An Investigation into Collaborative Scanners

Manually detecting and tracking collaborative scanners' behaviour over a prolonged period

Matyáš Kollert[1]

Supervisor(s): Georgios Smaragdakis[1]**, Harm Griffioen**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Matyáš Kollert
Final project course: CSE3000 Research Project
Thesis committee: Georgios Smaragdakis, Harm Griffioen, Kubilay Atasu

Abstract

Port scanning is a technique often used by adversaries to detect vulnerable services running on a machine. There are defense mechanisms in place that can detect fast, single-source port scanning, but one of the ways to remain hidden is to distribute the scan between multiple hosts. These distributed groups of machines can divide the address space and collaboratively scan the whole Internet within minutes and remain relatively hidden.

This paper proposes a simple method to detect these collaborative scanners based on the TCP/IP header and demonstrates its efficiency. It also tracks these scanners for a longer period and describes their behavior and how they develop over time. This includes the infrastructure they utilize, the specific ports they target, and additional relevant details. This perspective has not been previously explored in the academic literature and we find it to be important such that defenders get a better understanding of the threats they are facing.

1 Introduction

Making a service publicly available on the Internet requires the device running it to be connected to a public IP address. Once the device connects, the port that it is running on will start receiving unsolicited internet traffic almost immediately [6]. This traffic includes backscatter from Distributed Denial of Service (DDoS) attacks and, more alarmingly, port scanning activities.

Port scanning is a technique employed by individuals or groups to determine if services are running on a machine. This is achieved by sending packets to designated IP addresses and awaiting their responses. If the machine responds, adversaries will know that there is a service running on it and might either take note of it or even try to exploit it. When this activity extends across the entire IPv4 address space, it is known as internet-wide scanning [2]. Tools like ZMap [3] and Masscan [7] have been developed for this specific purpose. They allow interested parties to probe the whole internet in a matter of hours or less. Conducting these scans effectively requires either a powerful single machine with a high-speed internet connection or a network of coordinated machines.

Collaborative scanning involves multiple machines working together to scan large address spaces more efficiently. This method leverages the combined processing power and network bandwidth of several devices, enabling faster and stealthier scanning operations. By distributing the scanning tasks among multiple nodes, collaborative scanners can evade detection mechanisms that might otherwise block or rate-limit a single IP address. Investigating the methodologies of collaborative scanning is essential to develop more effective detection techniques, ensuring network security in the face of increasingly sophisticated scanning operations.

While port scanning can often be used for legitimate security purposes, it can also be exploited for malicious activities. According to various papers [10; 8; 2], attackers use port scanning to identify vulnerabilities in specific applications on a host. Once a vulnerability is discovered, it can be exploited to infect the host with malware, which in turn can spread to other systems. This can lead to data breaches, the deployment of ransomware, or becoming a part of a botnet used for further nefarious activities.

In this paper, we aim to manually identify collaborative internet-wide scanners using network telescope data, and describe the patterns of their behavior over time. First, we determine, how well can the /24 sub-net, Autonomous System (AS) and temporal patterns in the groups' probing traffic identify collaborative scanners. Second, we investigate what trends or changes in the behavior of collaborative scanners can be observed over an extended period of almost a year.

This paper makes two main contributions:

- It combines the proposed method of grouping hosts by their /24 sub-net with also using their AS and request timing to detect collaborative scanners and evaluate how effectively this relatively simple approach works.
- It describes the behavior of said scanners over an extended period. More specifically, it details how they choose the ports that they scan, how many IP addresses they use for the scans, and the changes in their methodologies over time.

1.1 Related Work

One of the methods to detect coordinated scanners was proposed by Gates [5], where she used the Set Cover algorithm to identify a set of source IP addresses that scan the entire IP address space together. While the Set Cover algorithm is theoretically sound, it does not scale well with the amount of data points. As the dataset size increases, the computational time increases exponentially since it belongs to the NP-Hard class. This makes it challenging to apply the Set Cover algorithm to large-scale networks or environments with high traffic volumes, requiring more scalable detection methods.

Staniford et al. [11] propose a machine learning approach that aims to detect stealthy scanners that would not be detected by traditional intrusion detection systems (IDS). The algorithm creates clusters based on the likelihood of a packet being an anomaly. This is quite effective, but it needs the traffic to be marked as anomalous and it was mostly designed for single host scans, leaving the distributed case as future work.

An approach based on fingerprinting the TCP/IP headers was proposed by several other studies [4; 8]. These fingerprints capture unique characteristics of network packets, enabling the identification of suspicious patterns and behaviors in the traffic. By analyzing these fingerprints, different types of scanning patterns can be detected and categorized, enhancing the accuracy of intrusion detection systems.

One of the simpler methods assumes that an entire scan originates from the same /24 sub-net was outlined in two other papers [10; 13]. This was shown to be effective but may not hold for large, sophisticated attackers who could distribute their scanning activities across multiple sub-nets to stay undetected. Additionally, the approach in [13] does not consider scanners consisting of fewer than 5 IP addresses which could result in missing smaller but significant scanners.

Therefore, while these methods provide valuable insights into detecting coordinated scans, they also highlight the need for more flexible and comprehensive detection strategies that account for diverse scanning behaviors and patterns.

