ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/339820038

Inferring and Investigating IoT-Generated Scanning Campaigns Targeting A Large Network Telescope

Article *in* IEEE Transactions on Dependable and Secure Computing · March 2020

DOI: 10.1109/TDSC.2020.2979183

CITATIONS
7

READS
319

6 authors, including:

Elias Bou-Harb
University of Texas at San Antonio
99 PUBLICATIONS 1,313 CITATIONS
SEE PROFILE

Chadi Assi
Concordia University Montreal
416 PUBLICATIONS 7,412 CITATIONS
SEE PROFILE

Elmouatez Karbab
Concordia University Montreal
24 PUBLICATIONS 428 CITATIONS
SEE PROFILE

Amine Boukhtouta
Ericsson
18 PUBLICATIONS 378 CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

UAV communications View project

Audit Ready Cloud View project

All content following this page was uploaded by Elias Bou-Harb on 15 March 2020.

The user has requested enhancement of the downloaded file.

Inferring and Investigating IoT-Generated Scanning Campaigns Targeting A Large Network Telescope

Sadegh Torabi, Elias Bou-Harb, Chadi Assi, ElMouatez Billah Karbab, Amine Boukhtouta, and Mourad Debbabi

Abstract—The analysis of recent large-scale cyber attacks, which leveraged insecure Internet of Things (IoT) devices to perform malicious activities on the Internet, highlighted the rise of IoT-tailored malware/botnets. These malware propagate by scanning the Internet for vulnerable, exploitable IoT devices that could be utilized for further malicious activities. In this paper, we devise a multi-level methodology to investigate Internet-scale reconnaissance activities generated by infected IoT devices. We leverage the `Shodan` IoT search engine and over 6TB of passive network traffic from a large network telescope (darknet) to infer compromised IoT devices and characterize the generated scanning campaigns. The results highlight a distinctive characteristic of IoT malware/botnets, represented by the targeted ports/services over the analysis interval. Furthermore, while these ports/services are mainly associated with well-known IoT malware/botnets (e.g., `Mirai` and `Satori`), we uncovered newly targeted ports, which indicate emerging IoT malware/botnet. Finally, by comparing two instances of analyzed IoT-generated scanning campaigns, we highlight the persistence and evolution of IoT malware/botnets (e.g., `ADB.Miner` and `Fbot`), which exploit existing, and in some cases, possibly new vulnerabilities.

Index Terms—Network telescope (darknet), compromised IoT devices, IoT malware/botnet, scanning campaigns, clustering analysis.

1 Introduction

Internet of Things (IoT) devices have been widely adopted in various parts of our daily activities. These Internet connected devices, which tend to have limited functionalities and resources, are mainly used to facilitate efficient data collection, monitoring, and information sharing. Despite the benefits of using IoT devices and their wide spread adoption, the increasing number of IoT-driven cyber-attacks illustrate the rise of IoT-tailored malware, which aim at exploiting vulnerable IoT devices that will be utilized within coordinated botnets to perform further malicious activities [1–3]. In fact, these IoT malware/botnets have gained much popularity among adversaries due to the insecurity of the IoT paradigm and the wide range of existing vulnerabilities. In addition, adversaries have been utilizing compromised IoT devices as effective attack enablers, which can be leveraged to evade detection while performing large-scale malicious activities (e.g., `Mirai` [4]).

In order to mitigate and prevent large-scale IoT-driven cyber attacks, there exists an utmost need to detect and characterize emerging IoT malware/botnets, which tend to spread over the Internet by searching for vulnerable IoT devices that could be exploited for future use. This cannot be done without possessing an Internet-scale perspective of IoT devices and their unsolicited activities over a period of time, which is indeed a challenging task as it requires addressing the following problems: (i) the lack of empirical data related to the widespread deployment of IoT devices [5], and (ii) the insufficient knowledge about compromised IoT devices and their underlying malicious activities [6].

To this end, an effective approach to gain Internet-wide cyber threat intelligence is to study passive measurements gathered using designated sensors or traps that collect traffic from the Internet [7, 8]. These sensors collect one-way traffic targeted towards routable, yet unused Internet Protocol (IP) addresses, which are known as darknets or network telescopes [9]. Characteristically, traffic destined to these inactive hosts is likely to represent suspicious and unsolicited activities. Moreover, a large portion of traffic captured at the darknet represents Internet reconnaissance activities [10, 11]. Therefore, motivated by the fact that IoT malware/botnets heavily rely on coordinated scanning activities to propagate through the Internet [4, 12], in this paper, we leverage macroscopic, empirical passive network telescope data to execute a multi-level methodology for inferring malware-infected IoT devices and investigating their generated scanning activities. In addition, we leverage data mining methods to unveil common scanning objectives among compromised IoT devices, which reflect the targeted ports/services. More importantly, we demonstrate a meaningful approach for identifying scanning campaigns by clustering correlated IoT devices based on their scanning objectives and similarities in their scanning behaviors over time.

We leverage over 6TB of passive darknet data with IoT device information from `Shodan`, and obtain about 172M TCP-SYN scanning packets generated by 8,444 compromised IoT devices over 11 days. Our initial data analysis revealed emerging IoT malware/botnets, illustrated by 18 clusters of correlated compromised IoT devices with similar characteristics of the underlying scanning campaigns. The majority of these IoT botnets (12 out of 18), were found to

- *S. Torabi, C. Assi, E. Karbab, and M. Debbabi are with the Concordia Institute for Information Systems Engineering, Montreal, QC H3G 1M8, Canada.*
  E-mails: {sa_tora, assi, e_karbab, debbabi}@encs.concordia.ca
- *E. Bou-Harb is with the Cyber Center For Security and Analytics, University of Texas at San Antonio, San Antonio, USA. E-mail: elias.bouharb@utsa.edu*
- *A. Boukhtouta is with the Ericsson Research Group, Montreal, QC H4R 2A4, Canada. E-mail: amine.boukhtouta@ericsson.com*

Manuscript received April 19, XXX; revised August 26, XXX.

utilize IoT devices to target short lists of commonly used ports/services, which are associated with known vulnerabilities (e.g., Telnet/23). Moreover, our results shed light on an emerging IoT malware/botnets, represented by a large scanning campaign towards a distinctive destination port range (e.g., 19328–19622), which to the best of our knowledge, are not associated with any known vulnerabilities. In addition, by analyzing and comparing two instances of IoT scanning traffic that were collected on well-separated time periods (13 months), we highlight the persistence of few well-known IoT malware/botnets, especially those targeting Telnet and HTTP. Finally, we highlight the evolution of IoT-generated scanning campaigns towards targeting new, or previously uncommon vulnerabilities, which indeed corroborate on the evolutionary nature of IoT malware/botnets.

In this context, we frame the contributions of this paper as follows:

- We extend our previous work [13] by introducing a stratified methodology, which utilizes passive darknet data for investigating emerging IoT malware/botnets through inferring compromised IoT devices and characterizing their underlying scanning campaigns.
- We demonstrate a meaningful approach for uncovering IoT-generated scanning campaigns, which is based on frequent pattern analysis to identify common scanning objectives (targeted ports) and unsupervised clustering of correlated IoT devices with similar behavioral characteristics over a period of time.
- We explore the persistence of IoT-generated scanning campaigns by analyzing and comparing two instances of collected data over a course of one year. We also corroborate the evolutionary nature of IoT malware/botnets by highlighting newly targeted destination ports, which tend to include a larger set of possibly vulnerable destination ports/services.

The remainder of the paper is organized as follows. Section 2 reviews the recent literature on IoT threats and vulnerabilities. Section 3 provides details on the adopted methodology in terms of data collection and analysis. The IoT-generated scanning campaign detection methodology and characterization results are presented in Section 4. Finally, the main outcomes of this work are discussed in Section 6, followed by concluding remarks in Section 7.

