within 3,920 seconds (1.08 hours) in June. Therefore, we expire scans that do not send any packets after 3,921 seconds, with 99.9% confidence. For February, a scanner under the above settings will appear in our darknet with 99% confidence within 965 seconds (approximately 16 minutes) and with 99.9% confidence within 1,447 seconds (around 24 minutes). Thus, we expire scans that do not send any packets after 1,448 seconds, with 99.9% confidence.

4.3. Data Refinement

Captured scanning traffic from the IBR dataset needs to be refined before conducting the main analysis. We perform the following steps:

Scanning Traffic. We discard non-TCP SYN traffic according to the IBR definition by Wustrow et al. [77].

Bogon Filtering. We identify and remove bogon IP addresses. Bogons are phony (false) IP addresses on the public Internet that include Martian packets and IP addresses that do not fall inside any range allocated or delegated by the Internet Assigned Numbers Authority (IANA) or a delegated Regional Internet Registry (RIR) and are permitted for public Internet use. According to RFC 1812 [8], a Martian packet is an IP packet that is visible on the public Internet and has a source or destination address that is set aside for special purpose by IANA. Such a packet cannot be sent via the public Internet, or it has a fake source address and cannot originate as stated. We detect one bogon IP address for Jun 23 and 721 bogon IP addresses for Feb 24.

Internal Traffic. We remove traffic data from within the Darknet i.e. originating in the three network ranges and the organization's Virtual Private Network (VPN) service. In June, we observe approximately 1.13 million TCP SYN packets and a total of 3.5 million packets (with various flags) from 96 internal source addresses to 52,690 destination addresses of our darknet. In February, we observe approximately 1.60 million TCP SYN packets and a total of 1.94 million packets (with various flags) from 8,297 internal source addresses to 19,969 destination addresses of our darknet. Additionally, we observe scanning activity from four IPs in June and 703 IPs in February, that originate in the organization's VPN network, suggesting a potential infection or an insider performing (Internet) scanning. Since internal traffic is usually innocuous and irrelevant for identifying and analyzing external internet-wide scanners, it is excluded from the dataset. That leads to lower noise levels in data, enhancing the effectiveness and precision of identification and analysis of remote scanners.

4.4. Spoofed Traffic

IP spoofing is the process of crafting packets using a fake source IP address in order to impersonate another computer system or conceal the identity of the sender. Given the passive operation of our network telescope, darknet addresses do not reply to received packets. Therefore, we cannot usually presume that incoming traffic packets do not contain spoofed source IP addresses. However, since legitimate source addresses are needed for effective scanning in order to get results, we rely on the assumption that these IP addresses are authentic. This is also supported in similar studies e.g. in [24]. Besides, we remove easily identifiable traffic such as bogon IP addresses (and internal traffic which might contain spoofing) as described in the previous section.

4.5. Ethical Considerations

Collecting network traffic for Internet and network measurement studies requires responsible data management. Our Darknet collects passively IBR traffic and does not reply to any probe. The data are stored and processed within the organization. We do not examine OSI Layer-2 data (e.g., device MAC addresses) or possible packet payloads embedded in the incoming TCP packets. Data do not reveal confidential or private information about individuals (i.e. employees and users within the organization or external entities). Therefore, the collection of the data in this study poses minimal risk to causing tangible harm to any person.

4.6. External Datasets

To facilitate our analysis and validate our results we make use of the following external datasets:

E1. Maxmind GeoLite2 Database: We use the Maxmind GeoLite2 ASN & City database [47] for IP Geolocation and classification of Autonomous System (AS). We employ the database of 8 December 2023 to analyse the telescope data from June 2023 and the database of 6 February 2024 to analyse the traffic in February 2024.

E2. IPinfo IP to Country and ASN Database: We use the non-commercial IP-to-country and IP-to-ASN databases by IPinfo [43] to cross-reference the results obtained by Maxmind database. The purpose is to validate the results obtained by the Maxmind datasets, being extensively used further in this study. In particular, we compare the IPinfo database - updated daily - from 8 February 2024 to the Maxmind GeoLite2 database of February 6th - updated weekly. All queried IP addresses belong to both databases. Mapping IP addresses to respective Autonomous System (AS) numbers yields the same result for 98.97% of the queried IPs. Likewise, mapping of IP addresses to ISO 3166-1 alpha-2 two-letter country code and country name matches 98.44% and 98.41% for the two databases. Lastly, only 56 IP addresses differ in both the AS number and country classification.

5 Scanner Characterization

In this section, we analyze collected traffic of two months: June 2023 and February 2024 distinctively. All measurements are obtained after data refinement and refer to each one-month period. Our analysis encompasses an IP address and port analysis, study of the AS ecosystem, tracking bad (scanning) Internet neighborhoods, scrutiny of aggressive hitters and lastly identification and analysis of anonymous actors using Tor.

5.1. IP Address and Port Analysis

In June 2023, the telescope observes a total of 1,561,228 distinct source IP addresses with a corresponding total traffic volume of 10,040,830,238 packets. The majority of traffic, at 40.30%, has an IP identification (IPId) of 54321, associated with Zmap; IPId 0 represents 1.45% of the traffic, while the remaining IPIds are below 0.03%. The average TCP header length is 43.08 bytes.

In February 2024, the telescope records 1,276,019 distinct source IP addresses, along with a total traffic volume of 9,998,016,067 packets. This signifies a decrease of approximately 18.25% in the number of considered IP addresses, whereas the amount of considered traffic decreases only 0.42% suggesting similar scanning volumes for these two months. Similar to June, the majority of traffic has a Zmap-related IP identification (IPId) of 54321 at 37.60%, while IPId 0 accounts for 1.26% of the traffic and the remaining IPId values lie below 0.03%. The average TCP header length is 42.71 bytes.

Tables 5.1 and 5.2 show basic statistics for our refined Darknet dataset for both months. Scanning traffic appears on daily basis. In total, 65,321 destination addresses are scanned in June 2023 (33.22% of a 3 /16 network) and 176,473 destination addresses in February 2024 (89.75% of a 3 /16 network). Netblock #1 attracts the highest total traffic volume compared to the rest netblocks for both months, with around 6.40 billion scanning packets per month. Although traffic volume remains similar for both months, however we observe variations in the distribution of incoming traffic per telescope netblock. On the one hand, netblock #1 is adequately scanned in June (>60% coverage), whereas netblocks #2 and #3 are scanned by at most one fifth of the size. On the other hand, all netblocks are extensively scanned in February (>86% coverage). Lastly, the number of scanned destination ports for both months appear consistent (65,522-65,523). The change in the percentage of destination IPs scanned can be attributed to a change of routing policies within the AS, given the traffic volume remains similar.

Table 5.1: Basic statistics for our Darknet dataset (Jun 23)

Netblock	% of DstIPs Scanned	Total Packets	# of DStPorts	Dates
#1	62.01%	6.37 billion	65,523	01-30 Jun 23
#2	20.85%	1.99 billion	65,523	01-30 Jun 23
#3	16.79%	1.67 billion	65,522	01-30 Jun 23

Table 5.2: Basic statistics for our Darknet dataset (Feb 24)

Netblock	% of DstIPs Scanned	Total Packets	# of DStPorts	Dates
#1	86.11%	6.48 billion	65,523	01-29 Feb 24
#2	96.44%	1.81 billion	65,523	01-29 Feb 24
#3	86.71%	1.69 billion	65,522	01-29 Feb 24

Traffic per Day. Analysing the traffic per day for June, we note the highest traffic days - with over 750 million packets - on the 1st, 2nd, and 10th of June. Conversely, the lowest traffic is observed on the 21st with approximately 270 million packets. On average, approximately 401.63 million packets reach the telescope per day. Comparing to the number of unique source IP addresses, the highest number is recorded on the 1st and 29th of June with around 200,000 IP addresses. The lowest number observed on 9th of June, with roughly over 96,000 IPs. On average, approximately 150,147 distinct source IPs hit the telescope space each day.

Traffic analysis for February shows similar trends. The peak traffic volume is observed on 9th Feb with 445 million packets and the lowest on 26th with around 287.67 million packets. Overall, the telescope receives around 344.75 million packets per day, representing a 14.1% decrease comparing to the average value for June. Comparing to the number of unique source IP addresses, our Darknet receives from 119,447 (on 21th Feb) to around 170,954 IP addresses (on 18th Feb) with an average value of 143,429 addresses per day. Hence, it is clear that the distribution of scanning probes on a daily basis is not uniform in terms of traffic volume and distinct source IP addresses for both months.

Device/OS Fingerprinting. Analyzing the Time-To-Live (TTL) values, we find that a considerable percentage of traffic has a TTL value of 250 (23.93% in Jun and 21.30% in Feb). In June, there are some noticeable variations at intervals of 60 (8.3%), 246 (7.4%), 245 (7.32%), and 43 (7.12%), while similar distribution is followed for February with peak intervals at 246 (14.08%), 60 (7.55%), 249 (6.14%), and 251 (5.63%).

Figures 5.1 and 5.2 illustrate the distribution of traffic volume over different TTL values for both months. Notably, the distributions show three separate peak regions, approximated as follows: 20-60, 100-123 and 230-251. 73.78% of the total traffic in June falls within the TTL range of 230-251. By comparison, TTL values ranging from 100-123 and 20-60 contribute significantly lower percentages, at 2.61% and 23.57%, respectively. Trends for February appear identical with the percentage of traffic for the three ranges 230-251, 100-123 and 20-60 to be 74.56%, 2.31% and 23.07%, respectively. Windows 98, NT 4.0, 2000 pro, XP, Vista, 7 and 10 are known to have default TTL value for TCP at 128 (some older versions such as Windows for Workgroups and Windows 95 have lower default TTL values for TCP at 32) [64, 63].

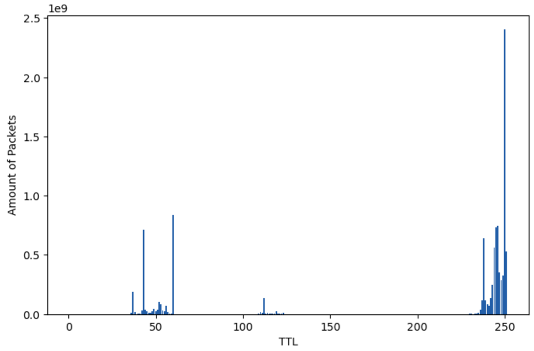


Figure 5.1: Amount of traffic per TTL value (Jun 23)

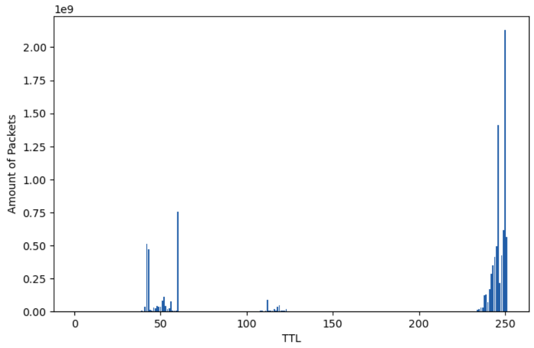


Figure 5.2: Amount of traffic per TTL value (Feb 24)

Furthermore, many *nix (Linux/Unix) distributions have a default TTL value for TCP at 64 and Solaris/AIX operating systems at 254. Given, the widespread application of Linux-based implementations for general purpose OSes and IoT devices, we assume a classification of remote hosts into Windows and non-Windows groups.

Source Ports. Source ports of incoming traffic range from 0 to 65535. It's reasonable to anticipate an even distribution of source ports per packet over a sufficiently large sample size, since a device usually selects a random available source port from the entire range of possible (or ephemeral) ports. Nonetheless figures 5.3 and 5.4 demonstrate that some ports are favoured (e.g., TCP/47288 in Jun, TCP/53249 in Feb). Considering the standard range for Linux ephemeral ports, which spans from 32768 to 61000 [28], it's noteworthy that a substantial portion of the traffic, amounting to 88.68% in June and 88.81% in February, falls within this Linux-defined range.

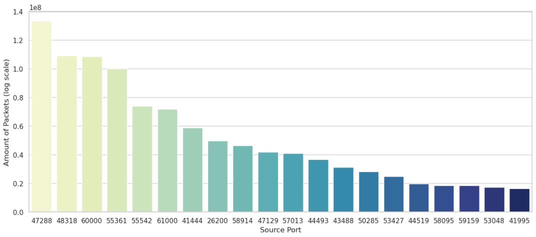


Figure 5.3: Top 20 source ports sorted by traffic volume (Jun 23)

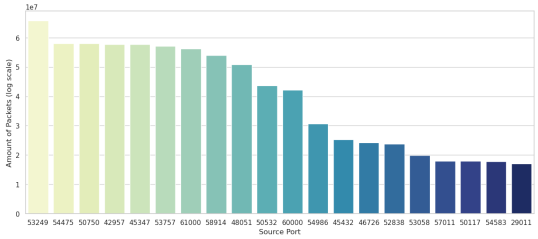


Figure 5.5: Top 20 source ports sorted by traffic volume (Feb 24)

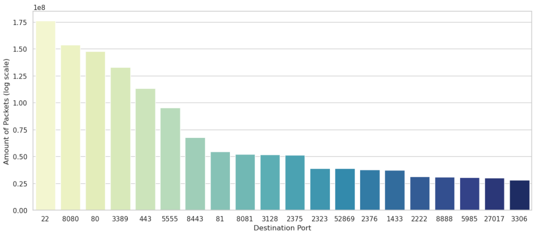


Figure 5.4: Top 20 destination ports sorted by traffic volume (Jun 23)

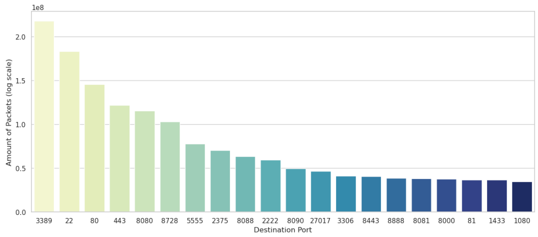


Figure 5.6: Top 20 destination ports sorted by traffic volume (Feb 24)

Destination Ports. 13 ports[1] - two of which are blocked by policy - do not receive any traffic in any of the investigated months. ECDF shows that the top-10% ports receive 76.25% of total traffic, while the top-1% - 656 ports - receives 44.40%, in June. The corresponding top-1% and top-10% ports for February reach 50.82% and 80.19% accordingly. Figures 5.4 and 5.6 illustrate the top 20 most popular destination ports per traffic volume for both months. The top five ports with the highest incoming traffic volume remain consistent in both months: 22, 80, 8080, 443 and 3389. Port TCP/22 attracts the most scanning activity in June, with 1.75% of total traffic. The rest top-5 ports fluctuate between 1% and around 1.5% of the total traffic, whereas the remaining ports account for less than 1% of traffic each. Delving further into the 10 most popular ports per day for June, we observe that not all port numbers appear consistently in the daily the top-10 selection. Ports TCP/9200, TCP/5038, TCP/3306, TCP/2323 are included only in one day, whereas ports TCP/22, TCP/80, TCP/8080, TCP/3389, TCP/5555, TCP/443 appear all days. Rest top ports fluctuate between two and 23 days.

Analysis on February data shows that TCP/3389 receives the most traffic (2.17% of total traffic) and is followed by TCP/22, TCP/80, TCP/443 and TCP/8080 which fluctuate between 1.15% and 1.86%. Similarly to June, we observe that not all port numbers appear consistently in the daily the top-10 selection. Ports TCP/1433, TCP/1900 and TCP/21 are included only in one day, whereas ports TCP/3389, TCP/80, TCP/22, TCP/2375, TCP/8080 and TCP/443 appear all days. Rest top ports fluctuate between two and 28 days. Additionally, TCP/8728 now receives approximately 1% of the total traffic from 0.007% in June, appearing in the top-10 daily ports of 27 days. This behavior signifies an 14,353% increase

[1] Jun 23 & Feb 24: 23 (blocked), 42, 111, 135, 137, 138, 139, 161, 162, 427, 445 (blocked), 524, 593

of the absolute amount of traffic packets. The port serves as the default API port for MikroTik RouterOS.

Mapping port number with known services and protocols, TCP/80 and TCP/8080 serve as main and alternative ports for HTTP, TCP/22 is the known port of Secure Shell Protocol (SSH) and TCP/443 is the known port for HTTPS. TCP/3389 is used by Microsoft Remote Desktop Protocol (RDP) linked with a history of vulnerabilities that can lead to information disclosure and denial of service attacks [49, 48, 76, 60]. Remote API of Docker Engine is running by default on TCP/2375 port when enabled. The port has been linked in the past with known CVEs [23]. Also, if used via unprotected TCP socket it can be exploited to allow the escape from the container-jail [57].

Daily Port Scan Strategy. Analysing the number of distinct destination ports scanned within a day by each source IP address in June, we observe that 50% (50th percentile) of IP addresses target daily up to 1.04 ports (arithmetic mean), whereas 75% and 90% target daily up to 2.12 and 5.08 ports, respectively. In February, 50% of IP addresses target daily up to 1.93 ports, whereas 75% and 90% target daily up to 5.44 and 30.41 ports, respectively. Thus, most of scanners target daily only a small fraction of destination ports. Lastly, February shows more extensive port scanning and higher outliers compared to June.

Daily IP Address Scan Strategy. Analysing the number of distinct destination IP addresses scanned within a day by each source IP address, in June, we observe that 50% of IP addresses target daily up to 10.52 Darknet addresses (arithmetic mean), whereas 75% and 90% target daily up to 44.16 and 196.68 Darknet addresses. In February, 50% of IP addresses scan daily up to 10.68 Darknet addresses, whereas 75% and 90% target daily up to 47.89 and 253.27 Darknet addresses. Hence, the majority of scanners target only a small fraction of destination addresses, approximately 0.30% (Jun) and 0.14% (Feb) of our Darknet space.

Figure 5.7 shows the number of destination addresses per source address observed in our Darknet during each one-month period. Visual inspection yields that two months are highly similar. Approximately 359,100 IPs (23% of total IPs) in June scan only one destination address. The equivalent amount for February is approximately 337,800 IPs (26.47% of total IPs). Around 60% of the total source addresses target from one to 10 Darknet addresses in both months (59.98% in Jun, 60.32% in Feb).

Lastly, almost 6,000 source addresses (0.38% of total source IPs) correspond to the highest 10% of the source IPs which target the most destination addresses in June i.e. each one scans at least 58,909 Darknet addresses in a one-month period or at least 90% of the Darknet. February trends show that the 90th percentile corresponds to 3,319 source address (0.26% of total source IPs) which individually target at least 62,230 Darknet addresses in the one-month period, or at least 35.15% of the Darknet space. Overall, there is a small fraction of source IPs targeting almost one entire /16 network (upper-left corner of the plots), whereas the majority of scanners target few Darknet addresses (lower-right corner of the plots).

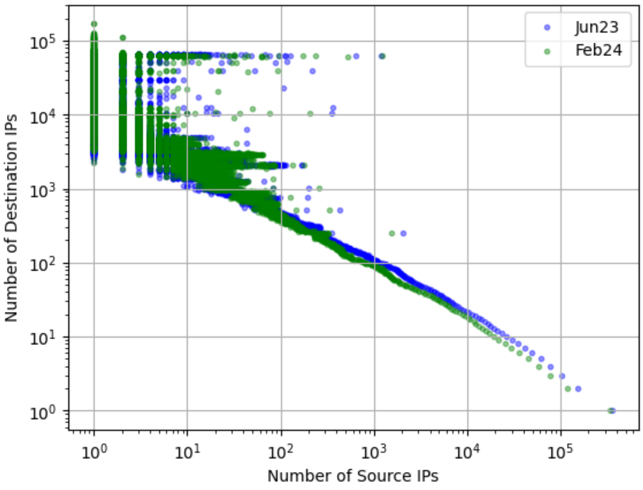
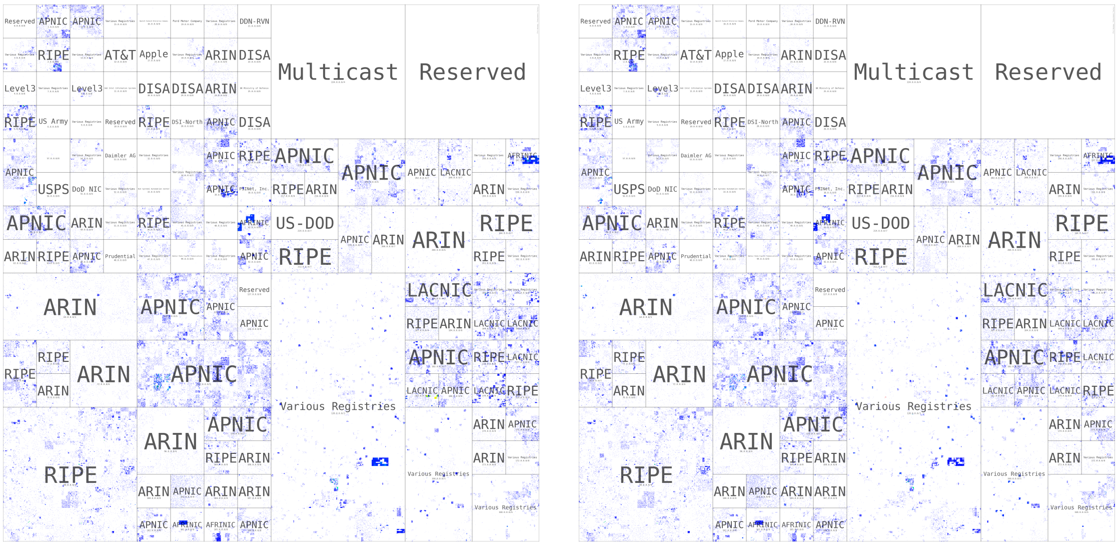


Figure 5.7: Number of Destination addresses scanned by each Source address

5.2. Autonomous Systems (AS) Profiling

Our research includes traffic originating from 12,352 Autonomous Systems (ASes), spread on 55,224 network prefixes and spanning 220 countries for June 2023. For February 2024, our study encompasses traffic sourced from 13,758 ASes, distributed across 60,190 network prefixes and extending across 225 countries. We note that 15,318 (3,134 in Feb) IPs cannot be mapped to an AS and 24 (14 in Feb) IPs cannot be mapped to a country. Cross-reference between the months shows 127,260 common IP addresses, corresponding to 8.15% and 9.97% of June and February datasets. Although the above percentages may seem insignificant, our perspective changes when we undertake an analysis of the common ASes and network prefixes. This is attributed to the fact that many IP addresses are expected to have changed due to DHCP, expiration of lease time in cloud hosting etc. In fact, there are 36,214 common network prefixes, which correspond to 65.57% for June and 60.16% for February. Additionally, we find 8,871 common ASes which map to 71.81% for June and 64.47% for February. Spatial visualization of Hilbert IPv4 maps for both months corroborates the above observation (fig. 5.8); maps exhibit congruence, wherein corresponding prefixes and Internet neighbourhoods align closely with each other.



(a)

(b)

Figure 5.8: Hilbert curve of IPv4 address space of scanner origins for (a) Jun 23 and (b) Feb 24

Our examination employs a two-fold approach: firstly we examine the distribution of IPs per AS and country and secondly we assess quantitatively the traffic volume from each AS and country.

Total IPs per AS. The majority of scanning IP addresses is concentrated to few major ASes. More precisely, figure 5.9 presents the top 20 ASes hosting the most IP address scanners observed in June 2023. TE Data (AS8452) - an Egypt Telecom subsidiary - holds the highest share of IP addresses per AS at 14.83%, followed by Chinese entities such as Chinanet (AS4134) at 9.91% and CHINA UNICOM China169 Backbone (AS4837) at 6.50%. National Internet Backbone (AS9829) of India, Telecomunicaciones MOVILNET (AS27889) and Iran Telecommunication Company PJS (AS58224) host 9.15%, 8.67% and 5.17% of the source addresses respectively. Combined, these six ASes provide 54.23% of the unique source addresses observed in June 2023.

Similar patterns are observed in data from February. Chinese entities such as CHINA UNICOM China169 Backbone and Chinanet are the top hitter ASes hosting together 22.78% of the total source IP addresses (fig. 5.10). Iran Telecommunication Company PJS, DIGITALOCEAN-ASN, National Internet Backbone and TE Data host 7%, 6.77%, 6.40% and 5.58% of total IPs respectively. Each of the rest 13,752 ASes hosts below 1.60% of the total IPs. Overall, 48.53% of the total source addresses reaching our Darknet is attributed to these six ASes. Notably, Telecomunicaciones MOVILNET plunged this month hosting only 99 scanning IP addresses.

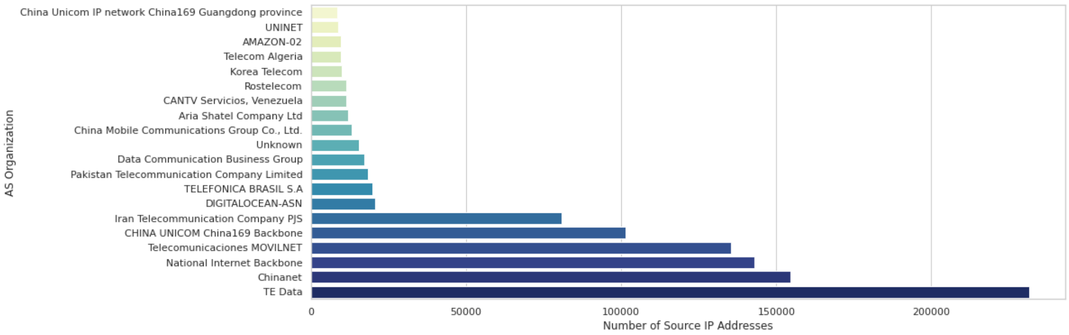


Figure 5.9: Top 20 ASes hosting the most IP address scanners observed in Jun 23

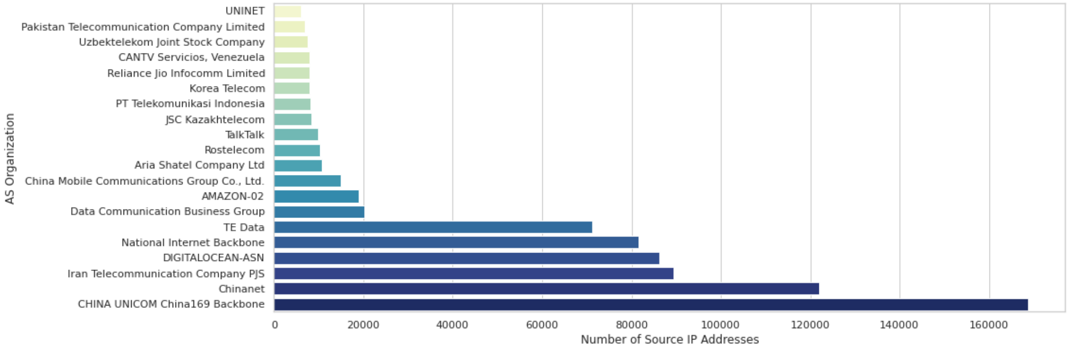


Figure 5.10: Top 20 ASes hosting the most IP address scanners observed in Feb 24

Total IPs per Country. The majority of scanning IP addresses is concentrated to few countries (fig. 5.11). Analysis from June 2023 shows that China leads with 21.08%, meaning that approximately 1 out of 5 IP addresses originate from China. Following closely are Egypt (14.97%), India (12.29%), Venezuela (9.79%) and Iran (8.23%). Rest countries contribute individually around 3.30% or less. Therefore, these five countries collectively account for 66.36% of all unique source IP addresses.

Regarding February 2024, China hosts the most scanning IP addresses (28.04%), followed by Iran (10.36%), India (9.87%), United States (6.51%) and Egypt (5.79%). The aforementioned five countries contribute 60.57% of the total IPs for this month. Overall, we note a recurrence and stability in the size of scanning IPs from the top hitter countries, particularly China, Egypt, Iran and India.

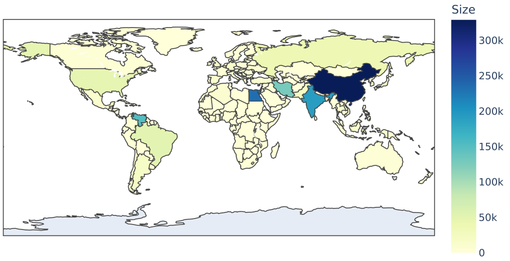


Figure 5.11: World map of source IP address concentration per country in Jun 23

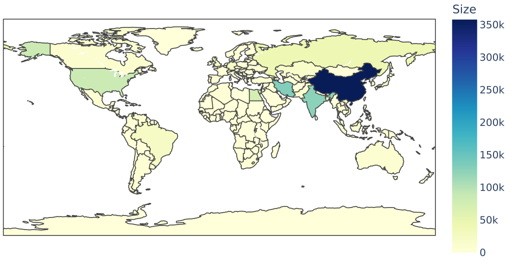


Figure 5.12: World map of source IP address concentration per country in Feb 24

Traffic Volume per AS. Examination of the traffic volume per AS reveals different AS numbers compared to the previous analysis. The top five ASes accounting for the highest traffic volumes remain the same for both months. GOOGLE-CLOUD-PLATFORM (AS396982) is one of the prominent ASes, accommodating 20.62% of the total traffic for June and 19.22% of the total traffic for February, respectively. IP Volume Inc. (AS202425), DIGITALOCEAN-ASN (AS14061), CENSYS-ARIN-01

(AS398324) and Chang Way Technologies Co. Limited (AS57523) follow Google, representing in June 15.91% (5.21% in Feb), 10.07% (9.78% in Feb), 7.27% (7.20% in Feb) and 6.96% (4.09% in Feb). Hence, Google significantly outpaces other ASes in terms of traffic volume. Yet, these five ASes comprise 60.83% of the total traffic in June and 45.51% for February.

It is notable that IP Volume has attracted negative publicity attributed to facilitating cyber crime, malware and operating as a bulletproof host. The Dutch Investigation Service for Financial and Tax Crime (FIOD) raided the company's data center in September 2020 [40]. On the other hand, CENSYS-ARIN-01 belongs to Censys Inc., known security industry company [31]. Thus, examination of Internet-wide scanning provides us with useful insights about the operation of ASes.