These papers propose various methods for detecting collaborative scanners, which were found to be effective despite their limitations in their assumptions or complexity. However, they do not focus on the long-term behavior and development of these scanners which could be important for several reasons. First, it gives us insight into the capabilities and strategies of adversaries, revealing what their intentions might be. Second, it helps us to predict their behavior in the future, allowing us to be proactive in creating new detection techniques. Not being able to predict the trends might leave current IDS outdated and unprepared, making networks far more vulnerable to more sophisticated attacks.

2 Methodology

This section will describe the proposed methodology used to detect collaborative scanners and track their behavior over time. The stepts taken will follow Figure 1. Firstly, it will describe how the data were collected and how to filter out unrelated data. Secondly, it will define a collaborative scan and a collaborative group. Lastly, it will describe how the groups will be tracked over an extended period to inspect their behavior.

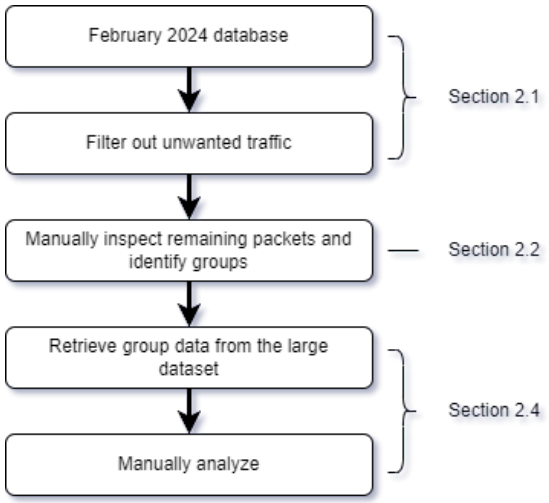


Figure 1: Research methodology divided into 5 distinct steps

2.1 Dataset

To perform this analysis, we will be using a database of TCP/IP headers from traffic collected by a Network Telescope provided by TU Delft. This telescope consists of over 60,000 IP addresses from multiple /24 sub-nets. These addresses are completely passive so they do not send out any packets themselves, only listening to incoming internet traffic. Even though this telescope cannot collect incoming traffic from the whole IPv4 address space, since it is of limited size, it is still excellent for detecting internet-wide scans.

In addition to the TCP/IP headers, the database also stores more information about each source IP address. This includes the AS, Country, City, among others. Since the location of an IP address can change over time, the location was checked on each day that the IP address appeared in the dataset. This ought to make sure that the information is correct and the scanners are identified accurately.

The telescope collects all TCP, IPD, ICMP, and other unknown traffic. This experiment will be focusing on TCP traffic since it makes up over 87% of the dataset and is therefore the most often used protocol. To correctly detect a probe, only headers with the TCP SYN flag are considered [8]. More specifically, it will focus traffic that most likely originated from ZMap as it is the most common tool identified from the data. This is possible because unmodified ZMap sets the Identification field of the IP header to 54321 [8]. Using ZMap packets also filters out unwanted traffic caused by DDoS backscatter and others. After applying these conditions, 4.5 billion packets were left from the original 12.6 billion TCP packets. Traffic from UDP and other scanning tools was also not considered since it would not be feasible for the scope of this research.

2.2 Defining a Collaborative Scanner

Collaborative scanners operate by coordinating multiple IP addresses to probe the internet together. Unlike single-source scanners, which originate from a single IP address, collaborative scanners distribute their requests among various IP addresses to possibly speed up each scan, avoid detection, and circumvent rate-limiting mechanisms. They might choose to start scanning from all of the hosts at the same time, one after the other, or in other patterns based on their goals. The speed-up comes from each host only needing to probe probe a portion of the IPv4 address space.

Probe packets cannot be easily distinguished from normal traffic which is the reason why they cannot be labeled as such. This presents a difficulty where we can never be certain that a set of packets forms a coordinated group. Therefore, a definition of it needs to be created in a manner that is identifiable from the data. A collaborative scan is a set of requests where the following conditions are satisfied:

1. **Same port:** Every request targets the same destination port. We only consider such scans because otherwise they would not have an internet-wide view of the ports, but would only know about a portion of it.
2. **Bursts:** The requests are sent in bursts that are shorter than 24 hours during a single calendar day. This is so that we can determine when the scan is over and correctly evaluate the data. A single calendar day was used because having more complex time intervals would not be feasible in the amount of time given to this research.
3. **Collaboration:** All were sent by a set of two or more IP addresses where none of them scanned the whole telescope by themselves. This behavior would make them a standalone actor and therefore not relevant to the research questions.
4. **Almost exact cover:** Together they reach all the IP addresses in the telescope N times where N is a natural

number. The thresholds that specify the telescope size are discussed in Section 2.3.

5. **Location:** All Source IP addresses belong to the same /24 sub-net or under the same Autonomous System. More complex IP locations are outside of the scope of this project.
6. **Temporal Patterns:** All Source IP addresses sent the traffic in a clear pattern such as in parallel or sequentially one after another.

Once all scans are identified, they can be aggregated by certain patterns:

- **Same IP addresses:** If multiple scans share the same IP addresses, we can consider them as a single group that targets multiple ports, noting down the ports.
- **Similar request timing:** If multiple scans send out requests in a similar temporal pattern, they could also be considered as a single collaborative scanner.
- **Telescope partitioning:** Lastly, if multiple scans split the telescope IP space in the same manner e.g. each member scans 1/M IP addresses where M is the size of the group, they could be considered to belong together.

