I Know Who You Scanned Last Summer: Mapping the Landscape of Internet-Wide Scanners

Julian Mayer*, Markus Schramm*, Lukas Bechtel*, Nils Lohmiller*,
Sabrina Kaniewski*, Michael Menth‡, Tobias Heer*
*Esslingen University of Applied Sciences, Germany, ‡University of Tübingen, Germany

***Abstract*—In today's Internet, a plethora of services are, intentionally or unintentionally, publicly accessible. Internet-wide port scanning is a widely used method to discover such services. Malicious actors as well as researchers use port scans to detect and analyze exposed Internet services. Vulnerabilities in Internet of Things (IoT) devices and their associated services have been a target of widespread malicious activity involving botnets, such as the Mirai botnet. As a result, IoT protocols are of particular relevance to vulnerability scanning and Internet security. This development raises the question of who is scanning the Internet, particularly IoT ports, and their intentions. In this work, we analyze the traffic received over the period of 8 months on 6 honeypots located across different continents. On each honeypot, we provide selected IoT services, namely AMQP, CoAP, MQTT, and OPC UA. We identify scanners that scan these honeypots and categorize their scanning behavior, e.g., the scanning range and frequency. We identified 43 organizations responsible for 30.3 % of overall traffic, of which 21 scan at least one of the selected IoT services. We further analyze the scanning tools used. We identified traffic generated by scanning tools such as Masscan, ZMap, and the Mirai botnet.**

***Index Terms*—Honeypot, IoT, Port Scan, Scanning Tools**

I. Introduction

The Internet of Things (IoT) is becoming increasingly prevalent. Today, many devices are connected to the Internet. Specifically in the manufacturing industry, with concepts such as Industry 4.0 and Industrial IoT (IIoT), Internet connectivity is increasing, resulting in a large number of connected devices and publicly accessible services. This public accessibility attracts different user groups on the Internet, i.e., actors with both benign and malicious intentions, especially in the case of vulnerable services. For example, during observations, we logged traffic from the Mirai botnet [1], which exploits vulnerabilities in IoT devices to launch orchestrated DDoS attacks.

The goal of this work is to identify scanners that scan the Internet and categorize their form of interaction with different services. We define a scanner as a host that sends packets to one of the honeypots, for example, as part of an Internet-wide port scan. To analyze scanning traffic, we deployed 6 honeypots using AWS-hosted instances on different continents. We have operated these instances since Summer 2023. For this work, we use data collected from Jun. 1 to Jan. 31, 2024. The honeypots provide interactive services for the IoT protocols AMQP, CoAP, MQTT, and OPC UA, as well as a logging system for incoming traffic on all TCP/UDP ports. Using this architecture, we identify scanners and their tools used to conduct scans. We further analyze their scanning behavior for different services, e.g., scanning range and frequency, and present criteria to categorize them. Within these observations, we searched for scanners with benign intents, such as the well-known search engines Censys and Shodan, and scanners with malicious intents, e.g., botnets. We aim to provide a better understanding of scanning behavior on the Internet for researchers and service operators of Internet-connected IoT services. Therefore, we open-source a sample dataset and code for the analysis pipeline and evaluation[1].

The remainder of this paper is structured as follows. In Section II, we present related work on the analysis of scanners and scanning behavior. In Section III, we present the honeypot and analysis infrastructure. We describe the process we established for identifying scanners and how we associate packets with specific scanning tools and botnets in Section IV. In Section V, we present identified organizations and categorize their scanning behavior, focusing on organizations that scan the selected IoT services. We conclude this work in Section VI.

II. Related Work

Internet-wide scanning is a broad research area with different aspects to focus on. We aim to identify Internet-wide scanners and provide an in-depth analysis of their scanning behavior. In this section, we present related work and discuss how this work differs from previous studies.

The landscape of Internet-wide scanning is constantly evolving, necessitating regular studies to map the current picture. Many studies cover the subject of identifying scanners and analyzing their behavior, utilizing traffic data captured by, e.g., network telescopes, or honeypots. Durumeric et al. [2] presented one of the first studies on Internet-wide scanning, analyzing scanners performing large horizontal scans, protocols targeted, and scanning tools used. Mazel et al. [3] and Heo et al. [4] studied scanning trends, classifying scanners' behavior based on spatial and temporal structure and horizontal and vertical spread, respectively. Richter et al. [5] studied differences in scanning characteristics of Internet-wide and localized scans. Hiesgen et al. [6] focused on two-phase scanners, discussing scanning patterns, targeted services, and

We thank *IPinfo* who support our research by IP lookups free of charge.

ISBN 978-3-903176-63-8 © 2024 IFIP

[1]Our Code: https://github.com/hs-esslingen-it-security/hses-honeypot

scanning origins. Several studies [7]–[12] utilized honeypots for their analyses. Chen et al. [7] and Bennett et al. [8] collected scanning traffic but mainly focused on the scanning behavior of Shodan and Censys. The studies [9]–[11] focused on the scanning of ICS protocols. In this context, Ferretti et al. [11] discussed different scanning behaviors, e.g., recurrent and occasional scanners, and campaigns over time. Torabi et al. [12] characterized scanning activities by compromised IoT devices. They presented scanning objectives based on the number or range of destination ports scanned, presenting the classes of range, strobe, and wide scans. Compared to these studies, we identify more scanners, providing an up-to-date list of 43 distinct scanning organizations. We provide insights into their scanning behavior and present a granular classification scheme considering scanning range, scanning schedule, and scanning picture. With a focus on the scanning of IoT services, we provide a current picture of the intentions of Internet-wide scanners aiming at IoT devices.

Ghiëtte et al. [13] and Tanaka et al. [14] present strategies to identify traffic generated by different scanning tools using encodings in TCP/IP header fields. Further, Tanaka et al. present fingerprints to detect scanning by botnets, such as Mirai and Hajime. We use these fingerprints to identify scanning tools and combine them with the identification of organizations to gain deeper insights into scanning strategies.

III. Honeypot and Analysis Infrastructure

In this section, we outline the infrastructure used. We describe the honeypots and IoT services that we set up and detail the logging and analysis mechanisms.

Our infrastructure consists of 6 Amazon Web Services (AWS) EC2 instances running Ubuntu 22.04. The instances are located in the AWS regions Cape Town, London, Mumbai, Northern California, São Paulo, and Tokyo. By distributing the honeypots over multiple regions, we collect information about differences in scanning behavior related to scanned destinations, i.e., scanners that only scan specific geographic regions.