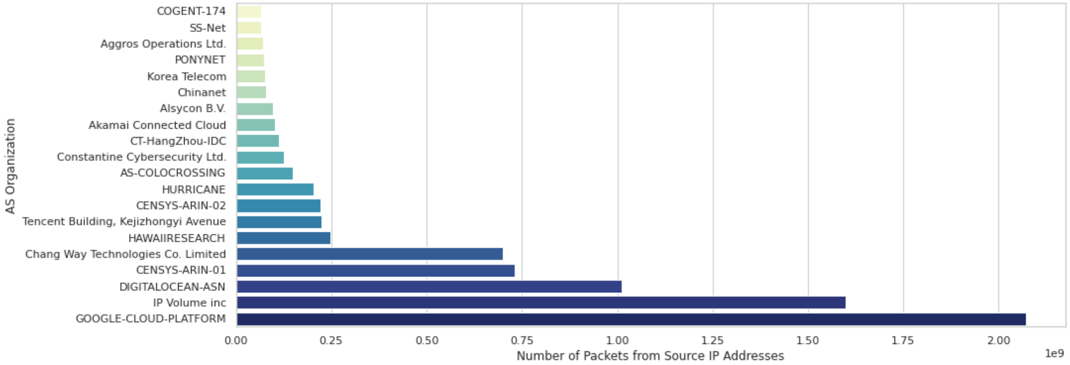


Figure 5.13: Top 20 ASes sending the most traffic to the Darknet in Jun 23

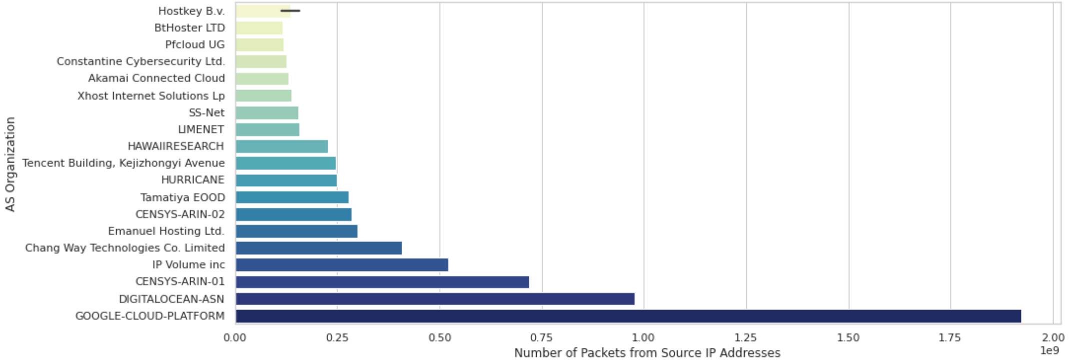


Figure 5.14: Top 20 ASes sending the most traffic to the Darknet in Feb 24

Traffic Volume per Country. Analysis of the traffic distribution per country highlights that the majority of traffic originates from few countries, with the top four heavy hitter countries being identical between the two months. Specifically, the United States (USA) constitutes the largest source of the overall traffic, contributing a significant percentage of 41.97% in June (fig. 5.15). Along with the Netherlands at 19.11%, these two countries contribute 61.08% of all incoming traffic. The respective percentages reach 41.02% and 13.06% in February, meaning that 54.08% of all incoming traffic traces back to these two countries (fig. 5.16). On the other hand, traffic from the rest countries is below 10% each, for both months. The United Kingdom and Russia contribute 9.62%, 8.40% in June and 8.74%, 8.02% in February. The combined traffic volume of these four countries accounts for 79.10% of the overall traffic in June, and reaches 83.20% if we include China (4.10%). The combined traffic volume of the remaining 215 countries is 16.80%. Accordingly for February, the above four countries contribute 70.84% of the total traffic and combined with Bulgaria (6.99%), China (4.39%) and Germany (4.07%) constitute a 86.29% of all incoming traffic.

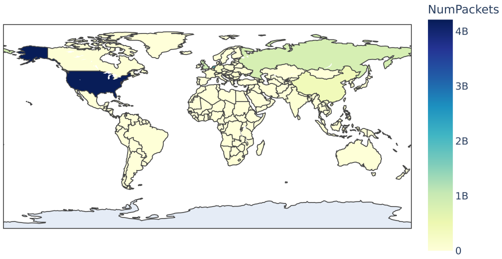


Figure 5.15: World map of traffic volume per country in Jun 23

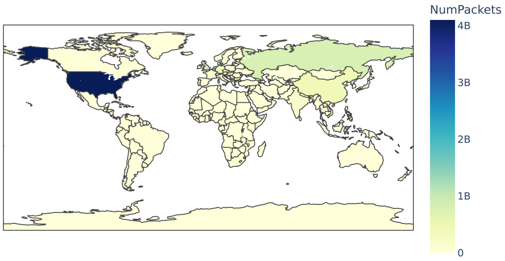


Figure 5.16: World map of traffic volume per country Feb 24

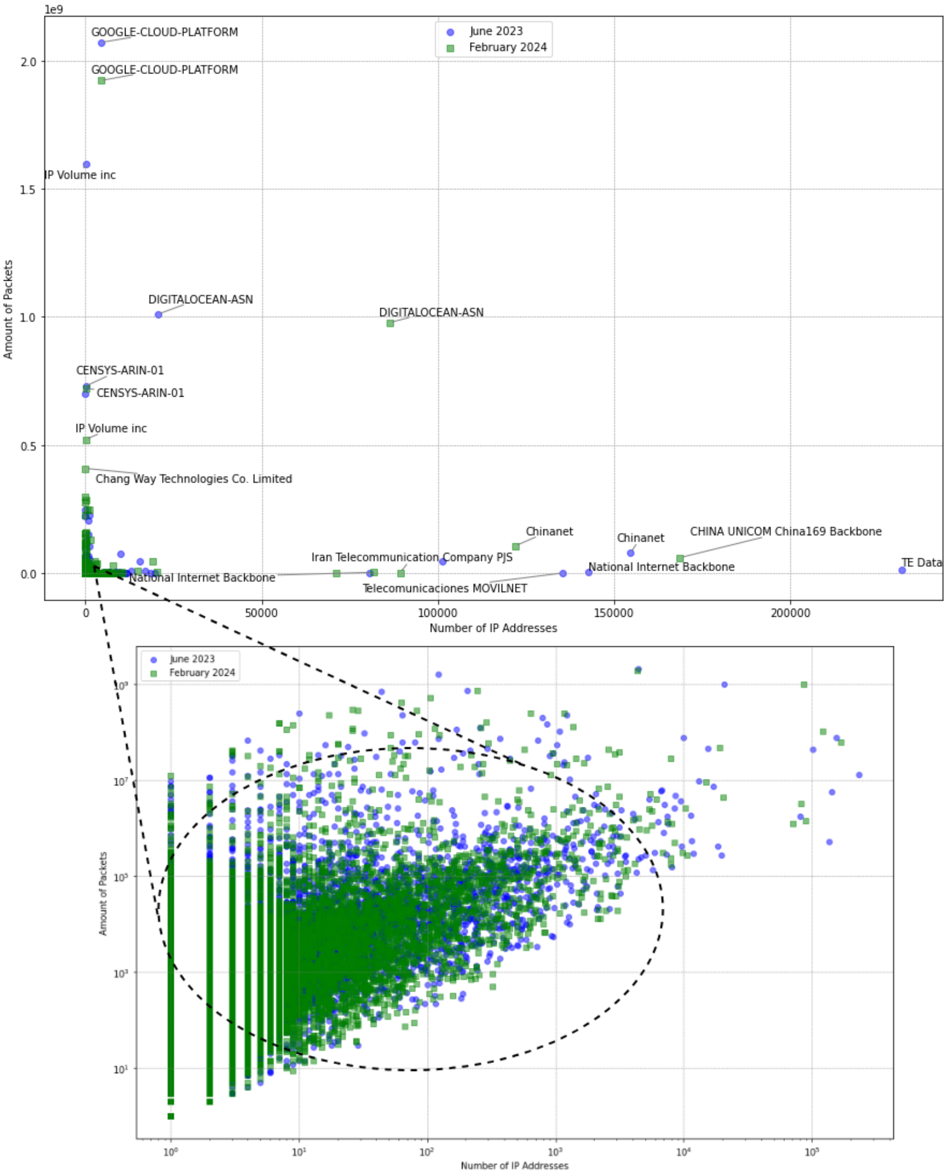


Figure 5.17: Distribution of ASes per number of IP addresses and traffic volume for Jun 23 and Feb 24

Comparison & Conclusion. In summary, our dataset for both months includes scanning traffic from over 12,000 ASes and over 200 countries. The majority of scanning IP addresses is concentrated to few major ASes and countries. Six ASes (or on average 0.046% of total ASes) host approximately

50% of the total source IP addresses. One would expect that the size of ASes is proportional to the traffic they generate. However, examination of the traffic volume per AS reveals different AS numbers compared to the previous analysis. The majority of traffic originates from few ASes and countries, with the top four heavy hitter countries being identical between the two months. Traffic in February appears more distributed across ASes, partially contributed to their larger amount comparing to June. The distribution of ASes per number of IP addresses and traffic volume for Jun 23 and Feb 24 (fig. 5.17) demonstrates similar results between the two months. This suggests recurring and consistent behavior. In relation to China, even though a large amount of Chinese IP addresses scan our Darknet, nonetheless, the amount of total traffic from China constitutes only approximately 4%. Conversely, the US exhibit the opposite behaviour, with four to seven times smaller IP pool size and 10 times larger traffic (around 41%).

5.3. Bad Internet Neighborhoods

Previous studies have demonstrated that malicious hosts are not distributed uniformly over the IPv4 address space [22, 20]. The reasoning for this concept is that networks and ASes have varying security policies. Hence, badly managed or tolerant networks are more likely to be abused than well-managed ones, increasing the concentration of malicious activity. According to Moura et al. [54], an Internet Bad Neighborhood is characterised as a collection of IP addresses grouped based on an aggregation criterion (i.e. network prefix) wherein a portion of the IP addresses engage in specific activities (i.e. scanning) for a predetermined duration (i.e. June 23/Feb 24). Furthermore, the likelihood of a particular IP address misbehaving increases when neighboring IP addresses of the same network prefix also misbehave [55].

Density Ratio. Based on the above definition, we can calculate the *density ratio* of each network prefix as follows:

$$\text{Density Ratio} = \frac{\text{\# of scanning IPs}}{2^{(32-n)}}$$

We consider the total network prefix size according to the CIDR notation i.e. including the network name and broadcast address. In order to identify bad scanning neighborhoods, we need to filter out those neighborhoods belonging to benign scanners such as research institutions, academia and the security industry. These entities may employ their own infrastructure (AS, advertised network prefixes etc.) or lease third-party hosting services.

Tables 5.3 and 5.4 list the top 10 prefixes ranked by density ratio. In June, 75% of the network prefixes present a density ratio up to 0.0027 (0.0029 in Feb), while the 99th percentile correspond to a density ratio of 0.0762 (0.0654 in Feb). Therefore, 99% of the observed network prefixes have a density ratio equal to or lower than 7.62%. It becomes clear that most network prefixes are sparse i.e. the number of scanning IP addresses per network prefix, lies below 8% for 99% of the networks. Therefore, we select as threshold the 99.9th percentile which corresponds density ratio of 0.45 (0.48 in Feb) and focus on the top-ranked highly-dense prefixes. Scrutiny of the top-0.1% of prefixes, for both months, reveals traffic from at least five known (benign) scanners: Alpha Strike Labs, Internet Census Group, Shadowserver, Driftnet and Cortex-Xpanse by Palo Alto Networks. On the other hand, we identify 70,687 IP addresses from 10 prefixes in June, and 1,166 IP addresses from nine prefixes in February which send exclusively traffic with the original Mirai signature. Highly dense Mirai prefixes correspond to six ASes in June and seven ASes in February. Three ASes are common between the two months: Iran Telecommunication Company PJS, China Mobile communications corporation, and China Mobile Communications Group Co. Ltd. The rest ASes for June are Telecomunicaciones MOVILNET (Venezuela), Iran Telecommunication Company PJS (Iran) and Tikona Infinet Ltd. (India). For February, the rest prefixes belong to Pishgaman Toseeh Ertebatat Company (Iran), Ipxo Limited (Germany), COGENT-174 (India) and ONERED JWG532 SRL (Dominican Republic).

Table 5.3: Top 10 prefixes ranked by ratio of number of scanning IPs per prefix by size of prefix (Jun 23)

Network	# of IPs	Network Size	**Density Ratio (%)**	AS Organization	ASN
125.164.21.248/31	2	2	100	PT Telekomunikasi Indonesia	AS7713
194.187.176.0/22	1,022	1,024	99.80	Alpha Strike Labs GmbH	AS208843
45.83.64.0/22	1,021	1,024	99.71	Alpha Strike Labs GmbH	AS208843
104.164.161.0/24	255	256	99.61	AS-WAVE-1	AS11404
104.232.39.0/24	255	256	99.61	VIVIDHOSTING	AS64200
104.165.105.0/24	254	256	99.22	GTT Communications Inc.	AS3257
39.144.17.0/24	254	256	99.22	China Mobile Communications Group Co.	AS9808
39.144.14.0/23	507	512	99.02	China Mobile communications corporation	AS56040
182.138.158.0/24	253	256	98.83	Chinanet	AS4134
193.163.125.0/24	252	256	98.44	Constantine Cybersecurity Ltd.	AS211298

Table 5.4: Top 10 prefixes ranked by ratio of number of scanning IPs per prefix by size of prefix (Feb 24)

Network	# of IPs	Network Size	**Density Ratio (%)**	AS Organization	ASN
140.99.12.0/24	256	256	100	AMAZON-02	AS16509
46.8.100.0/22	1,024	1,024	100	AMAZON-02	AS16509
140.99.52.0/24	256	256	100	AMAZON-AES	AS14618
140.99.53.0/24	256	256	100	AMAZON-02	AS16509
140.99.74.0/24	256	256	100	AMAZON-02	AS16509
140.99.61.0/24	256	256	100	AMAZON-02	AS16509
140.99.63.0/24	256	256	100	AMAZON-02	AS16509
140.99.221.0/24	256	256	100	AMAZON-02	AS16509
192.177.58.0/24	256	256	100	EGIHOSTING	AS18779
178.236.226.0/24	256	256	100	AMAZON-02	AS16509

Volume Ratio. To facilitate the analysis, we also consider a *volume ratio* calculated as the number of packets per network prefix divided by the total considered amount of packets for June 2023 (around 10 billion packets) and February 2024 (around 9.98 billion packets) respectively. Tables 5.5 and 5.6 list the top five prefixes ranked by volume ratio. In June, 99% of network prefixes presents a volume ratio up to 0.01% of the total traffic, while the 99.9% of prefixes has a volume ratio up to 0.23% (0.27% in Feb). Hence, the vast majority of network prefixes generates a small portion of the total traffic. Therefore, we set as threshold the 99.99th percentile, which corresponds volume ratio of 2.83% of total traffic, and study the top-ranked traffic-intense prefixes. Analysis of the top-0.01% reveals at least five known (benign) scanners in June - Cortex-Xpanse, Recyber.net, Censys, Shodan and Criminal IP - and at least three in February: Cortex-Xpanse, Censys and Academy for Internet Research LLC. Censys and Academy for Internet Research LLC employ their own AS, whereas the rest route their scanning traffic through cloud and hosting providers. On the other hand, we identify four Mirai-fingerprinted IP addresses in each month belonging to GOOGLE-CLOUD-PLATFORM from the US and India. It should be highlighted that the AS also contains (non-overlapping) traffic from one benign scanner.

Consequently, when anticipating attacks from unknown IP addresses, information on the concentration and intention of scanning hosts can enhance Cyber Threat Intelligence. Known scanners only collect security data without attempting to exploit potential vulnerabilities. Therefore, network administrators and SOC analysts can use this information for alert triage and to avoid triggering false-positive alerts. Accordingly, they can blacklist known bad scanning neighbourhoods where botnet and other malicious activity appears to take place consistently throughout time.

Table 5.5: Top 5 prefixes ranked by amount of traffic and ratio per prefix (Jun 23)

Network	# of Packets	Volume Ratio (%)	AS Organization	ASN
162.216.148.0/22	761,130,872	7.58	GOOGLE-CLOUD-PLATFORM	AS396982
35.200.0.0/14	756,599,249	7.54	GOOGLE-CLOUD-PLATFORM	AS396982
89.248.160.0/21	750,062,818	7.47	IP Volume inc	AS202425
94.102.48.0/20	727,715,777	7.25	IP Volume inc	AS202425
176.111.174.0/24	560,726,341	5.58	Chang Way Technologies Co. Limited	AS57523

Table 5.6: Top 5 prefixes ranked by amount of traffic and ratio per prefix (Feb 24)

Network	# of Packets	Volume Ratio (%)	AS Organization	ASN
162.216.148.0/22	672,155,667	6.72	GOOGLE-CLOUD-PLATFORM	AS396982
35.200.0.0/14	659,605,929	6.60	GOOGLE-CLOUD-PLATFORM	AS396982
79.110.62.0/24	298,907,023	2.99	Emanuel Hosting Ltd.	AS215766
162.142.125.0/24	298,830,073	2.99	CENSYS-ARIN-01	AS398324
198.235.24.0/24	291,251,089	2.91	GOOGLE-CLOUD-PLATFORM	AS396982

5.4. Aggressive Hitters (AH)

This section identifies and analyses the Aggressive Hitters observed by our Darknet. Internet scanners that exhibit excessive and persistent activity are known as Aggressive Hitters (AH) or Heavy Hitters (HH). They routinely scan the Internet for unsecured and public hosts. According to Anand et al. [3], these scanners can be divided into 1) benign research-oriented scanners used for Internet measurement and 2) malicious actors who search for vulnerable targets. Furthermore, due to their persistent and intrusive scanning efforts, these AH present higher chances of succeeding in identifying vulnerabilities at their target hosts.

We employ three definitions considered also by Anand et al. [3] to identify AH in our dataset: 1) address dispersion 2) total traffic volume and 3) number of distinct destination ports. Each definition yields a different result set, comprising approximately 10-15 thousand IP addresses.

5.4.1. Aggressive Hitters by Address Dispersion

We consider a source IP address appearing in our Darknet as aggressive if it targets at least 10% of the Darknet. For June, we identify 10,473 distinct IP addresses which contribute around 10.04 billion packets for a one-month period. In other words, 0.67% of total addresses accounts for 93.45% of the total traffic. For February, we identify 11,346 distinct IP addresses which contribute around 9.99 billion packets for a one-month period. In other words, 0.88% of total addresses accounts for 92.98% of the total traffic. Figures 5.18 and 5.19 present the distribution of AH IPs and traffic volume per AS and country comparatively for both months.

AH-1 for June 2023. The two leading ASes are DIGITALOCEAN-ASN (AS14061) and GOOGLE-CLOUD-PLATFORM (AS396982) hosting 0.18% and 0.11% of total source IP addresses. 85 addresses cannot be mapped to ASes and one of those cannot also be geolocated. In terms of traffic volume, we observe that Google produces more than 2.06 billion scanning packets, thereby becoming the AS that generates the most AH traffic (20.56% of total packets). IP Volume follows with more than 1.58 billion packets (15.75% of total traffic). Remaining ASes are below one billion packets for this one-month period.

Given the prominence of Google Cloud Platform combined with other US-based AHs, the United States hosts 0.33% of total observed IPs and attributes 39.92% of the total traffic, considered as a high-size high-volume AH country. The United Kingdom, China and The Netherlands follow next, hosting 0.06%, 0.05% and 0.03% of total source IPs and producing 9.10%, 8.22% and 18.84% of the total traffic volume respectively.

Analysing port scanning trends for June, ports TCP/443 and TCP/80 are targeted by over 5,000 AH IPs, while TCP/8080, TCP/8443, TCP/3389, TCP/8000 and TCP/8888 are targeted by over 4,000 AH IPs. In total, 65,523 ports are scanned by this set of AHs. In terms of traffic volume, top six ports that receive individually more than 80 million packets are the following: TCP/8080 (131.93 million), TCP/3389 (129.77 million), TCP/22 (126.82 million), TCP/80 (111.24 million), TCP/443 (90.60 million) and TCP/5555 (84.34 million).

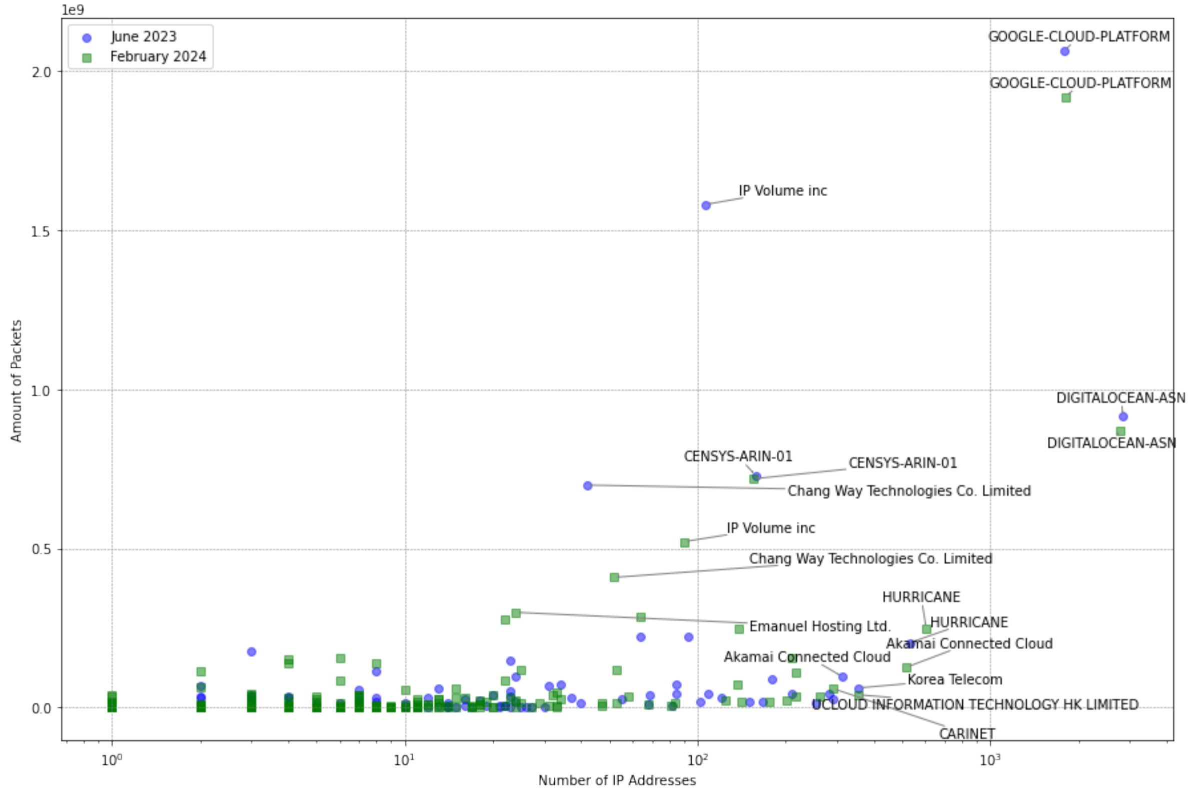


Figure 5.18: AH Definition 1. Distribution of ASes per number of aggressive IP addresses and traffic volume

AH-1 for February 2024. Similarly to June, the two leading ASes are DIGITALOCEAN-ASN and GOOGLE-CLOUD-PLATFORM hosting 0.22% and 0.14% of the total source IP addresses. The above ASes produce about 871.77 million and 1.92 billion scanning packets which correspond to 8.72% and 19.21% of total traffic. Remaining ASes host up to 0.05% of total IPs and contribute at most up to 7.19% of total packets for this one-month period. We note that 32 addresses cannot be mapped to ASes and one of those cannot also be geolocated.

The United States stands out in the number of hosted IP addresses, hosting 0.46% of total source IP addresses (51.18% of the AH subset) and attributing 38.55% of the total traffic. The United Kingdom and China follow next hosting 0.08% and 0.07% of total IPs, and contributing 8.39% and 3.38% respectively. Although each of the remaining countries hosts less than 0.04% of the total IPs, traffic analysis shows that the Netherlands (12.90%), Russia (7.31%) and Bulgaria (6.98%) account for a considerable amount of traffic. In fact, they compose the top-five heavy hitter countries in traffic volume along with the US (38.55%) and the UK (8.39%). The remaining AH countries lie below 4%.

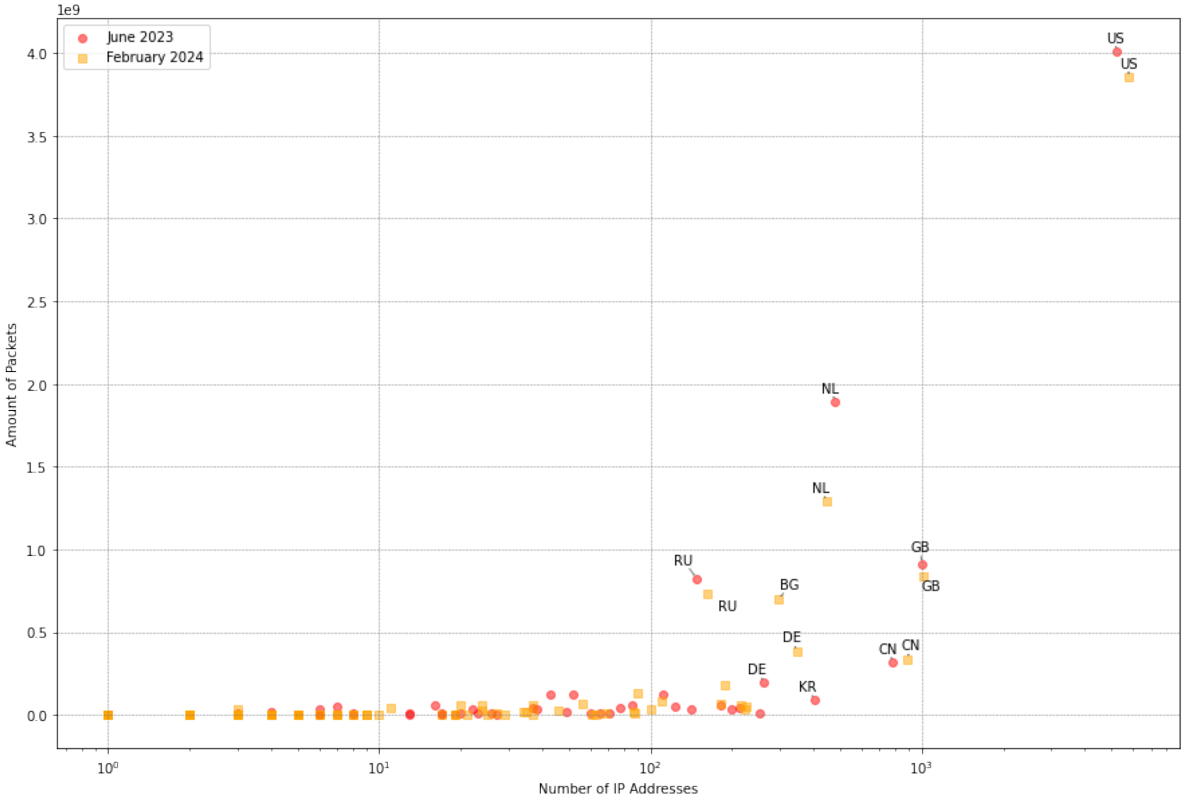


Figure 5.19: AH Definition 1. Distribution of countries per number of aggressive IP addresses and traffic volume

Port scanning trend analysis for February demonstrates that port TCP/443 is targeted by more than 6,000 IPs and TCP/80, TCP/8080, TCP/8443 and TCP/3389 are targeted by over 5,000 AH IPs (top-five). In total, 65,523 ports are scanned by this set of AHs. In terms of traffic volume, TCP/3389 receives more than 201.52 million packets in a one-month period, whereas ports TCP/22, TCP/80, TCP/8728 and TCP/443 receive over 100 million packets, composing the top-five ports.

5.4.2. Aggressive Hitters by Traffic Volume

In order to identify heavy hitters by traffic volume, we compile the Empirical Cumulative Distribution Function (ECDF) for the number of packets sent per IP address for a one-month period. Then, we consider the IPs of the 99th percentile as Aggressive Hitters. The subset of June corresponds to the top-1% IPs and contains 15,613 addresses that contribute around 9.77 billion to the total telescope traffic. In other words, 1% of total addresses accounts for 97.38% of the total traffic. Similarly for February, we identify 12,761 distinct IP addresses which contribute over 9.66 billion packets during the one-month period. In other words, 1% of total addresses accounts for 96.65% of the total traffic. Figures 5.20 and 5.21 present the distribution of AH IPs and traffic volume per AS and country comparatively for both months.

AH-2 for June 2023. The two leading ASes are DIGITALOCEAN-ASN and GOOGLE-CLOUD-PLATFORM hosting 0.28% and 0.12% respectively. Each of the remaining ASes hosts less than 0.04% of total IP addresses. 96 addresses cannot be mapped to ASes and one of those cannot also be geolocated. Although DigitalOcean hosts the most AH addresses, Google produces more than 2.06 billion scanning packets (20.59% or one fifth of total traffic), thereby becoming the AS that produces the most scanning traffic. IP Volume follows with more than 1.59 billion packets (15.91% of total traffic). The rest ASes generate below one billion packets during the investigated period.

Geolocation analysis shows that the United States hosts 0.41% of the total source IP addresses and attributes 41.67% of the total traffic. China and the United Kingdom follow next - hosting 0.11%

and 0.08% of the total IPs - and the rest countries host at most 0.05% of total IP addresses. In terms of traffic, the Netherlands (19.05%), the UK (9.57%), Russia (8.33%) and China (3.41%) along with the US formulate the top-five heavy hitter countries in traffic volume. The remaining AH countries generate less than 2.10% of the total traffic.

Examination of port scanning trends for June yields similar results to Definition-1. Ports TCP/443 and TCP/80 are targeted by over 6,000 AH IPs, while TCP/8080, TCP/22, TCP/8443 and TCP/8000 are targeted by over 5,000 AH IPs. In total, 65,523 ports are scanned by this set of AHs. In terms of traffic volume, the top four ports that receive individually more than 100 million packets are the following: TCP/22 (155.08 million), TCP/8080 (135.71 million), TCP/3389 (131.64 million) and TCP/80 (119.25 million).