The more of these patterns that multiple groups share, the more certain we can be of them acting together. This allows for investigating more complex behavior which would not be possible otherwise.

Most of these points need to be inspected visually as otherwise it would require complex algorithms such as the Set/Exact Cover which is not a part of this approach. This also means that the resulting scans/groups might differ slightly between different researchers.

2.3 Network Telescope Size

As discussed above, we need to know if a scanner targets the whole telescope. This is not so simple since the amount of IP addresses changes. For most of the days during February, the amount of IP addresses that received at least one packet was 61 000 +- 2.3%. On the 6th and 8th of February, there was an anomaly, where for a short period, the amount of IP addresses increased to 172 111 and 101 786 respectively but otherwise was also in the same range. For this reason, the minimal threshold was set at 60 000 IP addresses while the maximal threshold was set at 62 500. Afterward, these thresholds were used on some known scanners [9] to see whether they would be detected each day, which proved to be the case.

2.4 Tracking

Once all the collaborative scanners are identified, we will use the whole year dataset to track their behavior over time. This dataset goes from the 14th of April 2023 until the 15th of February 2024 which means that each group will be tracked backwards in time. For each group, multiple characteristics will be considered:

- **Longevity:** How long do the groups stay active for? This could indicate whether one can expect them to be scanning in the future or not. It is also telling in regards to how long the IP addresses are reused.
- **Group Timing:** Are the groups active every day/week/month? Do they prefer a certain day of the week or the month? Do they scan more near the end of the year or the beginning of a new one?
- **Ports:** What ports do the groups target? How are these ports chosen? Do they prioritize certain ports over others? Do they scan a port only once or multiple times?
- **Destination IP split:** Does each IP address always cover the same portion of the telescope? Do the IP addresses rotate the amount of IP addresses that they scan? Does each IP address scan its fair share or are some scanning more than others?
- **Scan Timing:** Do the groups scan one after another? Do they scan in parallel? Does the pattern of scanning change over time? Are there multiple patterns that repeat?
- **Size:** Do the groups always scan with the same amount of IP addresses? Is the size strictly increasing, decreasing, or oscillating between multiple values?

3 Experimental Setup and Results

This section will first describe how we used the proposed methodology to find collaborative scanners within the data of February 2024. Secondly, it will present the findings regarding the behavior and development of these scanners from March 2023 to February 2024.

3.1 February Database

To find collaborative scanners, a database of all the traffic collected by the network telescope during February 2024 was used. As mentioned in Section 2.1, we filtered out protocols other than TCP and made sure that backscatter and the Mirai botnet were not included.

The main database table stores the TCP/IP header fields, out of which we will be focusing on the source IP (SrcIP), destination IP (DstIP), destination port (DstPort), and Timestamp of each request. In addition, this table will be joined with another, which stores the AS for each SrcIP address for each day that the IP address appears in the main table. This allows us to confidently assign an AS to each SrcIP address.

To find candidates for the collaborative scanners, two experiments were performed. One where the data points are grouped by the DstPort, day, and /24, and another where they are grouped by the DstPort, Day, and the AS. These experiments expand on the method proposed in [13] allowing us to determine whether the method works and if the search space can be extended by also looking at the AS compared to only the /24 sub-net.

For each experiment, all the possible collaborative scans were listed and then manually checked each /24 sub-net or AS separately to ensure that there were no errors. This included visually checking the data and creating simple functions that helped us with the identification and validation.

After performing these two experiments, we observed, that there was not a single scan that was detected by the /24 sub-net but not by the AS. This meant that each detected sub-net was always contained within one AS. This observation made

us focus solely on the AS since it would remove possible duplicates and not lose any scans. There were also many scanners identified by the AS that were not identified by the /24 sub-net.

3.2 Significant groups

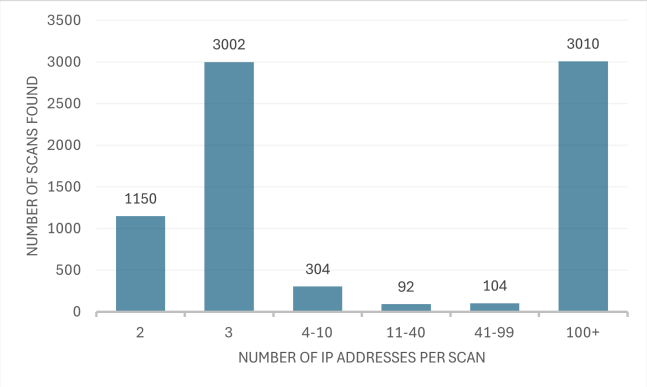


Figure 2: Distribution of scans based on the amount of IP addresses belonging to them

During February 2024, we have identified 7662 collaborative scans. Almost 40% of them were performed by 100+ IP addresses all belonging to the HURRICANE AS. Those that consisted of 3 IP addresses were the second most observed with also just under 40%. Most of these scans originated from DIGITALOCEAN-ASN. The third most common were sets of 2 IP addresses which account for 15% of the observations. A large portion of these came from Akamai Connected Cloud AS. The rest of them, which were formed by 4 to 100 IP addresses, account for only just above 6.5%. The whole distribution can be seen in Figure 2. Different sized were aggregated to make the figure more readable. These scans account for 17.5% of the whole dataset.