All honeypot instances run services for the IoT protocols AMQP, CoAP, MQTT, and OPC UA. Table I shows the software used to set up the services as well as the protocols and ports on which the services run. The services emulate IoT servers or brokers to simulate small building automation systems. We modified the services to log interaction with them in detail and offer basic interaction with simulated variables. If authentication is required, we use the default credentials.

On each honeypot, we log every connection attempt to any TCP/UDP port using iptables log rules. This enables an overview of the traffic received for different services. For log processing, we use Logstash, Elasticsearch, and Kibana (ELK stack). Logstash parses log lines from the honeypots, Elasticsearch saves the log data to query it, and Kibana provides an interface to submit queries and visualize data.

TABLE I: The selected IoT services set up on the honeypots.

Service	Software	Protocol	Port
AMQP	RabbitMQ (v3.11.16)	TCP	5,672
CoAP	aiocoap Python package (v0.4.3)	UDP	5,683
MQTT	HiveMQ (hivemq-ce-2021.3)	TCP	1,883
OPC UA	opcua-asyncio Python package (v0.9.95)	TCP	4,840

IV. Identification of Scanners and Tools

We aim to identify the organizations behind scans and their tools used. In the following, we present the process of identifying scanners and their scanning tools.

In a first step, we retrieve information for every IP address that scans one of the honeypots using the *IPinfo.io* API [15]. In particular, we combine the provided information from the *AS (Autonomous System) name*, *hostname*, and *company name* fields, if given, to name the organizations behind the scan. Within the retrieved information, we search for keywords. For example, we search for keywords such as '*scan*', '*measurement*', and '*security*' to find well-intended scanners, i.e., scanners announcing their activities. This way, we identify, e.g., *Censys* (∗.*censys-scanner.com*), *driftnet.io* (∗.*internet-measurement.com*), and *IPIP.net* (*scan-*∗.*security.ipip.net*) by *hostname*. We further search for '*university*', '*institute*', and '*research*' to identify academic institutions. We identify, for example, *Georgia Institute of Technology* and *Berkeley Research Scanning* based on the full name of the corresponding university in the *AS name* or *company name*. We look up identified organizations to find further information, such as the IP ranges used to conduct scans, to associate all packets originating from the same organization. Lastly, we manually review the remaining data to identify further scanners that use hosting providers, in which case they can only be identified when reviewing all provided information combined.

Scanners typically use tools that automate the scanning of numerous IP addresses and ports, such as ZMap, Masscan, Unicorn, and Nmap. We aim to identify scanning tools by fingerprinting, i.e., observing the information they encode in fields like the IP-ID [16] or the TCP sequence number [17]. For example, ZMap sets a hard-coded IP-ID value of 54,321 [13]. Masscan sets the IP-ID to the value obtained by XORing the destination address, destination port, and TCP sequence number [13]. Unicorn calculates the TCP sequence number using a static session key, the source IP address, source port, and destination port [13]. Nmap has several scan options, which result in different combinations of values in TCP fields [18]. We further aim to identify packets originating from botnets, namely Mirai [1] and Hajime [19]. Packets sent by the Mirai botnet can be detected by checking if XORing the IP destination address and the TCP sequence number results in 0 [14]. We also check if the packets are destined to the Telnet port 23 or port 2,323, i.e., the ports typically targeted by Mirai [20]. As we could not identify any packets associated with Hajime, we did not include this botnet in the following analyses and only focus on Mirai.

Since the scanning tools use similar TCP/IP fields, false positives or an association of multiple scanning tools are inevitable. We handle such errors by conducting manual checks to ensure that the errors do not invalidate the results. If multiple scanning tools are associated with one packet, we

associate the packet to the tool with the more error-proof fingerprint. For example, the identification of Masscan is less error-prone than the identification of Mirai as the fingerprint of Masscan involves more TCP/IP fields. Further, UDP scans cannot be assigned to any tool due to the lack of distinct header fields. Fingerprinting can also be bypassed by changing the TCP/IP parameters set in the source code of the tools. If an organization uses a widely used scanning tool but changes its fingerprint, we cannot identify the scanning tool.

V. Evaluation

In the following sections, we present identified organizations and the scanning tools they used. We further discuss their scanning behavior and present criteria to classify them. For the evaluation, we use data collected on the deployed IoT honeypots from Jun. 1, 2023 to Jan. 31, 2024. During this period, we received 26,159,934 TCP and UDP packets. These packets originated from 465,251 unique IP addresses. First, we identify organizations scanning the honeypots. Second, we examine which tools scanners and observed organizations use to perform scans. Third, we inspect TCP and UDP traffic received on selected IoT ports in more detail, discussing and classifying the scanning behavior of identified organizations. Lastly, we examine which organizations perform application layer scans next to simple port scans on the selected IoT ports. Through these analyses, we aim to provide a comprehensive understanding of scanning behavior on the Internet.

Since the honeypots run in the AWS cloud, there is a potential bias in the results. Amazon is only one of many hosting providers. Thus, we cannot distinguish scanners that solely scan the AWS address space from those that scan the entire or a large part of the IPv4 address space. We also cannot detect scanners that avoid scanning the AWS range.

A. Organizations Conducting Scans

In this section, we present identified organizations that scan the honeypots and classify them based on their intent. We further discuss differences in their scanning volume across the honeypot locations. This analysis provides an overview of who is currently scanning the Internet and to what extent.

By applying the identification process described in Section IV, we identified 47 different scanning routines. Some scanning routines belong to the same organization: *Team Cymru* has two scanning routines for different services (DNS and NTP), the *Technical University of Munich (TUM)* has three scanning routines associated with them, and the *University of Twente* hosts two scanning projects, namely the *Internet Transparency Research Project* and the *DACS Research Group*. Overall, we identify 43 distinct organizations, which account for 30.3 % of traffic received on the honeypots. We divide the identified organizations into three categories, based on their scanning intentions: i) universities or organizations that conduct scans for research (*Research*, R), ii) companies or organizations that scan the Internet as part of their security services (*Commercial*, C), and iii) organizations with other intentions or whose intent is not clear (*Other*, O). In further analyses, we summarize all these under the term *organizations*.