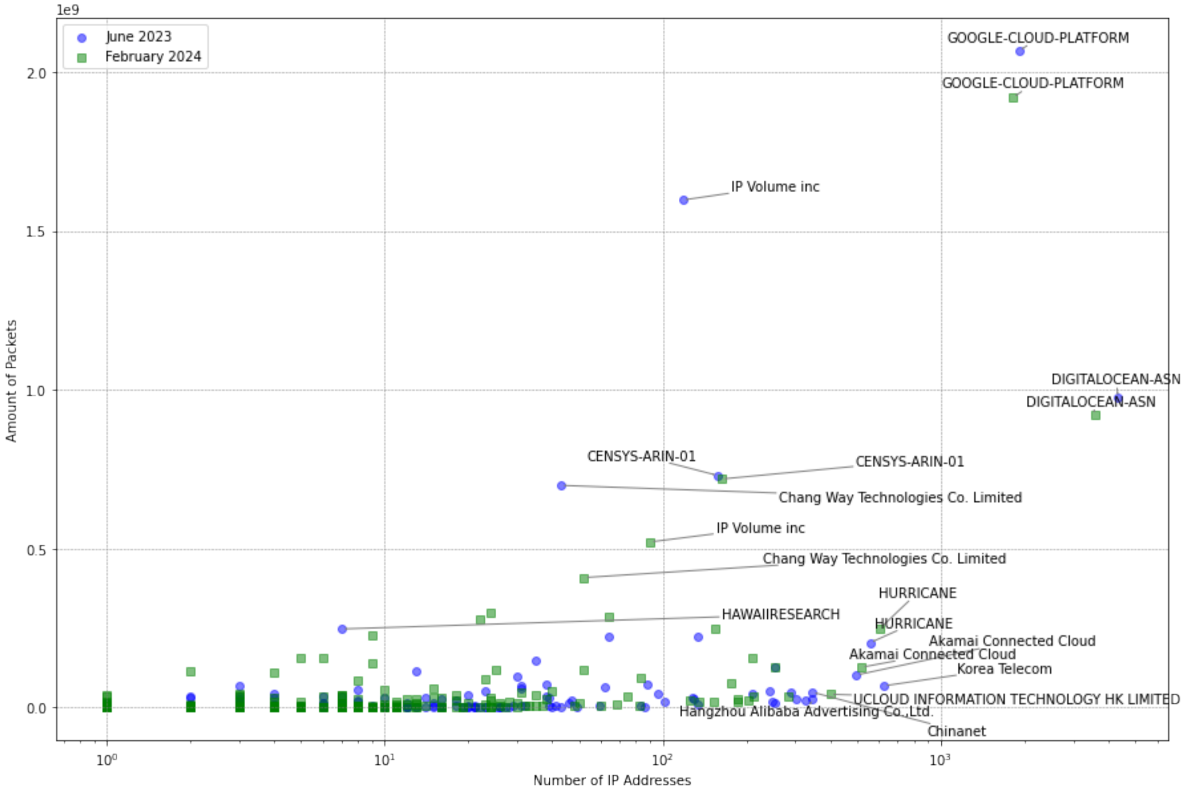


Figure 5.20: AH Definition 2. Distribution of ASes per number of aggressive IP addresses and traffic volume

AH-2 for February 2024. Similar to June, the two leading ASes are DIGITALOCEAN-ASN and GOOGLE-CLOUD-PLATFORM hosting 0.28% and 0.14%. 35 addresses cannot be mapped to ASes and one of those cannot also be geolocated. In terms of traffic volume, we observe that Google produces about 1.92 billion scanning packets (19.20% of the total traffic), thereby becoming the AS that generates the most scanning traffic. Remaining ASes are below one billion packets for the given time period, with DIGITALOCEAN-ASN the second highest at 920.49 million packets or 9.21% of the total traffic.

Geolocation analysis yields similar results to June. In particular, the United States hosts 0.47% of total IP addresses and attributes 40.41% of the total traffic. Although the United Kingdom and China follow next hosting 0.09% and 0.08% of total IPs, nonetheless they contribute considerably less traffic; 8.66% and 3.56%. The rest countries host at most 0.04% of total IP addresses. In terms of traffic, the Netherlands (12.95%), Russia (7.94%) and Bulgaria (6.98%) along with the US and the UK constitute the top-five heavy hitter countries in traffic volume, accounting for 76.94% of the total traffic. The remaining AH countries generate up to 4% of the total traffic.

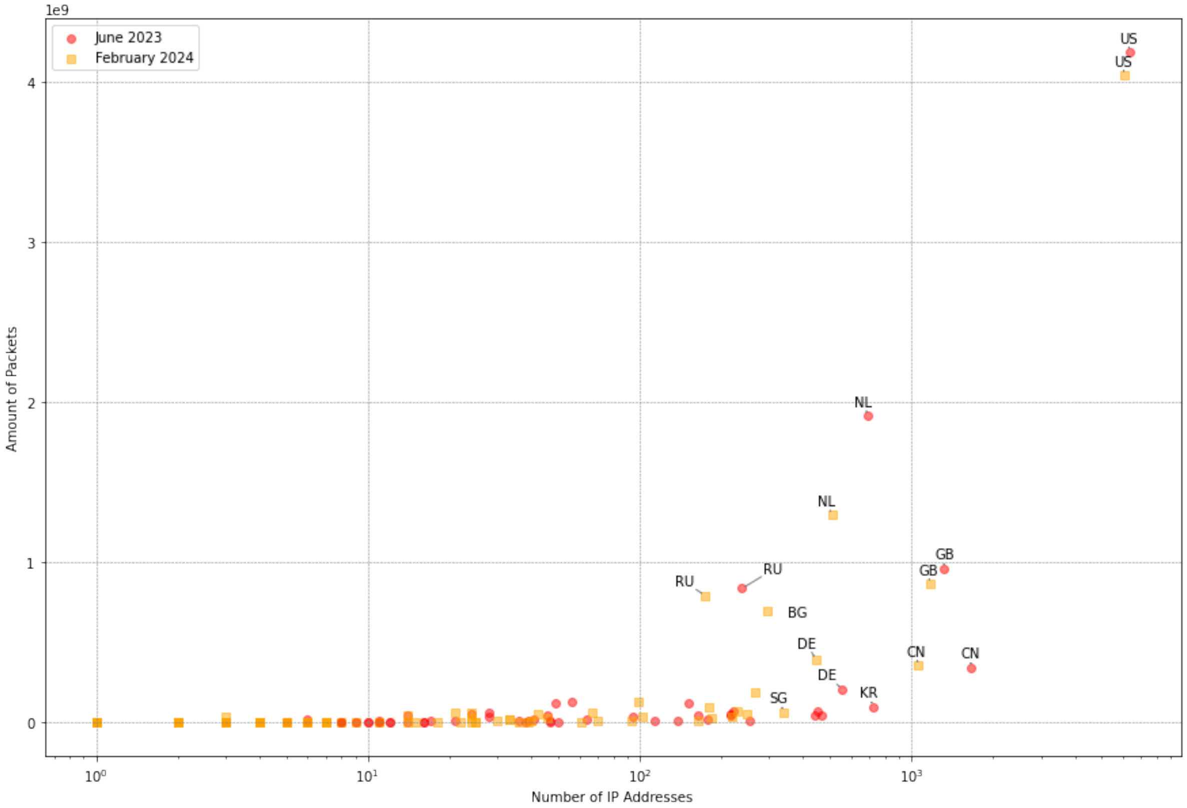


Figure 5.21: AH Definition 2. Distribution of countries per number of aggressive IP addresses and traffic volume

Port scanning trend analysis for February demonstrates that ports TCP/443, TCP/80, TCP/8080, TCP/8443 and TCP/4444, TCP/3389 and TCP/8000 are targeted by over 5,000 AH IPs (top-seven). In total, 65,523 ports are scanned by this set of AHs. Results are similar for Definition-1 in terms of traffic volume. TCP/3389 receives more than 202.43 million packets in a one-month period, whereas ports TCP/22, TCP/80, TCP/443 and TCP/8728 receive over 100 million packets, composing the top-five ports.

5.4.3. Aggressive Hitters by Number of Distinct Destination Ports

Similarly to the Definition-2, we compile the Empirical Cumulative Distribution Function (ECDF) for the number of destination ports targeted by each source IP address for the one-month period. We consider the IPs of the 99.5th percentile as Aggressive Hitters. The subset of June contains 9,601 addresses that contribute around 7.35 billion to the total telescope traffic. In other words, 0.61% of total addresses accounts for 73.29% of the total traffic. Our definition identifies 6,411 distinct IP addresses, in February, which contribute over 3.42 billion packets during the one-month period. In other words, 0.50% of total addresses accounts for 34.24% of the total traffic. Figures 5.22 and 5.23 present the distribution of AH IPs and traffic volume per AS and country comparatively for both months.

AH-3 for June 2023. Apart from the two known prominent ASes, DIGITALOCEAN-ASN (0.10%) and GOOGLE-CLOUD-PLATFORM (0.09% of the total traffic), Alpha Strike Labs GmbH appears to host the most AH IPs under this definition (0.13% of the IPs). Alpha Strike Labs is a German security research firm that specializes in Cyber Open Source Intelligence and scans the Internet to identify attack surfaces [2]. Each of the remaining ASes hosts up to 0.03% of total IP addresses (or about 500 IPs). Four addresses cannot be mapped to ASes. Although Alpha Strike Labs hosts the most AH addresses, Google produces more than 2.03 billion scanning packets (20.31% of total traffic), thereby becoming the AS that produces the most scanning traffic. IP Volume follows with more than 1.47 billion packets (14.73% of total traffic). The rest ASes generate below one billion packets for this one-month period.

Given the prominence of Google and DigitalOcean, the United States hosts 0.21% of the total IPs and attributes 31.96% of the total traffic. Germany and the United Kingdom follow next, - hosting 0.14% and 0.07% of total IPs - while the rest countries host at most 0.03% of total IP addresses. In terms of traffic, the Netherlands (15.77%), the UK (9.05%) and Russia (7.41%) constitute along with the US the top-four heavy hitter countries in traffic volume. Each of the remaining AH countries generates below 1.70% of the total traffic.

Examination of port scanning trends for June yields similar results to Definition-1. Ports TCP/80 and TCP/8080 are targeted by over 7,000 AH IPs, while TCP/443, TCP/8443, TCP/3389, TCP/8081 and TCP/8888 are targeted by over 6,000 AH IPs. In total, 65,523 ports are scanned by this set of AHs. When analysing the distribution per traffic volume, the top-six ports are identical to Definition-1 except for TCP/8443. Each port receives at least 15.98 million packets.

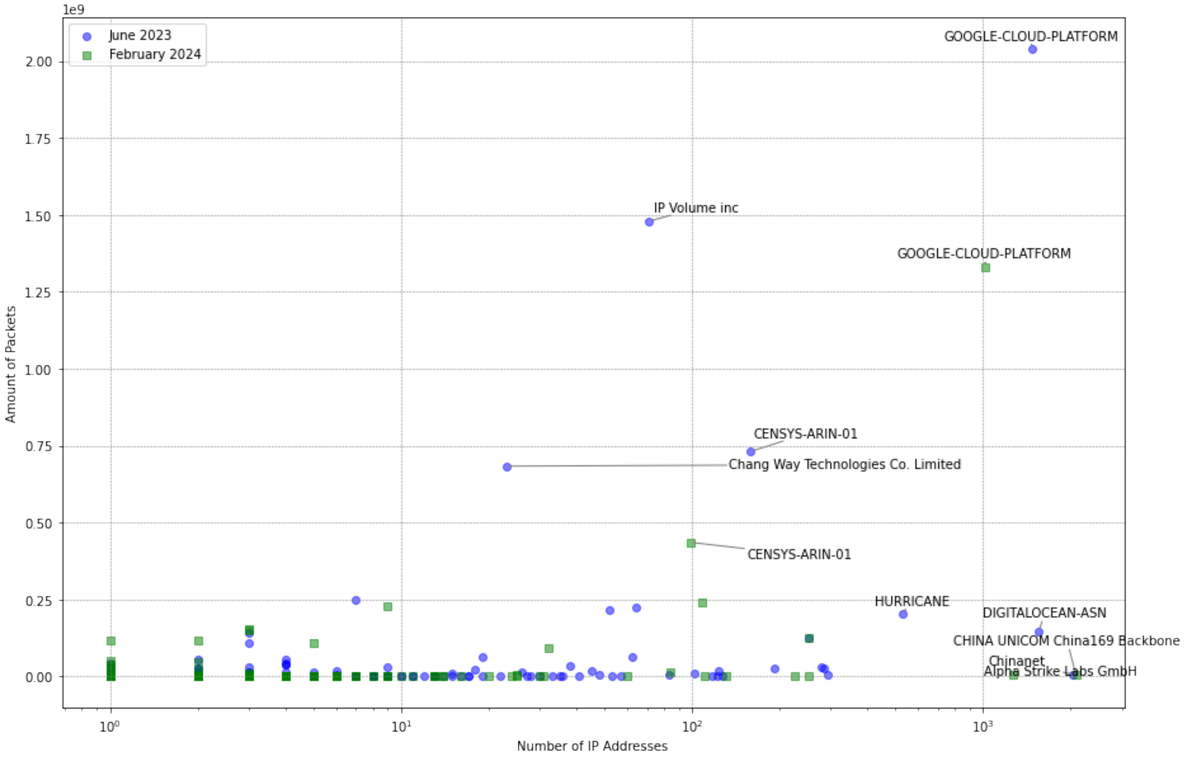


Figure 5.22: AH Definition 3. Distribution of ASes per number of aggressive IP addresses and traffic volume

AH-3 for February 2024. Results obtained under this definition exhibit considerable disparity regarding the distribution of AH IPs per AS. Two Chinese ASes, CHINA UNICOM China169 Backbone (AS4837) and Chinanet (AS4134) exhibit the highest ratio (and amount) of AH IP addresses per AS, collectively hosting 0.26% of total addresses. Considering the rest smaller Chinese ASes, China accommodates 0.33% of total addresses. Next, GOOGLE-CLOUD-PLATFORM hosts 0.08% of the total IP addresses and accounts for 13.31% of the total traffic. The rest ASes host below 0.02%.

Geolocation analysis shows that China is followed by the US and the UK in the number of hosted aggressive addresses, each hosting 0.06% of total IPs and attributing 15.06% and 7.87% of the total traffic. The rest countries host at most 0.01% of total IP addresses and contribute at most around 3% of the total traffic. It should be highlighted that even though China hosts the highest amount of aggressive IP addresses under this definition, nonetheless it generates lower aggressive traffic (0.18% of total packets) comparing to other heavy hitting countries.

Port scanning trend analysis for February demonstrates that the top-five ports per number of IP ad-

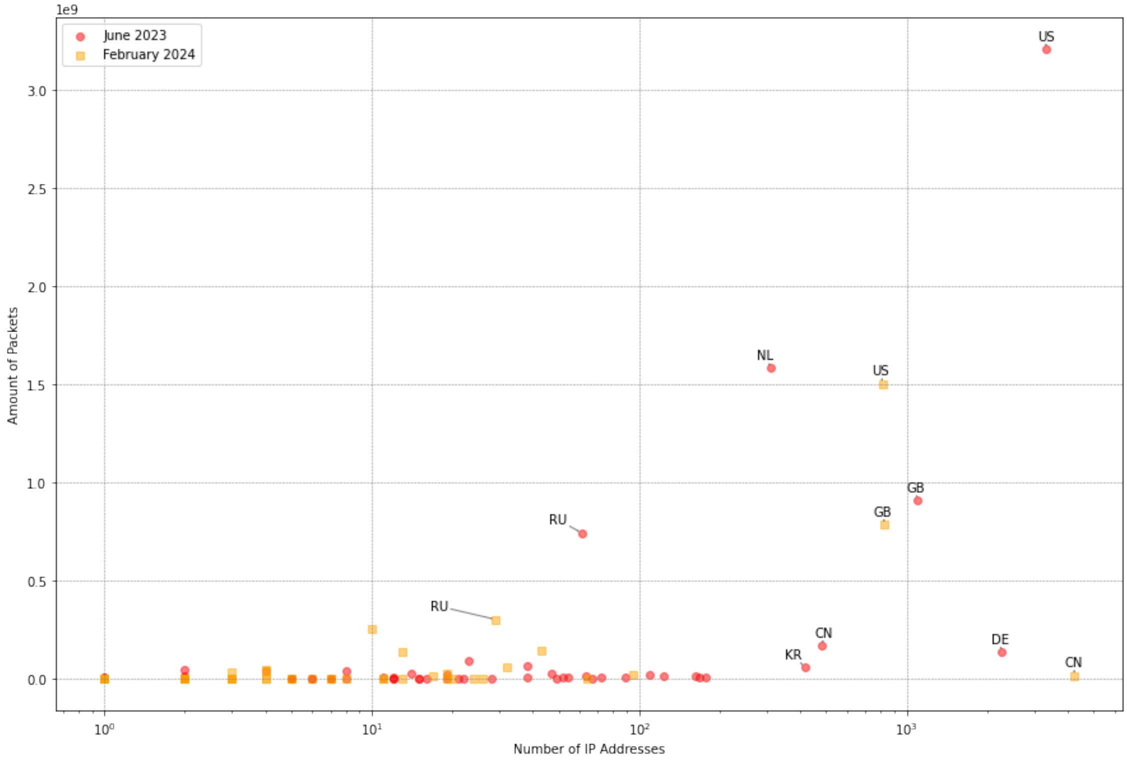


Figure 5.23: AH Definition 3. Distribution of countries per number of aggressive IP addresses and traffic volume

dresses are TCP/80, TCP/443, TCP/8080, TCP/2323 and TCP/9527. In total, 65,523 ports are scanned by this set of AHs. Results are similar for Definition-1 in terms of traffic volume. Ports TCP/443 and TCP/80 receive more than five million packets in a one-month period, while ports TCP/8080, TCP/7777, TCP/8888 and TCP/4444 receive over two million packets, composing the top-six ports.

5.4.4. Comparison of monthly AH Datasets

Figures 5.24 and 5.25 illustrate Venn diagrams of the AH subsets under each definition for both months. Examining the data from June, all AHs under Definition-1 fall also under Definition-2 .i.e. all aggressive scanners that target at least 10% of the Darknet, also generate the highest volume of traffic. Similar behavior is observed in February, where Definition-1 and Definition-2 have in common 9,080 IP addresses. On the other hand, it appears that identified AH IP addresses under Definition-3 present substantial differences compared to the other two definitions.

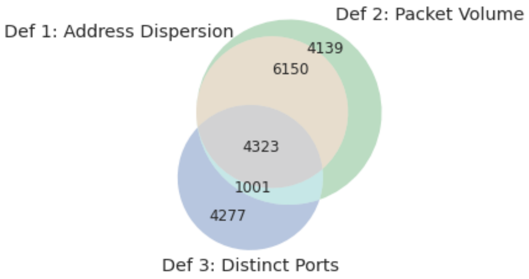


Figure 5.24: Venn diagram of AH subsets based on the three definitions (Jun 23)

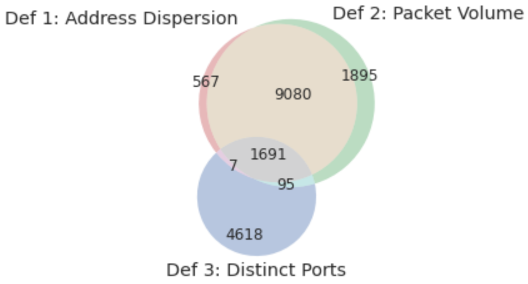


Figure 5.25: Venn diagram of AH subsets based on the three definitions (Feb 24)

For the AH subset under Definition-1, we obtain 4,163 common IP addresses, which constitute 39.74% of 10,473 IPs in June and 36.69% of 11,346 IPs in February. We observe 326 common ASes between the two months, which represents 50.15% of 650 ASes in June and 40.34% of 808 ASes in February. Furthermore, there are 774 common network prefixes (37.40% of the 2,069 prefixes in June and 34.15% of 2,266 prefixes in February). Therefore, the two months have on average 38.22% common IP addresses, 45.25% common ASes and 35.78% common network prefixes, under Definition-1.

Regarding Definition-2, we observe 4,377 common IP addresses, which constitute 28.03% of 15,613 IPs in June and 34.29% of 12,761 IPs in February. Also, we find 456 common ASes, which constitutes 41.68% of 1,094 ASes in June and 51.23% of 890 ASes in February. There are 1,060 common network prefixes (or 29.48% of the 3,598 prefixes in June and 41.67% of 2,546 prefixes in February). Hence, the average percentage of common IP addresses between the two months, under Definition-2 is 31.16%; the average percentage of common ASes is 46.45%; and finally the average percentage of network prefixes is 35.58%.

Lastly, we obtain 1,581 common IP addresses for Definition-3, which constitute 16.46% of 9,601 IPs in June and 24.66% of 6,411 IPs in February. We observe 100 common ASes which maps to 20.92% of 478 ASes in June and 46.94% of 213 ASes in February. The number of common network prefixes is 134 for the two months (10.49% of 1,277 prefixes in June and 23.10% of 580 ASes in February). Thus, the average percentages of common IP addresses, ASes and network prefixes between the two months, under Definition-3, are 20.56%, 33.93% and 16.79%, respectively.

5.4.5. Who are the Aggressive Hitters?

This section serves as a introduction to our research on known scanners and fingerprint-related actors that is presented in the next chapters. Based on the three definitions we obtain a total of 19,890 distinct source IP addresses in June 2023 and 17,953 distinct source IP addresses in February 2024. Each address has been filtered under different criteria, so not all addresses exhibit the same behaviour e.g. an IP scanner might target many ports but generating low volume of scanning traffic comparing to other aggressive hitters.

We cross-reference the AH IP addresses with those of known scanners and fingerprint-related actors and profile 16,670 IP addresses in June (83.81% of total AH IPs). We identify with high confidence 1) 5,562 IP addresses belonging to 33 known scanners 2) 2,117 Mirai scanners 3) 5,088 Zmap IP addresses of unknown intention and 4) 31 Masscan IP addresses of unknown intention. Furthermore, we identify with medium confidence 5) 126 IPs possibly infected with Mirai 6) 1,350 IPs possibly employing Zmap 7) 1,422 IPs possibly employing Masscan. These are IP addresses which send at least one fingerprinted packet but not all packets of a time-bounded and IP-specific scanning event exhibit the signature. Lastly, we have 8) seven IPs sending both Mirai and Zmap-fingerprinted traffic 9) five IPs sending both Mirai and Masscan-fingerprinted traffic and 10) 962 IPs sending both Zmap and Masscan-fingerprinted traffic. One hypothesis is that the above IP addresses lie behind the Network Address Translation (NAT) protocol and, therefore, there are two devices scanning. For example, the router is infected with Mirai and there is a user actor behind that who performs scanning. In case 10, one more hypothesis is that actors employ proprietary software that combines both scanning tools.

Regarding February, we profile 12,336 IP addresses (68.71% of total AH IPs). Specifically, We identify with high confidence 1) 5,589 IP addresses belonging to 33 known scanners 2) 397 Mirai scanners 3) 2,538 Zmap IP addresses of unknown intention 4) 37 Masscan IP addresses of unknown intention and 5) 56 IPs where all the packets of the scan contain both Mirai and Zmap signatures. Furthermore, we identify with medium confidence 6) 66 IPs possibly infected with Mirai 7) 1,461 IPs possibly employing Zmap 8) 1,256 IPs possibly employing Masscan. Lastly, we have 9) 13 IPs sending both Mirai and Zmap-fingerprinted traffic 10) three IPs sending both Mirai and Masscan-fingerprinted traffic and 11) 920 IPs sending both Zmap and Masscan-fingerprinted traffic.

5.5. Anonymous Actors via Onion Routing

The Tor Project and Onion Routing are technologies aiming to improve Internet privacy. Data is encrypted and routed through a number of network nodes - onion routers - before arriving at its final destination. Each node in the network only knows the previous and next nodes in the chain, guaranteeing that no node sees both the sender and recipient of data. This procedure anonymizes a user's online activities and makes it challenging for recipients to identify the original source. We obtained an IP list of Tor exit nodes (of 19 Apr 24) [72] and cross-referenced with source IPs appearing in our telescope to detect Tor traffic.

June 2023. Analyzing Tor traffic from June identifies 685 IPs (or 0.04% of total IPs) that account for 1.93 million packets (0.01% of total traffic). 44.59% of Tor traffic targets port 443, followed by port 80 (21.92%), 8006 (7.44%) and 8080 (6.17%). Rest ports receive less than 4.4% of the total Tor traffic. The majority of scanning activity targets a limited number of ports and IP addresses. More precisely, 50% of Tor IPs scan up to 1.48 ports (arithmetic average), 75% up to 2.20 ports and 90% up to 2.64 ports. Respectively, 50% of Tor IPs target up to 16.54 Darknet IP addresses, 75% up to 45.40 IPs and 90% up to 97.38 IPs. Detection of scanning events involves identifying partial-Masscan and partial-ZMap events and a combination of fingerprints in the same event. Nonetheless, the percentage of fingerprinted packets in each scanning event reaches a maximum of 0.31% for ZMap and 0.23% for Masscan and does not enable any further efficient tool characterization.

AS and geolocation analysis shows that the German Zwiebelfreunde e.V. (AS60729) is the most prominent AS, hosting 22.19% of Tor IPs and generating 31.79% of Tor traffic from the following domains: *for-privacy.net, relayon.org, artikel10.org, cccs.de*. PONYNET (AS53667) follows next hosting 10.07% of IPs and accounting for 5.78% of Tor traffic. Lastly, SURF B.V. (AS1101) contains 8.18% of IPs and contributes 15.20% of Tor traffic. Rest ASes host less than 3.50% of Tor IP addresses and generate less than 5.4% of Tor traffic. Mapping AS to countries, Germany, the Netherlands and the US host 30.36%, 14.89% and 13.58% of Tor IPs and produce 41.09%, 22.29% and 11.71% of Tor traffic. Hence, 58.83% of Tor IPs and 75.09% of Tor traffic originates in three countries.

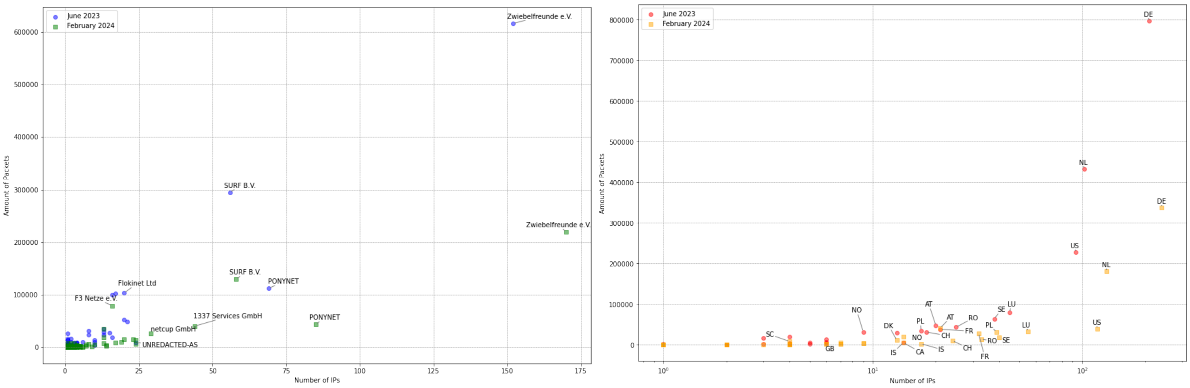


Figure 5.26: Distribution of (a) ASes and (b) countries per number of Tor IP addresses and traffic volume

February 2024. We identify 886 Tor IPs (or 0.06% of total IPs) that account for 806,123 packets (<0.01% of total traffic). 58.32% of Tor traffic targets TCP/443 port, followed by port 80 (19.61%). Rest ports receive at most 5% of the total Tor traffic. The majority of scanning activity targets a limited number of ports and IP addresses. More precisely, 50% of Tor IPs target up to 1.10 ports (arithmetic average), 75% up to 1.79 ports and 90% up to 2.13 ports. Respectively, 50% of Tor IPs scan up to 8.68 Darknet IP addresses, 75% up to 19.31 IPs and 90% up to 31.8 IPs. Detection of scanning events involves identifying partial-Masscan and partial-ZMap events and a combination of fingerprints in the same event. Again, the percentage of fingerprinted packets in a scanning events is insignificant (<0.28%) and does not any provide sufficient evidence for scanning tool characterization.

AS and geolocation analysis shows similar results to June. Zwiebelfreunde e.V. (AS60729) is the most prominent Tor AS, hosting 19.19% of Tor IPs and generating 27.23% of Tor traffic. PONYNET (AS53667) follows next hosting 9.59% of IPs and accounting for 5.39% of Tor traffic. Lastly, SURF B.V. (AS1101) contains 6.55% of IPs and contributing 16.06% of Tor traffic. Rest ASes host less than 5% of Tor IP addresses and generate less than 10% of Tor traffic. Furthermore, Germany and the Netherlands host 27.09% and 14.79% of Tor IPs and produce 41.84% and 22.39% of Tor traffic. These two countries account for 64.23% of Tor traffic and, when combined with the US (13.32% of Tor IPs), they host 55.2% of Tor IPs.

Comparison & Conclusion. Scrutiny indicates that the percentage of source IPs originated in Tor exit nodes constitutes a minor fraction, ranging between 0.04% to 0.06%. June data encompasses 36 countries, while data from February includes 47 countries, representing a wider range. Despite the breadth, Tor traffic remains negligibly small and accounts for merely 0.01% of the overall Darknet traffic for each month. Interestingly, there is a 96.35% recurrence rate in IPs from June to February. Common ASes constitute 93.18% of the 132 ASes in June and 73.21% of the 168 ASes in February. This indicates that Tor exits used for scanning that reach our telescope are in principle stable. Comparatively, three ASes and three countries stand out for hosting the bulk of Tor IPs and producing the largest amount of Tor traffic (fig. 5.26). Germany stands out as the most prominent with respectable IP size and traffic volume. Lastly, we identify signature evidence supporting the use of ZMap and Masscan tools.

6 Known Scanners and Bots

In this section, we study scanning entities which engage with the community, hereby referred to as known scanners. We extend the notion to also include industry bots that hit our Darknet. First, we explain our methodology and sources of data collection and aggregation. Then, we study known scanners and known bots separately.

6.1. Introduction

Known scanners are organizations and entities with internet-wide scanning activity that is believed to be non-hostile. These organizations usually host a public website describing their goals, listing their IP addresses or network prefixes and providing an opt-out or abuse complaint method [21]. The organizations might be related to the security industry (e.g. Censys, Shodan), non-profit organizations (e.g. Shadowserver) and academic institutions (University of Michigan, TU Munich etc.). We extend the notion of known scanners by including in our analysis known (good) bots. Search engine bots routinely scan the web and for new services, assisting search engines such as Google and Bing to provide accurate and speedy search results. Website monitoring bots regularly analyze website performance and identify updates or failures (Uptimerobot, Better Stack etc.). Furthermore, bots are used to collect data for market research allowing businesses to examine trends (DataForSEO Link Bot, SEOkicks etc.). In this chapter, we first describe the process of data collection and aggregation, then we analyse known scanners and finally we present our findings on known bots.