2 Related Work

IoT device vulnerabilities have been discussed in the literature from different perspectives. For instance, Cui et al. [14], performed large-scale Internet scans of IoT devices and provided quantitative evidence on the vulnerable devices. They found over half a million publicly accessible embedded devices configured with factory default root passwords. Interestingly, this vulnerability was in fact one of the main reasons behind the large-scale outbreak of the Mirai botnet in late 2016 [4]. Considering the impact of vulnerability analysis in identifying and addressing IoT threats, Sachidananda et al. [15] deployed a testbed of IoT devices in an experimental setting and demonstrated a preliminary effort towards building a feasible and usable platform for IoT vulnerability analysis and testing.

From a different perspective, Costin et al. [16] provided an extensive assessment of IoT device firmware. Similarly, FIRMADYNE was proposed by Chen et al. [17] to automatically analyze Linux-based firmware images and identify vulnerabilities. A noticeable number of IoT security research work has been dedicated to synthesizing IoT context-aware permission models. For instance, Yu et al. [5] proposed a policy abstraction language that is capable of capturing relevant environmental IoT contexts, security-relevant details, and cross-device interactions, to vet IoT-specific network activities. Along the same research direction, Jia et al. [18], proposed ContextIoT, a system that is capable of supporting complex IoT-relevant permission models through efficient and usable program-flow and runtime taint analysis. Fernandes et al. [19] proposed a similar program-flow tracking approach that used taint arithmetic to detect policy violations and restrict traffic generated from exploited IoT applications. In the context of protocol vulnerabilities, Ur et al. [20] studied numerous types of home automation IoT devices and unveiled various insights with regards to the security and usability of the implemented access control models. Ronen and Shamir [21] demonstrated information leakage attacks by instrumenting a set of IoT smart lights.

Passive network traffic analysis is introduced as an effective approach towards studying Internet-wide cyber threats. For instance, passive DNS data, which consist of historic replicas of DNS queries and responses, was utilized to detect various threats associated with DNS abuse/misuse [22]. Furthermore, given the rareness of IoT-relevant empirical data, several recent efforts were proposed to create honeypots for collecting, curating, and analyzing IoT data. The first IoT-tailored honeypot, coined, IoTPOT, was designed and deployed by Pa et al. [23]. IoTPOT emulates Telnet services of various IoT devices running on different CPU architectures. In alternative work, Guarnizo et al. [24] presented the Scalable High-Interaction Physical Honeypot platform for IoT devices (SIPHON). The authors demonstrated how by leveraging worldwide wormholes and a few physical devices, they were able to mimic various IoT devices on the Internet and to attract significant malicious traffic. Luo et al. [25] implemented a machine learning approach to create an intelligent honeypot that automatically learns the behavioral responses of IoT devices through active scanning in order to mimic realistic interactions with attackers. U-Pot was introduced by Hakim et al. [26] as an interactive open-source framework for emulating IoT devices that support Universal Plug and Play (UPnP) protocols/services. In addition to the promising evaluation results in terms of emulating real IoT devices, the usability of the framework and its ability to automatically create honeypots from device description documents is worth noting. In a recent work, Vervier et al. [12] deployed a honeypot that captured a wider range of emerging IoT threats as compared to previous honeypots (e.g., IoTPOT). They used 6 months of collected data along with multiple sources of cyber-intelligence to explore current IoT malware and their emerging behavioral characteristics.

In addition to IoT-tailored honeypots, passive network telescope or darknet data, which represents one-way network traffic collected at unused IP addresses, has been adopted to analyze cyber activities and obtain cyber-intelligence [7, 8]. The idea of leveraging darknet to unused IP addresses for security purposes was first to light in the early 1990's by Bellovin for AT& Lab's Internet-connected computers [27, 28]. Since t focus of network telescope studies has shifted sever closely following the volatile nature of new adv More importantly, with the rise of IoT-driven cyber passive network telescope data was leveraged to and analyze unsolicited IoT scanning activities. For i Fachkha et al. [8] presented a probabilistic model itizing network telescope data and inferring orch probing campaigns towards Cyber-Physical System Furthermore, Antonakakis et al. [4] used unique Mir signatures to capture Mirai-related scans at the telescope for further analysis of the botnet.

This paper complements previous contributions by extending network telescope research to address the problem of detecting IoT threats, which propagate by identifying and exploiting vulnerable targets through Internet-scale scanning campaigns. To this end, the paper extends our previous work [13] by following a stratified methodology, which leverages data mining methods and unsupervised learning approaches, to detect coordinated scanning campaigns generated by compromised IoT devices. In fact, the analysis of these IoT-generated scanning campaigns unveil unique characteristics of the underlying threats/malware, which can leverage the detection and mitigation of various IoT threats. The paper also sheds light on the persistence of IoT threats over time while exploring the emergence of new, previously undocumented threats.

3 Approach

In this research, we aim at answering the following main research question: *How can we leverage passive network measurements to identify exploited IoT devices and infer distinctive characteristics of the underlying scanning campaigns induced by IoT-tailored malware/botnets?*

To answer the above question, we follow a multi-stage approach (Figure 1), which consists of two main components. First, we correlate IoT device information with darknet traffic to identify exploited devices and their scanning traffic (Section 3.3). Second, we identify IoT-generated scanning campaigns and investigate their characteristics by: (i) performing first-level clustering of compromised IoT devices using frequent pattern analysis and association rules mining to group devices that have similar objectives in terms of targeted ports/services (Section 4.1.1), and (2) implementing unsupervised learning techniques to perform second-level clustering of the grouped devices by leveraging a set of aggregated flow features (Section 4.2). Finally, while the outcomes represent characteristics of the IoT-generated scanning campaigns, we explore the persistence and evolution of these campaigns over time by analyzing newly collected IoT traffic and comparing results with our initial findings (Section 5). Further details on the used methodology is provided in the following sub-sections.

3.1 Data Collection

We follow a multifaceted data-driven approach, which involves analyzing data collected from different sources:

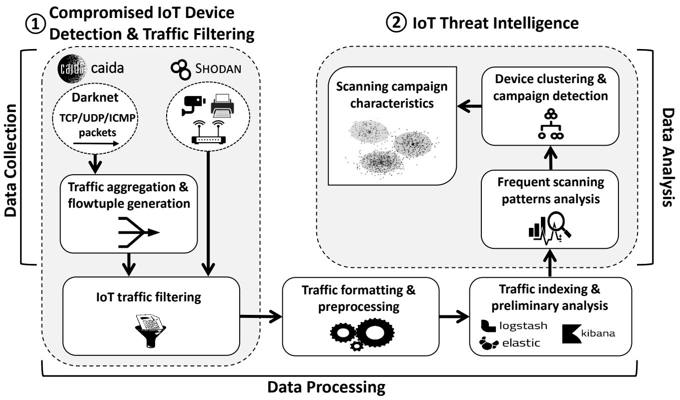

① Compromised IoT Device Detection & Traffic Filtering
② IoT Threat Intelligence
Data Collection
caida
SHODAN
Darknet
TCP/UDP/ICMP packets
Traffic aggregation & flowtuple generation
IoT traffic filtering
Traffic formatting & preprocessing
Traffic indexing & preliminary analysis
logstash
elastic
kibana
Frequent scanning patterns analysis
Device clustering & campaign detection
Scanning campaign characteristics
Data Analysis
Data Processing

Fig. 1. The overall approach for detecting and characterizing IoT threats.

IoT Device Information. We leverage a near real-time IoT database provided by `Shodan` [29]. This service executes large-scale active measurements to identify and index Internet-facing IoT devices. To this end, we obtained information related to 331,000 IoT devices from `Shodan`. These IoT devices were deployed in more than 200 countries. In this paper, we focus our analysis on a subset of all the IoT devices, namely stand alone devices that are deployed in the consumer realm. We obtained information related to approximately 181,000 IoT devices, including routers (46.9%), printers (29.1%), IP cameras (18.3%), and network storage media (4.6%). The remaining consumer IoT devices equate to only 1.1% of the total devices. These devices were deployed across 202 countries, with the U.S. hosting over 47,000 (24%) IoT devices, representing the country with the largest number of deployed IoT devices in our data. Both the U.K. and Russia followed the U.S. by a significantly less number of hosted devices (about 16,000), representing about 8% of all devices in each country.