As can be seen, a vast majority of the scans originated from an AS, belonging to hosting service providers. The most common being Hurricane, DigitalOcean, and Akamai. Fewer scans came from those belonging to Amazon and CARInet. All of these companies offer access to their virtual or physical servers for a monthly fee, allowing the occupant to use the server according to their terms and conditions. A set of scans came from the security company SecurityTrails which are open about their scanning activity. A considerable portion of the scans belonging to the 11-40 IP category were those coming from GEMNET LLC which is the primary backbone Mongolian Internet provider.

These could be grouped into 41 groups where each scan in the group shared at least two of the patterns discussed in Section 2.2.

The following list describes groups that were observed in both February of 2024 and also at least one other month using the larger dataset. These were chosen since there is enough data to describe their long-term behavior to some extent. The rest of the groups are described in less detail below this list. Each group is given its name by the Autonomous System that they originated and a unique identifier is added to those where further distinction was necessary.

- **AKAMAI-01 (AK01):** This group makes up almost 99% (1120 scans) of all scans originating from Akamai Connected Cloud AS. Each of their scans came from two IP addresses where one covered 5/6^{ths} and the other covered 1/6th of the telescope. This pattern repeats throughout the whole month, but the same two IP addresses seldom scan together more than once. This means that a set of 133 IP addresses was found, where each scan is performed by a different permutation of size two.
- **DIGITALOCEAN-MAIN (DOM):** Over 60% (3002 scans) from DIGITALOCEAN-ASN came from three IP addresses where one of them scanned exactly 39476 DstIPs, another scanned exactly 11998 DstIPs and the third scanned the rest of the telescope. This pattern also repeats for the whole month but unlike **AK01**, two scans were never performed by the same set of IP addresses. We found 2014 IP addresses, where each scan was performed by a different permutation of size three.
- **DIGITALOCEAN-01 (DO01):** Small group coming from the DIGITALOCEAN-ASN with a request timing pattern very similar to that of **HUR**. It can be characterized by 10 distinct bursts interleaved with 9 periods of rest. Each scan was performed by 8 or 9 IP addresses that did not show any clear split of the address space.
- **DIGITALOCEAN-25 (DO25):** One of the most interesting groups scans from the DIGITALOCEAN-ASN. This group consists of up to 72 IP addresses but not all of them are used at the same time. The main observed pattern is the equal split between all the hosts scanning at the time. Notably, this group scans in two distinct temporal patterns. No matter which of these they use, they always cover the whole telescope and end the scan at a similar time. These patterns can be seen in Figure 3 where each graph shows how each IP address that was observed sends the packets. Figure 3a reveals 60 IP addresses where each probes in tiny bursts and they all do so in parallel. Figure 3b is quite similar in terms of the amount of IP addresses and the parallelism but each host probes continuously until they are done.
- **AS62904 (AS6):** A group of 7 IP addresses scanned the whole telescope on the 29th of February 2024. Each host covered 1/7th of the telescope and they all started and ended the scan at almost the same time. All the IP addresses belonged to the AS62904 AS.
- **HURRICANE (HUR):** The largest collaborative scans came from the Hurricane AS and all belong to a single group. The size of the group varied but always exceeded 150 IP addresses. Each of the IP addresses covered 1/N IP addresses where N was the size of the group. The request timing was also very indicative of many scans belonging to the same group. Figure 4 shows 3 distinct timing patterns for the group taken as a whole. The patterns in Figures 4a and 4b have distinct periods of probing and periods of rest. The pattern in 4c is slightly different as there are more bursts towards the end.

- **GEMNET (GEM):** The largest group that does not come from a hosting provider. It consists of 37 IP addresses belonging to the Mongolian AS GEMNET LLC. Every time they perform a scan, all the IP addresses probe at the same time, and each covers $1/37^{th}$ of the telescope. Together they always create an exact cover.
- **CARINET-01 (CR01):** The first group of two IP addresses from the CARINET AS. One of the few where the hosts scan one after another, not at the same time.
- **CARINET-02 (CR02):** The second group consists of two IP addresses each from the CARINET AS which scans the entire telescope together. The timing of every scan seems random, but they split the IP addresses in two distinct patterns with the approximate ratios being 5:1 and 29:23.
- **SECURITYTRAILS (STS):** Two sets of nine consecutive IP addresses all originating from the SecurityTrails LLC AS. All the hosts probe the telescope perfectly in parallel and each takes $1/9^{th}$ of the telescope IP addresses.

These groups accounted for over 99% of all the scans. In contrast, twenty groups from DIGITALOCEAN-AS were identified, each responsible for only one scan. None of these groups shared even two of the three criteria that needed to be grouped together. Additionally, we discovered three other groups from the same provider, each conducting between two to three scans. Two groups from Akamai Limited Cloud also appeared on a single day and were never seen again. Similarly, one group each from AMAZON-02 and AMAZON-AES conducted three scans each and then did not reappear. Lastly, there was a single group in the PONYNET AS which targeted port 433 on two different days during February but did not share any other patterns.