6.2. Data Collection & Aggregation

In order to identify known scanners and known bots reaching our Darknet, we collect and aggregate data from diverse sources. To integrate data, we employ a three-phase data warehousing and analytics process, called ETL (Extract, Transform, Load), which transforms data from several sources into a target database or data warehouse. The data is extracted from multiple resources, transformed by cleaning and filtering and finally stored into the data warehouse.

6.2.1. Data Sources

Data is gathered from the following sources:

E1. **Acknowledge Scanner Repository (ASR)**: The Acknowledge Scanner repository is a public data repository listing organizations and entities with internet-wide scanning activity that is believed to be non-hostile. These organizations usually host a public website describing their goals, listing their IP addresses or network prefixes and providing an opt-out or abuse complaint method. The repository contains lists of scanning IP addresses from 41 organizations and is based on the work by Collins et al. [21].

E2. **GreyNoise**: GreyNoise deploys distributed honeypot sensors across several cloud providers globally and offers a platform and data feed service to assist in identifying and filtering Internet

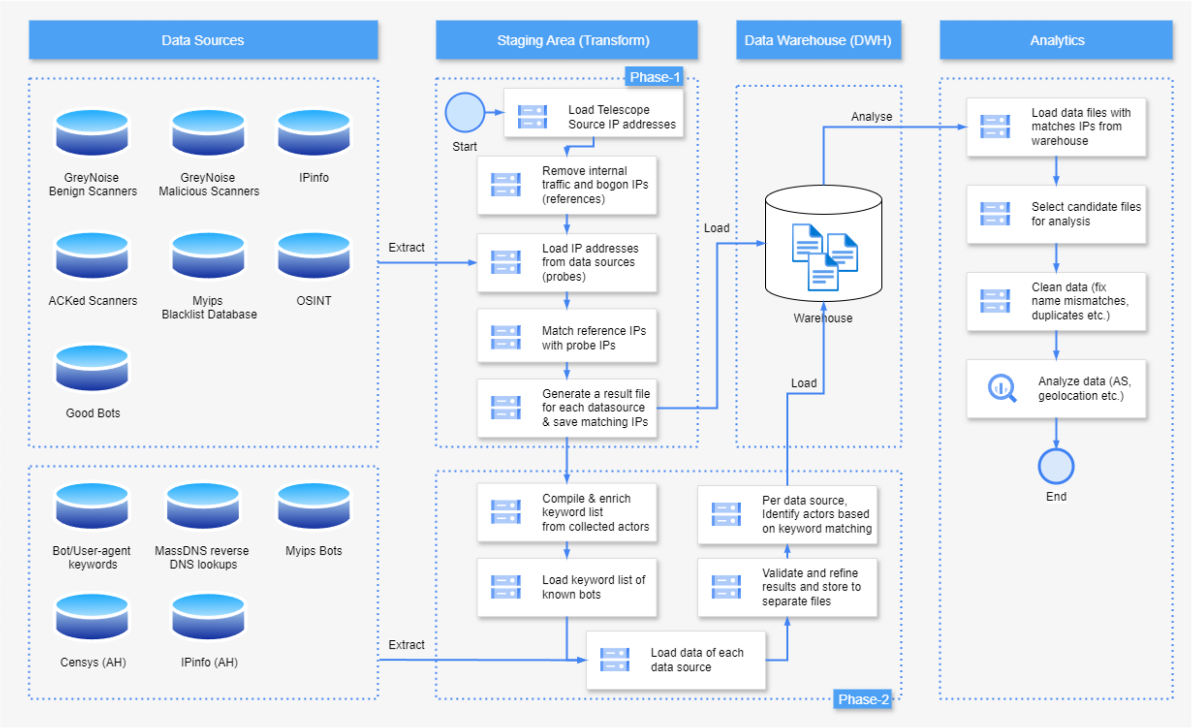

Data Sources
Staging Area (Transform)
Data Warehouse (DWH)
Analytics
GreyNoise Benign Scanners
GreyNoise Malicious Scanners
IPinfo
ACKed Scanners
Myips Blacklist Database
OSINT
Good Bots
Bot/User-agent keywords
MassDNS reverse DNS lookups
Myips Bots
Censys (AH)
IPinfo (AH)
Extract
Extract
Phase-1
Start
Load Telescope Source IP addresses
Remove internal traffic and bogon IPs (references)
Load IP addresses from data sources (probes)
Match reference IPs with probe IPs
Generate a result file for each datasource & save matching IPs
Load
Analyse
Warehouse
Load
Compile & enrich keyword list from collected actors
Load keyword list of known bots
Per data source, Identify actors based on keyword matching
Validate and refine results and store to separate files
Load data of each data source
Phase-2
Load data files with matches IPs from warehouse
Select candidate files for analysis
Clean data (fix name mismatches, duplicates etc.)
Analyze data (AS, geolocation etc.)
End

Figure 6.1: ETL (Extract, Transform, Load) process for data integration on known scanners

background noise. Identified IPs are categorized as benign, malicious or unknown. The platform also lists know actors per each category. We obtain a list of benign actors used for OSINT and lists of sample IPs for actors whose IPs cannot be identified through OSINT.

E3. MyIP.ms: MyIP.ms [56] is a website that provides tools and statistics on IP addresses, domains, and hosting providers. It keeps a publicly accessible real-time blocklist of IP addresses connected with malicious activity, which users may utilize to strengthen their cybersecurity defenses. We obtain the Blacklist IP Addresses Live Database of 138,628 IPs from 4 Jan 24 and validated against our Darknet data. Additionally, the site maintains a live list of well-known web bots and web spiders used by websites and search engines (e.g. Googlebot, Bingbot, GPTbot etc.). The list matches the User-Agent request header of each bot with respective IP addresses. We obtain the list from 13 Mar 24.

E4. GoodBots: GoodBots [62] is a public GitHub project containing IP addresses whitelists of known bots and crawlers (e.g., Bingbot, Googlebot etc.). The lists are automatically updated on a daily basis by scraping and aggregating the IP address lists that bot owner companies publish on their websites. We obtain the lists from 1 to 30 June 2023 (excluding the outage period) and 1 to 29 February 2024 and match against our Darknet data for each day.

E5. Censys API: Censys.io [31] continuously scans the IPv4 address space and maintains a Universal Internet Dataset of all publicly accessible hosts with metadata regarding the employed services, software etc. We obtain data of February 2024 referring to the subset of aggressive IP hitters identified in our Darknet. The limitation of this dataset is that data of Feb 24 are also used to profile aggressive hitters of Jun 23.

E6. IPinfo API: We use the IPinfo API [43] to cross-reference the actors extracted from Censys and validate the identified bots from MyIP.ms.

E7. IPinfo AS: Some known scanners maintain their own AS. This allows us to obtain the list of network prefixes for each identified organization from IPinfo [43].

E8. **Reverse DNS**: For each source address reaching our Darknet, we employ reverse DNS (rDNS) to resolve an IP address back to a domain name returning a DNS Pointer (PTR) record that yields the domain name linked to the IP address. We performed reverse DNS using MassDNS [16], a high-performance DNS stub resolver, on 19 - 21 January 2024 and collected 130,358 PTR records. Then we extract the top level domain and matched against a list of 44 keywords from known scanners. The list is a refined and updated version of the initial keyword list created by Anand and et al. [3].

E9. **OSINT**: We perform exhaustive open source intelligence (OSINT) online search to identify ASes, IP addresses and network prefixes for those actors labelled as benign by GreyNoise. The research involves lookups on scanner websites and forward DNS lookups. OSINT proves to be a useful method to identify known scanners.

E10. **Bot/User-Agent Keywords**: To facilitate the identification of known bots, we use a list of user-agents used by robots, crawlers, and spiders from a public GitHub repository [50]. The list contains regular expressions and metadata (URL, instances) of known bots with open contributions. We obtain the JSON list of 29 Mar 24 that contains data for 568 known bots.

6.2.2. Extract, Transform, Load (ETL)

The first step in ETL refers to the extraction of data from each data source. Staging area includes the transformation step and comprises two phases; IP-based matching (Phase-1) and IP-keyword-based matching (Phase-2). Collected data contain in principle lists of IP addresses with the respective entity (known scanner) they belong to. However, some of the data we collect are either indirect or metadata linked to IPs scanning our Darknet; thus we need to scrap them and extract meaningful information. For example, Censys provides network information and service headers but not explicitly relate to the actor/owner behind each IP address. Hence, we perform keyword matching in order to match IP addresses with corresponding entities/actors. Our keyword list is composed of i) known scanner keywords extracted from actors during IP-based matching in Phase-1, enriched with manual additions and ii) a list of known bots (source E10).

During IP-based matching (Phase-1), we match the source IP addresses appearing in the Darknet with those of the data sources. Phase-2 requires customized data processing for each data source. Having compiled the keyword list, we search for keywords in four datasets: Censys API (source E5) , IPinfo (E6), reverse DNS (E8) and MyIP.ms bots (E3). We extract the following fields from Censys data: WHOIS network handle, network name, organization name, WHOIS admin and abuse emails, response header location, forward and reverse DNS names and service banners. Fields are ordered from the most important to the least important one. Next, we extract domain names from IPinfo and massDNS and match with the keyword list. Lastly, the Myips.ms live web crawlers list contains IP addresses from bots/crawlers and their User-agent header. We match the IP address and then we match the User-agent header with our keyword list. To reduce potential false positives, we perform cross-validation with IPinfo records matching the identified keyword (actor) against the hostname or AS name of the examined IP address. Then, we manually inspect and fine-tune the results, given that some actors have different names than their bots e.g. Microsoft owns Bingbot. We keep only validated results. Having transformed the datasets and matched the IP addresses, we load the result files into the warehouse. Next, we launch an analytics phase where all transformed data (except MyIP.ms blacklist database and the GreyNoise malicious scanners which are left for future study) are selected for analysis. Results are presented in the remainder of this chapter.

6.3. Who are the Known Scanners?

The aggregated subset of matched IPs in June 2023 contains 5,689 distinct addresses from 36 organizations that contribute over 5.15 billion packets to the total telescope traffic. In other words, 36 organizations correspond to 0.36% of the total source IP addresses and account for 51.31% of the total telescope traffic in Jun 23. Accordingly for February 2024, there are 8,035 IP addresses which belong to 40 organizations and contribute approximately 5.08 billion packets to the total telescope traffic.

In other words, 40 organizations correspond to 0.62% of the total source IP addresses and account for 50.86% of the total telescope traffic in Feb 24.

We can divide the known scanners into three categories: i) security industry ii) non-profit organizations iii) academic/research institutions. Below, we list the known scanners appearing in both months:

- **Security Industry**: Adscore, Alpha Strike Labs, BinaryEdge, Bit Discovery (by Tenable), Bufferover.run (`https://newtls.bufferover.run/`), Censys, Cortex Xpanse (by Palo Alto Networks), Criminal IP, CyberResilience.io, Driftnet.io, ESET, Internet Census Group (by BitSight Technologies), InterneTTL, Intrinsec, IPIP, leakIX, Onyphe, Rapid7 (Project Sonar), SecurityTrails LLC, Shodan, Stretchoid, Threatsinkhole
- **Non-profit Organizations**: The Shadowserver Foundation
- **Academic/Research Institutions**: Academy for Internet Research LLC, Arbor Observatory (NETSCOUT), Recyber.net, Ruhr-Universität Bochum, RWTH Aachen University, Technical University of Munich, Stanford University, University of Michigan (UMich), Internet Transparency research project by University of Twente

Additionally, we discover the following known scanners appearing only in Jun 23:

- **Academic/Research Institutions**: UC Berkeley, FH Münster, University of California San Diego (UCSD), University of Colorado Boulder (CU Boulder)

Lastly, we discover the following known scanners appearing only in Feb 24:

- **Security Industry**: DataGrid Surface, Hadrian.io, IPinfo, Leitwert.net
- **Academic/Research Institutions**: Georgia Tech (`cc.gatech.edu`), Inter-University Computation Center (IUCC), Project 25499, Research Scanner (`research-scanner.com`)

Scanning Objectives. Known scanners employ Internet scanning in order to achieve certain goals and deliver specific products and services. Firstly, Stretchoid focuses on identifying online services of organizations. A search engine for Internet-connected devices is provided by Shodan and Censys. Large-scale Internet measurements are carried out by Internet Census Group to assess security performance and trends across industries. LeakIX scans and indexes web services monitoring for leaks. Intrinsec offers vulnerability management and Cyber Threat Intelligence among its cybersecurity services. A framework for retrieving and examining DNS data is provided by bufferover.run. Palo Alto Networks provides an attack surface management solution via Cortex Xpanse. Adscore's goal is to classify website traffic that is originally generated or purchased by their client companies. CyberResilience.io provides insights into security flaws. Driftnet.io offers footprint discovery so their client companies can assess the level of their services' exposure to the Internet. Rapid7 is a cybersecurity company running Project Sonar to facilitate security research. SecurityTrails LLC offers a broad spectrum of services such DNS history, brand protection, threat hunting etc. Alpha Strike Labs performs global scans and collaborates with governmental agencies and national Computer Emergency Response Teams (CERTs). Bit Discovery is part of the Tenable Attack Surface Management service for cyber risk management. Criminal IP offers an OSINT-based search engine for Cyber Threat Intelligence, and an attack surface management tool. Leitwert.net, Hadrian.io and DataGrid Surface offer Threat Intelligence Data as a Service. Non-profit organizations like The Shadowserver Foundation and academic institutions such as UCSD, U Michigan etc. focus on Internet security and Internet measurement research aiming to improve security. U Michigan mentions that scanning assists computer scientists in researching the implementation and setup of network protocols and security solutions and enables scientists to quantify the worldwide Internet and examine trends in technology adoption and security [75].

Scanning Intention. Several known scanners adhere to the best practices: they provide reverse DNS PTR records and abuse contacts, announce IP ranges publicly and describe their purpose (scope of experiment). However, not all scanners offer comprehensive documentation about their purposes. More precisely, Recyber.net supports researchers and academic institutions [70], and Academy for Internet Research LLC is a team of security researchers who scan the web and check for vulnerabilities [1]. However, no affiliated research institutions or references to published studies are provided by these two known scanners. In fact, SURF [15] and Microsoft [58] have received abuse complaints for Recyber.net due to excessive scanning. Stretchoid facilitates the identification of existing Internet services. Although it provides an opt-out page, nonetheless, there is no further information regarding the scanner's activities [68]. Furthermore, GreyNoise marks as malicious over 1,700 IP addresses by Stretchoid (as of 01 Mar 2024) [67]. Threatsinkhole searches for computers that have been infected with Winnti malware and provides an opt-out form [71]. However no further information regarding its owner, funding, software etc. is provided. InterneTTL continually scans all hosts on the Internet, giving IT and security teams with real-time insight into active servers [42]; yet there is no claimed ownership, the website has been deactivated and GreyNoise marks all (17) InterneTTL IPs as malicious (as of 11 Apr 24). Lastly, `research-scanner.com` searches for SSH host keys and JARM hashes in publicly-facing servers, nonetheless, there is no information about the intention of this scanner or the management of collected data. GreyNoise lists all the IPs of this scanner as benign (as of 11 Apr 24). Consequently, a sufficient documentation of scanning intentions and a clear data management plan can help security teams and organizations identify benign scanners and avoid false positive alerts.

6.4. Characterisation of Known Scanners

6.4.1. Scanner Composition and Infrastructure

Not all known scanners have the same size, generate similar traffic volume or can be classified as persistent. Figure 6.2 presents the distribution of known scanners per number of hosted IP addresses and traffic volume for both months. Inspecting traffic from June 2023, Cortex Xpanse by Palo Alto Networks contributes over 2.03 billion packets, or 20.23% of the total traffic. The rest top seven known scanners per traffic volume include Censys (9.49%), Criminal IP (7.12%), Recyber.net (6.68%), Academy for Internet Research LLC (2.46%), The Shadowserver Foundation (2.03%) and Driftnet.io (1.26%). Rest organizations contribute less than 1%. Essentially seven organizations - mapping to 0.17% of the total source IP addresses - are responsible for 49.27% of the total traffic in June 2023.

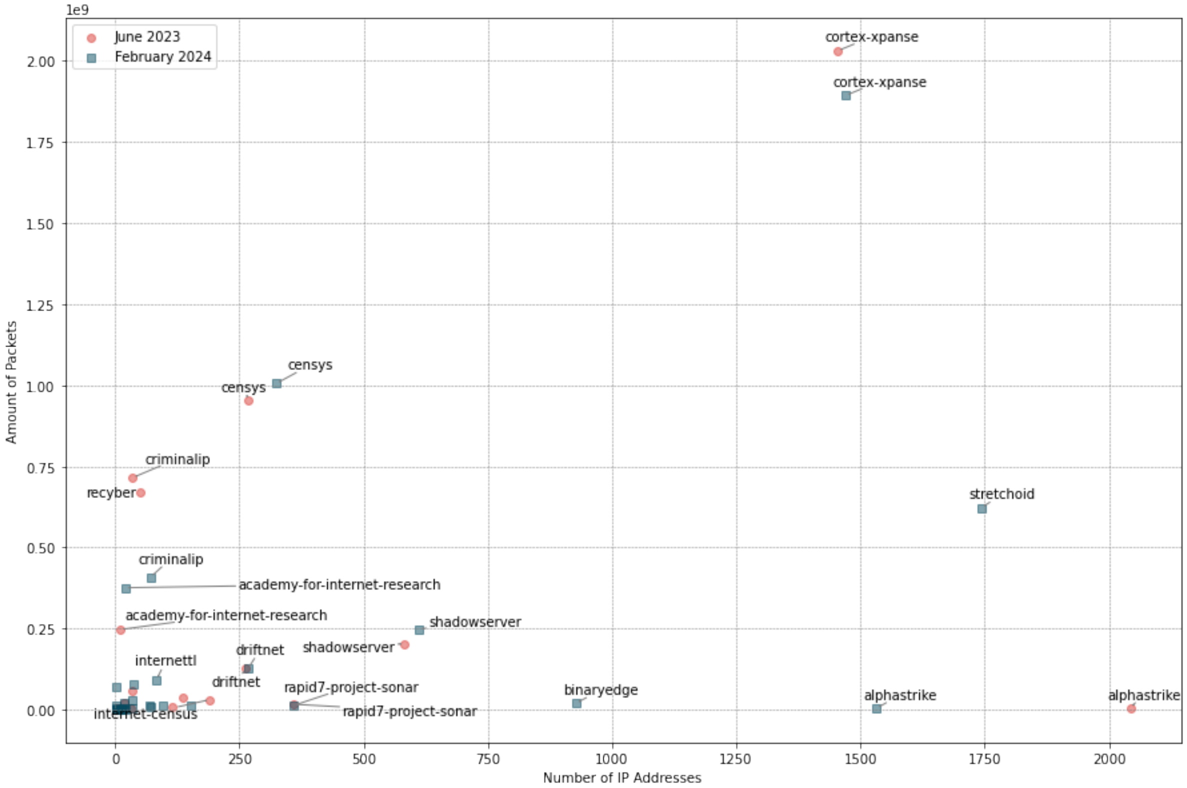


Figure 6.2: Distribution of known scanner organizations per number of scanning IP addresses and traffic volume

Similarly for February 2024, Cortex Xpanse generates 1.89 billion packets, or 18.94% of the total traffic. The rest top seven known scanners per traffic volume include Censys (10.06%), Stretchoid (6.22%), Criminal IP (4.08%), Academy for Internet Research LLC (3.76%), The Shadowserver Foundation (2.48%), Driftnet.io (1.26%), interneTTL (0.91%) Shodan (0.79%) and Recyber.net (0.70%). Rest organizations contribute less than 0.3%. Essentially 10 organizations - mapping to 0.36% of the total source IP addresses - account for 49.20% of the total traffic in February 2024.

Analysing the number of IPs per known scanner, we observe that Cortex Xpanse, The Shadowserver Foundation, Rapid7 and Censys appear in top positions for both months hosting from 268 to 2043 IP addresses each (322 to 1532 IPs). Notably, the amount of detected IPs of BinaryEdge and Stretchoid increases considerably from Jun 23 to Feb 24, demonstrating a percentage increase of 5,041.67% (18 to 927 IPs) and 1,191.11% (135 to 1743 IPs).

6.4.2. Scanning Frequency

Inspection of the daily traffic can offer insights regarding the recurrent activity of known scanners. Not all known scanners are persistent. Figures 6.3 - 6.6 demonstrate a daily timeline per known scanner for each one-month period. Organizations and scanning actors may scan from multiple IP addresses to avoid detection and blacklisting or achieve better load balance. Therefore we consider as a scan whatever scanning probe is reaching our telescope from any IP address identified as part of a known scanner organization. In June, 17 organizations scan on a daily basis. Research universities maintain a low scanning frequency of one to 12 days per month, except for the Arbor Observatory which generates scanning traffic almost every day. In February, 22 out of 40 organizations perform continuous daily scans. Again, research institutions maintain low frequencies up to eight days, except for the Arbor Observatory and TU Munich which scan at least 20 days.

6.4.3. Scanning Recurrence

All timestamps discussed below refer to the time our telescope observes the first and last packet from a known scanner. The possible event start times are described by a geometric distribution and correspond to less than or equal of single packet detection times. Based on the methodology presented in section 4.2.3 and 99% confidence, events of June begin no earlier than 43 minutes, and events of February begin no earlier than 16 minutes.

Looking into scanning patterns, Intrinsec, Onyphe, Shadowserver, Censys, Cyberresilience, Driftnet.io, IPIP, SecurityTrails and Ruhr-Universität Bochum (only in Feb) send packet volumes of similar size during each one-month period, exhibiting a stable volume recurring behaviour. More interestingly, U Michigan, TU Munich, UCSD appear to scan with a specific frequency. UCSD deploys two large-volume scans two days apart in CEST time. In particular, approximately 60 thousand packets are sent on 12 and 14 June 2023, continuing until morning hours of 15 June (00:30 CEST). The scans are repeated after every seven calendar days from the beginning of the previous scan: on 19 and 21 June 2023 (until morning hours of 22nd) and again on 26 and 28 June 2023 (until morning hours of 29th). Establishing the pattern, the first scan takes place on evening hours, from 18:00 to 20:00 CEST and the second scan commences two days later on 02:00 and lasts until 00:30 CEST of the next day. The scans appear to be automated. Converting timestamps to local time in San Diego, CA, USA, we need to consider the nine-hour time difference between the Central European Summer Time (CEST) zone (Amsterdam) and the Pacific Daylight Time (PDT) zone (San Diego) at that time. Thus, the first scan corresponds to 09:00 - 11:00 PDT and the second scan lasts from 17:00 PDT of the next day until around 15:00 PDT of they day after. Therefore, scans are one day apart, but appear in our telescope to be two days apart due to the timezone change. Furthermore, assuming the first scan in June 2023 ends June 1st CEST (it started on May 29th), then we can estimate the next scans to take place on 5 and 7 June 2023, CEST time. These days fall within the outage period and no traffic is collected from our telescope. However, the seven-day pattern holds since then next scan is detected on 12 June CEST, seven calendar days after the beginning of the presumed scan.

Similar behavior is observed by TU Munich. This academic scanner appears to send 50 to 70 thousand packets on 13-15 June (from 07:00 of 13th until 03:00 morning hours of 15th, CEST). Seven days after the initiation of the previous scan - at June 20th - a new scan commences, lasting from 07:00 of

20th until 15:00 CEST morning hours of 22nd and followed again by a seven-day gap until 27 June. The scans appear to be automated. We estimate that additional scanning took place at 6-7 June during the outage period. There is no timezone change. In February, although we detect one low-volume additional scan from 9 Feb (09:54) to 17 Feb (05:54), nonetheless the seven-day pattern appears to hold as well. We observe scan bursts in similar hours at 6, 20, 27 Feb and a considerable greater traffic volume at the expected scanning dates (13-15 Feb) which overlap with the above out-of-order scan.

Lastly, U Michigan performs scans every seven calendar days from the start of the previous scan. In particular, scans are observed at 2-3, 9-10, 16-17, 23-24 and 30 June, CEST. During each campaign, traffic starts reaching our telescope at around 7:10 CEST with a deviation on 19 June when traffic is first detected at 10:40 CEST. Converting the timestamps to local time in Ann Arbor, MI, USA, we need to consider the six-hour time difference between the Central European Summer Time (CEST) zone (Amsterdam) and the Eastern Daylight Time (EDT) zone (Ann Arbor) at that time. Trying to establish a pattern, U Michigan scans span from 01:00 EDT morning hours until 03:00 to 05:00 EDT morning hours of the next day. The precise pattern also appears in February data. The rest known scanners exhibit randomized behavior.

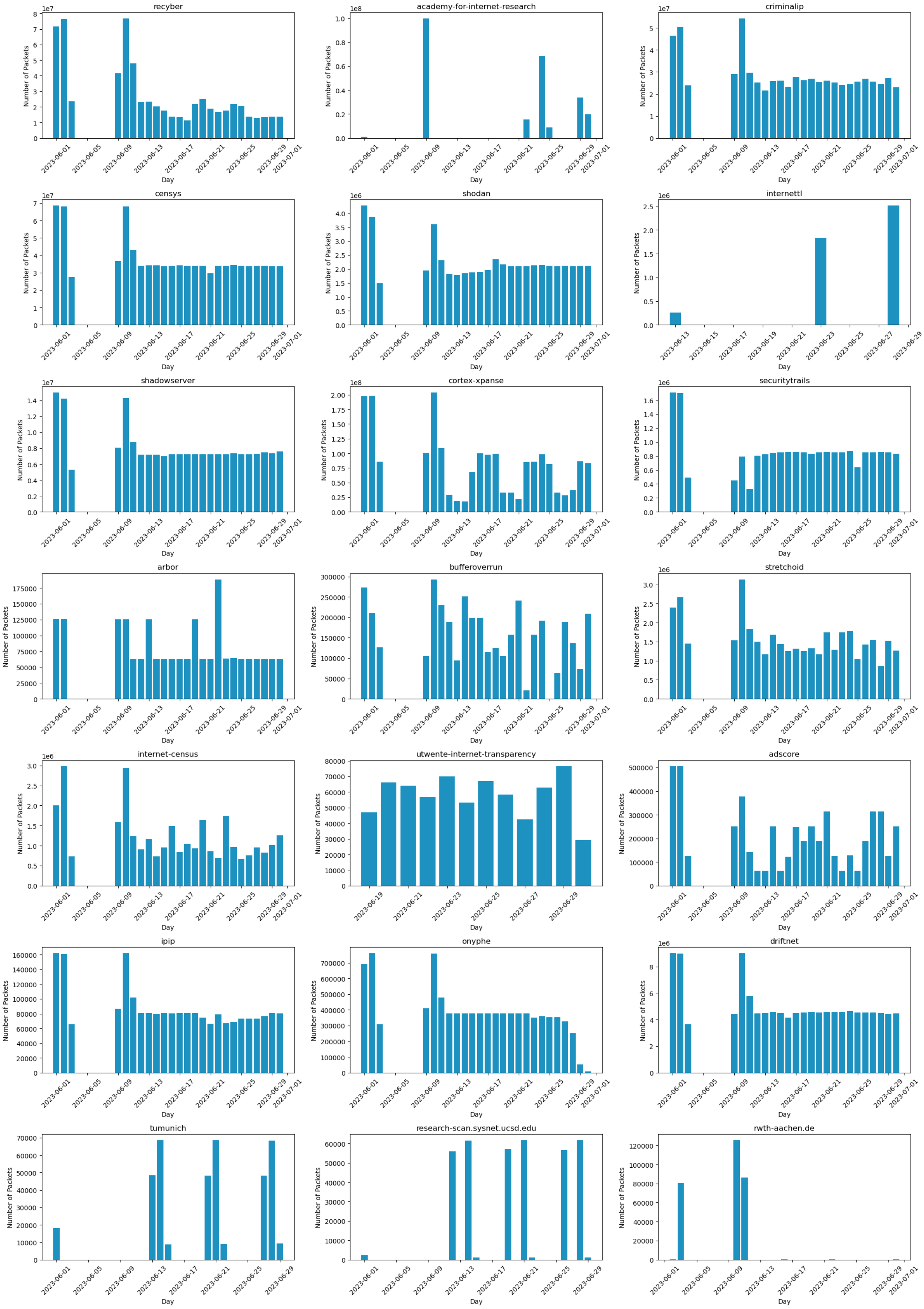


Figure 6.3: Daily Traffic composition per known scanner in Jun 23 (part 1)

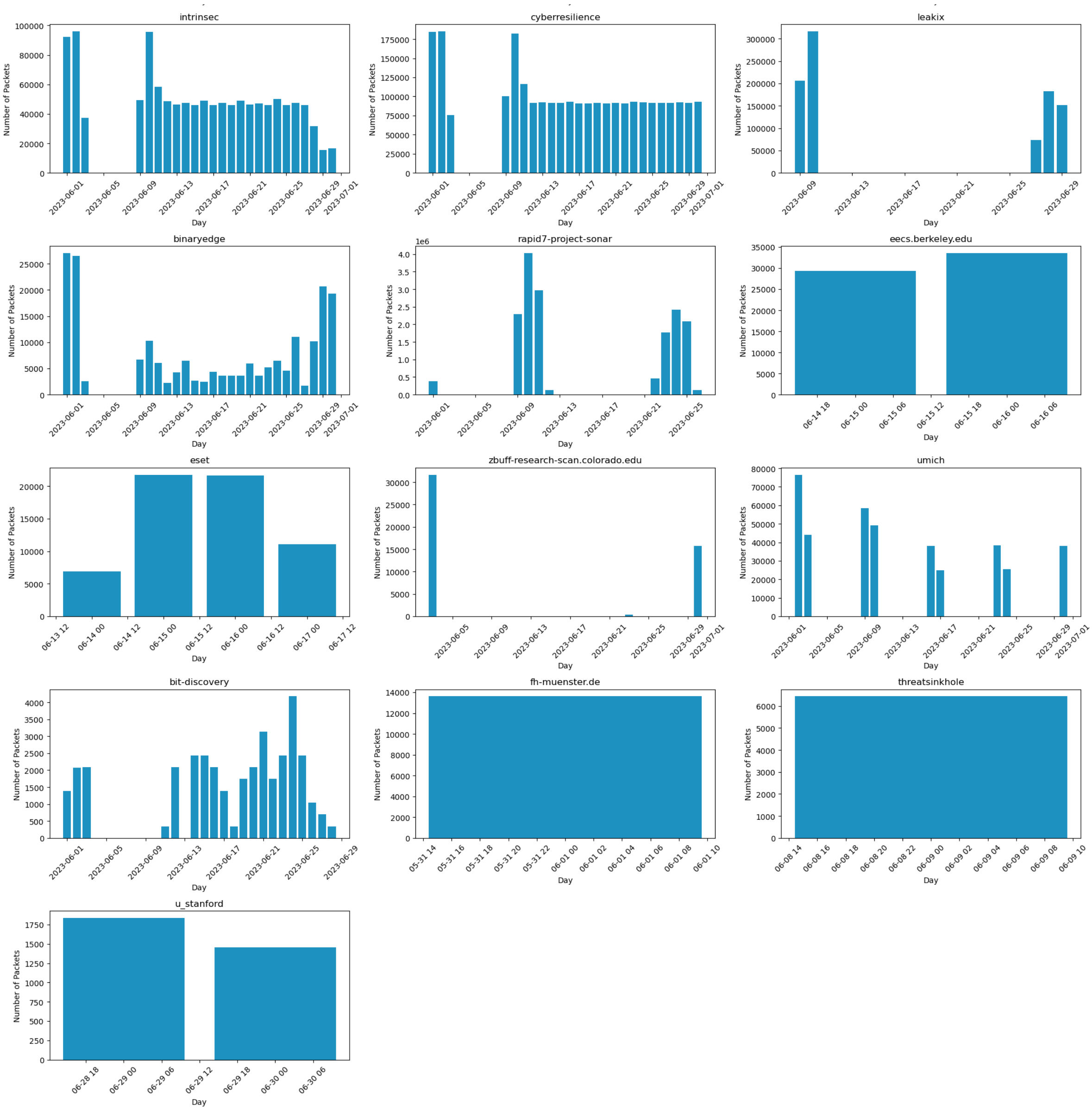


Figure 6.4: Daily Traffic composition per known scanner in Jun 23 (part 2)