Network Telescope Data (Darknet). Darknet data consists of one-way traffic targeted towards routable, allocated yet unused IP addresses (dark IP addresses). Since these IP addresses are not bound to any services, any traffic targeting them is characteristically unsolicited [9, 30]. Typically, darknet data consists of scanning and backscatter activities, in addition to other less common traffic such as misconfiguration and reflection attacks [9, 30–32]. In this paper, we initially explored 6 days of passive darknet traffic between April 12-17, 2017, representing 143 hours of darknet data (over 50 GB of hourly traffic). The darknet traffic is obtained from the UCSD real-time network telescope maintained by the Center for Applied Internet Data Analysis (CAIDA) [33]. It is one of largest available sources of passive darknet traffic with about 16.7 million globally routed destination IPv4 addresses that capture over a billion packets every hour. We processed about 3TB of darknet data to obtain more than 65M IoT-generated packets that were captured at the darknet during the initial analysis interval (April 2017). In addition, we utilized the darknet to collect new IoT-generated traffic over 5 days in May, 2018 (Section 5). Overall, more than 6TB of darknet data was processed during both analysis intervals, resulting in capturing approximately 172M IoT-generated packets.

3.2 Data Processing

The packets captured at the darknet are processed using the `Corsaro` tool, which is a software suite for performing large-scale analysis of trace data [34]. We used `Corsaro` to obtain hourly "flowtuple" files, representing information about incoming flows towards the darknet. Each flow illustrates incoming packets from a source IP to a darknet IP address during one minute time intervals, encompassing the following flow information: source/destination IP addresses and used ports, protocol, Time To Live (TTL), TCP flags, IP length, and total number of packets (per minute). To infer compromised IoT devices, we executed a correlation algorithm that leverages IP header information IoT device information with darknet data to filter out IoT-generated traffic. Finally, the acquired hourly traffic via filtering is prepared in tabular format (flowtuple files) and fed into the search and analysis engine (Figure 1), which is implemented using the `ELK Stack` [35]. More specifically, we used `Logstash` for importing data into `Elasticsearch`, which is utilized for flow indexing and analysis. We also used the `Kibana` visualization and navigation tool to run queries and generate corresponding data sets that are used for further analysis throughout the paper (Figure 1). In what follows, we provide further information on our methodology and obtained results.

3.3 Preliminary Analysis (Initial Data Set)

We identified 15,299 unsolicited IoT devices that were correlated with the darknet during the initial analysis period (April 2017). The identified devices generated different types of traffic towards the darknet [13], among which about 80% were TCP-SYN flows. While there are several ways for scanning the Internet, in this paper, we focus our analysis on TCP-SYN scans, as they represent the most prominent method of scanning [10, 36]. It is also important to understand that ICMP Echo requests, which are also commonly used for network scans, are excluded from further analysis due to their negligible magnitude in the overall data (0.23%). In addition, the stateless UDP packets (about 8%), which require further investigation of the packet payload to identify their nature (e.g., scanning vs. non-scanning), are also excluded from further analysis throughout the paper.

Compromised IoT Devices. The analysis of recent large-scale cyber attacks caused by the `Mirai` botnet and its later variants [1, 4], demonstrated the role of malware-infected IoT devices within coordinated botnets, which are used for scanning the Internet for vulnerable hosts. Given that a benign IoT device has no justifiable reason for scanning the Internet, from here onwards, we label these unsolicited devices as "compromised" or "exploited" IoT devices. Accordingly, we identified 6,797 compromised IoT devices that generated about 54.6M TCP-SYN scanning packets towards the darknet, as illustrated in Figure 2. In general, these exploited devices scanned less than 200 unique destination ports per hour, except at interval 119, where we noticed an abrupt increase in the total number of scanned ports (Figure 2). Further analysis at interval 119 showed that a single compromised IP camera located in the Dominican Republic was performing a typical vertical scan of over 7,400 ports on 55 destination addresses.

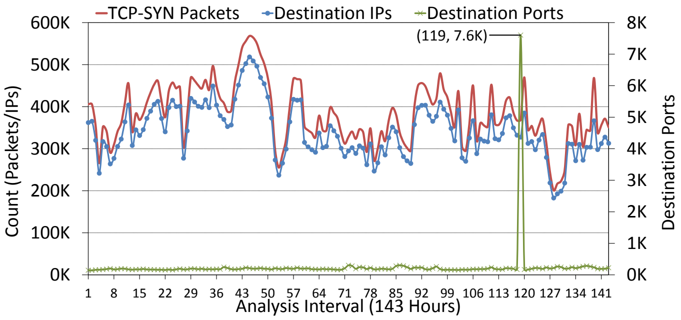


Fig. 2. The distribution of all TCP-SYN scanning packets generated by compromised IoT devices during the analysis interval (143 hours) [13].

TABLE 1
Top 15 scanned services/ports (CP=98.7%). "Src." and "Dst." IP counts represent the number of IoT devices and scanned IP addresses.

Service/Port	Packets		IP count	
	(M)	%	Src.	Dst. (M)
Telnet/23	**29.88**	**54.71**	640	12.75
HTTP/80	5.62	10.29	1,223	4.25
Unassigned/81	2.61	4.83	1,079	2.40
Kerberos/88	2.64	4.78	889	2.40
SSH/22	2.59	4.74	64	2.32
WSDAPI-S/5358	2.39	4.37	89	2.18
CWMP/7547	2.01	3.7	169	1.88
Alt. Telnet/2323	1.84	3.37	199	1.69
MS-SQL-S/1433	1.21	2.21	7	0.71
SMB/445	1.13	2.06	51	0.67
iRDMI/8000	0.66	1.21	875	0.65
HTTP/8080	0.66	1.20	1,053	0.61
EthernetIP/2222	0.28	0.52	53	0.28
RDP/3389	0.24	0.44	39	0.12
FTP/21	0.13	0.24	21	0.06

Scanned Ports and Services. The analysis of the number of scanned destination ports indicates that the majority of the compromised IoT devices (90.6%) tend to scan less than 10 unique destination ports. In fact, about half of all devices were found to scan no more than 2 ports, while on the other hand, only about 5% of all IoT devices scanned more than 20 ports. Indeed, this behavior reflects a unique characteristic of the majority of the compromised consumer IoT devices, which were utilized to scan a handful of known ports/services, something that was different in comparison to other IoT devices in the CPS [13]. In addition, the analysis of the top 15 scanned destination ports, which contribute towards 98.7% of all scanning packets, indicates that Telnet/23 was scanned by the highest number of packets (about 54.7%), as presented in Table 1. Despite that, we notice that some ports such as 80, 81, 88, 8000, and 8080, which received smaller number of scanning packets, were in fact targeted by a relatively larger number of compromised devices, as compared to Telnet. These differences in terms of the number of involved IoT devices in scanning certain ports, along with the scanning rate and total targeted destinations, might indicate distinctive characteristics of the malware-infected IoT devices and their generated scanning activities, which are investigated further throughout the paper.

3.4 Limitations

The generalizability of our findings might be hampered by the nature of our data (IoT device information and darknet). Considering the limited empirical data on existing IoT devices, we used our resources to sample data from `Shodan` by focusing on consumer IoT devices [29]. Nevertheless, while our sample (about 300,000 IoT devices) is not representative of the overall population of the IoT devices on the Internet, it serves well in supporting our research objectives and generating insights that could be used as a basis for future research towards understanding the activities of malicious/compromised IoT devices. In addition, the darknet represents one-way traffic captured at a slice of the entire IPv4 Internet address space. Despite that, the UCSD network telescope data used in this research provides about 16.7 Million destination IP addresses (/8 address space), which is one of the largest available sources of darknet data for research purposes [33].