Figure 1 visualizes the scanning volume, i.e., the number of TCP and UDP packets received, across the honeypot locations for the identified scanning routines. For each identified routine, the label includes the associated category. As can be seen from the labels, *Research* organizations account for the largest share with 24 distinct scanning routines, followed by 19 *Commercial* organizations. Among the observed organizations, the number of received packets differs greatly. While organizations such as *Censys* or the *Recyber Project* sent more than 10^5 packets per location, other organizations, such as *Team Cymru*, sent less than 10 packets per location. The number of packets received from a single organization across the honeypot locations is equally distributed, with only a few exceptions. For the DNS scanner of *Team Cymru*, *research.knoq.nl*, the *University of Colorado*, and the *Georgia Institute of Technology* (*Georgia Tech*), we see greater deviations. The DNS scanner of *Team Cymru* scanned the São Paulo honeypot only once during the observation period. For *research.knoq.nl*, no scans were seen on the Mumbai honeypot. The number of packets received from the *University of Colorado* and *Georgia Tech* increased at Cape Town and Tokyo, respectively. A possible explanation for the lack of scans on some instances is provided by Wan

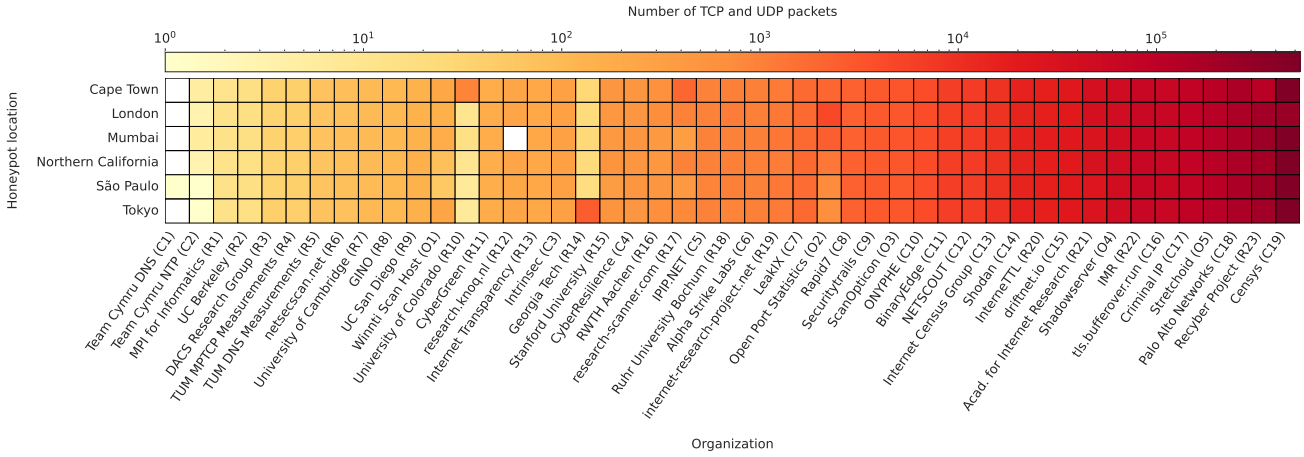


Fig. 1: Identified organizations with the number of packets received across the honeypot locations.

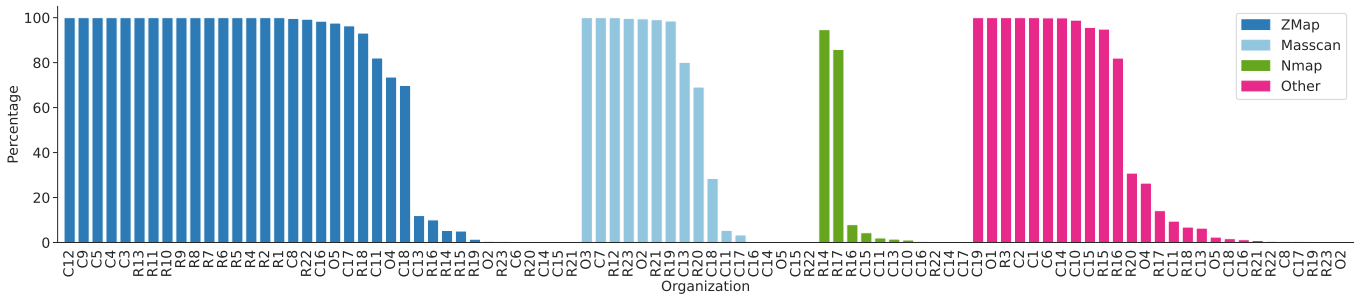


Fig. 2: Percentage of packets sent with the respective scanning tool for each identified organization. Labels as defined in Fig. 1.

et al. [21]: Scanners reach significantly fewer hosts on the Internet with single-origin scans. Transient network problems also prevent scanners from scanning all instances.

In summary, while we observe large differences in the scanning volume, we observe that most organizations scan (at least the AWS address space) evenly, with only a few exceptions. These exceptions possibly stem from reachability problems caused by the scanning location.

B. Use of Scanning Tools

In this section, we discuss which scanning tools are used for scanning, and which ports are often scanned. We introduced the process of associating tools using fingerprints in Section IV. We provide an overview of the popularity of various tools and their use by different types of organizations. Such an analysis does not yet exist to this extent, i.e., collecting data over 8 months using 6 honeypot locations.

We first analyze which scanning tools are used by which of the identified organizations. For all TCP and UDP packets received, we observe the following distribution of used tools: 43.0 % ZMap, 30.4 % Masscan, 2.1 % Unicorn, 0.7 % Nmap, 2.5 % Mirai, and 21.3 % other tools. For the identified organizations, Figure 2 shows the traffic shares for the scanning tools ZMap, Masscan, and Nmap. Packets that cannot be associated with a listed tool belong to the other tools share. In each of the four groups, a bar represents one organization that uses the tool at the given percentage. If an organization uses multiple tools, each covered group includes a bar proportional to its use. This representation allows us to observe if an organization uses one scanning tool exclusively or combines it with one or multiple other scanning tools. Note that a low share of a scanning tool may indicate false positives during the identification process. For instance, it is unlikely that the *Academy for Internet Research* (R21) sent only one packet using ZMap. We still included possible false positives in Figure 2, as we cannot reliably determine them. The scanning tool identification shows that ZMap is the most popular scanning tool, used by 37 organizations. Out of these organizations, 25 use ZMap as the main scanning tool, i.e., they generate more than 50 % of the packets with it. ZMap is a popular scanning tool due to its good performance. A single computer with Gigabit Ethernet can scan the entire IPv4 address space within 45 minutes using ZMap [22]. Masscan is used by 17 organizations. Masscan is also suited for Internet-wide scans but performs slightly worse than ZMap [23], [24]. We identify the use of Nmap by 11 organizations. The overall low share of Nmap is to be expected as Nmap is not designed to scan the entire Internet, but rather to probe many ports on a small number of hosts. Hence, it is more often used as an addition to other tools for follow-up and protocol-specific scans on the application layer of selected hosts [22]. For example, 7.9 % of the packets received from *RWTH Aachen* were Nmap, the rest ZMap, and other tools. Filtering possible false positives, there are 23 organizations that use at least two different scanning tools. None of the organizations used Unicorn. Of all packets associated with Unicorn, 61.9 % originate from a single unidentified scanner, which resides in the network of *ChinaNet*. Packets sent by the Mirai botnet were also not associated with any identified organization. However, as Mirai accounts for 2.5 % of the packets received, there are still infected hosts on the Internet.