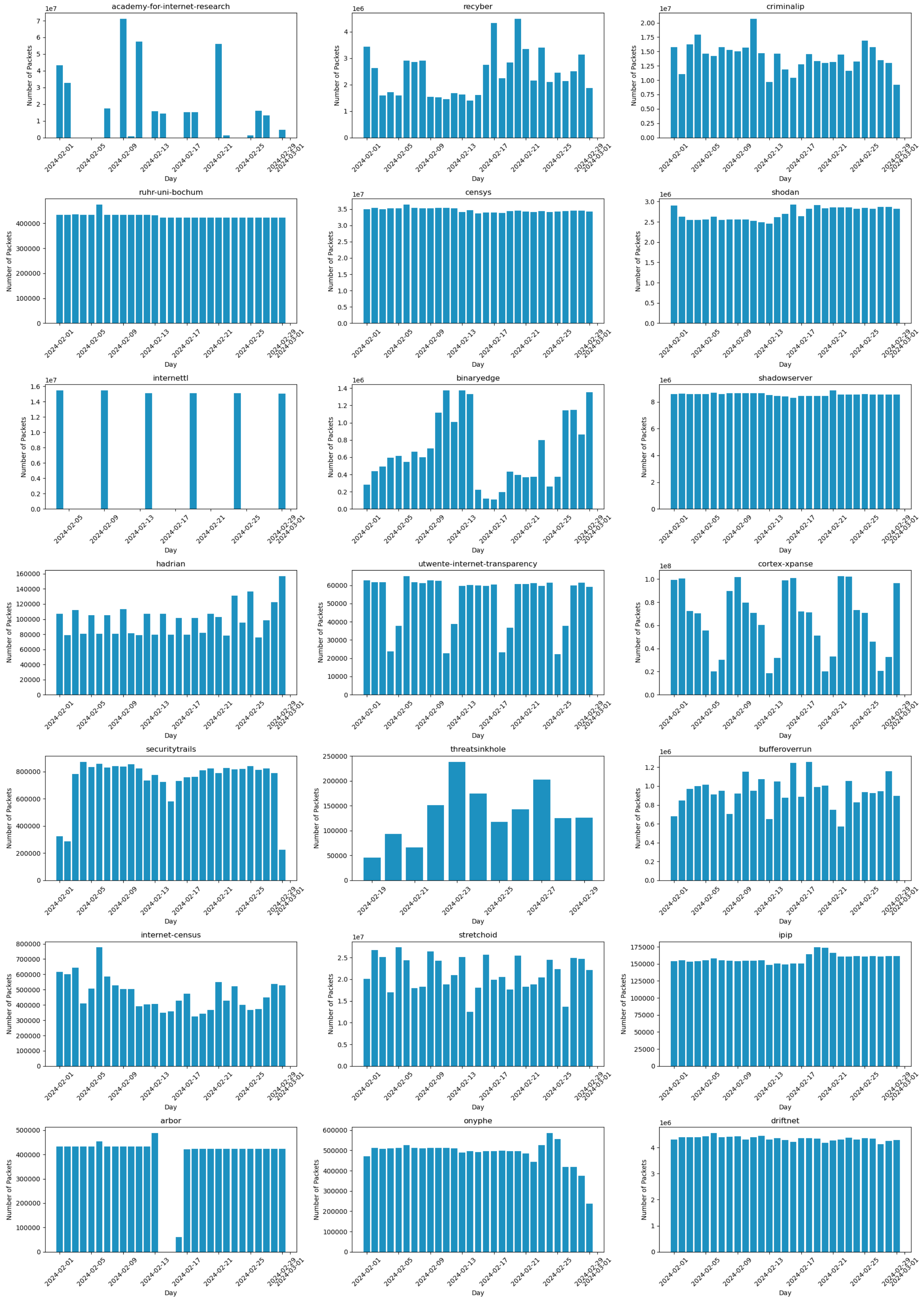


Figure 6.5: Daily Traffic composition per known scanner in Feb 24 (part 1)

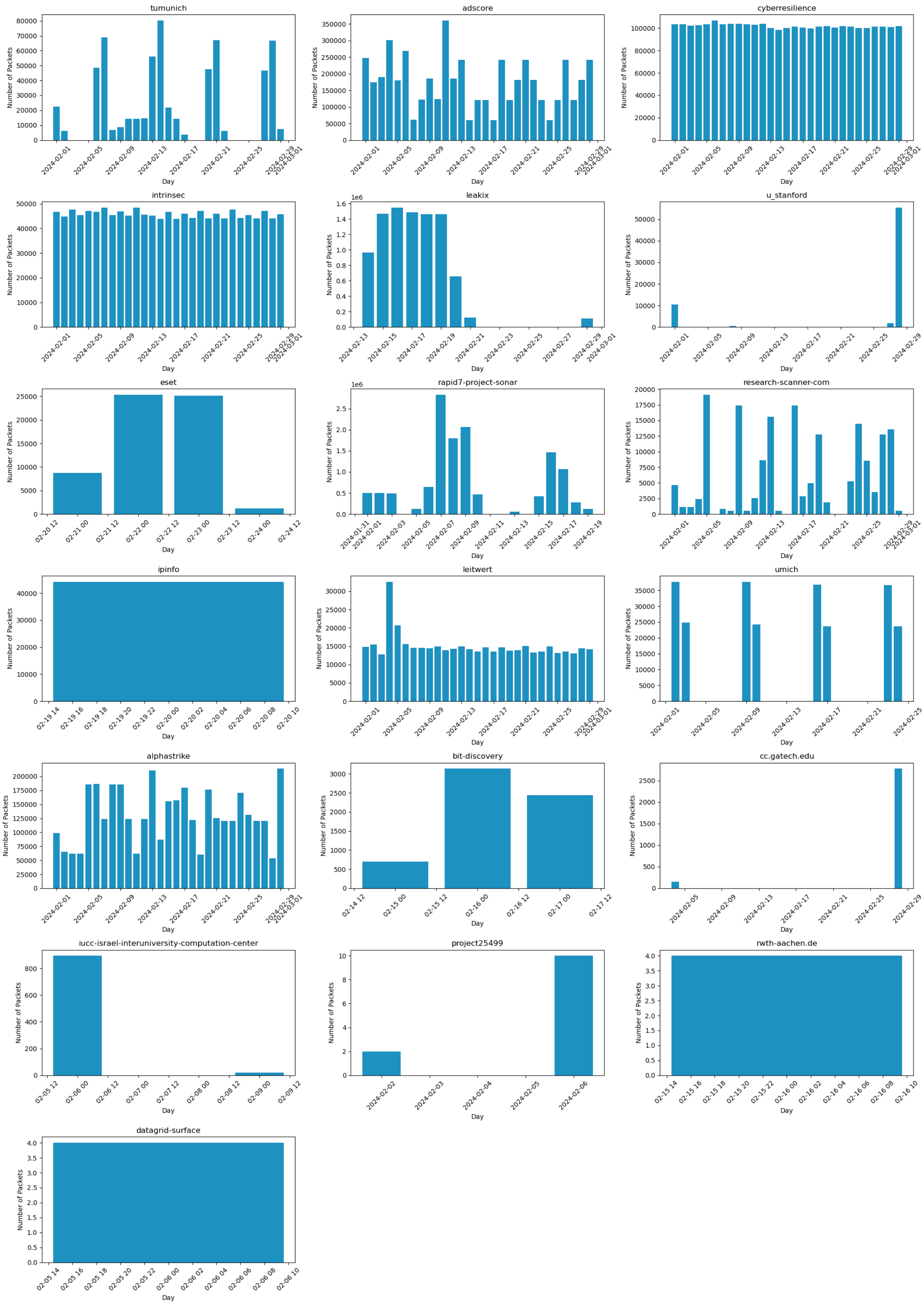


Figure 6.6: Daily Traffic composition per known scanner in Feb 24 (part 2)

6.4.4. Destination Port Scan Strategy

Internet scanners utilize a variety of port scanning strategies, each characterized by distinct patterns. Some scanners concentrate on a small number of ports, whilst others methodically probe a broad range of ports. These various strategies provide insights on the goals and intents of each scanner and highlight the specific targets and information they seek for. For each destination port we provide a list of known (or probable) services run on the port, using a public database [65].

June 2023. Figure 6.7 presents the distribution of known scanners per scanned destination port in June 2023. Specifically, the Arbor Observatory runs scans on TCP ports 443 (HTTPS) and 3128 (Squid HTTP Proxy). Adscore scans ports 1080 (SOCKS proxy protocol), 1723 (Point-to-Point Tunneling Protocol VPN), 3128 (Squid HTTP Proxy) and 3389 (RDP - Remote Desktop Protocol). IPIP focuses the scanning activities on ports 21 (FTP), 53 (DNS), 80 (HTTP), 110 (POP3) and 808 & 8000 (HTTP alternatives commonly used for web proxies). Furthermore, Intrinsec scans ports 80, 443, 873 (rsync Protocol), 9200 (Elasticsearch REST API) and 27017 (MongoDB Database Server). Cyberresilience on ports 21, 22, 53, 80, 443, 2000 (Cisco SCCP - Skinny Client Control Protocol), 5060 (SIP - Session Initiation Protocol), 7547 (TR-069 Protocol used by ISPs for remote management of customer-premises equipment), 8080 & 8085 (HTTP alternative commonly employed as a proxy and caching port) and 8089 (used for communication between Splunk components).

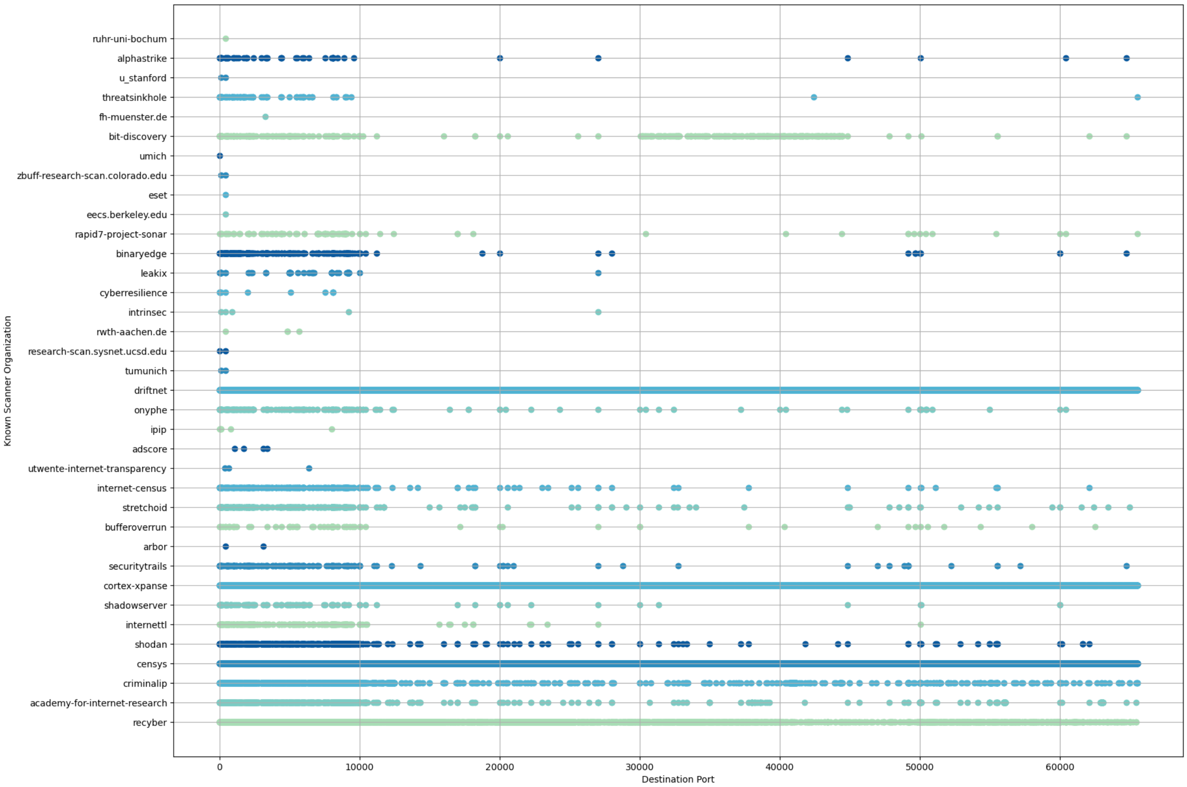


Figure 6.7: Distribution of known scanners per scanned destination port (Jun 23)

Academic entities demonstrate different patterns than the industry. UCSD performs scans on TCP ports 22 and 443 and RWTH Aachen on ports 443, 4840/4843 (OPC UA TCP Protocol) and 5671 (AMQP - Advanced Message Queuing Protocol). UTwente focuses on ports 389 (LDAP - Lightweight Directory Access Protocol), 636 (LDAP over SSL) and 6360 (MetaEdit+ Multi-User). Likewise, FH Münster scans port 3268, commonly used by Microsoft Global Catalog LDAP server. UC Berkeley, Ruhr-Universität Bochum and ESET probe only port 443, whereas CU Boulder, TU Munich and Stanford conduct scanning on ports 80 and 443. On the other hand, U Michigan pings only port 7, used for diagnostic purposes by Echo Protocol.

Lastly, several organizations perform wide-range scanning in a one-month period. In particular, Censys and Driftnet.io are the two companies scrutinizing the greatest number of ports with 65,522 and 65,521 ports respectively. Recyber.net scans 20,017 ports, Cortex-Xpanse 8,794 ports, Criminal IP 3,328 ports, Academy for Internet Research LLC 1,471 ports and Shodan 1,214 ports. Rest organizations scan up to around 360 ports.

February 2024. Figure 6.8 presents the distribution of known scanners per scanned destination port in February 2024. Specifically, the Arbor Observatory runs scans on TCP ports 80, 443, 2000 (Cisco SCCP - Skinny Client Control Protocol), 5678 (used by Linksys and Cable/DSL for router remote administration; used by MikroTik Neighbor Discovery protocol) 8291 (MikroTik RouterOS API) and 9001 (Tor ORPort; Cisco-XRemote router configuration; Citrix video redirection service). IPIP focuses its scanning activities on ports 21, 22, 80, 143 (IMAP), 443, 3306 (MySQL database server), 3389 (Remote Desktop Protocol), 5900 (VNC - Virtual Network Computing), 7777 (applications such as Oracle 9i Portal, iChat server etc.), 8000 (HTTP alternative), 8001 (squid HTTP Proxy server scan), 9000 (applications such as AltaVista HTTP Server, Buffalo LinkSystem Web access, DBGp, SqueezeCenter web server, Cisco WebEx, ManageEngine AssetExplorer etc.) and 9999 (Abyss web server remote web management interface). Project 25499 performs scans on common HTTP(S) ports 80 and 443.

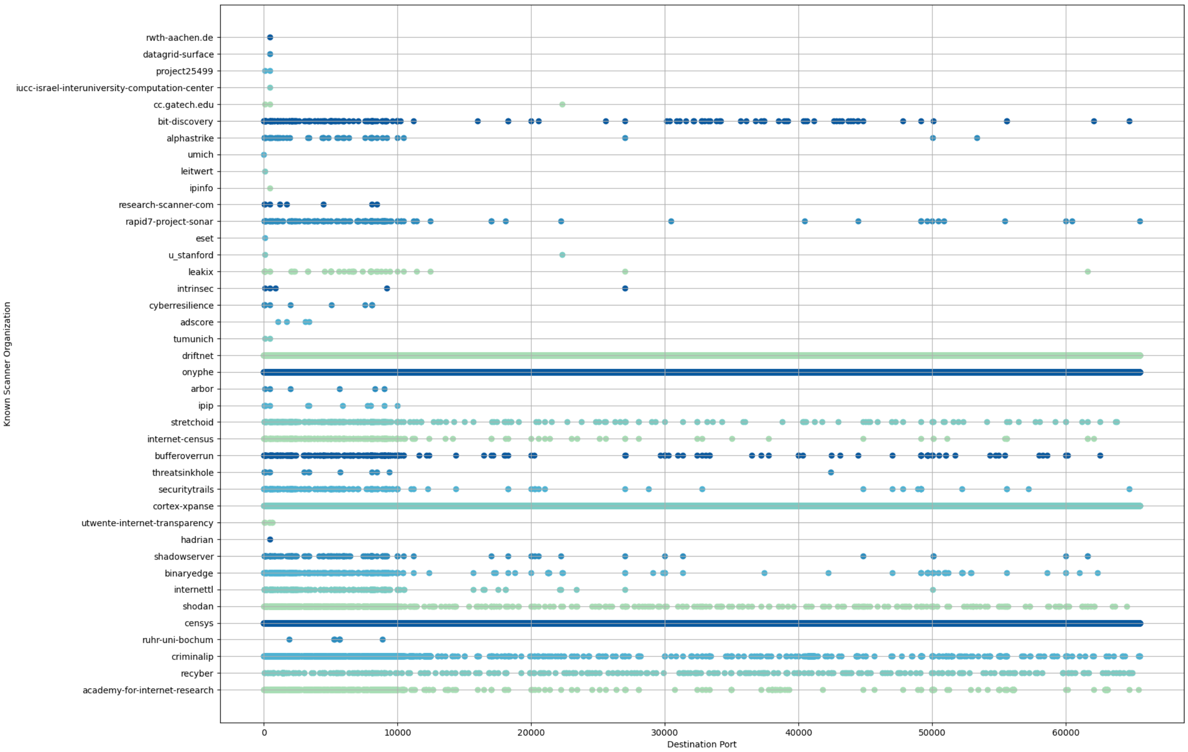


Figure 6.8: Distribution of known scanners per scanned destination port (Feb 24)

UTwente expands the range targeting ports 53, 80, 443, 587 (e-mail message submission SMTP over SSL/TLS), 389 (LDAP) and 636 (LDAP over SSL). Likewise, Ruhr-Universität Bochum probes ports 1883 (MQTT - Message Queuing Telemetry Transport), 5222 (XMPP - Extensible Messaging and Presence Protocol), 5269 (XMPP server-to-server communication), 5672 (AMQP - Advanced Message Queuing Protocol), 5683 (CoAP - Constrained Application Protocol - for IoT), 5684 (CoAP - Constrained Application Protocol - over DTLS) and 8883 (MQTT over TLS). Stanford conducts scanning on ports 80 and 22323, while Georgia Tech scans additionally port 443. Notably, port 22323 is scanned only in February and only by the aforementioned universities and BinaryEdge. ESET and Leitwert.net probe port 80, while Israel InterUniversity Computation Center, IPinfo, RWTH Aachen, DataGrid Surface and Hadrian.io scan only port 443. research-scanner.com focuses on ports 22, 80, 443, 1194 (OpenVPN), 1723 (Point-to-Point Tunneling Protocol VPN), 4443/8443 (HTTPS alternative) and 8080 (HTTP alternative). Lastly, rest known scanners remain on same ports as in Jun 23, along with wide-range hitters who scan the same number of ports in both months.

Figures 6.9 and 6.10 present a timeline with daily average number of ports scanned per IP address. Focusing on the outliers, each Censys IP addresses targets daily an average of 10,920 ports and each driftnet IP addresses targets daily an average of 2,226 ports. Furthermore, we observe that IP addresses of eight scanners[1] target daily more than 100 distinct ports each, nine scanners[2] lie below 100 and above 2 and lastly 17 scanners probe only one or two ports daily. In February, the most extensive scanning in terms of amount of targeted ports is performed again by Censys: each Censys IP addresses targets daily an average of 22,468 ports. Driftnet.io follows with an average of 2,220 ports scanned per IP address. IP addresses of six[3] target daily more than 100 distinct ports each, 11 scanners[4] lie below 100 and above 2 and lastly 21 scanners probe only one or two ports daily.

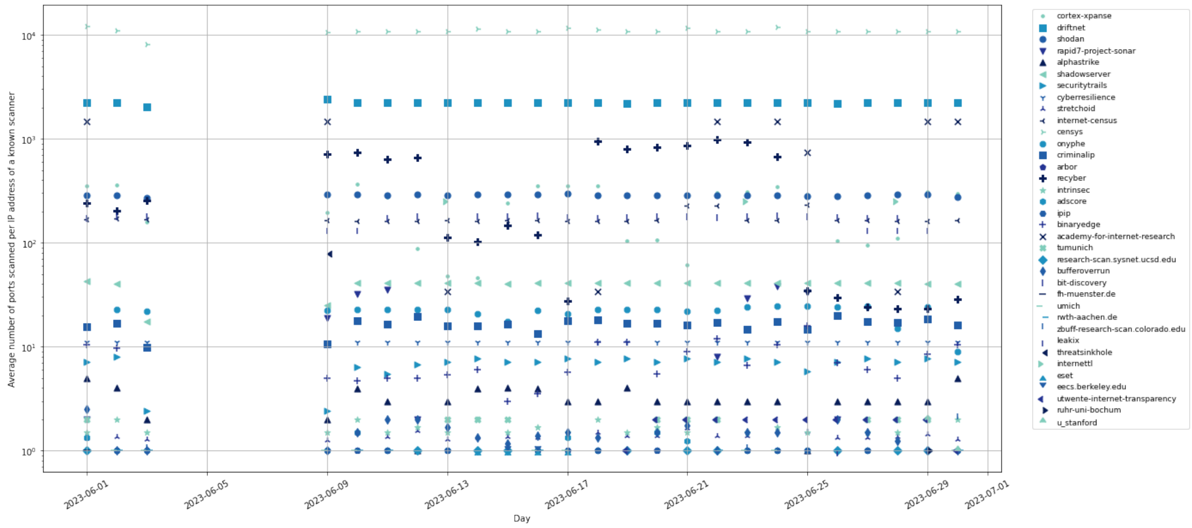


Figure 6.9: Timeline with daily average number of ports scanned per IP address of each known scanner (Jun 23)

6.4.5. Network Topological & Geographical Placement

Several known scanner organizations employ their own AS, whereas others use hosting or cloud infrastructure provided by third companies. An AS can contain multiple netblocks geolocated in various countries. Further, a scanner may distribute its scanning activities into multiple ASes. We study the network scanning architecture per each known scanner. The employed MaxMind GeoLite2 database offers acceptable precision, however some data centers or company headquarters may be less accurately geolocated.

15 organizations scan from their proprietary AS based in a single country. SecurityTrails (US), The Arbor Observatory (US), UCSD (US), Georgia Tech (US), RWTH Aachen (DE), UC Berkeley (US), CU Boulder (US), U Michigan (US), Academy for Internet Research LLC (US), Israel InterUniversity Com-

[1] Academy for Internet Research LLC (960 ports), Recyber.net (405 ports), Shodan (286 ports), interneTTL (250 ports), Cortex-Xpanse (225 ports), Internet Census (131 ports), Bit Discovery (175 ports), leakIX (116 ports)

[2] Threatsinkhole (79 ports), The Shadowserver Foundation (39 ports), Onyphe (21 ports), Rapid7 (20 ports), Criminalip IP (16 ports), Cyberresilience.io (11 ports), BinaryEdge (7 ports), SecurityTrails LLC (6 ports), AlphaStrike Labs (3 ports)

[3] Academy for Internet Research LLC (536 ports), Shodan (299 ports), interneTTL (250 ports), Cortex-Xpanse (231 ports), Internet Census (222 ports) and Bit Discovery (176 ports)

[4] leakIX (66 ports), The Shadowserver Foundation (45 ports), Onyphe (39 ports), Recyber.net (30 ports), BinaryEdge (18 ports), Rapid7 (14 ports), Threatsinkhole (14 ports), Cyberresilience.io (10 ports), Criminal IP (8 ports), SecurityTrails LLC (7 ports), Ruhr-Universität Bochum (7 ports)

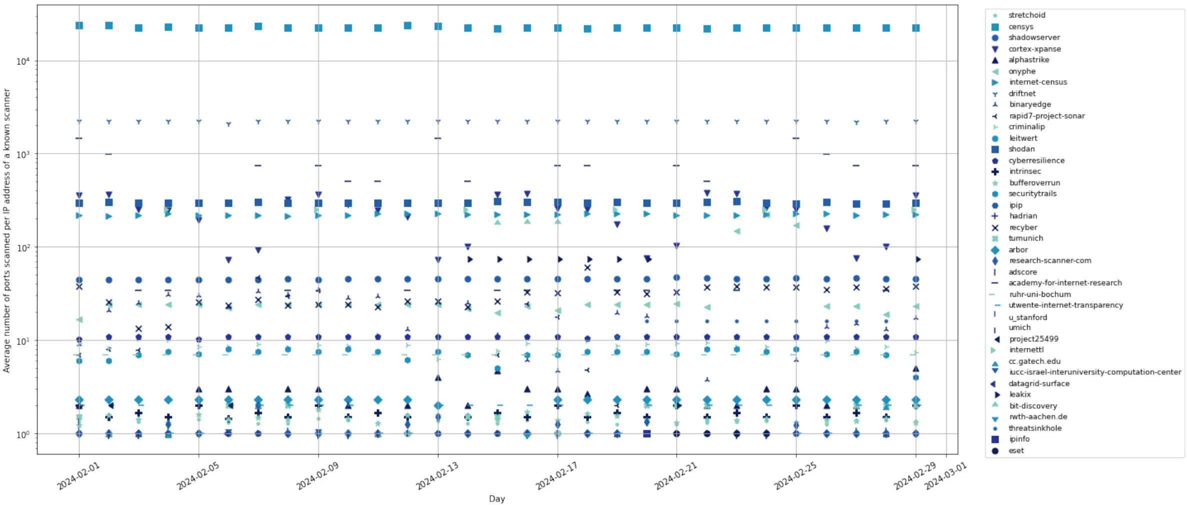


Figure 6.10: Timeline with daily average number of ports scanned per IP address of each known scanner (Feb 24)

putation Center (Israel), Stanford (US) and AlphaStrike Labs (Germany) use proprietary ASes to scan the Internet. Additionally, some universities use infrastructure provided by their partnered national research and education networks i.e. FH Münster and Ruhr-Universität Bochum, in Germany, traffic their scanning trough the German Research Network DFN-Verein, UTwente through the Dutch SURF B.V. TU Munich scans via the German Leibniz Supercomputing Center (Leibniz-Rechenzentrum) in both months and additionally via Technische Universitaet Muenchen (AS209335) in Feb 24. Lastly, Censys scans from proprietary ASes based on the US and Germany in June. Hence, the majority of known scanners employing proprietary ASes are research institutes and universities. Utilising such infrastructure renders scanners easily detectable to their targets, allowing administrators and security teams to whitelist these ASes and avoid false positives. On the other hand, it can potentially downgrade the quality of scanning results, if scanning targets block the IP ranges of known scanners.

On the other hand, several organizations lease third-party cloud platforms to conduct scanning originating in a single country. Criminal IP and Recyber.net both use IP Volume inc (AS202425) from the Netherlands. InterneTTL traffics through RETHEMHOSTING (AS14987) in the US, Shadowserver via HURRICANE (AS6939) in the US, and Cortex-Xpanse through GOOGLE-CLOUD-PLATFORM (AS396982) located in three different countries (US, UK, Brasil). Hadrian.io uses Scaleway S.a.s. (AS) located in France. Furthermore, Threatsinkhole and DataGrid Surface route scans through Akamai Connected Cloud (AS63949) located in Germany, whereas Bufferover.run uses the same AS from the US. ESET, Project 25499 and Stretchoid use DIGITALOCEAN-ASN (AS14061) from Canada and the US, respectively. Cyberresilience.io also uses DigitalOCean with almost equal traffic volume from three different countries (Germany, US, the Netherlands). Intrinsec partners with OVH SAS (AS16276) geolocated in France and Bit Discovery via CARINET (AS10439) based in the US.

Lastly, a number of known scanners adopts a diverse approach, distributing scanning to multiple cloud infrastructure located in multiple countries. Shodan scans the Internet from six different ASes in June 2023, located in four different countries (US, the Netherlands, Austria, Romania). These are CARINET (AS10439), COGENT-174 (AS174), SINGLEHOP-LLC (AS32475), IP Volume (AS202425), M247 Europe SRL (AS9009) and DIGITALOCEAN-ASN. In February 2024, AMAZON-02 (AS95221) replaces M247 Europe SRL. Onyphe uses mainly OVH SAS (AS16276) with almost equal traffic volume from France and Canada. Secondarily, they partner with ZEN-ECN (AS21859) based in Hong Kong and DIGITALOCEAN-ASN, based in Singapore. Analysis from Feb 24 also reveals Onyphe traffic from Akamai Connected Cloud (US). LeakIX scans from two distinct ASes and four countries: DIGITALOCEAN-ASN from the US, Canada, Netherlands and the UK, and Akamai Connected Cloud originating in Canada. Similarly, BinaryEdge utilizes the same ASes and their IPs are geolocated at the US, UK, Austria and Canada (additionally Germany and the Netherlands in Feb 24). Internet Census uses two distinct ASes: Sistemas Informaticos S.A. (AS211680) based in Portugal and ZEN-ECN

(AS21859) based in the US and the Netherlands. Regarding IPIP, the majority of its scanning traffic passes through Beijing Tiantexin Tech. Co., Ltd. (AS136180) in China and a smaller amount through Akamai Connected Cloud from Japan. Rapid7 uses CARINET (US), UK-2 Limited (AS13213) at the US and Hosting Services Inc (AS29302), ordered by descending amount of traffic. Driftnet.io uses primarily the AS of a partnered UK-based company Constantine Cybersecurity Ltd. (AS211298) with a minor scanning activity from DIGITALOCEAN-ASN located in the Netherlands (and US in Feb 24). Lastly, Leitwert.net employs both a proprietary AS Leitwert GmbH (AS29108) located in Germany and leases cloud infrastructure in: Tehnologii Budushego LLC (AS37327) based in Ukraine, RELIABLESITE (AS33731) from the US and CITYNET (AS30032) located in Egypt.