Another limitation of the work is that the initial data was collected in April 2017, and some of the compromised IoT devices might have been already cleansed. Furthermore, due to DHCP churn [37], the associated IP addresses to those IoT devices might have changed over time. Nevertheless, a comparison of the new list of IoT device information from `Shodan` with the initial IoT device information shows that about 99% of the devices in our initial data were still actively connected to the Internet (on May 2018). Finally, the identification of the exact IoT device type is a challenging task as some of these IoT devices are assigned with dynamic IP addresses. Further, it is common to have IoT devices operating behind a gateway or router (using port forwarding), and therefore, while the associated IP addresses might depict an IoT device, they might be also representing the public IP address of the gateway.

4 IoT-Generated Scanning Campaigns

In this paper, we propose an approach for detecting malware-infected IoT devices and characterizing the underlying IoT-generated scanning campaigns. The assumption is that compromised IoT devices are likely to perform similar malicious reconnaissance activities within orchestrated scanning campaigns [4, 12, 36, 38]. Given our initial data set (April 2017), we follow a multi-stage clustering/classification approach to identify groups of IoT devices that tend to behave in a similar manner. Our aim is twofold. Firstly, to identify scanning objective(s) by finding unique sets of scanned ports, thus contributing towards campaign intent analysis. Secondly, given groups of compromised IoT devices with common objectives, we perform subspace clustering using a set of raw and aggregate flow features to identify compromised IoT devices with similar objectives and behavioral characteristics.

4.1 Scanning Objective(s)

We identify scanning objective(s) by exploring the targeted destination port sets by compromised IoT devices. This is considered as the first step towards inferring scanning campaigns, as discussed in the next sub-section. Furthermore, given that IoT-tailored malware are likely to target a small number of vulnerable ports/services, identifying the scanning objectives is key to attributing the inferred scanning campaigns to known IoT malware, as discussed in Section 4.4. Scanning objectives are identified as follows:

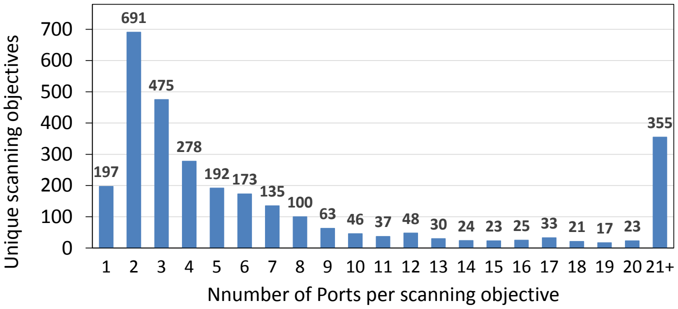


Fig. 3. The distribution of unique scanning objectives over the number of scanned ports.

Let $D = \{d_1, d_2, \ldots, d_N\}$ be a set of N identified compromised IoT devices that sent TCP-SYN scanning packets to the darknet during the analysis interval E. Let $P = \{p | 0 \leq p \leq 65535\}$ be a set of all TCP ports. For every compromised IoT device $d_i \in D$, we determine scanning objective S_i as a set of all scanned ports $P_{Si} \subseteq P$. Note that these port sets do not account for the order in which the ports were scanned. Let $S = \{S_1, S_2, \ldots, S_N\}$ be a set of N identified scanning objectives for all compromised IoT devices. Given that IoT devices infected by the same malware are likely to produce similar scanning objectives, we define $S_{unique} = \{S_1, S_2, \ldots, S_k\}$ as a set of all distinct scanning objectives ($S_{unique} \subseteq S$).

We identified $k = 2,986$ combinations of targeted destination ports, representing unique scanning objectives. As shown in Figure 3, the distribution of the unique scanning objectives over the number of targeted ports indicates that compromised IoT devices are likely to target a small number of vulnerable ports/services, with about 88% of all scanning objectives containing less than 21 destination ports. This is an interesting characteristic of consumer IoT devices, especially when compared to other IoT devices, such as those deployed in the CPS, which tend to target a larger number of ports/services [13]. It is also important to understand that each scanning objective may correspond to the behavior of one or more compromised IoT devices. In fact, we found that about 60% of all compromised IoT devices produced 227 scanning objectives that were common among two or more devices. On the other hand, while each one of the 2,759 remaining IoT devices was associated to its unique scanning objective, these scanned port sets were very similar, and in many cases they represented subsets/supersets of other scanning objectives. This provides yet another indication that many IoT devices are in fact following similar scanning behaviors in terms of the targeted ports/services throughout the analysis interval.

4.1.1 Scanning Classes

We examined the identified scanning objectives S_i to find similarities in the behavior of compromised IoT devices. The results reflect three classes of mutually exclusive scanning behaviors, as described in the following sub-sections:

Range Scans. This class represents the behaviors of IoT devices that targeted destination ports within distinctive ranges. For instance, the analysis revealed about 4,536 (66.7%) compromised IoT devices that were mainly scanning ports within the following ranges: 19328–19622 and 36224–36582. Given the distinctive behavior in terms of scanning these uncommon port ranges, it is highly likely that the involved compromised IoT devices were in fact driven by similar IoT malware/botnet. To the best of our knowledge, these port ranges are not associated with known IoT malware/botnets, and therefore, the behaviors of the compromised IoT devices might indicate an emerging IoT malware that targets new, or uncommon vulnerabilities.

Moreover, only a handful of known services are registered within the identified port ranges [39]. For instance, TCP ports 19410–19412 are associated with HP services, while the Java Control Panel (JCP) Client is registered on port 19541. Interestingly, ports 19539–19540 are associated with Silex wireless and USB drive adapters [40], which enable wireless connection and network sharing capabilities on many devices such as printers, scanners, and disk drives, to name a few. These adapters use the "SX-Virtual Link" software developed by Silex Technologies to add sharing capabilities on different operating systems (e.g., Windows and Linux), in addition to other embedded devices (e.g., wireless routers). While we do not have conclusive evidence on the actual targets of the scanning campaigns as they targeted different ports within the specified ranges, these findings may shed light on possible intentions of the emerging IoT malware and its targeted devices/vulnerabilities.

Furthermore, about 5% of the IoT devices within this class were also scanning other known ports, with about half of them scanning one or more of the following ports: HTTP/80/8080, Unassigned/81, Kerberos/88, iRDMI/8000, and HTTPS/443. In addition, a small number of devices (16) scanned Telnet/23 along with other known services such as Alternative Telnet/2323, SSH/22, WSDAPI-S/5358, CWMP/7547, and EthernetIP/2222. It is worthy to note that these ports are associated with known IoT malware/botnet (e.g., Mirai [4] and Hajime botnets [41]). Nevertheless, having these ports scanned along with the specified port ranges in this scanning class gives us a clear indication of an evolving IoT malware/botnet, which is targeting new vulnerabilities. However, proving this requires further investigation, which is beyond the scope of this paper and will be considered in future work.

Strobe Scans. The analysis of recently discovered IoT malware/botnets showed that compromised IoT devices were utilized to scan a relatively small number of vulnerable ports/services. We classify these scanning behaviors as strobe scans. In line with that, about 31.5% (2,144) of the compromised devices within our initial data were performing strobe scans, targeting less than 7 ports. In general, these devices targeted 40 different ports/services, among which, HTTP/80 was scanned by the largest number of exploited devices (54.5%). In addition, as illustrated in Figure 4, almost all of the top scanned ports are associated with known services that run on IoT devices to enable common operations such as information sharing (e.g., HTTP/80/8080), remote login (e.g., Telnet/23/2323), and communication (e.g., SSH/22), to name a few. It is also clearly observed that a significantly larger number of compromised devices were scanning the first six destinations ports (80–23), as compared to the remaining destination ports.