We now take a look at the use of scanning tools based on the category of organization, i.e., *Research*, *Commercial*, and *Other*. Figure 3 shows the share of scanning tools for each of the three categories, where each organization contributes the same proportion, regardless of traffic volume. This prevents organizations with high scanning activity, such as *Censys*, from being overrepresented. We observe that *Research* organizations mainly use ZMap and Masscan to conduct scans, i.e., ready-to-use scanning tools that are flexible to configure. We further observe that *Commercial* organizations mainly use ZMap or other (customized) tools that meet the requirements for the offered services. For example, *Censys* developed their scanning tool to fit their needs [25]. Organizations outside the *Research* and *Commercial* categories use Masscan and ZMap more frequently than other tools.

Lastly, we examine whether a scanning tool is used to scan a wide range of ports or only a few selected. Figure 4 shows the number of scans per port for ZMap, Masscan, Nmap, and other tools. For readability, we highlighted the ports 1,024 and 49,152, i.e., the end of well-known and beginning of ephemeral ports, using red lines. All histograms show that the lower port ranges, on which common services

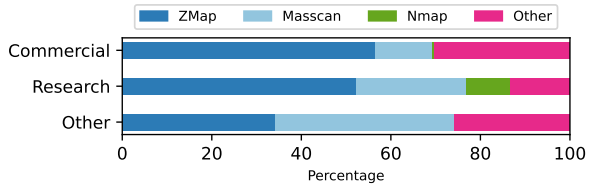


Fig. 3: Scanning tools spread per category of organization.

TABLE II: Number of TCP and UDP packets per scanning tool for identified and unidentified scanners.

	Nmap	ZMap	Masscan	Other
Identified scanners	17,102	2,880,136	1,793,695	3,242,305
Unidentified scanners	160,038	8,379,853	6,157,861	2,331,663

run, are scanned more often. There is also a focus on scanning activity at the beginning of the ephemeral ports. ZMap and Masscan are used to scan the full port range while Nmap is rather used to scan selected ports, see the top 1,000 default Nmap ports [26]. Unexpectedly, the port most probed using Nmap is port 6,379, which is registered for Redis, an open-source database. An explanation for this activity could be the exploitation of vulnerabilities [27]. Scanners using other scanning tools mostly scan lower port ranges, focusing on ports of common services such as Telnet, HTTP, and SSH. We omit diagrams for Unicorn and the Mirai botnet as only a few ports were scanned using both tools. Unicorn is an efficient scanning tool, so we expected that it would be used to scan a large number of ports [13]. However, only a few scanners used Unicorn, scanning 11 ports in total. The mentioned scanner originating from the network of *ChinaNet* scanned the ports 8,563 and 443. We did not find a reason for scanning port 8,563, as no IANA-registered or other widely-used services are running on it. The Mirai botnet scans the ports 23 and 2,323 to identify victim devices with an open telnet server running. If a victim responds, Mirai launches a brute-force login attack with default credentials [20].

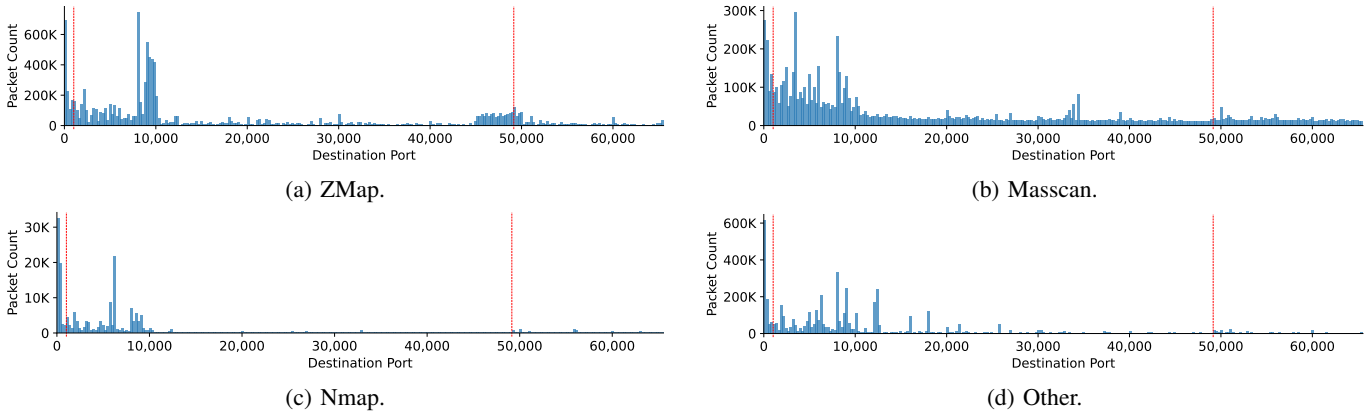
Depending on the intentions of an organization, the used tools vary. In Table II, we show how packets are distributed between tools for identified scanners, i.e. organizations, and unidentified scanners. Most packets from scanners are associated with ZMap and Masscan. Hence, scanners, independent of the associated type of organization, show great interest in tools suited for full-range scans on a large number of hosts.

C. Scanning Behavior

In this section, we analyze the scanning behavior of identified organizations, focusing on organizations interested in the selected IoT ports. First, we provide an overview of the scanning behavior. Then, we present criteria to classify behavior, i.e., scanning ranges, scanning schedules, and scanning pictures. These criteria can be applied by other researchers and operators to improve their understanding of scanners.

1) Overview: In total, 1.4 % of the TCP and UDP traffic addressed the selected IoT ports. 21 identified organizations scanned at least one selected IoT port, which accounts for 53.1 % of the traffic received on these ports. In total, the AMQP port received the most traffic, followed by OPC UA, MQTT, and lastly CoAP. Table III shows insights into the scanning behavior for each of the 21 organizations scanning at least one of the selected IoT ports. The second column presents the number of TCP and UDP ports an organization scans. The third column shows the number of packets received on all TCP and UDP ports, as well as the share of IoT packets. The fourth column holds the percentage of days with at least one scan, which helps to discuss specific scan patterns. We present the average scan interval in the fifth column. The scan interval denotes the time between two scans of the same port by the same organization. We calculate the scan interval for all ports scanned at least twice per organization and honeypot location so that the interval is independent of the location. We present the average scan interval of all scanned ports on the left and of the scanned IoT ports on the right. As the scan interval is an average calculated across all ports, we cannot determine solely from this value whether an organization with a scan interval of, e.g., two days scans all ports on the same day or distributes a scan over multiple days. The full range of ports could be scanned every second day, or half on the first day and the other half on the second. The respective ratio of days scanned is then either 50 % or 100 %. The last two columns contain histograms to visualize the number of scans across all ports and the observation period. For readability, the histograms contain red lines for characteristic port ranges on the x-axis, and for 1,000 or 50 packets received on the y-axis.