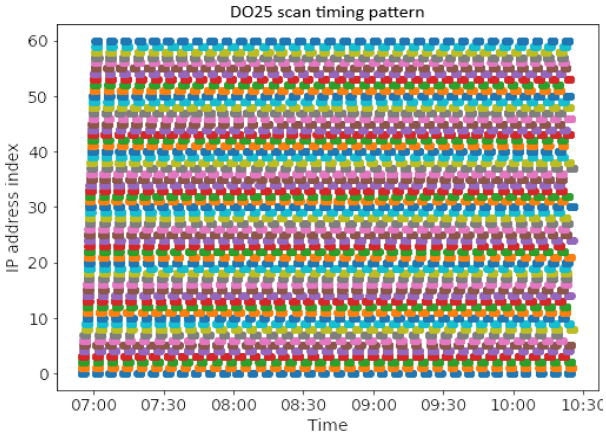
As there is very limited data and they were not detected in any other months, we were not able to make any conclusions about these groups. They were investigated with as much scrutiny as others, but the lack of data is why they are not mentioned further.

3.3 Tracing the groups for a longer period

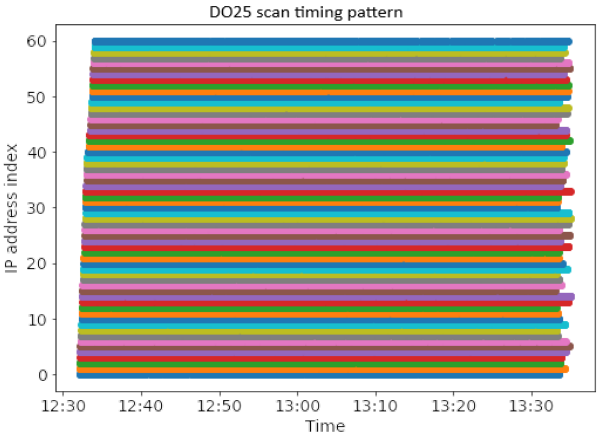
From the 41 groups found, the set of source IP addresses that we then used to filter the whole year dataset was extracted. Only the SrcIP, DstIP, DstPort, and Timestamp of each packet were considered to do further research. For each group, we filtered the data points belonging to it and inspected them on their own. Those that were identified on at least 4 different occasions, out of which at least one was outside of February 2024, are described in the following paragraphs.

AK01 kept increasing their daily scans slightly over the whole period as can be seen in Figure 5. Each of their scans was performed by two IP addresses as described above. They mostly targeted ports 8080, 8081, 7443, 1194, 8443, 22, 2222, 50050, 442, and 80 which together accounted for 2750 (30%) out of all the 8933 scans.

DOM was identified throughout the whole dataset with light activity averaging 20 scans per day until 20-07-2023 where a sharp increase to over 250 scans per day can be seen



(a) Per IP address timing of packets with gaps between short bursts



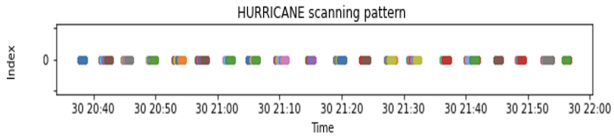
(b) Per IP address timing of packets with no gaps

Figure 3: Two repeating patterns used by the DO25 collaborative scanner

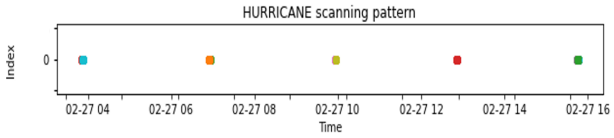
in Figure 6. Other temporal patterns were quite neutral with no clear preferences. As for the ports, out of 54042 performed scans, they focused on ports 443, 9200, and 8443 with 1570, 1112, and 1039 scans respectively while the other ports were scanned at most 532 times with their average being just over 75.

DO01 did not show any clear patterns. Altogether, they scanned 17 times, and each time they chose a different port. Since there were so few scans, we were not able to determine how they chose the targeted ports or even the days that they scanned on. The first scan occurred on 11-12-2023 and since then they were active quite sporadically until the end of the period.

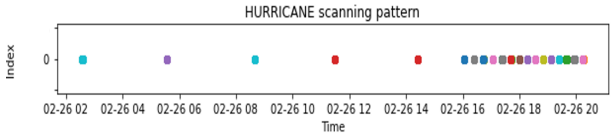
DO25 did not show many clear patterns in their 1661 scans either. While they sometimes went up to 10 days without appearing, overall their distribution over the week remained quite constant with a slight preference for Monday, Thursday, and Saturday. As for the monthly distribution, days 4, 9, 23, and 31 were the least used, averaging under 30 scans per day. And days 8, 21, and 24 were the most scanned averaging over



(a) Scanning pattern with 20 equally spaced bursts



(b) Scanning pattern with 5 equally spaced bursts



(c) Scanning pattern with lots of bursts near the end

Figure 4: Packet timing of scans origination from the HUR collaborative scanner

75 scans per day.

AS6 was identified 13 times. Out of these 13 scans, 5 of them focused on port 20256 which has no clear assignment. Three scans targeted port 10001 and the rest focused on ports 443, 9092, 3002, 59876, and 1801. Most of their activity was during December 2023 with two of them going as far back as 14-02-2023. As for the group size, 5 scans were performed by 7 IP addresses and the other 7 by 6 IP addresses. There was no clear reason for this difference and the telescope coverage did not change.