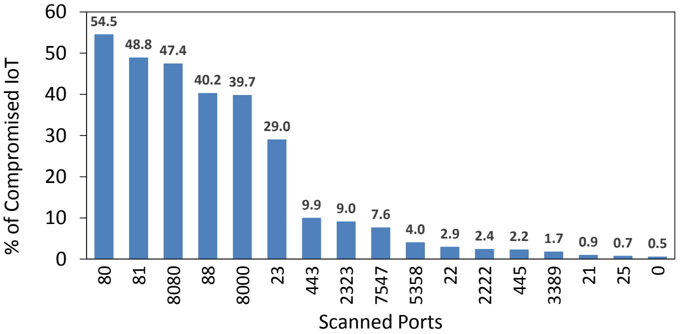


Fig. 4. Top 17 scanned destination ports by the highest number of IoT devices within strobe scans.

In addition, the analysis highlights 89 unique scanning objectives within the identified strobe scans (S_{strobe}). A summary of the most frequent scanning objectives within S_{strobe} is presented in Table 2 . It is worth noting that the first scanning objective was common among a large number of IoT devices (38.3%), followed by a relatively less number of IoT devices that targeted the remaining port sets. At this stage, the analysis of the frequent scanning objectives highlights specific intentions of the compromised IoT devices and their targeted ports/services, which represent unique characteristics of the underlying IoT malware/botnets. Furthermore, it is clearly observed that some of the scanned ports were likely to be scanned together as they appeared in several scanning objectives (e.g., HTTP ports 80 and 8080). To explore the correlations between the identified scanned ports (S_{strobe}), we perform association rules mining [42]. Let X and Y be two scanned port sets that belong to $S_i \in S_{strobe}$. An association rule X→Y describes the probability of port set Y being scanned given that port set X was probed. The *Support* of a rule is the count of the patterns in S_{strobe} that contain X∪Y. The *Confidence* of a rule is the support of the rule divided by the number of patterns that contain only X.

As shown in Table 3, we provide a sample of the association rules related to the frequent scanning objectives (S_{strobe}) identified in Table 2. As described through rules 1 to 5 in Table 3, there is a high correlation ($Conf. > 99\%$) between scanned ports within S_1 (ports 80 81 88 8000 8080). Furthermore, association rules 6 and 7 show a strong correlations between both ports 7547 and 2323, and port 23 (Table 3). This means that if either ports 7547 or 2323 is probed, there is a high chance that port 23 is also going to be probed. Nevertheless, the opposite rules (e.g., 23→7547) were not significant ($Conf. < 85\%$), which means that having port 23 scanned does not necessarily require scanning ports 7547 and/or 2323. The remaining association rules presented in Table 3 corroborate the high correlation among scanned ports within the scanning objectives presented in Table 2. Therefore, we may conclude that these frequent scanning objectives S_{strobe}, which represent the targeted destination ports/services by compromised IoT devices over a period of time, could reflect unique characteristics of the underlying scanning campaigns. We will elaborate on this in Section 4.2.

Wide Scans. In contrary to the range and strobe scanning classes, the remaining identified scanning activities were targeting a variable number of destination ports and IP ad-

TABLE 2
Top 20 frequent scanning objectives within S_{strobe} generated by about 94% of all devices in strobe scans class.

S_i	# Devices	%	Scanned Ports
1	**821**	**38.3**	80 81 88 8000 8080
2	187	8.7	23
3	160	7.5	81
4	152	7.1	23 7547
5	139	6.5	23 2323
6	110	5.1	80 443 8080
7	82	3.8	80 443
8	80	3.7	23 5358
9	74	3.5	80
10	43	2.0	445
11	40	1.9	22 23 2222 2323
12	32	1.5	3389
13	23	1.1	80 8080
14	21	1.0	80 81 8080
15	14	0.7	21
16	13	0.6	25
17	12	0.6	443
18	8	0.4	0
19	7	0.3	81 88 8000 8080
20	7	0.3	8080

TABLE 3
Association rules related to the scanned ports identified in Table 2.

ID	Rule	Support	Conf. (%)
1	81 88 8000 8080 → 80	829	99.2
2	80 88 8000 8080 → 81	829	99.9
3	80 81 8000 8080 → 88	829	99.5
4	80 81 88 8080 → 8000	829	99.5
5	80 81 88 8000 → 8080	829	99.3
6	7547 → 23	161	98.8
7	2323 → 23	193	99.5
8	443 8080 → 80	113	98.3
9	443 → 80	198	93.0
10	5358 → 23	83	96.5
11	23 2222 2323 → 22	44	100.0
12	22 2222 2323 → 23	44	97.8
13	22 23 2323 → 2222	44	100.0
14	22 23 2222 → 2323	44	97.8
15	8080 → 80	997	98.1
16	80 → 8080	997	85.4
17	81 8080 → 80	859	99.0
18	80 8080 → 81	859	86.2
19	80 81 → 8080	859	98.7
20	88 8000 8080 → 81	836	99.9
21	81 8000 8080 → 88	836	99.5
22	81 88 8080 → 8000	836	99.4
23	81 88 8000 → 8080	836	98.8

dresses. We classify these scans as wide scans, as they tend to target a large number of randomly scanned destination ports over the analysis interval. Furthermore, the scanned ports span over all existing ports, including reserved well-known ports that are assigned to widely used services (0–1023), other less commonly used registered ports (1024–49151), and dynamic ports (49152–65535).

We identified 117 IoT devices that implemented different strategies to perform wide scans. It is clearly observed that utilizing exploited IoT devices to perform wide scans is not very likely, as illustrated by the significant difference in the number of involved IoT devices when comparing wide scans with other scanning classes. Despite that, we detected an IP camera from the Dominican Republic that scanned more than 7,000 ports during a short period of time (interval 119-Figure 2). These typical port scanning behaviors (e.g., vanilla or sweep scans) might be easily detected by existing defensive measures as they tend to target a large set of ports and IP addresses. On the other hand, adversaries try to evade detection by implementing a combination of scanning techniques in a randomized and stealthy manner. For instance, a printer located in Taiwan scanned 1,122 ports on 1,132 destination IP addresses throughout the analysis intervals. In fact, almost all exploited devices within this class (except the IP camera from the Dominican Republic) were performing scans with a relatively small average scanning rate (about 88 packets per hour). This however, might reflect the behaviors of the majority of compromised IoT devices that were performing wide scans, as they were active (undetected) for a relatively long period of time.

4.1.2 Involved IoT Devices

The analysis of the involved IoT devices per scanning class illustrates that range and strobe scans contribute to the largest number of compromised IoT devices, with about 66.7% and 31.6% of all devices, respectively. More importantly, the distribution of the IoT device types per scanning classes highlights a noticeable difference between range and strobe scans, with range scans to contain a significantly larger number of routers and printers, while strobe scans containing a relatively larger number of IP cameras, as illustrated in Figure 5. From a different perspective, while Russia hosted the largest number of overall compromised IoT devices (2,169 devices), it also contributed to the largest number of devices that performed range (46%) and wide scans (29%), respectively. As illustrated in Figure 6, it is also clearly observed that the majority (96.4%) of the devices hosted in Russia belong to range scanning class. Similarly, the majority of devices located in China (93.3%), S. Korea (85.4%), and the Philippines (84.4%), were performing range scans, while the behaviors of most of the devices hosted in Thailand (80.3%) and Singapore (75.7%) were classified as strobe scans. Indeed, the distribution of IoT devices per scanning classes, device types, and hosting countries (Figures 5–6), reveals differentiating characteristics of the underlying scanning activities generated by compromised IoT devices. Nevertheless, while it is difficult to find the exact reason for such dominant scanning behaviors in different contexts, the analysis shed light on important characteristics of the underlying IoT malware in terms of the targeted vulnerable device types (or services), and the countries in which these devices are deployed the most. Confirming this assumption requires further investigations, which is considered for future work.