2) Scanning Range: We use the number of ports and packets as well as the scans per port displayed in Table III to discuss different scanning ranges that emerge from the data.



(a) ZMap.

(b) Masscan.

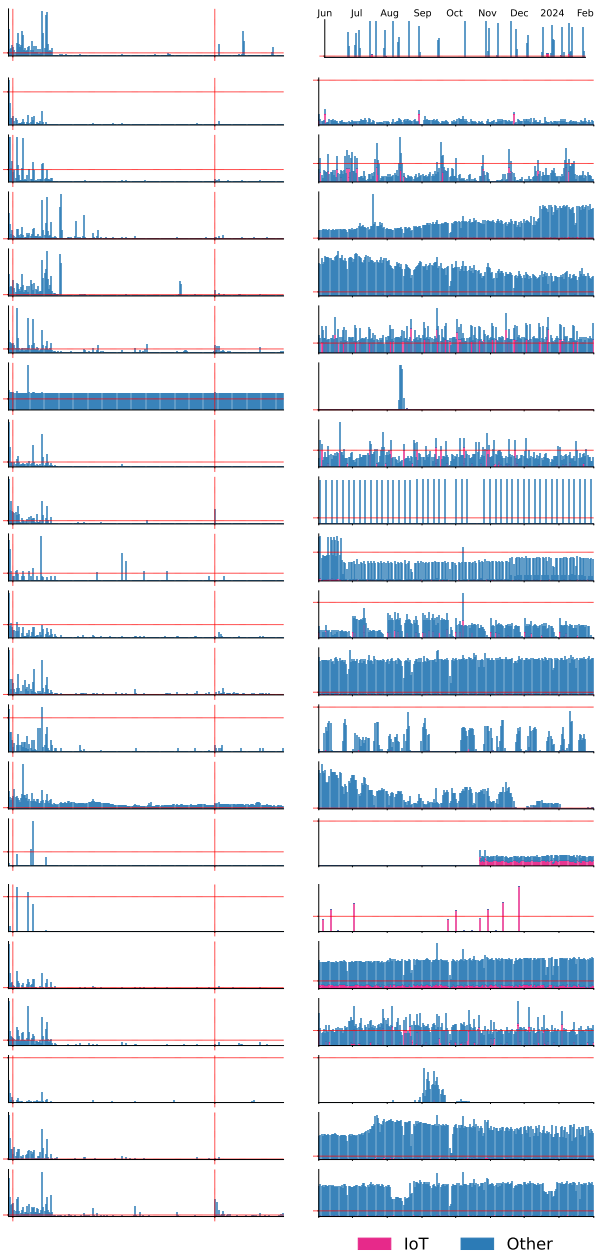
(c) Nmap.

(d) Other.

Fig. 4: Scans per port using different scanning tools; histograms with 256 bins; the red lines mark the end of well-known (1,024) and beginning of ephemeral ports (49,152).

TABLE III: Scanning behavior characteristics of organizations that probe at least one of the selected IoT ports (AMQP, CoAP, MQTT, and OPC UA). Organizations marked with * probed all four IoT ports. We define the percentage of days with at least one scan as days scanned. Scans per port shows the scans across all ports using a histogram with 256 bins; the red lines on the x-axis mark the end of well-known (1,024) and the beginning of ephemeral ports (49,152); the red line on the y-axis at 1,000 packets allows to compare scanning volume between organizations. Scans over time shows the scans of IoT and non-IoT ports across the observation period on the London honeypot using a histogram with one bin per day; the red line on the y-axis at 50 packets allows to compare the scanning volume between organizations.

Organization	#Ports		#Packets		Days Scanned		Avg. Scan Interval (days)		Scans per Port	Scans over Time
	TCP	UDP	All	IoT	All	IoT	All	IoT		
Acad. for Internet Research*	1,467	1	191,169	1.2 %	15.2 %	13.0 %	11.0 d	10.5 d		
Alpha Strike Labs	103	57	6,057	4.2 %	97.7 %	2.5 %	45.4 d	91.2 d		
BinaryEdge	1,161	16	35,847	10.0 %	98.2 %	9.6 %	41.7 d	17.8 d		
Censys*	65,422	787	2,838,080	2.1 %	100.0 %	99.9 %	57.0 d	1.2 d		
Criminal IP*	14,852	660	493,934	0.4 %	100.0 %	30.7 %	12.0 d	26.5 d		
driftnet.io	5,573	1	130,066	19.3 %	100.0 %	27.7 %	41.0 d	10.1 d		
IMR*	65,534	0	389,200	0.7 %	4.1 %	1.2 %	17.2 d	5.0 d		
Internet Census	303	38	53,532	9.0 %	100.0 %	20.9 %	15.2 d	14.7 d		
InterneTTL	306	112	102,920	0.5 %	19.4 %	18.8 %	5.5 d	5.3 d		
NETSCOUT	11	67	40,648	2.9 %	99.3 %	52.6 %	8.3 d	1.9 d		
ONYPHE	214	15	23,866	5.8 %	96.1 %	16.8 %	15.8 d	14.0 d		
Palo Alto Networks	607	18	954,426	2.0 %	100.0 %	94.1 %	1.1 d	1.1 d		
Rapid7	144	22	13,905	1.1 %	57.7 %	6.2 %	19.2 d	29.8 d		
Recyber Project*	65,535	3	1,154,785	0.0 %	100.0 %	6.1 %	52.5 d	53.4 d		
Ruhr-University Bochum	16	2	5,996	46.6 %	42.7 %	41.0 %	1.6 d	1.0 d		
RWTH Aachen*	11	3	3,403	92.2 %	23.1 %	5.2 %	54.9 d	69.4 d		
Shadowserver*	167	38	267,462	10.8 %	100.0 %	99.7 %	1.2 d	1.0 d		
Shodan*	1,251	77	91,176	10.3 %	100.0 %	30.8 %	32.4 d	10.5 d		
Stanford University	738	0	2,770	1.4 %	14.2 %	0.7 %	5.5 d	10.0 d		
Stretchoid	752	67	648,398	3.0 %	100.0 %	96.1 %	4.7 d	1.2 d		
tls.bufferover.run	1,189	0	393,786	1.7 %	100.0 %	45.8 %	5.1 d	2.3 d		



Focused Scanners: We observe several organizations that scan a small number of ports. For example, 12 organizations scanned less than 10 ports. We define *focused scanners* as scanners that send packets to less than 30 ports, i.e., they are interested in about 5 to 10 services and their variants, e.g., port 8,443 is a variant of port 443. With about three variants, focused scanners scan around 30 ports in total. The *focused scanner* applies to 17 identified organizations.