6.4.6. Device/Scanner Fingerprinting

In order to determine the OS used by known scanners we rely on qualitative analysis of TCP/IP fields such as the source port and the TTL value. We stress out that these are not definitive metrics to fingerprint the OS, yet they provide useful insights to understanding the scanners' underlying infrastructure. Since many organizations scan from different countries, ASes and prefixes, it might be reasonable to expect variations in TTL values from the same organization. Source port numbers can be defined by multiple actors such as the underlying OS, the scanning software, or malware in case of a infected device scanning for vulnerable devices. The default TCP dynamic ports for Linux range from 32768 to 61000 [28], while Microsoft has expanded the dynamic client port range for outbound connections, from 1025 - 5000 [27] to 49152 - 65535, to meet the suggestions by the Internet Assigned Numbers Authority (IANA). 32 organizations[5] appear to use source ports only in the Linux range 32768 - 61000. Cross-reference with TTL results shows that 22 organizations[6] (nine in both months, 10 in Jun and three in Feb) appear to operate solely Unix/Linux systems.

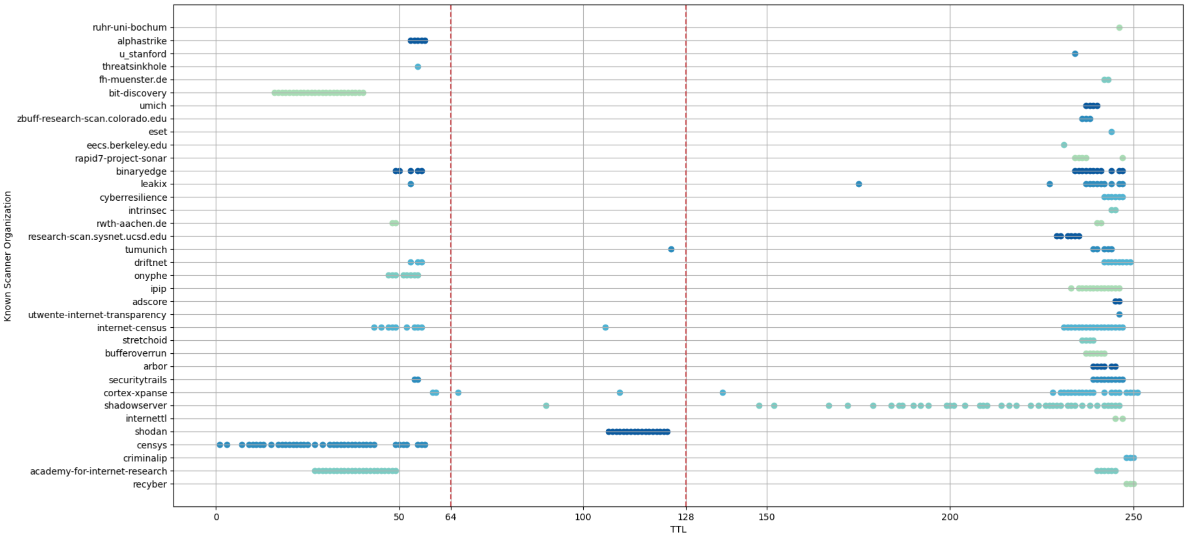


Figure 6.11: Distribution of known scanners per TTL value (Jun 23)

[5] 1) Appearing In both months: Adscore, Arbor, Bufferover.run, Criminal IP, ESET, InterneTTL, Intrinsec, IPIP, leakIX, Recyber.net, Ruhr-Universität Bochum, RWTH Aachen, SecurityTrails, Stretchoid, Threatsinkhole, TU Munich, U Stanford, U Michigan, UTwente 2) Appearing only in Jun: Academy for Internet Research LLC, Cortex-Xpanse, UC Berkeley, FH Münster, UCSD, Shadowserver, CU Boulder 3) Appearing only in Feb: Georgia Tech, DataGrid-Surface, IPinfo, Leitwert.net, Project 25499, research-scanner.com

[6] 1) Both months: Criminal IP, The Arbor Observatory, ESET, U Michigan, Stanford, Ruhr-Universität Bochum, UTwente, Adscore, Intrinsec 2) Jun 23: Recyber.net, IPIP, UCSD, UC Berkeley, CU Boulder, Bufferover.run, SecurityTrails, Stretchoid, RWTH Aachen, FH Münster 3) Feb 24: research-scanner.com, Georgia Tech, TU Munich

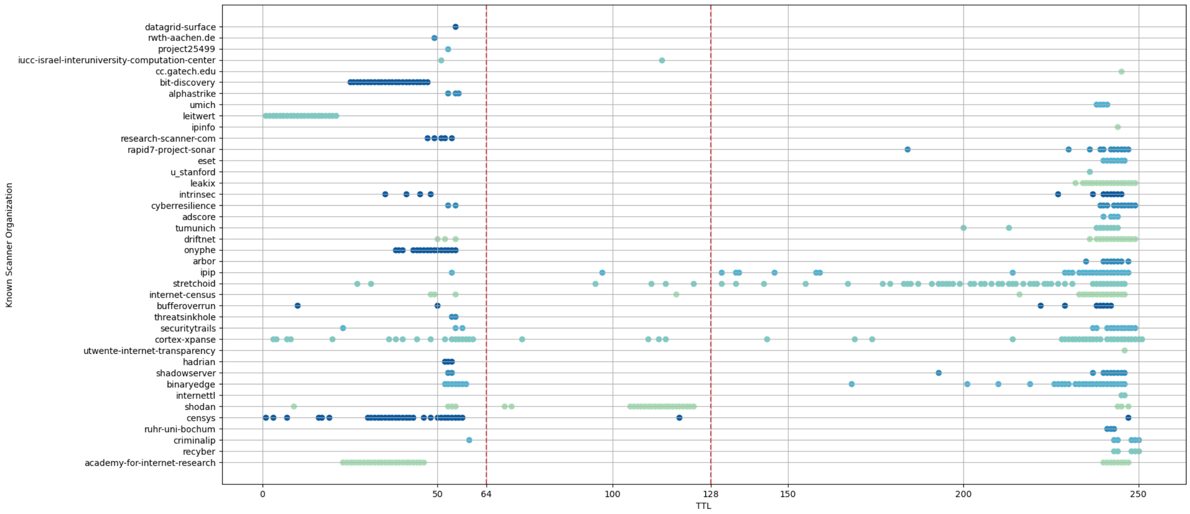


Figure 6.12: Distribution of known scanners per TTL value (Feb 24)

6.5. Who are the Known Bots?

The aggregated subset of matched IPs in June 2023 contains 1,846 distinct addresses from 19 bots that contribute around 49,000 packets, or below 0.01% of the total telescope traffic. Accordingly for February 2024, there are 3,311 IP addresses which belong to 27 bots and contribute 214,587 packets to the total telescope traffic (<0.01%). Therefore, known bots appear to contribute negligible traffic volume. Below, we list the known bots appearing in both months:

- **Search Engine Bots/Crawlers**: Baiduspider, PetalBot, Naver Yeti Spider, AhrefsBot, Applebot, Bingbot, Googlebot, MojeekBot
- **SEO and Web Analysis Bots**: DataForSEO Link Bot, SEOkicks web crawler
- **Social Media Bots**: FacebookBot
- **Specialized Bots**: TurnitinBot, BLEXBot Crawler

Additionally, we discover the following known bots appearing only in Jun 23:

- **Search Engine Bots/Crawlers**: AppEngine-Google, SemrushBot, Yandex Bot, netEstate Crawler
- **Web Archive Bots**: ArchiveBot
- **SEO and Web Analysis Bots**: DomainStatsBot

Lastly, we discover the following known scanners appearing only in Feb 24:

- **Search Engine Bots**: AdsBot-Google, SeznamBot, Sogou, Coc Coc Bot, DuckDuckBot
- **E-commerce Bots**: AmazonBot
- **Image Crawlers**: ImagesiftBot
- **Web Monitoring**: Better Stack Uptime (formerly Better Uptime), UptimeRobot
- **Chat and Messaging Bots**: Telegram Bot, Twitterbot
- **Other Bots**: ChatGPT-User, GPTBot, Bytespider

6.6. Characterisation of Known Bots

6.6.1. Destination Port Scan Strategy

Bots utilize a variety of port scanning strategies, each characterized by distinct patterns. Most bots concentrate their scanning activities to known HTTP(S) ports.

June 2023. Most known bots scan main or alternative HTTP(S) ports. In particular, ArchiveBot, BLEXBot, PetalBot, Baiduspider, Naver Yeti Spider, MojeekBot, DomainStatsBot and SEOkicks web crawler scan only port 80. Additionally, AppEngine-Google, Yandex Bot, FacebookBot, TurnitinBot, AhrefsBot, netEstate Crawler focus on both ports 80 and 443. Bingbot scans ports 80, 443, 8000, 8080 and 8500 (applications such as Adobe ColdFusion and Consul), whereas SemrushBot and DataForSEO Link Bot scan ports 80, 443 and 8000. Googlebot focuses on eight ports: 21, 80, 443, 2323, 4505 (SaltStack), 8000, 8080 and 37215 and Applebot on three ports: 80, 8000 and 42069.

February 2024. Telegram Bot, SEOkicks and DuckDuckBot scan only port 80. AdsBot-Google, Better Uptime, SeznamBot, Twitterbot, ChatGPT-User scan only port 443. Furthermore, 14 bots[7] conduct scans in both ports. Conversely, some bots adopt a wide-range approach. First, Amazonbot scans ports 80, 443, 4505, 8000 and 8080, while Googlebot scans ports 21, 80, 443, 4505, 8000, 8080. Bingbot scans ports 80, 443, 8080, 9443 (HTTPS over SSL/TLS), and DataForSEO Link Bot scans five ports - 80, 443, 8000, 8080 and 8500. Applebot focuses on ports 80, 443 and 3306 (MySQL database server).

6.6.2. Scanning Frequency

Inspection of the daily traffic offers insights regarding the recurrent activity of known bots. Not all known bots are persistent. Figures 6.13 and 6.14 show the amount of packets per day and known bot for each one-month period. Similar to known scanners, we group in the same scan all scanning probes reaching our Darknet from any IP address identified as part of a known bot. PetalBot, Bingbot and Googlebot perform scans every day both in June and February. Furthermore, SemrushBot, AhrefsBot and DomainStatsBot/Yandex Bot are comparably persistent scanning 24, 23 and 22 days respectively in June 2023. Accordingly for February, Amazonbot sends scanning traffic 28 days followed by AhrefsBot with 21 days. Rest bots fluctuate from one to 15 days in any month.

6.6.3. Network Topological & Geographical Placement

Several companies use part of their own AS infrastructure to deploy their bots[8], whereas others scan through cloud/hosting services[9]. In June, an equal number of companies (nine each) use a proprietary AS and rent infrastructure in third parties. Additionally, a single company[10] adopts both approaches to perform scanning. Hosting services by Hetzner Online GmbH (AS24940) are particularly preferred when outsourcing scanning activities. In February, 16 bots scan from within their organization AS and 11 bots use third-party clouds. Consequently, we can make two remarks. First, it is usually trivial for

[7] ImagesiftBot, UptimeRobot, GPTBot, Bytespider, PetalBot, MojeekBot, Sogou, BLEXBot, Turnitin Bot, FacebookBot, Coc Coc Bot, Baiduspyder, AhrefsBot and Naver Yeti Spider

[8] Own Infrastructure - Appearing in both months: Bingbot (MICROSOFT-CORP-MSN-AS-BLOCK AS8075), Googlebot (GOOGLE AS15169, GOOGLE-CLOUD-PLATFORM AS396982), AppEngine-Google (GOOGLE-CLOUD-PLATFORM AS396982), TurnitinBot (TURNITIN AS46851), Applebot (APPLE-ENGINEERING AS714), Naver Yeti Spider (NAVER Cloud Corp. AS23576), Facebookbot (FACEBOOK AS32934)

Own Infrastructure - Appearing only in Jun: ArchiveBot (INTERNET-ARCHIVE AS7941), Yandex Bot

Own Infrastructure - Appearing only in Feb: Amazonbot (AMAZON-AES AS14618), ImagesiftBot (CSTL AS36321), ChatGPT-User & GPTBot (MICROSOFT-CORP-MSN-AS-BLOCK AS8075), AdsBot-Google (GOOGLE AS15169), DuckDuckBot (MICROSOFT-CORP-MSN-AS-BLOCK AS8075), SeznamBot (Seznam.cz, a.s. AS43037), TwitterBot (TWITTER AS13414), Telegram Bot (Telegram Messenger Inc AS62041)

[9] Cloud hosting - Appearing in both months: BLEXBot Crawler, DataForSEO Link Bot and SEOkicks web crawler use Hetzner Online GmbH (AS24940), PetalBot (HUAWEI CLOUDS ASv), Baiduspider (IDC, China Telecommunications Corporation AS23724, CHINA UNICOM China169 Backbone AS4837), MojeekBot (CustodianDC Limited AS50300)

Cloud hosting - Appearing only in Jun: AhrefsBot (OVH SAS AS16276), netEstate NE Crawler (euNetworks GmbH AS13237), DomainStatsBot (Hetzner Online GmbH AS24940)

Cloud hosting - Appearing only in Feb: Bytespider (CHINA UNICOM China169 Backbone AS4837, China Telecom AS141771), Sogou (China Unicom Beijing Province Network AS4808, IDC, China Telecommunications Corporation AS23724, Chinanet AS4134, China Unicom Guangdong IP network AS135061), Coc Coc Bot (VNPT Corp AS45899), UptimeRobot (LIMESTONENETWORKS AS46475, DIGITALOCEAN-ASN AS14061, AMAZON-02 AS16509, AMAZON-AES AS14618), Better Stack Uptime (Akamai Connected Cloud AS63949, Hetzner Online GmbH AS24940)

[10] Cloud & Own Infrastructure - Appearing only in Jun: SemrushBot (SEMrush CY LTD AS209366, GOOGLE-CLOUD-PLATFORM AS396982)

researchers and network administrators to identify the legitimacy of known bots when scanning from their owner's AS. Secondly, given the volatility of IP addresses for bots using the Cloud, it becomes difficult to track their evolution over time or block permanently scanning activities.

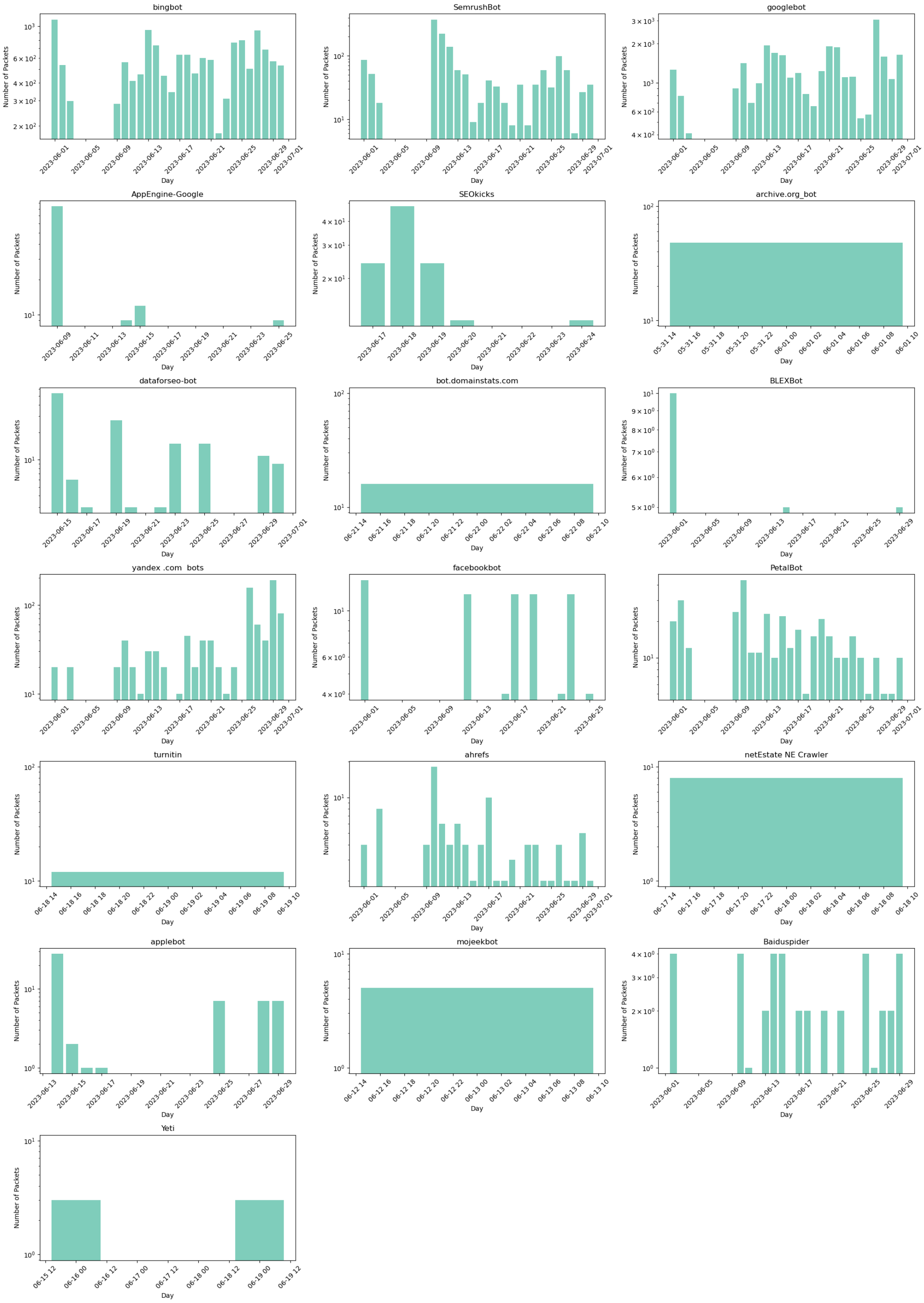


Figure 6.13: Daily Traffic composition per known bot (Jun 23)

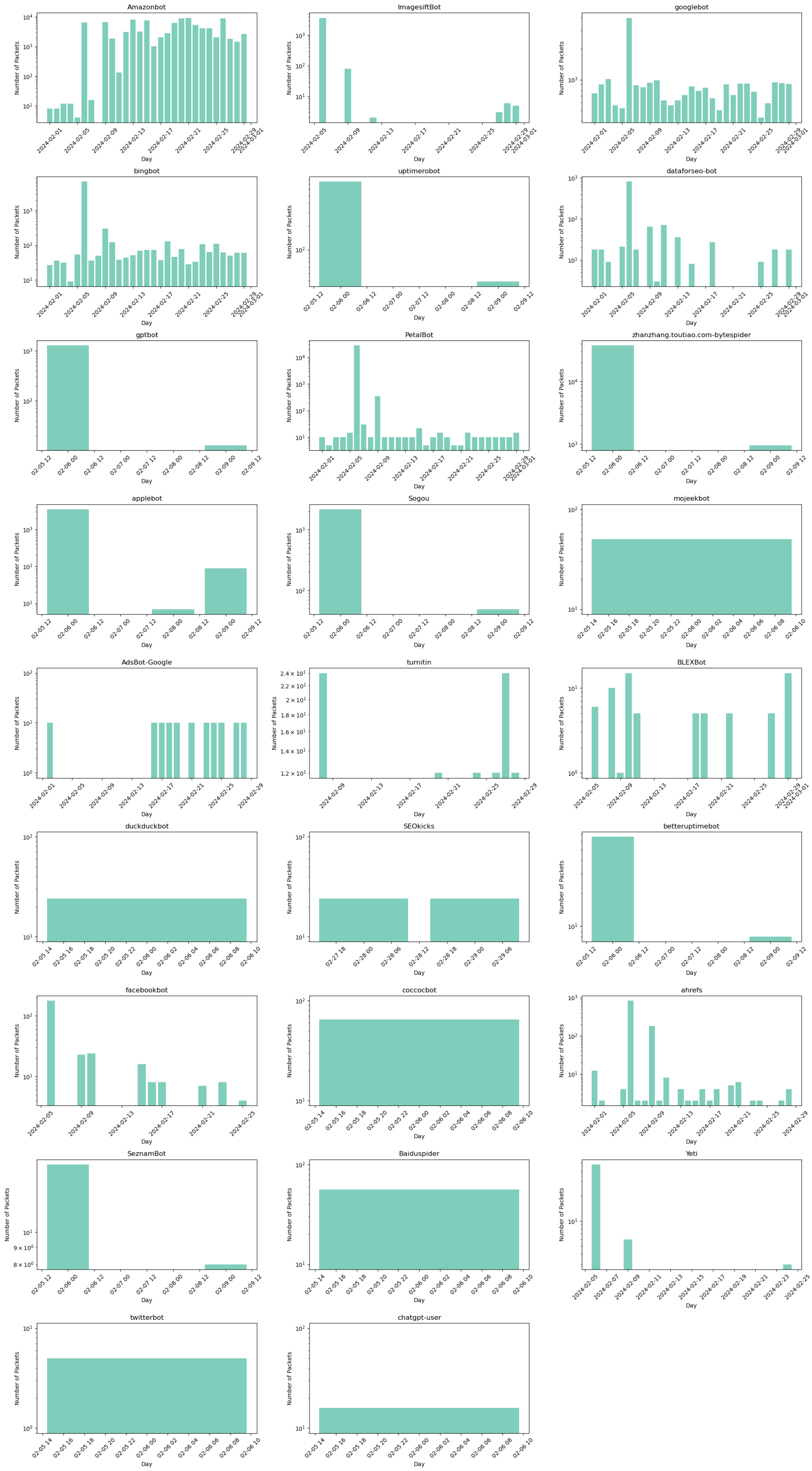


Figure 6.14: Daily Traffic composition per known bot (Feb 24)

7

Fingerprints

7.1. Scanning Events

In this section, we detail our approach to identify and attribute scanning events based on known fingerprint patterns. We focus on three prevalent fingerprints: ZMap, Masscan and Mirai. The detection process begins with compiling and analyzing scanning events based on the scan definition established in section 4.2.3. Scans are classified as high-confidence when *all* observed packets match the same fingerprint. It would be reasonable to expect for a bot or actor employing certain scanning tools, that all packets contain the related fingerprint. However, we acknowledge instances where only a subset of packets within a scanning event matches a known fingerprint. These unidentified known scans are further analyzed to detect medium-confidence scans. In such cases, at least one packet exhibits a known fingerprint, but not all packets adhere to the same pattern. This observation allows us to identify scenarios involving multiple actors operating behind Network Address Translation (NAT). For example, there is one Mirai-infected router and an end-point device behind it that performs scanning. Similarly, medium-confidence Masscan and ZMap scans could indicate actors employing a combination of scanning tools or a proprietary suite. Therefore, the occurrence of multiple signatures within a single scan necessitates such scans to be classified with medium confidence. Under the scan definition, we group 98.92% of the total scanning traffic in June 2023 which maps to 8.15% of total source IP addresses. In February, we group 98.49% of the total traffic, which maps to 6.32% of observed source IPs. Composition of scanning events is presented in table 7.1 below.

Table 7.1: Composition of scanning events in terms of detected fingerprints

. Confidence	Fingerprint	#Scans (Jun 23)	% of traffic (Jun 23)	#Scans (Feb 24)	% of traffic (Feb 24)
High	Mirai	282,813	1.22	71,001	0.56
	ZMap	162,524	25.35	789,162	23.86
	Masscan	108	0.01	2,405	0.07
	ZMap & Mirai	-	-	849	0.18
Medium	Possible Mirai	9,733	0.06	2,023	0.02
	Possible ZMap	47,865	20.97	78,178	19.88
	Possible Masscan	6,655	1.98	8,744	2.28
	Possible Mirai & ZMap (NAT)	121	0.03	133	0.03
	Possible Mirai & masscan (NAT)	77	0.01	71	0.03
	Possible Masscan & ZMap	12,662	36.36	15,539	35.95
	Other (unidentified)	204,046	14.01	374,336	17.14
	Total	726,604	100%	1,342,441	100%

7.2. Mirai Botnet

Mirai is a type of malware that targets Linux-based IoT devices and turns them into a network of bots that can be controlled remotely and used to launch DDoS attacks. It scans for vulnerable devices with default credentials and infects them. After its developers launched major attacks in 2016, the source code became public. Based on a feature of Mirai's stateless scanning, each probe carries a TCP sequence number that is identical to the destination IP address. This allows us to uniquely fingerprint (the original) Mirai traffic among other scanning activities. One limitation of our Darknet is that port TCP/23 is blocked by policy and therefore we cannot assess the amount of missed traffic.

Low-Confidence Fingerprinting. The probability of detecting the signature incidentally is $1/2^{32}$ and thus we would expect the signature to appear in 2.34 out of the 10.04 billion total packets of our dataset. In June, nonetheless, we observe 156.11 million packets from 752,556 distinct source IP addresses that contain the signature. In other words 48.20% of the total source IP addresses observed by our Darknet within the one-month period send at least one packet containing the Mirai signature (1.55% of total traffic). Further traffic decomposition shows that the percentage of Mirai IPs sending below 100 packets in the one-month period is 40.68% accounting for 0.09% of total traffic, whereas the percentage of Mirai IPs sending below 100 packets per day during the above period is 45.59% contributing 0.21% of total traffic.

Accordingly for February, we detect 101.75 million packets from 245,036 distinct source IP addresses that exhibit the Mirai fingerprint. In other words 19.20% of the total source IP addresses observed by our Darknet within the one-month period send at least one packet containing the Mirai signature (1.01% of total traffic). Furthermore, 15.86% of total IPs send below 100 Mirai packets during the examined period, accounting for 2.63 million packets (0.02% of total traffic), whereas 18.42% of total IPs send below 100 packets per day and contribute 10.57 million packets (0.10% of total traffic).

High-Confidence Fingerprinting. Although the probability of accidentally mislabelling a non-Mirai packets is small, however we stress out that the raw count of traffic packets and source IP addresses over time is a poor metric to estimate the botnet size and traffic volume of Mirai-infected IPs. Thus, we employ the scan definition, as described in section 4.2.3, to group traffic flow into logical scans. In June, we note 282,813 scanning events (scans) from 79,612 IP addresses (5.09% of total observed IPs) accounting for 121.24 million Mirai packets (1.20% of total traffic). Comparably to low-confidence fingerprinting, we maintain 77.66% of the Mirai-fingerprinted traffic and 10.57% of Mirai-related addresses. Our Darknet receives on average around 4.84 million Mirai packets per day.

In February, we detect 71,001 scanning events from 13,295 IP addresses (1.04% of total observed IPs) accounting for 55.20 million Mirai packets (0.55% of total traffic). We maintain 54.25% of low-confidence traffic and 5.42% of the related addresses. Our Darknet receives on average around 1.90 million Mirai packets per day.

7.3. Characterization of Mirai Botnet

In this section we perform an analysis and characterization of the Mirai botnet based on the subset obtained through high-confidence fingerprinting.

7.3.1. Device/OS Profiling

In June 2023, 99.77% of Mirai traffic is concentrated between TTL values of 40 and 60. Additionally, the TTL value of 52 exhibits the highest frequency, constituting 35.30% of Mirai traffic and appearing in 43.01% of total scanning events. Conversely, although the majority of Mirai traffic packets in February still falls within the TTL range of 40 to 60, nonetheless it comprises a lower percentage of 85.15%. Lastly, a notable proportion - 14.81% of Mirai traffic - is observed within the range 230 - 251 in this month.

7.3.2. Destination IP & Port Scan Strategy

In order to propagate, Mirai first enters a fast scanning phase during which it concurrently transmits TCP SYN probes to non-blacklisted pseudo-random IPv4 addresses on Telnet ports TCP/23 and TCP/2323 [4]. We observe traffic towards various ports. Figure 7.1 demonstrates the top 20 Mirai-targeted destination ports in June 2023. A scanning event may target one or multiple ports. Specifically, TCP/2323 is the most popular port among the Mirai scans, receiving traffic from 36.45% of scanning events. TCP/52869 follows next with 24.28% of scans. TCP/60023, TCP/8080 and TCP/22 complete the top five ports with 11.89%, 11.75% and 10.69% of hosts.

Among the analyzed scans in February (fig. 7.2), a few ports are particularly noticeable as hubs for network activity. TCP/8080 is the most popular port receiving traffic from 43.95% of scans. Furthermore, TCP/37215 follows next at 43.65%. This port is related to Huawei HG532 routers which allow directory traversal and remote code execution. TCP/80 attracts 28.41% of scans; Telnet TCP/2323 and TCP/8088 receive traffic from 27.52% and 23.43% of scanning events respectively.

In June, 50% of Mirai scans target daily up to 177.0 Darknet IP addresses. 75% of scans cover a broader range, reaching up to 303.81 IP addresses, while 90% extend up to 668.932 IP addresses. On February, a similar pattern emerges with a slight increase. 50% of Mirai scanning events target daily up to 329.36 destination IP addresses. 75% target up to 635.23 IP addresses, while 90% scan up to 947.64 IP addresses. Lastly, quantitative port inspection for June reveals that 50% of Mirai scans target daily up to a single destination port, while 75% and 90% target up to 1.56 and 3.04 ports respectively. Similarly for February, 50% of Mirai scanning events reach up to 2.51 ports, 75% scan up to 5.10 ports and 90% target up to 8.0 ports.

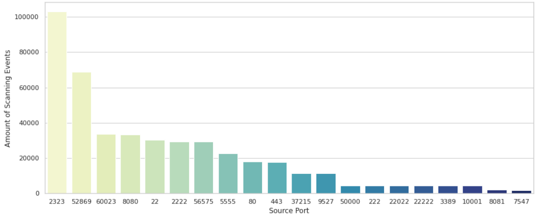


Figure 7.1: Top 20 Mirai-targeted destination ports sorted by amount of scans (Jun 23)

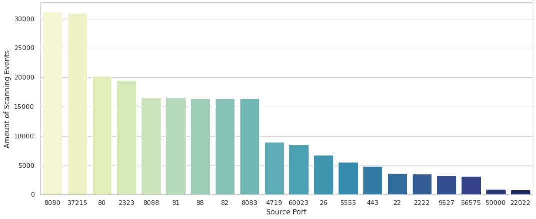


Figure 7.2: Top 20 Mirai-targeted destination ports sorted by amount of scans (Feb 24)

7.3.3. AS and Geolocation Analysis

Mirai originates from 2,459 Autonomous Systems (ASes), spread on 10,953 network prefixes and spanning 155 countries in June 2023. For February 2024, our study encompasses traffic sourced from 1,903 ASes, distributed across 6,044 network prefixes and extending across 134 countries. We note that 106 (10 in Feb) IPs cannot be mapped to an AS. Cross-reference between the two months highlights 2,754 common network prefixes, which correspond to 25.14% in June and 45.56% in February, and 988 common ASes which map to 40.17% in June and 51.91% in February. Notably, there are 617 common IP addresses signifying an 8-month persistent infection or multiple re-infections of the corresponding devices that remained undetected.