4.2 Campaign Detection

To detect scanning campaigns, which represent the behaviors of well-coordinated botnets operating “in the wild,” we leverage our knowledge on compromised IoT devices with similar scanning objectives and classes, and utilize their behavioral characteristics to perform clustering following an unsupervised learning approach. We leverage the Density Based Spatial Clustering of Application with Noise (DBSCAN) [43]. This algorithm is widely adopted as it does not require a priori knowledge about the number of

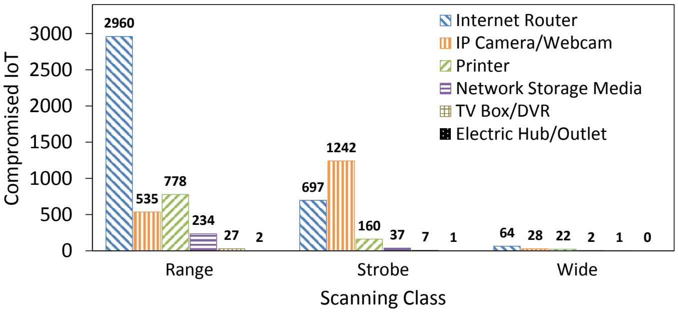


Fig. 5. Distribution of compromised IoT device types per scanning class.

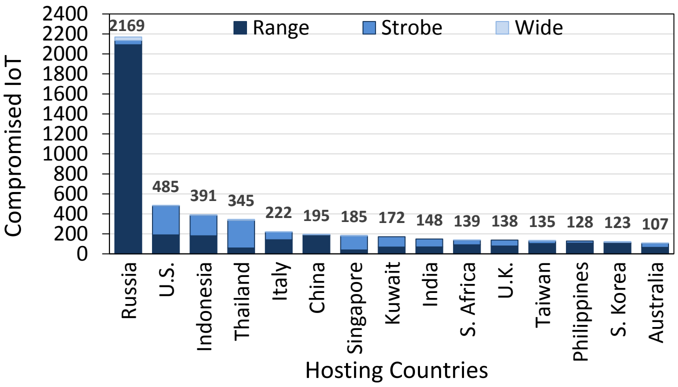


Fig. 6. Countries with the largest number of exploited IoT devices from each class (initial data–April 2018).

clusters, while it can detect arbitrary shaped clusters and outliers [44]. Given a set of points in a specified space, DBSCAN groups neighboring points if they form a cluster with a minimum number of points *MinPts* that are reachable within a predefined radius ϵ. DBSCAN can be used with any distance function, however, in this paper, we adopt the *Euclidean distance* for further analysis using R statistical analysis tools. It is worth noting that using DBSCAN requires adjusting the initial values of ϵ and *MinPts*, which is not a straight forward task as it requires extra measures to select the appropriate values in different settings. In what follows, we provide further information on the feature selection process and the results.

4.2.1 Flow Features

Features selection and extraction is a complicated part of unsupervised learning approaches, which has no unique prescribed solution. Let $F_d = \{f_{d1}, f_{d2}, ..., f_{dN}\}$ be a set of aggregate flows corresponding to N compromised IoT devices in the analysis time interval E. Each aggregate flow $f_{di} \in F_d$ is described by a set of β flow attributes or features. It is important to understand that when using unsupervised classification approaches, we can not apply standard feature extraction methods to validate the optimal number of required features. Therefore, we leveraged the literature to obtain a set of widely used traffic features (e.g., packet rate) [45, 46], along with raw and aggregate flow features from our data analysis. Our analysis resulted in selecting $\beta = 10$ features that are summarized in Table 4. These features are extracted from the raw flow information and aggregated throughout the analysis period, which represents 143 hourly intervals (6 days). Note that the list of features is not conclusive and we can always add or remove features to improve the clustering results (if necessary).

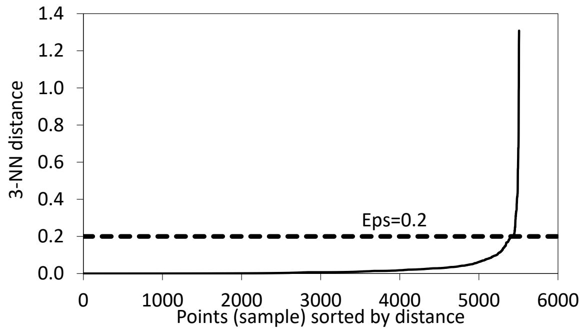


Fig. 7. An example of K-NN distance graph for IoT devices classified within strobe scans.

TABLE 4
The selected flow features for analysis using DBSCAN ($\beta = 10$).

β	**Selected Features**
1	number of active intervals (hours)
2	per hour packet rate
3	ratio of TCP-SYN packets to non-backscatter packets
4	per destination address packet rate
5	per source port packet rate
6	average number of used source ports per hour
7	average length of the IP packet (from IP header)
8	number of TCP-SYN packets
9	number of scanned destinations
10	number of scanned destination ports

4.2.2 Procedure

We use DBSCAN for inferring scanning campaigns within the identified scanning classes (range, strobe, and wide). To reduce noise and enhance the overall results, we filter out IoT devices that sent less than 10 packets to the darknet during the analysis period. The extracted features are then normalized and prepared to be used in DBSCAN by applying unitization with zero minimum ($x_{norm.} = (x - min)/range$). Moreover, we set $MinPts = 3$ as we assume that a campaign consists of three or more IoT devices that scan the Internet for certain vulnerabilities. To identify the values of ϵ, we perform the K^{th}-Nearest Neighbor (K-NN) distance analysis with $K = MinPts$. Given a sample of N points, we calculate the distances between every point and its K nearest neighbors. The resulting NxK calculated distances are then sorted in ascending order to illustrate the K-NN distance plot (Figure 7), with the Y-axis to represent the calculated distance values for all NxK data points (X-axis). Note that choosing a very small ϵ will cause a big portion of the sample to be unreachable via other points, and thus not clustered. On the other hand, choosing a very large ϵ will result in grouping the majority of the sample into a single cluster. Therefore, to ensure covering the majority of the data points in the clustering analysis, a reasonable value for ϵ is selected at the point where we observe the beginning of a sharp increase in the values of the calculated K-NN distances, as depicted by the provided example in Figure 7.

We set ϵ to be 0.15, 0.2, and 0.3 for range, strobe, and wide scans, respectively. Given the selected values of

MinPts and ϵ, we perform DBSCAN clustering analysis on devices within each scanning class and report results in the following sub-sections.

4.2.3 Cluster Evaluation

In contrary to supervised learning approaches, cluster evaluation or validation methods are not well developed for unsupervised learning approaches. There are a number of common approaches that are traditionally used to evaluate/validate the clustering results. Nevertheless, clustering evaluation is highly application dependent and thus, subjective. In this paper, we use "Internal Measures" such as cluster *cohesion* and *separation* to evaluate our clustering results. Cluster cohesion measures how closely related are objects in a cluster. It is represented by the average within distances among objects of clusters. Cluster separation, on the other hand, measures how distinct or well separated a cluster is from other clusters, and is presented as the average between distance among different clusters.

We analyzed the intrinsic characteristics of the clustering and summarized the evaluation results in Table 5 (Cluster Evaluation). Considering that the results are normalized (0.0–1.0), we want the average within distance to be as small as possible, while having a larger average between distance is always preferable. As summarized in Table 5, the resulting average within distances for all evaluated scanning classes is reasonable, with values equal to about 0.38, 0.46, and 0.36 for the three classes, respectively. In addition, the average between distances show that the resulting clusters are well distanced from each other in all classes, with an average of about 0.86 (range), 0.88 (strobe), and 0.89 (wide). Overall, while it is difficult to have perfect clustering, the evaluation of the resulting clusters in terms of *cohesion* and *separation* is reasonable. In what follows, we present detailed results in terms of the identified clusters and the underlying IoT-generated scanning campaigns.