Table III highlights *IoT-focused scanners* through the high ratio of packets aimed at the selected IoT ports among the total amount of packets received. Specifically, *RWTH Aachen* spends most of their effort in scanning IoT ports with more than 90 % of packets directed at the selected ports. *RWTH Aachen* states that they scan a selection of common ports on every public IP address daily [28]. We observe that they mostly scan the selected IoT services, but not daily. The *Ruhr-University Bochum* is another example of a focused scanner, scanning the AMQP, CoAP, and MQTT ports.

Wide-Range Scanners: The majority of identified organizations scan in a less focused way, i.e., cover wider port ranges. The entire port range can be divided into three sub-ranges: well-known ports (0 – 1,023), registered ports (1,024 – 49,151), and ephemeral ports (49,152 – 65,535). Well-known and registered ports are assigned by IANA [29]. We define *wide-range port scanners* as scanners that scan more than 30 ports and partly cover one or several of these ranges. In particular, the data show no clear differentiation between well-known, registered, and ephemeral port scanners. Rather, the 23 organizations scanning wide ranges cover parts of each range. For example, the *Academy for Internet Research* scans 1,467 distinct ports in total, of which 8.7 % are well-known, 79.0 % registered, and 12.3 % ephemeral. They announced to proactively scan for active vulnerabilities and open ports, comparing themselves with *Shodan* and *Censys* [30]. While the scanning range is comparable to *Shodan's*, differing in 238 ports of which most are ephemeral, *Censys* scans a much wider range of ports. *driftnet.io* state that they perform deep Internet scans [31]. With 5,573 distinct ports probed in total, out of which 4.3 % are well-known, 76.3 % registered, and 19.5 % ephemeral, *driftnet.io* scan a considerably wide range of ports. The selection of scanned port ranges can be based on different resources. For example, the list of IANA-assigned ports can be used as a reference to scan widely known services. Out of the 225 distinct ports probed by *ONYPHE*, 80.4 % are IANA assigned. The *Georgia Institute of Technology*, which uses Nmap to probe ports, covers 97.1 % of the top 1,000 service ports listed by Nmap [26].

Full-Range Scanners: We define *full-range scanners* as scanners that send packets to more than 99 % of all ports. We classify *internetmeasurementresearch.com (IMR)* and the *Recyber Project* as full-range port scanners, with the former showing equally distributed scans across all ports, with one spike for port 4,840 (OPC UA). *Recyber Project* scanned all ports but focused on lower port ranges. Similarly, *Censys* scanned the full port range at least once, indicated by the total number of scanned ports, but focused on specific ports for most of their scan probes. This behavior is in line with *Censys*' own information [32]. They scan the entire Internet with various scanning routines. On a daily basis, *Censys* scans 137 IANA-assigned ports and the 1,140 most popular ports on cloud providers, including AWS. Less popular ports are scanned every 10 days. The full IPv4 and port range is scanned at a low background rate.

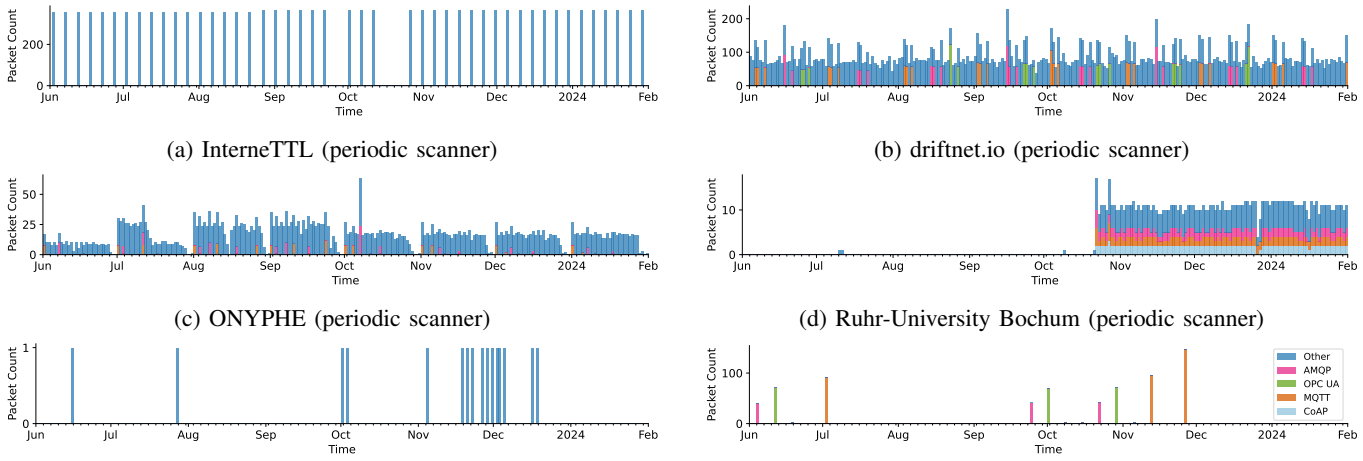
3) Scanning Schedule: Next, we classify scanning behavior with regard to the temporal aspect of scanning, e.g., the time interval of the scan and the resulting scanning schedule. We execute a Fast Fourier Transformation (FFT) on the scans over time of each organization. With the results of the FFT, we categorize the scanning schedule of organizations into *periodic*, *sporadic*, and *acyclic*.

Periodic: For 28 scanners, we identify a periodic scanning pattern, i.e., scans are conducted regularly in a similar manner to, e.g., capture trends over time. We identify the periodicity through the FFT by analyzing the scanning frequency for individual ports. If the frequency across multiple ports is the same, we call a scanner *periodic*. For example, *InterneTTL* shows a clear *periodic* pattern, scanning selected ports every 5 days, see Figure 5a. Between late September and the end of October, there are deviations from the pattern, resulting in some missing scans on the London honeypot. For *driftnet.io*, we also identify *periodic* patterns. Figure 5b presents the interval of 30 days for AMQP, OPC UA, and MQTT. For most other protocols, the FFT reveals that *driftnet.io* uses a scanning interval of 3 days. Figure 5c shows a third *periodic* scanning example of the organization *ONYPHE*. The diagram shows the scanning interval of 30 days. However, for CoAP and MQTT, *ONYPHE* uses an interval of 5 days.