HUR was the most active collaborative scanner of them all. They had no clear preferences for the day of the week or month apart from a slight dip in scans on the 31st of each month which can be attributed to half the months not having this day. Far more interesting was the sharp increase in scanning activity after 15-12-2023 where it increased seven-fold. This can be seen in Figure 7. There is also a high variance in the amount of IP addresses performing the scans. As can be seen in Figure 8, there are two clear peaks at 205 and 327.

GEM was the least consistent large group. While every one of their scans was performed by 37 IP addresses perfectly in parallel, they did not scan every day. There were no scans detected between 21-04-2023 and 16-04-2023 which was very unusual. They also did not scan evenly throughout the days of the week or month as can be seen in Figure 9.

CR01 was seen throughout the whole dataset. Their size remained constant for every scan they performed. As the only larger group, they had strong preferences for the days that they observed. As can be seen in Figure 10, most of their scans happen on Monday to Thursday with very few being on Friday and Saturday and none of them on Sunday. There are no other clear timing patterns.

CR02 appeared only 4 times. Port 989 on 19-04-2023, port 8983 on 12-10-2023, and ports 587 and 623 on 06-02-2024. For each, the two IP addresses scanned sequentially, and one scanned at least twice as much as the other as can be seen in Figure 11. This was its main distinguishing feature since no other groups scanned in this manner.

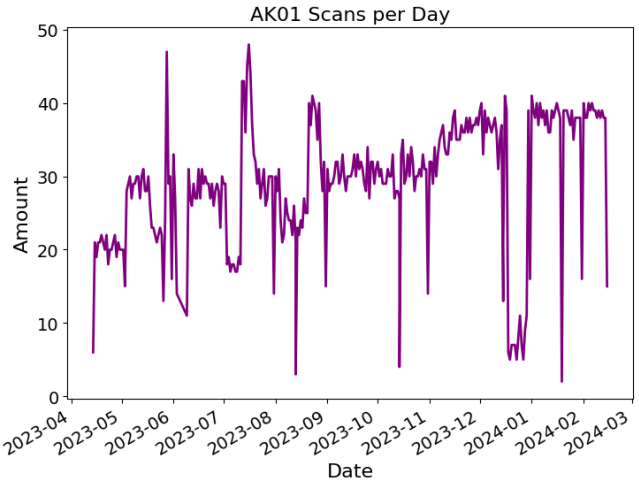


Figure 5: A graph showing the per day amount of scans for the AK01 collaborative scanner

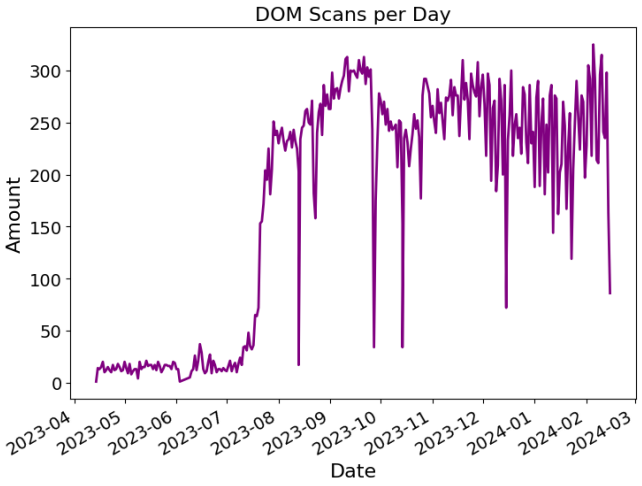


Figure 6: A graph showing the per day amount of scans for the DOM collaborative scanner

STS was seen between one and five times each day. Their packet timings remained the same each time and on the days where they scanned more than one port, it was done sequentially one right after the other with each scan taking approximately 3.5 hours. They had a clear preference for ports 5357, 3389, and 80 as they were scanned at least four times as often as any other port.

4 Responsible Research

This section outlines the responsible practices employed during this research. The Netherlands Code of Conduct for Research Integrity[1] served as a guiding framework, where our primary focus was on adhering to its principles and standards.

[1]https://www.nwo.nl/en/netherlands-code-conduct-research-integrity

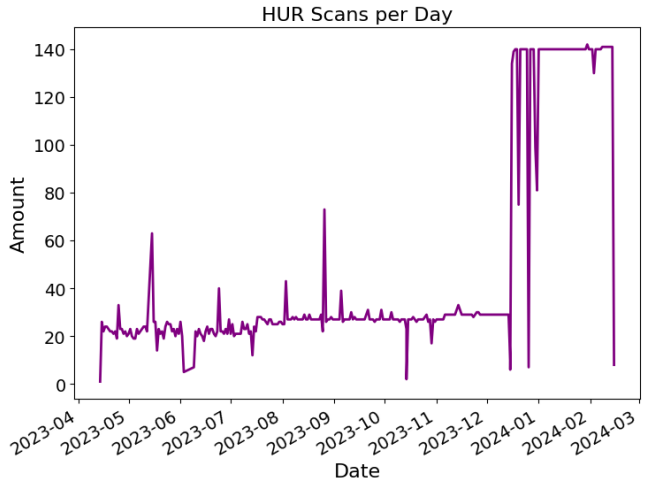


Figure 7: A graph showing the per day amount of scans for the HUR collaborative scanner