4.2.4 Clustering Results

As summarized in Table 5, we identified 18 clusters of exploited IoT devices that participated in scanning campaigns. These clusters, which represent groups of correlated IoT devices with similar scanning objectives and behaviors, are illustrated in Figures 8 (a)–(c). Note that the clustering is performed based on 10 feature (dimensions), among which features 3 and 4 (Table 4) were selected to illustrate the clusters. Therefore, although the clusters are mutually exclusive, they might look overlapping in the 2-dimensional Figures 8 (a)–(c). In addition, outliers, which were not grouped with any of the existing clusters, are represented as isolated black dots in Figures 8 (a)–(c). In what follows, we discuss the the characteristics of the identified scanning campaigns with respect to each scanning class.

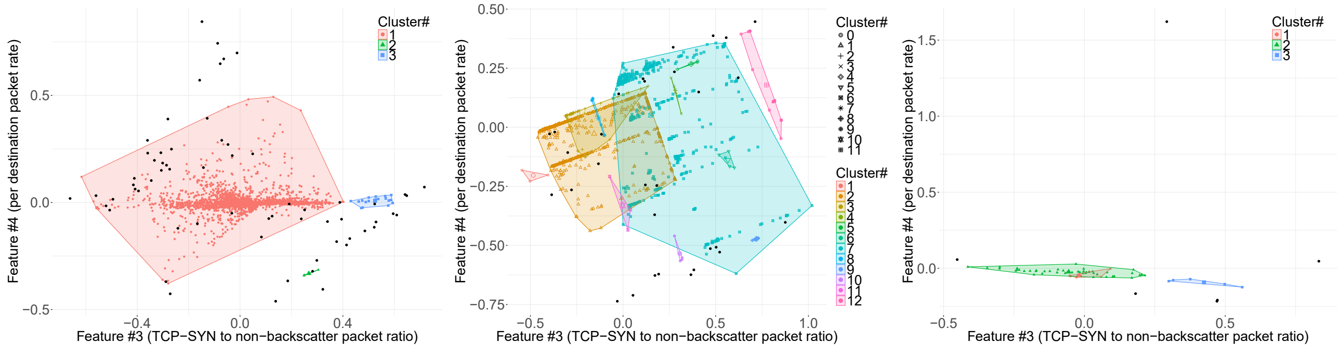
Range Scans. The majority of the exploited IoT devices (about 96.8%) within the range scanning class were correlated under cluster #1, as depicted by the largest cluster in Figure 8a. These flow similarities confirm our initial classification according to common scanning objectives (Section 4.1), which highlight the correlation among compromised IoT devices that target similar port ranges (range scans). Moreover, considering that these port ranges are not associated with commonly used services or targeted vulnerabilities, they may in fact reflect a unique characteristic of the underlying IoT malware/botnet. In addition, we noticed differences in the distribution of device types when comparing cluster #1 (Figure 9a), with clusters #2 (3 IP cameras) and #3 (about 70% IP cameras and 30% routers), respectively. Furthermore, the scanning behaviors were also found to be slightly different when comparing the clusters, with devices within clusters #2 and #3 to be mainly scanning objectives of known destination ports along with the identified port ranges (19328–19622). In fact, 13 out of the 21 devices within clusters #2 and #3 were scanning ports 80, 81, 88, 8000, and 8000, representing the first frequent scanning objective (S_1) from Table 2, while the remaining were scanning a combination of Telnet/23 and other ports. This however, gives us yet another clue about the characteristics of the underlying IoT malware/botnet, which behave differently, as reflected by the common scanning objectives within the campaigns. Another interesting characteristics that may differentiate between the identified clusters is the average ratio of TCP-SYN to non-backscatter packets, with a value of about 0.59 for cluster #1, and about 0.96 for clusters #2 and #3. This indicates that on average, devices within cluster #1 were involved in sending a noticeably higher ratio of non-backscatter packets, such as ICMP-REQ and/or UDP packets, as compared to clusters #2 and #3.

Strobe Scans. As summarized in Table 5, the analysis resulted in identifying 12 clusters within the strobe scanning class, with clusters #7, #2, and #3 having the largest populations, respectively. The initial analysis of the identified clusters in Figure 8b, showed that the clustering results are highly dependent on the number of scanned destination ports within the scanning objectives (feature 10). For instance, cluster #1 consists of IoT devices that scanned 6 destination ports, while devices in cluster #2 scanned 5 ports. Moreover, almost all clusters consist of devices that scanned equal number of ports, except for cluster #7, which contained devices with variable number of scanned ports (1–3 ports). Therefore, although the number of scanned ports specified in the scanning objective might not be a characterizing factor by itself, it can reflect an abstract view of the scanning behavior in terms of the total number of targeted ports/services, which is an important characteristic of the IoT-generated scanning campaigns.

In addition, the investigation of the targeted ports/services highlighted similar scanning objectives among a considerable number of the exploited devices within most of the identified clusters. For instance, the vast majority (99.3%) of devices within cluster #2 were only scanning ports 80, 81, 88, 8000, and 8080, represented by S_1 in Table 2. Furthermore, about 66% of IoT devices within cluster #3 scanned S_{11} (ports 22, 23, 2222, and 2323), while about 33% of the devices scanned combinations of ports that are subsets of S_1 (e.g., ports 80, 81, 8000, and 8080). On the other hand, devices within cluster #7, which represents the largest cluster within the strobe scanning class, generated over 30 different scanning objectives, among which, about 56% were associated with Telnet (e.g., ports 23 and 7547). These results indicate that despite the reasonable grouping of correlated IoT devices based on their

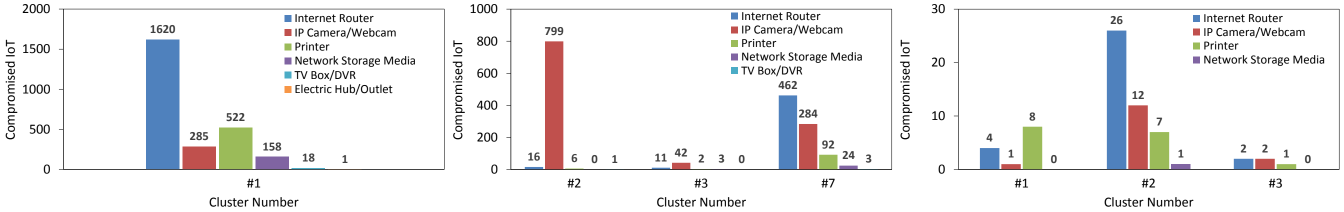
TABLE 5
Summary of the clustering results and evaluation (*MinPts*=3 and $\beta = 10$ features).

	Clustering Results					Cluster Evaluation	
Class	**Device**	ϵ	**Cluster**	**Cluster Size**	**Outliers**	**Within Distance**	**Between Distance**
Range	2,688	0.15	3	2604, 3, 18	63	0.379	0.859
Strobe	1,836	0.2	12	3, 822, 58, 4, 4, 4, 865, 11, 11, 4, 3, 17	30	0.457	0.879
Wide	71	0.3	3	13, 46, 6	6	0.363	0.889



(a) Range ($\epsilon = 0.15$, clusters=3) (b) Strobe ($\epsilon = 0.2$, clusters=12) (c) Wide ($\epsilon = 0.3$, clusters=3)

Fig. 8. Clustering results ($MinPts = 3$) for (a) range, (b) strobe, and (c) wide scanning classes. The x- and y-axis represent features 3 and 4.



(a) Range (b) Strobe (c) Wide

Fig. 9. The distribution of IoT device type in the largest clusters within (a) range, (b) strobe, and (c) wide scanning classes.

aggregate flow features, the clustering algorithm will not be always sufficient to detect distinctive scanning campaigns within strobe scanning class. Therefore, to overcome this limitation and group IoT devices into meaningful scanning campaigns, it is necessary to consider a combination of the clustering and common scanning objectives.