If a *periodic* scanner has an average scanning interval of 1 day, we call it *continuous*. For example, the *Ruhr-University Bochum* conducts *continuous* scans since late October, see Figure 5d. They scan all their selected ports daily. In total, we classify 14 scanners as *continuous*.

Sporadic: Unlike the *periodic* schedule, *sporadic* scans do not follow any temporal pattern. Instead, scans only appear for a short period and provide too little data for analysis. Scanners may follow a *sporadic* schedule, for instance, in the case of a specific interest at a specific point in time. None of the 5 organizations classified as *sporadic* is interested in scanning IoT protocols. For example, Figure 5e presents the scanning activities of *Berkeley Research Scanning*. They only scan individual ports occasionally, with no consistent pattern.

Acyclic: In comparison to the *sporadic* pattern, we classify 8 scanners that present a pattern among different ports and, thus, show some scanning routines, as *acyclic*. For example, we classify the scanning schedule for IoT protocols of the *RWTH Aachen*, see Figure 5f, as *acyclic*. They have an obvious pattern between the protocols AMQP, OPC UA, and MQTT. However, the initial start of this routine is not in a cyclic pattern. Interestingly, the *RWTH Aachen* scans all other ports



(a) InterneTTL (periodic scanner)

(b) driftnet.io (periodic scanner)

(c) ONYPHE (periodic scanner)

(d) Ruhr-University Bochum (periodic scanner)

(e) Berkeley Research Scanning (sporadic scanner)

(f) RWTH Aachen (acyclic scanner)

Fig. 5: Scans over time on the London honeypot for selected organizations; classified scanning schedule in brackets.

with an interval of 7 days. Due to low scanning volume, the intervals for these ports are difficult to read in Figure 5f.

4) Scanning Picture: Finally, we differentiate two pictures that scanners can capture through their activities. First, a scanner can create a *consistent snapshot* by scanning all ports of Internet hosts in a very short period of time. The data collected for the ports across hosts has a similar temporal meaning, relevant for, e.g., discussing the state of configured security policies across multiple systems. We include all continuous scanners, e.g., the *Ruhr-University Bochum*, in the group of *consistent snapshots*. Additionally, scanners like the IoT scanning routine of *RWTH Aachen* create a *consistent snapshot*. Second, a scanner can detect the landscape of Internet services with a *background scan*. Specifically for larger trends, e.g., the availability of specific services over the last year, multiple services and hosts are scanned over longer periods. We classify scanners as background scanners if the average scan interval is relatively large compared to the high scan frequency indicated by the number of days scanned. For example, as shown in Table III, *Censys* scans continuously, but only has an average scan interval of 57.0 days. Hence, they scan selected ports in the background as stated, see [32].

Classifying the behavior of scanners into the above categories allows us to conclude their intent. *Full-range scanners* aim to get an overview of all accessible services, without focusing on specific port ranges. *Wide-range scanners* scanning well-known and registered ports show an interest in widely used services and possible variants. Scanning ephemeral ports may be motivated by the discovery of hidden services. *Focused scanners* specifically decide what ports to scan. Depending on the scanning schedule, i.e., scanning in a *periodic*, *sporadic*, or *acyclic* pattern, organizations are further interested in capturing trends over time or recording observed services at specific points in time. The resulting scanning picture shows either a *consistent snapshot* of scanned services or *background scans*.

Of all scans, 69.7 % remain unidentified, i.e., we cannot assign the scans to an organization. The top 5 regions for unidentified scanners are: USA (30.8 %), Great Britain (19.4 %), Netherlands (11.3 %), China (7.8 %), and Bulgaria (7.6 %). For unidentified scanners, we cannot accurately analyze their intentions in scanning IoT devices. Blocking traffic from these regions may reduce potentially malicious scans.

D. Protocol Scans on IoT Services

Each honeypot provides IoT services that allow interaction. For example, the OPC UA server mimics a small building automation system, consisting of a temperature and a window contact sensor. A client can change sensor states and subscribe to events, such as variable changes. In this section, we address the interest of identified organizations and other scanners in protocol scans and discuss noticeable behavior.

Not every organization connects or interacts with the provided IoT services on the application layer. Out of 18 organizations that scanned the AMQP port, 16 organizations connected to the AMQP service. Out of 14 organizations that scanned the CoAP port, 7 connected to the CoAP service. Out of 17 organizations that scanned the MQTT port, 11 connected, and 6 also subscribed to the MQTT service. Out of 11 organizations that scanned the OPC UA port, only 2 organizations initiated a session. Further, not all traffic intended for an IoT port is traffic for the associated protocol. We observe that scanners scan without considering specific protocols. For example, we observe HTTP traffic in the AMQP logs for 11 of 16 interacting organizations.

In the MQTT log analysis, we discovered that scanners use different MQTT versions for their protocol scans, i.e., v3 and v5. The newest version v5 was standardized in 2019 and is not compatible with older versions. We expected scanners interested in MQTT to adjust their scanning tools to use the latest version. However, there are only 3 scanners that use v5, out of which we only identified 1, namely *RWTH Aachen*.

While we observe some extremes for application traffic volume from individual scanners, e.g., a case where a single unidentified scanner interacted with the MQTT service for three days, we observe overall only little interaction with IoT

services. There could be several reasons for the lack of interest in these services. As we host all honeypots on AWS, we are limited to the AWS IP address range. This address range may not be of interest to IoT scanners that search for vulnerable IoT devices in private or industrial buildings, but not in the cloud. Also, the IoT services we provide are simple and, thus, may not be of greater interest to scanners.

VI. Conclusion

In this paper, we analyzed the behavior of scanners conducting scans on deployed IoT honeypots over 8 months. We collected data to identify scanning organizations and their tools used. We identified 47 scanning routines that belong to 43 distinct organizations. We further presented criteria to classify their scanning behavior with respect to scanning ranges, scanning schedules, and the overall scanning picture. Among the identified organizations, we observed, for example, *full-range port scanners*, such as *Censys*, scanning the Internet *periodically* using *background scans*, and *focused scanners* that scan specific ports *acyclically* to create a *consistent snapshot*, e.g., *RWTH Aachen*. We identified the use of popular scanning tools, such as ZMap, Masscan, and Nmap, and combined this information with the identified organizations. ZMap and Masscan were the most popular tools used by research and commercial organizations almost exclusively for full-range scans. Nmap is rarely used for wider scans but rather for specific protocol scans.