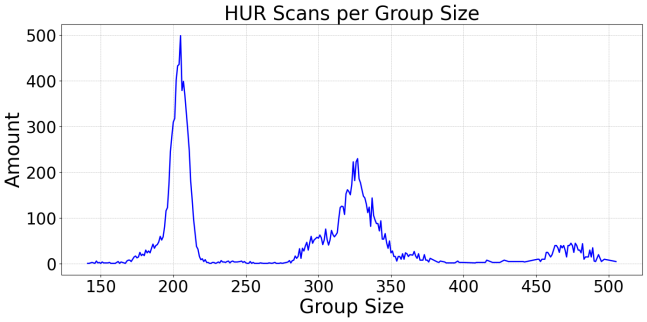
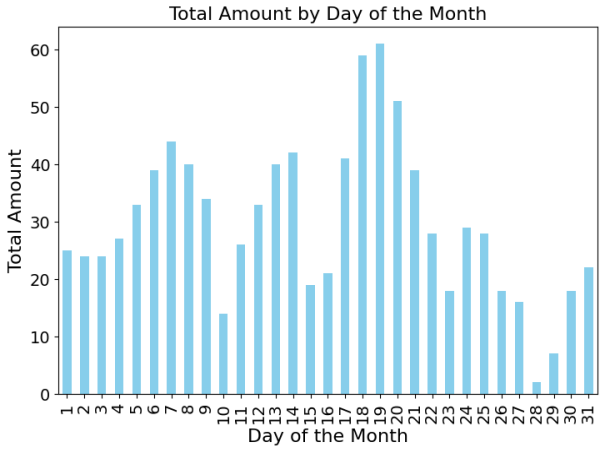
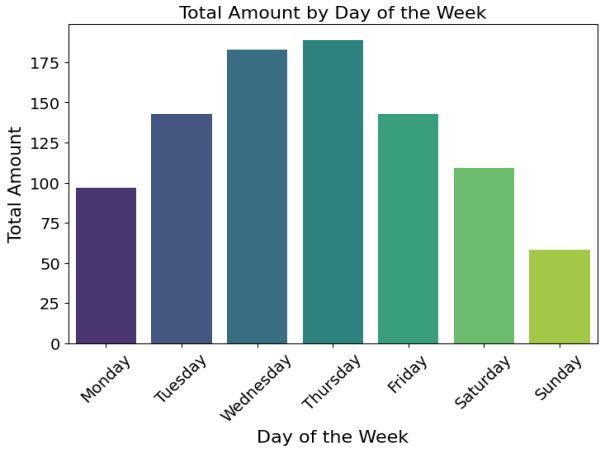


Figure 8: A graph showing the frequency of each observed group size of the HUR collaborative scanner



(a) Scan distribution per the day of the month



(b) Scan distribution per the day of the week

Figure 9: Long-term temporal patterns of the GEM collaborative scanner

4.1 Data Privacy

The data used in this research was provided solely by a Network Telescope, where all IP addresses were passive. This means that they were not in use and were not sending any traffic. Therefore, none of the captured traffic contained personal information and all of it is anonymous. Since the data captured are only the TCP/IP header of packets that came from the Internet, there is no personal data to be stored. In order to minimize the possibility of incorrectly identifying sets of IP addresses as malicious, we did not publish the specific IP addresses belonging to each group as the risk outweighed the benefits.

4.2 Unintended consequences

While this research is intended to help defenders better understand the possible threats, it could also provide an attacker with an in-depth understanding regarding detection techniques in use and how to bypass them. The sharing of details of knowledge on defense strategies, vulnerabilities, and mechanisms of detection may, in turn, inadvertently equip malicious actors with the tools to develop highly sophisticated methods of attack, these being able to stay under the radar of traditional detection systems. This dual-use dilemma is a critical ethical consideration in cybersecurity research, necessitating a careful balance between transparency for the sake of progress and caution to prevent potential abuse.

4.3 Reproducibility

We tried to describe the criteria for defining a collaborative scanner in as much detail as possible such that an independent researcher can reproduce the same results if they had access to the same dataset. Since visual inspection was used, it could be argued that a pattern that one person sees might not be recognized by another which might create small diversions from what was reported here. The dataset is not publicly available but could be obtained if needed. If the dataset was different, one should still be able to find similar patterns but not exactly the same ones.

5 Discussion

Based on the methodology, we have found 7662 scans which were then aggregated into 41 groups. Out of these groups, some were visible throughout the whole period while some

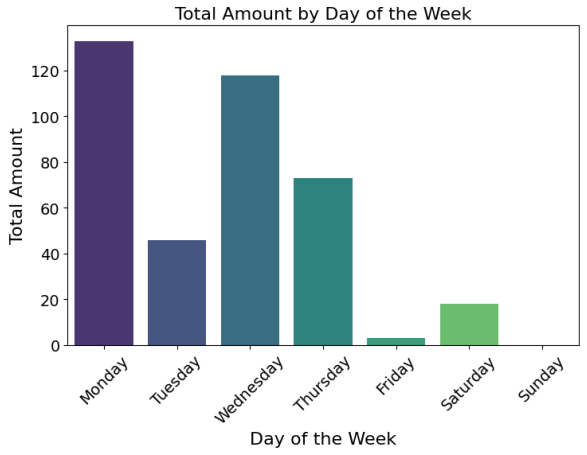


Figure 10: Week day distribution of scans performed by the CR01 collaborative scanner

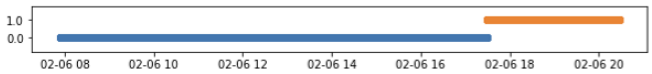


Figure 11: Per IP address timing of packets origination from the CR02 collaborative scanner

disappeared after one or two scans. In the long-term groups, we have managed to find some patterns regarding their targets and more.