The majority of scanning events originate in few major ASes. More precisely, Chinanet (AS4134) holds the highest share of scanning events per AS at 18.48% of total Mirai scans, followed by TE Data (AS8452), Korea Telecom (AS4766) and CHINA UNICOM China169 Backbone (AS4837) at 12.63%, 11.97% and 10.92% of Mirai events respectively. Therefore, four ASes host 54% of Mirai scans. Rest ASes host below 2.50% of the Mirai subset. Similarly, geolocation analysis shows that, along with the rest minor ASes, 35.56% of Mirai scans come from China, 16.26% from South Korea, 12.67% from Egypt and 5.89% from the US. Hence, four countries host 70.38% of Mirai scanning events.

Regarding February 2024, Chinanet (10.67%), Korea Telecom (9.64%) and China Mobile Communications Group Co., Ltd. (5.36%) produce the highest amount of Mirai scans, with the rest countries contributing below 5% of the total Mirai subset each. Mapping AS to countries, we note that China (24.90%), South Korea (13.04%), the US (6.96%), Russia (5.45%) and Taiwan (5.26%) are the five countries with the highest absolute number of scanning events and contribute a total of 55.61% of Mirai scans for this month. Overall, we note a recurrent and stable behavior from the top hitter countries, particularly China and South Korea.

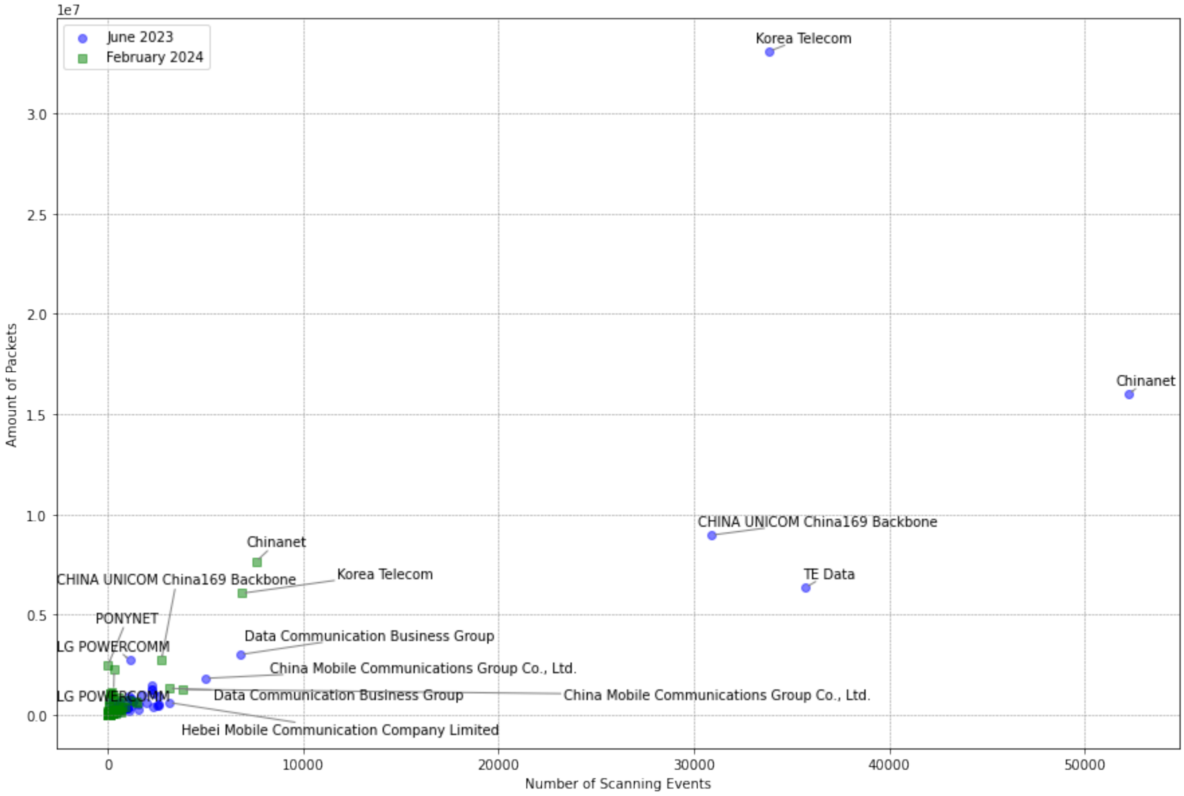


Figure 7.3: Distribution of ASes per number of Mirai scanning events and traffic volume

Examination of the traffic volume per AS reveals comparable results. The top three ASes with the highest traffic volume remain the same for both months. Korea Telecom is one of the prominent ASes, accommodating 27.28% of the total traffic for June and 10.98% of the total traffic for February respectively. Chinanet and CHINA UNICOM China169 Backbone represent 13.19% in June (13.83% in Feb) and 7.39% (4.84% in Feb) of the Mirai traffic subset. The majority of traffic originates from a few countries, with the top three heavy hitter countries being identical between the two months: South Korea, China and the US constitute the largest source of the overall traffic (65.73% in Jun). Along with Egypt at 5.25%, these four countries generate 70.98% of Mirai traffic. The respective percentage for the top three countries in February is 61.46%.

7.4. ZMap Scanners

The probability of detecting incidentally the original ZMap signature is $1/2^{16}$ and thus we would expect the signature to appear in 153,210 out of the 10.04 billion total packets of our dataset. However, we observe 4.04 billion ZMap-fingerprinted packets from 25,804 distinct source IP addresses in June, and 3.75 billion packets from 38,123 distinct source IP addresses in February. In other words, ZMap traffic corresponds to 40.30% and 37.60% of total traffic, in each month. Traffic analysis shows that the percentage of ZMap IPs sending below 100 packets per day is at most 0.02%.

In order to study ZMap scanners further, we employ high-confidence fingerprinting and group traffic flow into logical scans. In June, we detect 162,524 scanning events from 16,294 IP addresses (1.04%

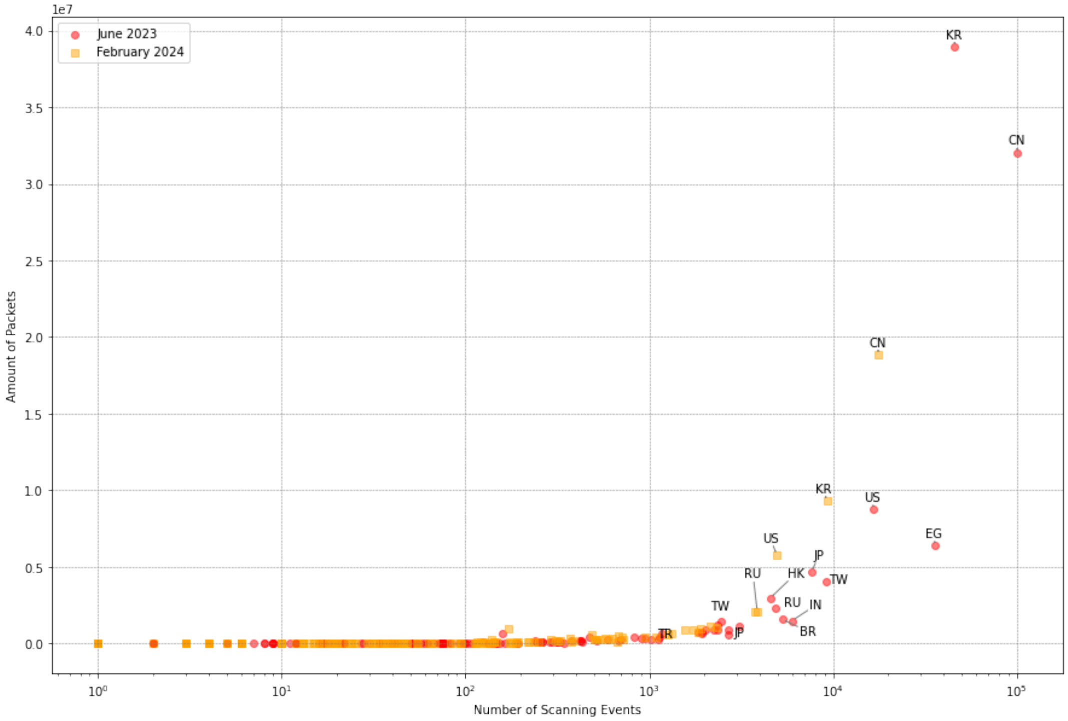


Figure 7.4: Distribution of countries per number of Mirai scanning events and traffic volume

of total observed IPs) accounting for 2.51 billion packets (25.07% of total traffic). Comparably to low-confidence fingerprinting, we maintain 62.20% of the ZMap-fingerprinted traffic and 63.14% of ZMap-related IP addresses. Our Darknet receives on average around 100.70 million ZMap packets per day. In February, we detect 789,162 scanning events from 28,859 IP addresses (2.26% of total observed IPs) accounting for 2.34 billion ZMap packets (23.50% of total traffic). Comparably to low-confidence fingerprinting, we maintain 62.50% of the ZMap-fingerprinted traffic and 75.69% of ZMap-related addresses. Our Darknet receives on average around 81.02 million ZMap packets per day.

7.5. Characterization of ZMap Scanners

In this section, we perform an analysis and characterization of ZMap scanners based on the subset obtained through high-confidence fingerprinting.

7.5.1. Profiling of Scanners

In June, 99.89% of Zmap traffic and 99.02% of scanning events are concentrated between TTL values of 230 and 251. 50% of scans target daily up to 2,053 Darknet IP addresses. 75% of scans cover a broader range, reaching up to 8,217 IP addresses, while 90% extend up to 23,598 IP addresses. Moreover, quantitative port inspection reveals that 50% of scans target daily up to 2.2 destination ports, while 75% and 90% target 8.04 and 17.58 ports respectively. Similarly for February, 99.82% of ZMap traffic packets and 99.10% of scanning events fall within the 230 - 251 TTL range. 50% of Mirai scanning events target daily up to 309.32 destination IP addresses. 75% expand up to 786.75 IP addresses, while 90% scan up to 2342.94 IP addresses. Lastly, 50% of Mirai scanning events reach up to 1.31 ports, 75% scans up to 2.58 ports and 90% target 4.86 ports.

Monthly port trends demonstrate similarities in targeted ports. In June, top selections are HTTP(S) ports TCP/80 and TCP/443. In February, ZMap actors target mostly TCP/22 (SSH) and TCP/3389 (RDP - Remote Desktop Protocol). Overall, the top 20 most targeted ports refer to a variety of services, such as i) HTTP(S) services (e.g., TCP ports 80, 443, 8080, 81) ii) web management/configuration interfaces (e.g., TCP ports 8001, 8090, 9001, 9090, 2031) iii) proxies (e.g., TCP ports 8001, 3128, 8123, 9001) iv) database servers (e.g., TCP/3306) v) remote access services (e.g., TCP/3389) and vi) known protocols and services (FTP at TCP/21, Elasticsearch at TCP/9200 etc.).

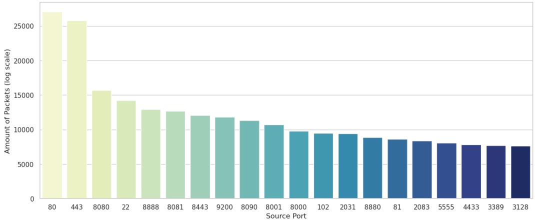


Figure 7.5: Top 20 ZMap-targeted destination ports sorted by amount of scans (Jun 23)

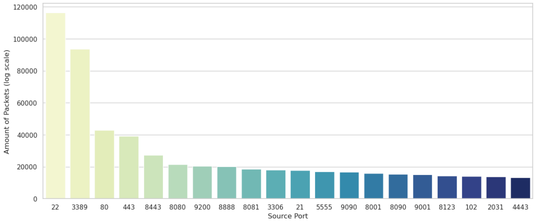


Figure 7.6: Top 20 ZMap-targeted destination ports sorted by amount of scans (Feb 24)

7.5.2. AS and Geolocation Analysis

ZMap originates from 411 Autonomous Systems (ASes), spread on 1,199 network prefixes and spanning 68 countries in June 2023. For February 2024, our study encompasses traffic sourced from 404 ASes, distributed across 1,115 network prefixes and extending across 86 countries. 86 (320 in Feb) IPs cannot be mapped to an AS and one of those cannot also be geolocated. Cross-reference between the two months highlights 469 common network prefixes, which correspond to 39.11% in June and 42.06% in February, and 195 common ASes which map to 47.44% in June and 48.26% in February.

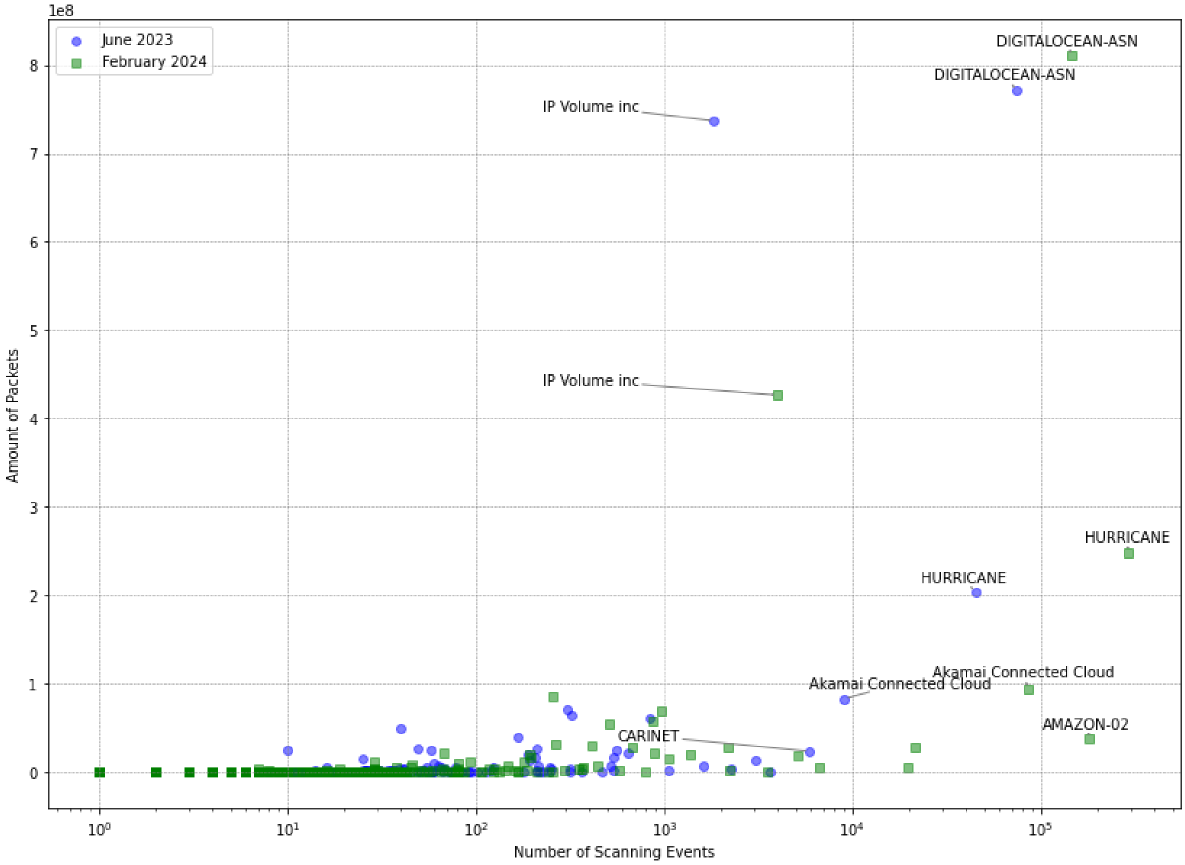


Figure 7.7: Distribution of known scanners per scanned destination port (Jun 23)

The majority of scanning events originate in few major ASes. More precisely, DIGITALOCEAN-ASN (AS14061) holds the highest share of scanning events per AS, at 63.88% of total ZMap scans, followed by GOOGLE-CLOUD-PLATFORM (AS396982), Akamai Connected Cloud (AS63949) and

HURRICANE (AS6939) at 7.46%, 4.62% and 3.59% of ZMap events in June. Therefore, four ASes generate 79.55% of ZMap scans. Rest ASes host below 3% of the ZMap subset. Furthermore, geolocation analysis shows that 40.95% of scans come from the US, 10.07% from the UK, 8.37% from Germany, 8.11% from Singapore and 7.51% from the Netherlands. Hence, five countries host 75.01% of ZMap scanning events.

Regarding February, DIGITALOCEAN-ASN (40.79%) and AMAZON-02 (32.97%) produce the highest amount of ZMap scans. Along with GOOGLE-CLOUD-PLATFORM at 6.96%, these three ASes constitute 80.72% of total ZMap scans. Rest countries contribute below 4% of the total ZMap subset each. Mapping AS to countries, we note that the US hosts 43.84% of ZMap scans and along with Japan (9.03%) and the Netherlands (8.42%) host 61.29% of total scans.

7.5.3. Known Scanners Employing ZMap

ZMap is extensively used by the research community and the security industry. Figures 7.8 and 7.9 showcase the distribution of IP Identification field values per known scanners for both months. Upon data inspection of June, we detect 25 known scanners utilising ZMap, from 2,389 IPs that contribute 1.02 billion packets, or 40.82% of ZMap traffic. In February, we detect 20 known scanners using ZMap, from 5,307 IPs that account for 1.40 billion packets, or 59.75% of ZMap traffic. Certain organizations generate entirely ZMap-fingerprinted traffic (*full-ZMap*). We are able to identify 11[a] research institutes/universities and six[b] (non-profit) security industry firms. Other scanners[c] contain only a subset of IPs from which we observe exclusively ZMap-fingerprinted scanning events. We can conclude that ZMap is partially used by these organizations (*partial-ZMap*). Additionally, two organizations[d] exhibit full-ZMap behaviour in June, but only partial-ZMap behaviour in February. Notably, an organization[e] employs a combination of high-confidence ZMap scans, medium-confidence possible ZMap scans and medium-confidence Masscan & ZMap scans in both months (*multi-ZMap*). Lastly, a number of scanners produce full-ZMap[f] or partial-ZMap[g] activity in June, but multi-ZMap behaviour in February. Hence, we assume they use multiple scanning tools or a proprietary suite throughout these months.

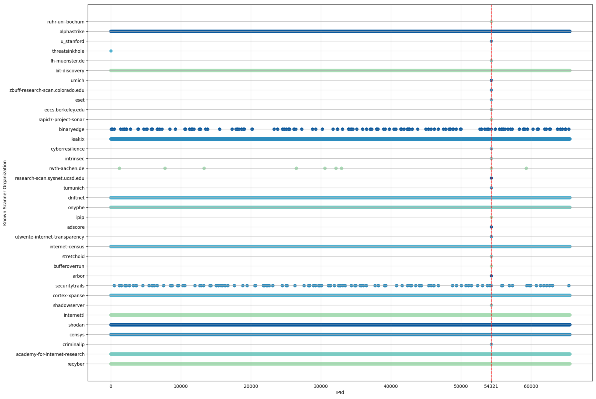


Figure 7.8: Distribution of IP Identification field per known scanner (Jun 23)

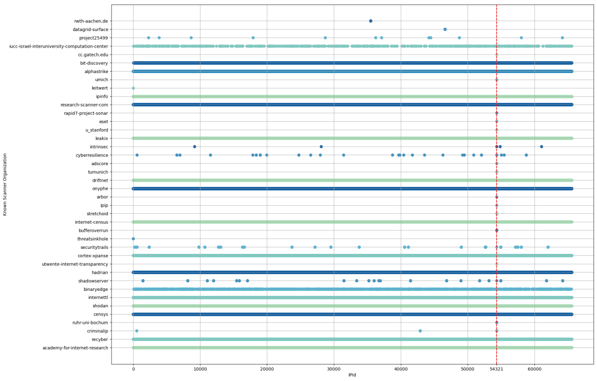


Figure 7.9: Distribution of IP Identification field per known scanner (Feb 24)

[a] Arbor, Ruhr-Universität Bochum, TU Munich, UTwente, U Michigan, Stanford, UC Berkeley, UCSD (only in Jun), CU Boulder (only in Jun), FH Münster (only in Jun), Georgia Tech (only in Feb)

[b] bufferover.run, Stretchoid, Adscore, IPIP, Rapid7 and ESET

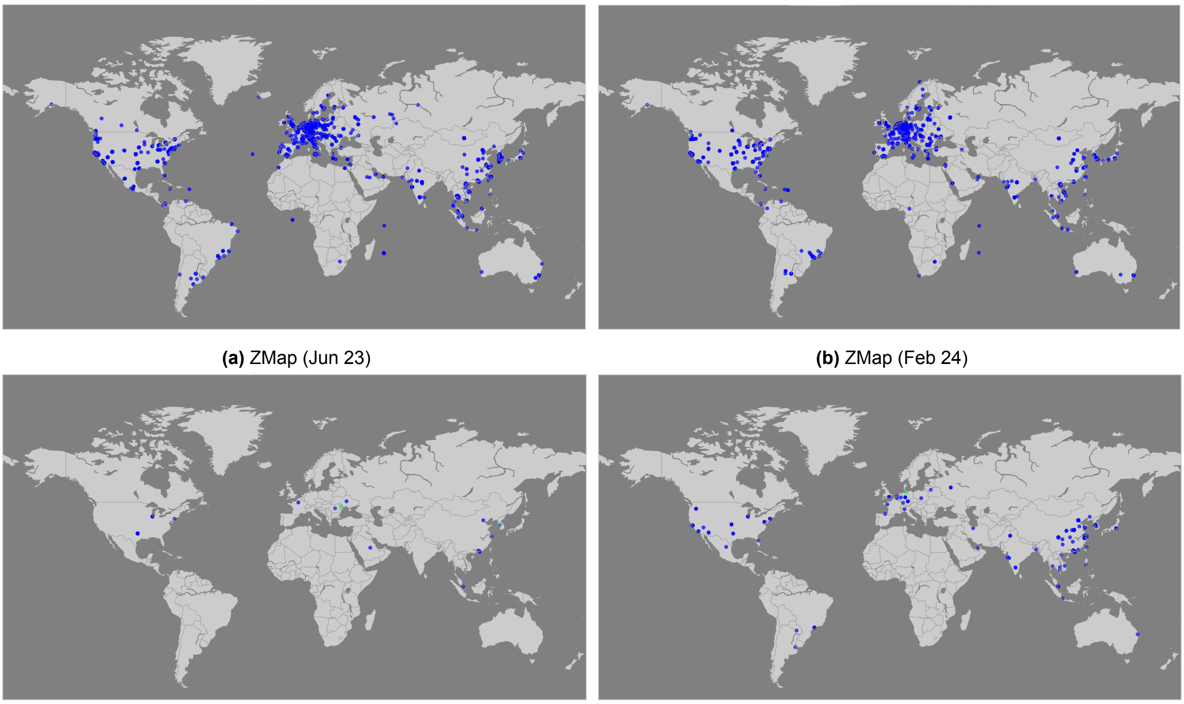
[c] SecurityTrails, Recyber.net (only in Jun) and RWTH-Aachen (only in Jun)

[d] Cyberresilience, Intrinsec

[e] Palo Alto Networks

[f] Shadowserver, Criminal IP

[g] BinaryEdge



(a) ZMap (Jun 23)

(b) ZMap (Feb 24)

(c) Masscan (Jun 23)

(d) Masscan (Feb 24)

Figure 7.10: World Map of (a) ZMap Jun 23 (b) ZMap Feb 24 (c) Masscan Jun 23 (d) Masscan Feb 24. Blue dots represent a low volume of traffic packets. Green dots denote a high volume of packets.

7.6. Masscan Scanners

We observe 845,867 packets from 9,035 distinct source IP addresses (0.57% of total IPs) that exhibit the Masscan signature. Traffic decomposition shows that 8,869 IPs (0.56%) send below 100 packets during the one-month period, whereas the percentage of Masscan IPs sending below 100 packets per day is 8,991 (0.57%). Accordingly for February, we detect 6.77 million Masscan packets (0.06% of total traffic) from 9,452 distinct source IP addresses. 8,386 IPs (0.65%) send below 100 Masscan packets during the examined period and 8,621 IPs send below 100 packets per day. Unlike ZMap, Masscan fingerprint appears sparsely in our dataset, less than 0.07% of each monthly traffic.

Grouping the traffic flow into logical scans (high-confidence fingerprinting) generates 108 scanning events from 41 IP addresses accounting for 730,077 packets. In February, we detect 2,405 scanning events from 882 IP addresses contributing 6.57 million packets. Our Darknet receives on average around 29,203 packets per day in June and 226,582 packets per day in February. None of the known scanners or known bots employ this particular scanning tool. Although, some scans from known scanners include a small portion of Masscan-fingerprinted packets, however it is insignificant comparing to the scan's traffic volume ($\leq$ 0.90%).

7.7. Characterization of Masscan Scanners

In this section we perform an analysis and characterization of actors using Masscan, based on the subset obtained through high-confidence fingerprinting.

June 2023. The distribution of Masscan traffic is exclusively concentrated between TTL values of 30 and 60. 50% of Mirai scans target daily up to 228.30 destination ports. 75% of scans cover a broader range, reaching up to 414.09 ports, while 90% extend up to 539.24 ports. Quantitative destination address inspection reveals that 50% of Masscan scans target daily up to 3,684 destination

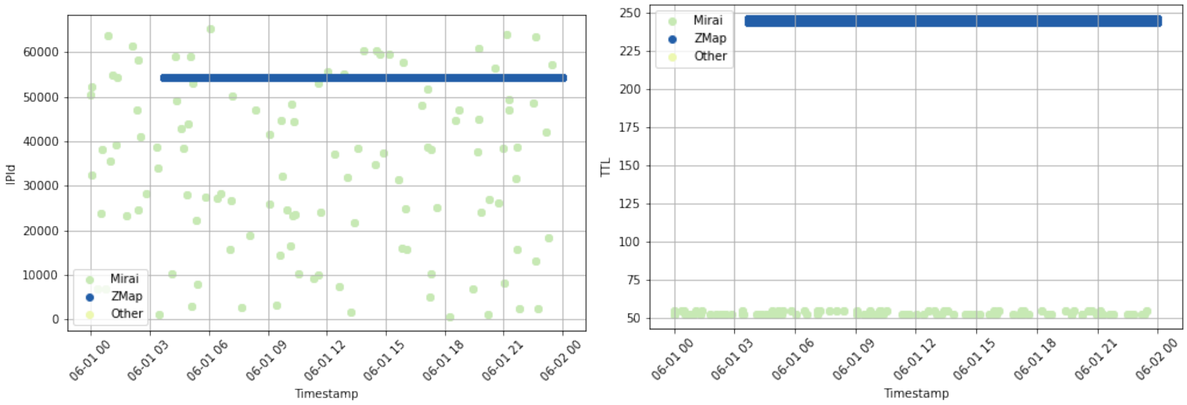
IPs, while 75% and 90% target around 5,814 and 8,407 ports respectively. Traffic originates from 11 Autonomous Systems (ASes), spread on 19 network prefixes and spanning 10 countries for June 2023 (fig. 7.10c). Datacamp Limited (AS60068) holds the highest share of scanning events per AS at 63.41% of total Masscan events, followed by Hangzhou Alibaba Advertising Co. Ltd. (AS37963) at 12.20%. Rest ASes generate less than 5% of the events. Traffic analysis shows that Stark Industries Solutions Ltd (AS44477) generate 33.73% of the total Masscan traffic, followed by DIGITALOCEAN-ASN (AS14061), AMAZON-02 (AS16509), and Datacamp Limited at 16.28%, 16.10% and 12.39% of Masscan traffic. Therefore, the majority of scans originate from the United States (63.41%) and combined with China (14.63%) they produce 78.04% of total Masscan events. Conversely, Moldova produces 33.73% of traffic followed by Singapore (17.20%), South Korea (16.10%) and the US (12.39%). Rest countries produce less than 7% of the total Masscan traffic.

February 2024. 92.11% of Masscan traffic lies in the TTL range 100-123. 50% of Mirai scans target daily up to 1.75 ports, whereas 75% (and 90%) target up to 2 ports. Additionally, 50% of scanning events target daily up to 1,166 destination IP addresses. 75% expand up to 3,017 IP addresses, while 90% scan up to around 3,810 IP addresses. Our study encompasses traffic sourced from 92 ASes, distributed across 279 network prefixes and extending across 32 countries (fig. 7.10d). Hangzhou Alibaba Advertising Co.,Ltd. (AS37963) hosts 38.78% of total Masscan events and produces 41.19% of traffic, followed by Alibaba US Technology Co., Ltd. (AS45102) hosting 24.26% of scans and generating 27.62% of Masscan traffic. Rest countries host below 10% of scans and generating less than 5.2% of traffic each. Mapping AS to countries, we note that 82.65% of events originate from China (62.81%) and Hong Kong (19.84%) and account for 80.93% of total Masscan traffic (56.75% and 24.18% respectively).