We analyzed the traffic received at the IoT services AMQP, CoAP, MQTT, and OPC UA in more depth. Of the identified organizations, 21 probe at least one of the selected IoT ports. In particular, we identify *RWTH Aachen* and *Ruhr-University Bochum* as scanners interested in snapshots of Internet-connected IoT services. Overall, there was little interaction from scanners with the selected IoT services.

We revealed that the selected IoT ports do not receive the same level of attention from scanners as other ports despite their growing prevalence and exposure to botnets or other malicious scanning activity. Organizations should incorporate popular IoT services in their scanning activities to identify existing vulnerabilities early and help improve Internet security.

References

[1] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman *et al.*, "Understanding the Mirai Botnet," in *USENIX Security Symposium*, 2017.

[2] Z. Durumeric, M. Bailey, and J. A. Halderman, "An Internet-Wide View of Internet-Wide Scanning," in *USENIX Security Symposium*, 2014.

[3] J. Mazel, R. Fontugne, and K. Fukuda, "Profiling Internet Scanners: Spatiotemporal Structures and Measurement Ethics," in *Network Traffic Measurement and Analysis Conference (TMA)*, 2017.

[4] H. Heo and S. Shin, "Who is Knocking on the Telnet Port: A Large-Scale Empirical Study of Network Scanning," in *Asia Conference on Computer and Communications Security*, 2018.

[5] P. Richter and A. Berger, "Scanning the Scanners: Sensing the Internet from a Massively Distributed Network Telescope," in *ACM Internet Measurement Conference*, 2019.

[6] R. Hiesgen, M. Nawrocki, A. King, A. Dainotti, T. C. Schmidt, and M. Wählisch, "Spoki: Unveiling a New Wave of Scanners through a Reactive Network Telescope," in *USENIX Security Symposium*, 2022.

[7] Y. Chen, X. Lian, D. Yu, S. Lv, S. Hao, and Y. Ma, "Exploring Shodan From the Perspective of Industrial Control Systems," *IEEE Access*, 2020.

[8] C. Bennett, A. Abdou, and P. C. van Oorschot, "Empirical Scanning Analysis of Censys and Shodan," *NDSS Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb)*, 2021.

[9] A. V. Serbanescu, S. Obermeier, and D.-Y. Yu, "ICS Threat Analysis Using a Large-Scale Honeynet," in *International Symposium for ICS & SCADA Cyber Security Research (ICS-CSR)*, 2015.

[10] A. Mirian, Z. Ma, D. Adrian, M. Tischer, T. Chuenchujit, T. Yardley, R. Berthier, J. Mason, Z. Durumeric, J. A. Halderman, and M. Bailey, "An Internet-Wide View of ICS Devices," in *Conference on Privacy, Security and Trust (PST)*, 2016.

[11] P. Ferretti, M. Pogliani, and S. Zanero, "Characterizing Background Noise in ICS Traffic Through a Set of Low Interaction Honeypots," in *Proceedings of the ACM Workshop on Cyber-Physical Systems Security & Privacy*, 2019.

[12] S. Torabi, E. Bou-Harb, C. Assi, E. B. Karbab, A. Boukhtouta, and M. Debbabi, "Inferring and Investigating IoT-Generated Scanning Campaigns Targeting a Large Network Telescope," *IEEE Transactions on Dependable and Secure Computing*, 2020.

[13] V. Ghiëtte, N. Blenn, and C. Doerr, "Remote Identification of Port Scan Toolchains," in *IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, 2016.

[14] A. Tanaka, C. Han, and T. Takahashi, "Detecting Coordinated Internet-Wide Scanning by TCP/IP Header Fingerprint," *IEEE Access*, 2023.

[15] IPinfo, "Ipinfo.io," https://ipinfo.io, accessed: 2023-08-25.

[16] J. Touch, "RFC 6864: Updated Specification of the IPv4 ID Field," 2013.

[17] J. Postel, "RFC 793: Transmission Control Protocol," 1981.

[18] S. Liao, C. Zhou, Y. Zhao, Z. Zhang, C. Zhang, Y. Gao, and G. Zhong, "A Comprehensive Detection Approach of Nmap: Principles, Rules and Experiments," in *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, 2020.

[19] S. Herwig, K. Harvey, G. Hughey, R. Roberts, and D. Levin, "Measurement and Analysis of Hajime, a Peer-to-peer IoT Botnet," in *Network and Distributed Systems Security Symposium (NDSS)*, 2019.

[20] C. Kolias, G. Kambourakis, A. Stavrou, and J. Voas, "DDoS in the IoT: Mirai and Other Botnets," *Computer*, vol. 50, no. 7, 2017.

[21] G. Wan, L. Izhikevich, D. Adrian, K. Yoshioka, R. Holz, C. Rossow, and Z. Durumeric, "On the Origin of Scanning: The Impact of Location on Internet-Wide Scans," in *ACM Internet Measurement Conference*, 2020.

[22] Z. Durumeric, E. Wustrow, and J. A. Halderman, "ZMap: Fast Internet-Wide Scanning and Its Security Applications," in *USENIX Security Symposium*, 2013.

[23] H. Kim, T. Kim, and D. Jang, "An Intelligent Improvement of Internet-Wide Scan Engine for Fast Discovery of Vulnerable IoT Devices," *MDPI Symmetry*, May 2018.

[24] D. Adrian, Z. Durumeric, G. Singh, and J. A. Halderman, "Zippier ZMap: Internet-Wide Scanning at 10 Gbps," *USENIX Workshop on Offensive Technologies*, 2014.

[25] B. Dickson, "Censys: How a University Project became a Major Commercial Security Platform," 2022. [Online]. Available: https://portswigger.net/daily-swig/censys-how-a-university-project-became-a-major-commercial-security-platform

[26] G. Lyon, "Nmap Network Scanning: Official Nmap Project Guide to Network Discovery and Security Scanning," 2008. [Online]. Available: https://nmap.org/book/

[27] "Redis Security Notice: Heap Overflow Vulnerabilities," 2023. [Online]. Available: https://redis.io/blog/security-notice-heap-overflow-vulnerabilties/

[28] RWTH Aachen, "Communication & Distributed Systems." [Online]. Available: http://researchscan.comsys.rwth-aachen.de

[29] M. Cotton, L. Eggert, J. Touch, M. Westerlund, and S. Cheshire, "RFC 6335: Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Port Number Registry," 2011.

[30] Academy for Internet Research. [Online]. Available: https://academyforinternetresearch.org

[31] driftnet.io, "Analyzing your Exposure Starts with one Click." [Online]. Available: https://driftnet.io

[32] Censys, "Censys Internet Scanning Intro," 2024. [Online]. Available: https://support.censys.io/hc/en-us/articles/360059603231-Censys-Internet-Scanning-Intro