The scans belonging to **HUR** can be contributed to the Shadowserver Foundation. This can be verified by inputting any of the IP addresses into the browser. Each of these IP addresses serves a website that describes the purpose of the scans. The company claims to be using the data collected to make the Internet safer for everyone but one can never be sure if everyone who has access to the data shares their goals.

As for the groups originating from hosting services, DigitalOcean states on their website [1] that their services are not to be used for "Accessing or using any System without permission, including attempting to probe, scan, or test the vulnerability of a System or to breach any security or authentication measures used by a System." As can be seen from groups **DOM** and **DO25**, they have not stopped these IP addresses from scanning even after several months, even though their IP addresses are publicly reported [9]. Similarly, as stated by Akamai Technologies, "attempting to probe, scan or test the vulnerability of a system or network" is prohibited [12] but they also have not stopped the **AK01** group even though they reuse their IP addresses regularly.

We were not able to find any public information about the groups apart from SecurityTrails and the Shadowserver Foundation. This might suggest that they want to stay anonymous and are likely not scanning for the benefit of the public. With their possible motivation being the exploitation of certain services or even just selling the gathered information to others, we should remain cautious.

The situation is less clear for the groups that were not described in detail. As they scanned only between 1 and 4 ports, there is not much information about their motivations. The first option is that they accomplished their goals with so few scans and had no reason to keep doing it. Another possibility is that they are too sophisticated to be tracked by our proposed approach. Either way, it is important to keep track even of these collaborative scanners as they might be better identified in the future.

6 Conclusions

In this paper, we propose an approach for detecting collaborative scanners based on information in the packet header and other publicly known information. This simple method was shown to be effective at detecting scanners not complex enough to span multiple Autonomous Systems. Having identified 41 of them, we also conducted a thorough examination for the purpose of allowing quicker detection in the future. We describe the methodology these groups use to target specific ports, their scanning patterns, and how long they are scanning for. We also note the size of the groups and how that develops over time.

This method successfully detects and tracks known scanners such as SecurityTrails or the Shadowserver Foundation but also unknown scanners that have not been publicly described before such as sophisticated groups using DigitalOcean or Akamai. It also managed to show certain patterns regarding the approach that these groups follow, be it the timing of their scans, the ports that they target most often, or even their size.

Many of the limitations of this project could be addressed given more time. These points should be explored more in-depth in the future:

- Expanding the domain to traffic not generated only by ZMap but also by other known tools such as Masscan and others. Since these tools are well known, others might try to create custom tools or modify these existing ones. By removing this limitation, we can improve our ability to identify an increasing number of these tools.
- Expanding the domain by looking at groups spanning multiple /24 sub-nets and multiple Autonomous Systems. It could be argued that the most sophisticated groups know these simple detection methods, therefore spreading their hosts as much as possible.
- Not limiting a scan to a single calendar day since there is no reason to believe that these groups do not scan across multiple days. Some of these scanners might be in different time zones which also does not work well with this model. Some actors that value their stealthiness above all else could try to slow down their scans as much as possible possibly resulting in one scan taking more than 24 hours as well.
- Each group could also be tracked more thoroughly, for example by replying to the packets that were received and waiting for more incoming traffic. This might show us whether they are only scanning the port or also trying to gain access to a service running on it.

References

[1] LLC DigitalOcean. Digitalocean acceptable use policy. https://www.digitalocean.com/legal/acceptable-use-policy#security-violations. Accessed: 2024-06-01.

[2] Zakir Durumeric, Michael Bailey, and J. Alex Halderman. An Internet-Wide view of Internet-Wide scanning. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 65–78, San Diego, CA, August 2014. USENIX Association.

[3] Zakir Durumeric, Eric Wustrow, and J. Alex Halderman. ZMap: Fast internet-wide scanning and its security applications. In *22nd USENIX Security Symposium (USENIX Security 13)*, pages 605–620, Washington, D.C., August 2013. USENIX Association.

[4] Carrie. Gates. Co-ordinated port scans: A model, a detector and an evaluation methodology., Jan 1970.

[5] Carrie Gates. Coordinated scan detection. 01 2009.

[6] 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)*, pages 1–5, 2016.

[7] Robert Graham. Masscan.

[8] 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*, pages 1–9, 2020.

[9] AbuseIPDB LLC. Abuseipdb official website. https://www.abuseipdb.com/about.html. Accessed: 2024-06-01.

[10] S. Robertson, E.V. Siegel, M. Miller, and S.J. Stolfo. Surveillance detection in high bandwidth environments. In *Proceedings DARPA Information Survivability Conference and Exposition*, volume 1, pages 130–138 vol.1, 2003.

[11] Stuart Staniford, James Hoagland, and Joseph McAlerney. Practical automated detection of stealthy portscans. *Journal of Computer Security*, 10:105–136, 01 2002.

[12] Akamai Technologies. Akamai acceptable use policy. https://www.akamai.com/legal/acceptable-use-policy. Accessed: 2024-06-01.

[13] Vinod Yegneswaran, Paul Barford, and Johannes Ullrich. Internet intrusions: Global characteristics and prevalence. 31, 05 2003.