7.8. NAT Remote Detection

7.8.1. Proof of Concept

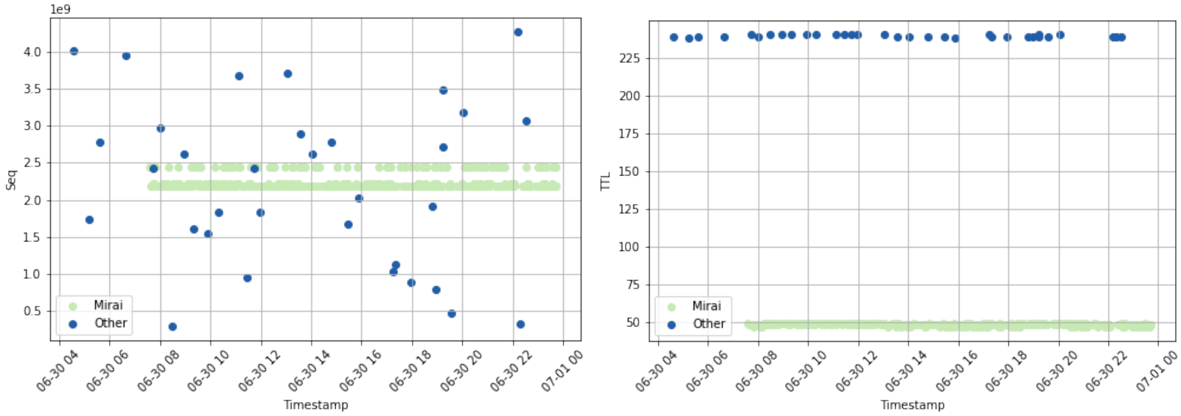
Different fingerprints within the same scanning event suggest the existence of a NAT IP address. NAT - standing for Network Address Translation - modifies the network address information of the IP headers as they are being transmitted through a traffic routing device - such as a router or firewall - in order to map an IP address space into another one. Thus, it enables multiple devices within a local network to share a single public IP address. We focus on identifying NAT based on the combination of Mirai and one of the following: ZMap, Masscan, no fingerprint. This allows a fast and reliable extraction of results. Figures 7.11 and 7.12 present two events as examples of the proof of concept. Event-1 refers to a medium-confidence possible Mirai & ZMap event from Vietnam observed at 1 Jun 23. Annotation of probes within the event reveals a ZMap actor (fig. 7.11; blue line) and a Mirai bot with random IPId values (green dots). Analysis of TTL values over time shows that actors scan concurrently and can be effectively distinguished based on TTL values. It is noted that we might observe more than two actors in some events (ZMap, Mirai bot and unidentified traffic).



(a) IPId over time

(b) TTL over time

Figure 7.11: *Event-1*: (a) IPId value over time and (b) TTL value over time for a medium-confidence possible Mirai & ZMap event at 1 Jun 23.



(a) TCP Sequence over time

(b) TTL over time

Figure 7.12: *Event-2*: (c) TCP Sequence value over time and (b) TTL value over time for a medium-confidence possible Mirai event at 30 Jun 23.

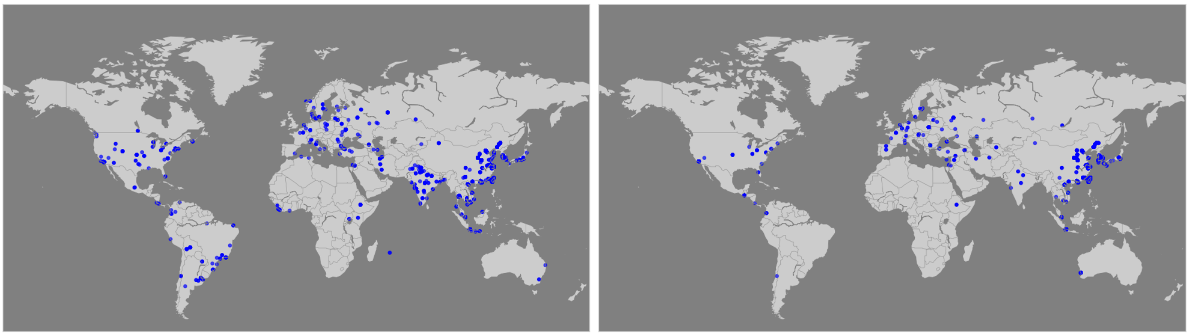
Event-2 refers to a medium-confidence possible Mirai event originating in Iran, logged at 30 Jun 23. Traffic analysis shows a Mirai bot scanning continuously (fig. 7.12a; green lines) and an unidentified actor with random TCP Sequence values (blue dots). One could state that the same device is infected by another malware or Mirai variant. Nonetheless, inspection of TTL values over time shows that Mirai and non-Mirai traffic fall within different TTL ranges that cannot originate from the same underlying OS of the infected device. Hence, we assume that at least two devices exist behind this public IP address, provided that neither actor modifies the OS behavior.

7.8.2. NAT Characterization

We study 9,931 events from 2,844 IPs contributing 10.60 million packets in June. 63.36% of NAT IPs are hosted in Chinanet (AS4134), followed by CHINA UNICOM China169 Backbone (AS4837) at 11.26% and PT. Pasifik Satelit Nusantara (AS9875) at 4.14%. The rest ASes host less than 3% of NAT IP addresses. Overall, the majority of NAT is located in China, or 80.68% of NAT IPs (fig. 7.13a). Indonesia follows with a considerable distance, hosting 5.56% of IPs. Rest countries host up to 2% of the NAT subset IPs.

Similar trends are observed in February (fig. 7.13b). We study 2,227 events from 505 IPs accounting for 7.26 million packets. 23.71% of NAT IPs are hosted in Chinanet (AS4134), followed by Data Communication Business Group (AS3462) at 19.17%. Ambit Systemy Informatyczne Sp. Z O.o. (AS197262), CHINA UNICOM China169 Backbone (AS4837), China Mobile Communications Group Co., Ltd. (AS9808) complete the top five ASes in terms of NAT IP size, hosting 9.03%, 7.32% and 6.78% respectively. Rest ASes host less than 3.5% of NAT IP addresses. Mapping ASes to countries yields a more distributed and diverse picture than in June. Overall, a significant number of NAT IPs are located in China (41.09%). Indonesia, Poland and South Korea follow at 19.71%, 9.03% and 7.54% of NAT IPs. Rest countries host up to 3% of the NAT subset IPs.

Traffic composition demonstrates that Mirai signature appears in up to 85.59% (13.39% in Feb) of total packets in 50% of scanning events, while it reaches up to 94.97% (86.95% in Feb) of traffic in 75% of scanning events and up to 97.52% (95.38% in Feb) of traffic in 90% of scanning events. On the other hand, ZMap and Masscan fingerprints are occasional (90% of scanning events include zero fingerprinted packets from these tools). Lastly, half of NAT scans contain up to 14.28% (63.43% in Feb) of unidentified packets, meaning that these packets do not match to any known fingerprint. The percentage increases up to 21.54% (96.47% in Feb) for 75% of scans, and up to 64.63% (97.75% in Feb) for 90% of scans. These packets possibly belong to Mirai variants or are generated by custom tools.



(a)

(b)

Figure 7.13: World Map of NAT locations at (a) Jun 23 and (b) Feb 24

8 Discussion

8.1. Main Findings

In chapter 1, we defined the research question of who scans the Internet, how we can classify Internet scanners based on their origin and what the discernible differences between scanning activities conducted by malicious actors and those associated with research organizations are. Having divided the main research question into four sub-questions, we discuss our findings per sub-question.

Q1. Can scanning probes be classified into specific categories based on their originating IP addresses, and if so, what are these categories?

Scanning probes be classified into specific categories based on their originating IP addresses:

- benign
- malicious
- unidentified

Additionally, we can classify actors based on their motives into the following six categories:

- security industry
- non-profit organizations
- academic/research institutions
- commercial bots
- malicious botnets
- other

Q2. Are there identifiable differences in the scanning methodologies and techniques used by malicious actors compared to those used by legitimate research entities?

There are discernible qualitative and quantitative differences in the modus operandi of malicious actors and entities with legitimate interest, in terms of traffic volume, destination address dispersion, pool size of source IPs, ports targeted, software used etc. For instance, Mirai comprises 1.55% of total traffic in June (1.01% in Feb), whereas institutional scanners generate 51.31% of the total traffic. On the other hand, 48.20% of total source IPs send at least one packet containing the Mirai signature (19.20% in Feb), while only 0.36% of the total source IPs is attributed to institutional scanners (0.62% in Feb).

Overall, the relationship between *IP address size* and *volume traffic* appears to follow the Pareto principle i.e. 1% of total addresses accounts for 97.38% of the total traffic in June 2023 and 96.65% in February 2024.

Among this top-1% of aggressive hitters, we identify 33 known scanners, Mirai botnets and ZMap/Masscan scanners of unknown intention. Conversely, 0.04% to 0.06% of source IPs originates in Tor exit nodes and Tor traffic remains negligibly small and accounts for merely 0.01% of the overall Darknet traffic for each month.

Port strategy differs per actor. Mirai targets TCP ports 2323, 52869, 8080, 37215 etc., commercial bots scan HTTP(S) ports such as 80 and 443, whereas some security industry actors scan the whole range of destination ports, and research institutions usually target only few ports.
Lastly, packet fingerprinting yields that 40.30% and 37.60% of total traffic belongs to ZMap in each month, whereas Masscan fingerprint appears sparsely, less than 0.07% of each monthly traffic. In fact, ZMap is extensively used by the research community and the security industry. 25 known scanners (20 in Feb) utilize ZMap in June or 40.82% of the identified ZMap traffic (59.75% in Feb).

Q3. What are the geographic distributions of scanners, and can geographic patterns help distinguish between malicious and non-malicious scanning activities?

Internet-wide scanners are highly distributed originating from over 12,000 Autonomous Systems (ASes), spread on more than 55,000 network prefixes and spanning over 220 countries. Even though the number of common IP addresses between the two months is low (8-10%), nonetheless we observe 60-65% common prefixes and 64-72% common ASes between the two months, suggesting a recurrent behavior.

Despite the global origin of traffic, the majority of scanning IP addresses is concentrated to few major ASes. Six ASes provide 54.23% of the unique source addresses in June and 48.53% in February (TE Data, Chinanet, CHINA UNICOM China169 Backbone, National Internet Backbone, Iran Telecommunication Company PJS; the sixth is Telecomunicaciones MOVILNET in Jun and DIGITALOCEAN-ASN in Feb). Furthermore, five ASes comprise 60.83% of the total traffic in June and 45.51% for February (GOOGLE-CLOUD-PLATFORM, IP Volume Inc, DIGITALOCEAN-ASN, CENSYS-ARIN-01, Chang Way Technologies Co. Limited).

The majority of scanning IP addresses is concentrated to few countries. Five countries collectively account for 66.36% of all unique source IP addresses (China, Egypt, India, Venezuela, Iran). 1-1.5 out of 5 IP addresses is hosted in China. Overall, we note a recurrence and stability in the size of scanning IPs from the top hitter countries, particularly China, Egypt, Iran and India. The combined traffic volume of four countries (USA, UK, NL, Russia) accounts for 79.10% of the total traffic in June and 70.84% in February.

Known benign scanners may employ proprietary ASes or use third-party cloud hosting. Since the former renders them easily identifiable, network administrators and SOC analysts can use this information for alert triage and to avoid triggering false-positive alerts. Accordingly, they can identify suspicious traffic by inspecting packets for known fingerprints or evaluating the sender's AS reputation, provided that, for instance, Mirai frequents in certain ASes more than others.

Q4. To what extent can historical data and datasets from disperse source nodes be used to predict the intent behind a scanning operation, whether research-driven or malicious?

Data from disperse sources and historical data (open source, academic repositories, commercial Threat Intelligence and Attack Surface Management platforms, IP blocklists etc.) prove to make a significant contribution when predicting the intent behind a scanning IP address.

We compile a comprehensive repository of known (benign) scanners and known bots based on diverse sources for each month separately. We collect and aggregate data for 36 organizations in June that correspond to 0.36% of the total source IP addresses and account for 51.31% of the total telescope traffic. We also identify 40 organizations in February, which correspond to 0.62% of the total source IP addresses and account for 50.86% of the total telescope traffic. In total, we find traces from 44 known scanners. On the other hand, known bots appear to contribute negligible traffic volume. We identify 19 bots in June and 27 bots in February, corresponding to 0.11% and 0.25% of total IP addresses and contributing below 0.01% of the total telescope traffic. In total, we identify traffic from 34 commercial bots.

Essentially seven organizations (Palo Alto Networks, Censys, Criminal IP, Recyber.net, Academy for Internet Research LLC, The Shadowserver Foundation, Driftnet.io) - mapping to 0.17% of the total source IP addresses - are responsible for 49.27% of the total traffic in June 2023. Similarly, 10 organizations (Palo Alto Networks, Censys, Stretchoid, Criminal IP, Academy for Internet Research LLC, The Shadowserver Foundation, Driftnet.io, InterneTTL, Shodan, Recyber.net) - mapping to 0.36% of the total source IP addresses - account for 49.20% of the total traffic in February 2024.

Results corroborate the emerging need for mutual sharing of threat intelligence among defenders (not just blocklisted IP addresses), due to the evolving cyber threat landscape and the fact that scanning is often a early indication of upcoming cyber attacks. Exchange of intelligence allows the creation of an up-to-date threat assessment, enables organizations to protect timely their digital infrastructure and, therefore, should be encouraged further.

8.2. Reflection and Limitations

In this section, we summarize all the limitations we have discussed throughout the study. First, our study is bounded by the characteristics of conventional network telescopes, which are described in section 2.3.3. Additionally, we only examine the IPv4 address space. We rely on the non-commercial Maxmind Geolite2 to map IP addresses to ASes and countries. As indicated in section 4.6, Maxmind offers acceptable precision, however some data centers or company headquarters may be less accurately geolocated. Traffic to ports TCP/445 and TCP/23 is blocked before reaching our telescope. Lastly, some datasets, e.g. the Censys API dataset, are used to extract results for both months due to time constraints in data access.

9 Conclusion

9.1. Conclusion

In this work, we have surveyed Internet-wide scanning traffic of two monthly snapshots in two different years (2023 & 2024) of approximately 10 billion packets each. Internet-wide scanning is used by security researchers and industry specialists to collect information and deliver security services and products, while hackers use it to target vulnerable devices. Internet scanning, as an early stage of a cyber attack, poses a significant threat to organizations due to its prevalence and the limited efficiency of preventive solutions. In order to answer the main research question of who scans the Internet, how can we classify Internet scanners based on their origin and what differentiates malicious scanning activities from those associated with research organizations, we performed an empirical analysis from a single vantage point using a network telescope. We have studied the behavior of botnets and shed light in the current use of scanning software such as ZMap and Masscan, and have provided a methodology for data collection and aggregation of known scanners.

9.2. Future Work

Possible future work directions involve extending the data collection period, due to the evolving nature of the scanning traffic. Another important aspect is the establishment of a pipeline with continuous discovery and integration of known scanners. This will allow researchers to exclude properly known (institutional) scanners and, thus, avoid misinterpretation of the actual malicious traffic. Lastly, extending the analysis to the IPv6 address space could offer useful insights to understanding scanning as a practice.

References

[1] *Academy for Internet Research*. URL: https://academyforinternetresearch.org/.

[2] *Alpha Strike Labs*. URL: https://www.alphastrike.io/en/.

[3] Aniket Anand et al. *Aggressive Internet-Wide Scanners: Network Impact and Longitudinal Characterization*. 2023. arXiv: 2305.07193 [cs.NI].

[4] Manos Antonakakis et al. "Understanding the Mirai Botnet". In: *26th USENIX Security Symposium (USENIX Security 17)*. Vancouver, BC: USENIX Association, Aug. 2017, pp. 1093–1110. ISBN: 978-1-931971-40-9. URL: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/antonakakis.

[5] MITRE ATT&CK®. *Active Scanning, Technique T1595 - Enterprise | MITRE ATT&CK®*. 2022. URL: https://attack.mitre.org/techniques/T1595/.

[6] MITRE ATT&CK®. *Internet Scan, Data Source DS0035 | MITRE ATT&CK®*. 2021. URL: https://attack.mitre.org/datasources/DS0035/.

[7] MITRE ATT&CK®. *Search Open Technical Databases: Scan Databases, Sub-technique T1596.005 - Enterprise | MITRE ATT&CK®*. 2021. URL: https://attack.mitre.org/techniques/T1596/005/.

[8] Fred Baker. *RFC 1812: Requirements for IP Version 4 routers*. URL: https://datatracker.ietf.org/doc/html/rfc1812.

[9] Tao Ban et al. "Behavior Analysis of Long-term Cyber Attacks in the Darknet". In: vol. 7667. Nov. 2012, pp. 620–628. ISBN: 978-3-642-34499-2. DOI: 10.1007/978-3-642-34500-5_73.

[10] Tao Ban et al. "Detection of Botnet Activities Through the Lens of a Large-Scale Darknet". In: *Neural Information Processing*. Ed. by Derong Liu et al. Cham: Springer International Publishing, 2017, pp. 442–451. ISBN: 978-3-319-70139-4.

[11] Tao Ban et al. "Towards Early Detection of Novel Attack Patterns through the Lens of a Large-Scale Darknet". In: *2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*. 2016, pp. 341–349. DOI: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0068.

[12] Karyn Benson et al. "Gaining insight into AS-level outages through analysis of Internet Background Radiation". In: *2013 Proceedings IEEE INFOCOM Workshops, Turin, Italy, April 14-19, 2013*. IEEE, 2013, pp. 447–452. DOI: 10.1109/INFCOMW.2013.6562915. URL: https://doi.org/10.1109/INFCOMW.2013.6562915.

[13] Karyn Benson et al. "Leveraging Internet Background Radiation for Opportunistic Network Analysis". In: *Proceedings of the 2015 Internet Measurement Conference*. IMC '15. Tokyo, Japan: Association for Computing Machinery, 2015, pp. 423–436. ISBN: 9781450338486. DOI: 10.1145/2815675.2815702. URL: https://doi.org/10.1145/2815675.2815702.

[14] Monowar H. Bhuyan, Dhruba Kumar Bhattacharyya, and Jugal Kumar Kalita. "Surveying Port Scans and Their Detection Methodologies". In: *Comput. J.* 54 (2011), pp. 1565–1581. URL: https://api.semanticscholar.org/CorpusID:2867949.

[15] Wim Biemolt. *2021-06-26: Recyber-net en Netfilter rootkit*. 2021. URL: https://wiki.surfnet.nl/display/SURFcert/2021/06/26/2021-06-26%3A+recyber-net+en+Netfilter+rootkit.

[16] Birk Blechschmidt. *A high-performance DNS stub resolver for bulk lookups and reconnaissance (subdomain enumeration)*. URL: https://github.com/blechschmidt/massdns.

[17] Elias Bou-Harb, Mourad Debbabi, and Chadi Assi. “Cyber Scanning: A Comprehensive Survey”. In: *Communications Surveys & Tutorials, IEEE* 16 (Jan. 2014), pp. 1496–1519. DOI: 10.1109/SURV.2013.102913.00020.

[18] N Brownlee. “One-way Traffic Monitoring with iatmon”. In: *Passive and Active Network Measurement Workshop (PAM)*. Mar. 2012. DOI: https://catalog.caida.org/paper/2012_one_way_traffic_iatmon.

[19] *CAIDA*. Jan. 2018. URL: https://www.caida.org/projects/network_telescope/.

[20] Z. Chen, C. Ji, and P. Barford. “Spatial-Temporal Characteristics of Internet Malicious Sources”. In: *IEEE INFOCOM 2008 - The 27th Conference on Computer Communications*. 2008, pp. 2306–2314. DOI: 10.1109/INFOCOM.2008.299.

[21] M. Patrick Collins, Alefiya Hussain, and Stephen Schwab. “Identifying and Differentiating Acknowledged Scanners in Network Traffic”. In: *2023 IEEE European Symposium on Security and Privacy Workshops (EuroSPW)*. 2023, pp. 567–574. DOI: 10.1109/EuroSPW59978.2023.00069.

[22] Michael Collins et al. “Using uncleanliness to predict future botnet addresses”. In: Oct. 2007, pp. 93–104. DOI: 10.1145/1298306.1298319.

[23] *CVE-2016-9223*. 2016. URL: https://www.cve.org/CVERecord?id=CVE-2016-9223.

[24] Alberto Dainotti et al. “Analysis of a ”/0” stealth scan from a botnet”. In: *Proceedings of the 2012 Internet Measurement Conference*. New York, NY, USA: Association for Computing Machinery, 2012, pp. 1–14. ISBN: 9781450317054. DOI: 10.1145/2398776.2398778. URL: https://doi.org/10.1145/2398776.2398778.

[25] Alberto Dainotti et al. “Analysis of country-wide internet outages caused by censorship”. In: *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*. IMC ’11. Berlin, Germany: Association for Computing Machinery, 2011, pp. 1–18. ISBN: 9781450310130. DOI: 10.1145/2068816.2068818. URL: https://doi.org/10.1145/2068816.2068818.

[26] Alberto Dainotti et al. “Extracting benefit from harm: using malware pollution to analyze the impact of political and geophysical events on the internet”. In: *SIGCOMM Comput. Commun. Rev.* 42.1 (Jan. 2012), pp. 31–39. ISSN: 0146-4833. DOI: 10.1145/2096149.2096154. URL: https://doi.org/10.1145/2096149.2096154.

[27] Dansimp. *TCP/IP port exhaustion troubleshooting - windows client*. URL: https://learn.microsoft.com/en-us/troubleshoot/windows-client/networking/tcp-ip-port-exhaustion-troubleshooting.

[28] *Default dynamic port range for TCP/IP for Linux*. URL: https://tldp.org/LDP/solrhe/Securing-Optimizing-Linux-RH-Edition-v1.3/chap6sec70.html.

[29] Zakir Durumeric, Michael Bailey, and J. Alex Halderman. “An Internet-Wide View of Internet-Wide Scanning”. In: *23rd USENIX Security Symposium (USENIX Security 14)*. San Diego, CA: USENIX Association, Aug. 2014, pp. 65–78. ISBN: 978-1-931971-15-7. URL: https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/durumeric.

[30] Zakir Durumeric, Eric Wustrow, and J Alex Halderman. “ZMap: Fast Internet-wide scanning and its security applications”. In: *22nd USENIX Security Symposium*. 2013.

[31] Zakir Durumeric et al. “A Search Engine Backed by Internet-Wide Scanning”. In: *22nd ACM Conference on Computer and Communications Security*. Oct. 2015.

[32] Glenn Fink. “Visual Correlation of Network Traffic and Host Processes for Computer Security”. In: (Oct. 2006).

[33] Carrie Gates et al. “Detecting Scans at the ISP Level”. In: (Apr. 2006), p. 46.

[34] Vincent Ghiette, Norbert Blenn, and Christian Doerr. “Remote Identification of Port Scan Toolchains”. In: *2016 8th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*. 2016, pp. 1–5. DOI: 10.1109/NTMS.2016.7792471.

[35] Luca Gioacchini et al. “DarkVec: automatic analysis of darknet traffic with word embeddings”. In: *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies*. CoNEXT ’21. Virtual Event, Germany: Association for Computing Machinery, 2021, pp. 76–89. ISBN: 9781450390989. DOI: 10.1145/3485983.3494863. URL: https://doi.org/10.1145/3485983.3494863.

[36] Robert Graham. *MASSCAN: Mass IP port scanner*. 2013. URL: `https://github.com/robertdavidgraham/masscan`.

[37] Harm Griffioen and Christian Doerr. “Discovering Collaboration: Unveiling Slow, Distributed Scanners based on Common Header Field Patterns”. In: *NOMS 2020 - 2020 IEEE/IFIP Network Operations and Management Symposium*. 2020, pp. 1–9. DOI: `10.1109/NOMS47738.2020.9110444`.

[38] Harm Griffioen and Christian Doerr. “Quantifying autonomous system IP churn using attack traffic of botnets”. In: *Proceedings of the 15th International Conference on Availability, Reliability and Security*. ARES ’20. Virtual Event, Ireland: Association for Computing Machinery, 2020. ISBN: 9781450388337. DOI: `10.1145/3407023.3407051`. URL: `https://doi.org/10.1145/3407023.3407051`.

[39] Raphael Hiesgen et al. *Spoki: Unveiling a New Wave of Scanners through a Reactive Network Telescope*. Oct. 2021.

[40] Carola Houtekamer and Rik Wassens. *Het Afvoerputje van het internet zit in een Noord-hollands dorp*. Apr. 2021. URL: `https://web.archive.org/web/20220223063404/https://www.nrc.nl/nieuws/2021/04/02/het-afvoerputje-van-het-internet-zit-in-een-noord-hollands-dorp-a4038329`.

[41] Félix Iglesias and Tanja Zseby. “Pattern Discovery in Internet Background Radiation”. In: *IEEE Transactions on Big Data* 5.4 (2019), pp. 467–480. DOI: `10.1109/TBDATA.2017.2723893`.

[42] *Internettl Research Project*. URL: `https://web.archive.org/web/20210921015530/http://www.internettl.org/`.

[43] *IPinfo*. URL: `https://ipinfo.io/`.

[44] Barry Irwin. “A baseline study of potentially malicious activity across five network telescopes”. In: *2013 5th International Conference on Cyber Conflict (CYCON 2013)*. 2013, pp. 1–17.

[45] Michalis Kallitsis et al. “Detecting and Interpreting Changes in Scanning Behavior in Large Network Telescopes”. In: *IEEE Transactions on Information Forensics and Security* 17 (2022), pp. 3611–3625. DOI: `10.1109/TIFS.2022.3211644`.

[46] FortiGuard Labs. *Intrusion prevention - TCP Bad Flags*. 2014. URL: `https://www.fortiguard.com/encyclopedia/ips/12145/tcp-bad-flags`.

[47] MaxMind. *Geolite2 Free Geolocation Data*. 2023. URL: `https://dev.maxmind.com/geoip/geolite2-free-geolocation-data`.

[48] *Microsoft Security Bulletin MS01-040*. 2003. URL: `https://learn.microsoft.com/en-us/security-updates/SecurityBulletins/2001/ms01-040`.

[49] *Microsoft Security bulletin MS02-051*. 2002. URL: `https://learn.microsoft.com/en-us/security-updates/SecurityBulletins/2002/ms02-051`.

[50] Martin Monperrus. *Monperrus | Syntactic patterns of HTTP user-agents used by bots/robots/crawlers/scrapers/spiders*. 2024. URL: `https://github.com/monperrus/crawler-user-agents`.

[51] David Moore et al. “Inferring Internet denial-of-service activity”. In: *ACM Trans. Comput. Syst.* 24.2 (May 2006), pp. 115–139. ISSN: 0734-2071. DOI: `10.1145/1132026.1132027`. URL: `https://doi.org/10.1145/1132026.1132027`.

[52] David Moore et al. “Inside the Slammer worm”. In: *Security Privacy, IEEE* 1 (Aug. 2003), pp. 33–39. DOI: `10.1109/MSECP.2003.1219056`.

[53] David Moore et al. “Network Telescopes: Technical Report”. In: (Aug. 2004).

[54] Giovane Moura, Ramin Sadre, and Aiko Pras. “Bad Neighborhoods on the Internet”. In: *Communications Magazine, IEEE* 52 (July 2014), pp. 132–139. DOI: `10.1109/MCOM.2014.6852094`.

[55] Giovane C. M. Moura et al. “Internet bad neighborhoods aggregation”. In: *2012 IEEE Network Operations and Management Symposium*. 2012, pp. 343–350. DOI: `10.1109/NOMS.2012.6211917`.

[56] *myips.ms*. 2024. URL: `https://myip.ms/browse/blacklist/Blacklist_IP_Blacklist_IP_Addresses_Live_Database_Real-time`.

[57] Martin Pizala. *Exploit Database | Docker daemon - unprotected TCP Socket*. July 2017. URL: https://www.exploit-db.com/exploits/42356.

[58] Steve Platti. *Excessive scanning*. 2021. URL: https://learn.microsoft.com/en-us/answers/questions/472441/excessive-scanning.

[59] *Port Scanning Techniques | Nmap Network Scanning*. URL: https://nmap.org/book/man-port-scanning-techniques.html.

[60] *Remote Desktop Client Remote Code Execution Vulnerability - Microsoft Security Response Center*. 2023. URL: https://msrc.microsoft.com/update-guide/en-US/vulnerability/CVE-2023-24905.

[61] Philipp Richter and Arthur Berger. "Scanning the Scanners: Sensing the Internet from a Massively Distributed Network Telescope". In: *Proceedings of the Internet Measurement Conference*. IMC '19. Amsterdam, Netherlands: Association for Computing Machinery, 2019, pp. 144–157. ISBN: 9781450369480. DOI: 10.1145/3355369.3355595. URL: https://doi.org/10.1145/3355369.3355595.

[62] Anton Röhm. *GoodBots: Updated lists of IP addresses/whitelists of good bots and crawlers*. URL: https://github.com/AnTheMaker/GoodBots.

[63] Albin Sebastian. *Default time to live (TTL) values*. Dec. 2009. URL: https://web.archive.org/web/20150206054041/http://www.binbert.com/blog/2009/12/default-time-to-live-ttl-values/.

[64] Subin Siby. URL: https://subinsb.com/default-device-ttl-values/.

[65] SpeedGuide. *SpeedGuide TCP/IP Ports database*. URL: https://www.speedguide.net/ports.php.

[66] Stuart Staniford, James Hoagland, and Joseph McAlerney. "Practical Automated Detection of Stealthy Portscans." In: *Journal of Computer Security* 10 (Jan. 2002), pp. 105–136. DOI: 10.3233/JCS-2002-101-205. URL: http://pld.cs.luc.edu/courses/intrusion/fall05/hoagland_spade.pdf.

[67] *Stretchoid | Greynoise visualizer*. URL: https://viz.greynoise.io/tags/stretchoid.

[68] *Stretchoid Opt-Out*. URL: https://stretchoid.com/.

[69] *TCP FIN, NULL, and Xmas Scans (-sF, -sN, -sX) | Nmap Network Scanning*. URL: https://nmap.org/book/scan-methods-null-fin-xmas-scan.html.

[70] *The Recyber Project*. URL: https://www.recyber.net/.

[71] *Threatsinkhole*. URL: http://winnti-scanner-victims-will-be-notified.threatsinkhole.com.

[72] *TOR Node List*. URL: https://www.dan.me.uk/tornodes.

[73] Roman Trapickin, Oliver Gasser, and Johannes Naab. "Who Is Scanning the Internet". In: 2015. URL: https://api.semanticscholar.org/CorpusID:55929767.

[74] Daniel Wagner et al. "How to Operate a Meta-Telescope in your Spare Time". In: *Proceedings of the 2023 ACM on Internet Measurement Conference*. IMC '23. <conf-loc>, <city>Montreal QC</city>, <country>Canada</country>, </conf-loc>: Association for Computing Machinery, 2023, pp. 328–343. DOI: 10.1145/3618257.3624831. URL: https://doi.org/10.1145/3618257.3624831.

[75] *Why am I receiving connection attempts from the University of Michigan?* 2024. URL: https://cse.engin.umich.edu/about/resources/connection-attempts/.

[76] *Windows Remote Desktop Protocol Security Feature Bypass - Microsoft Security Response Center*. 2023. URL: https://msrc.microsoft.com/update-guide/en-US/vulnerability/CVE-2023-35332.

[77] Eric Wustrow et al. "Internet background radiation revisited". In: *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*. IMC '10. Melbourne, Australia: Association for Computing Machinery, 2010, pp. 62–74. ISBN: 9781450304832. DOI: 10.1145/1879141.1879149. URL: https://doi.org/10.1145/1879141.1879149.