Wide Scans. The analysis of IoT devices within the wide scanning class, which involved a significantly fewer number of compromised devices (71), resulted in three correlated clusters (Table 5). These clusters of IoT devices, which were grouped based on similarities in their aggregate flow features, are illustrated in Figure 9c. It is worthy to note that the nature of the underlying scanning campaigns in terms of variable length of the scanning objectives, along with the randomness in the targeted destination ports, makes it extremely difficult to associate these IoT devices with unique IoT malware/botnet. Nevertheless, by analyzing the aggregate features with respect to the IoT devices within each cluster, we found a significant difference in the ratio of TCP-SYN packets to non-backscatter packets, with an average value of about 0.98, 0.60, and 0.20, for the three clusters respectively. In addition, while cluster #2, which represents the largest group of exploited IoT devices within the wide scanning class (about 65%), consist of a relatively larger number of routers and IP cameras, cluster #1 contained slightly more infected printers (50%), as illustrated in Figure 9c. These results corroborate that exploited devices from the same type are likely to generate similar scanning behaviors and therefore, forming clusters of correlated devices that operate within different scanning campaigns.

4.3 Results Summary

The analysis of the identified scanning campaigns generated by compromised IoT devices provides insights on the behavioral characteristics of the underlying IoT malware. For instance, the analysis revealed common scanning objectives that represent possibly vulnerable destination ports/services. Furthermore, the analysis highlighted behavioral similarities in terms of the aggregated flow features. Together, these similarities were effectively used to uncover groups of IoT devices that were likely to be infected by similar IoT malware, as reflected from their scanning activities on the darknet.

Moreover, the analysis revealed that the ratio of the generated TCP-SYN scanning packets to non-backscatter packets (e.g., ICMP-REQ and UDP) is in fact a differentiating feature when characterizing the behaviors of exploited IoT devices within different scanning classes/campaigns. For instance, a Kruskal-Wallis rank sum test with pair-wise comparison tests (Bonferroni adjustment $p_{bonf.} = 0.016$ showed statistically significant differences ($p < 0.0001$) means in the ratio of TCP-SYN to non-backscatter pack when comparing strobe scanning class with other classes, spectively. Indeed, devices within the strobe scanning cl were mainly sending TCP-SYN packets (average ratio about 0.99), and therefore, highlight a unique characteri that can distinguish them from other devices.

From a different perspective, the prevalence of cert IoT device types within the identified scanning campai determine a feature of the underlying IoT malware/botr which is tailored to exploit certain vulnerable devices. instance, clusters #2 and #3 within the strobe scanning cl consist of mainly IP cameras. Nevertheless, it is interest to see that devices withing these clusters targeted differ destination ports, with the majority of devices within clus #2 and #3 to target S_1 and S_{11} (Table 2), respectively. Wh we do not have concrete information on the actual malware/botnet that is generating these scanning campaigns, these behaviors can in fact illustrate the emergence and evolution of IoT-tailored malware, which tend to target multiple vulnerabilities on the targeted devices.

A main characteristic that differentiates between scanning campaigns is the scanning objective, which reveals the targeted ports that relate to existing vulnerabilities. More importantly, while these targeted ports are usually associated with known malware/botnets, the identification of scanning campaigns that target uncommon ports (e.g., range scanning class), which are not associated with known vulnerabilities can be utilized to predict and mitigate emerging IoT malware.

4.4 IoT Malware Attribution

To validate our approach in terms of detecting scanning campaigns based on common scanning objectives, we collected more than 9,000 real IoT malware executables and performed multiple experiments to extract real IoT malware traffic. Our objective herein is to corroborate findings from analyzing the darknet and attribute the identified scanning campaigns to known IoT malware/botnets. In what follows, we elaborate on the data collection methodology, experimental setup, and results.

4.4.1 Data Collection

We leveraged the data collected by an IoT-based honeypot (`IoTPOT` [23]) to acquire about 8,000 samples of IoT-specific malware. We also extracted about 1,000 samples of IoT-related Mirai and Bashlite malware executables from a generic online malware repository (VirusShare.com). It is important to realize that we performed a number of pre-processing steps to filter out corrupted malware samples from our experiments. Furthermore, due to our sandbox environment limitations, we had to discard malware samples that did not work on the used instruction set architectures (e.g., malware samples for SH4). Finally, given that malware family names might not be conclusive, we leveraged `VirusTotal` to obtain reliable malware family names/information, while excluding samples with unreliable/insufficient information.

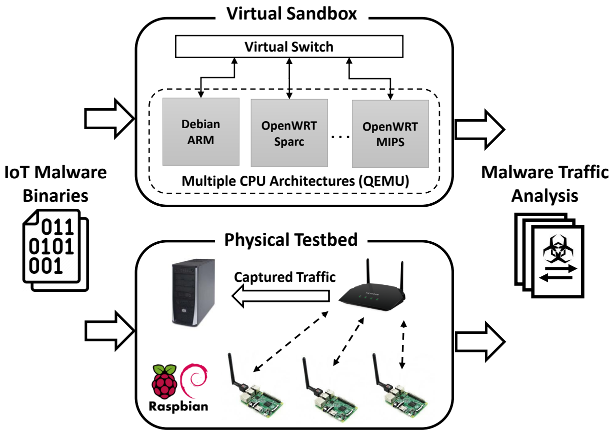


Fig. 10. The created environment for analyzing IoT malware.

TABLE 6
Analyzed IoT malware samples and their targeted ports.

IoT Malware MD5	Targeted Ports	Malware Family
807a15c2c87c7bb21d7660251e0db6f8	81	Mirai-Satori
05a8435816bb768761fdc893e79dc988	23 2323	Mirai-A
0540e803f1788f75369f434ace742346	445	Lightaidra
215e366b75e8998e214dcc2094f7c95d	443	Tsunami
67609e719aca8bfce3ac8c2500cfdacf	80 81 8080	Gafgyt-A
62a907378286e3fa431279dc2df948a4	23 80 8080	Mirai-G
d14d3483aac0032f37a9b3c42722e51a	5555	Mirai-B/ ADB.Miner
4cf9d9961da97c204b303bbfe874a035	2000	Bashlite

4.4.2 Experimental Setup

Given the collected IoT malware samples, we developed two experimental environments for executing and analyzing the malware binaries, as illustrated in Figure 10. First, considering the fact that IoT malware are found to target almost all existing CPU architectures, we setup a multi-architecture environment that emulates the most common CPU architectures using a virtual sandboxing environment on `Qemu` systems [47]. Second, we created an experimental testbed to mimic the behaviors of IoT devices connected to a wireless access point using three `Raspberry Pi3 (Model B+)` boards with Rasbian OS [48]. It is worth noting that the created testbed, which supports the execution of ARM-based malware only, was utilized to validate the actual behaviors of the IoT malware by testing for employed sandbox detection/evasion techniques. In fact, our analysis showed almost identical traffic generated by the tested malware on both environments (virtual and physical), which indicates the absence of employed evasion techniques. Finally, we utilized the created testing environments to execute IoT malware samples for thirty minutes each while capturing the exchanged traffic at the gateway using `TShark`.

4.4.3 Results

As summarized in Table 6, the experimental analysis resulted in identifying a number of IoT malware variants, which generated scanning campaigns towards ports similar to those identified in our initial darknet data (Tables 2). For instance, the `Mirai-Satori` was targeting port 81, which matches one of the common scanning objectives in our initial data set (S_3 from Table 2). In addition, the `Mirai-A` was found to be targeting ports identified within S_5 (Table 2), which corresponds to the behavior of 139 compromised IoT devices in our initial data. Furthermore, while the targeted ports by some of the analyzed IoT malware such as `Lightaidra` and `Mirai-G` were less prevalent among the identified scanning objectives in our initial data set, we identified a relatively larger number of exploited devices that scanned these ports when analyzing a recent sample of darknet data, as described next in Section 5. This might be justified by the evolving nature of IoT malware, which are tailored to target new combinations of ports that are associated with emerging vulnerabilities.

It is important to understand that our experimental results are bound to the limited number of analyzed malware samples, which do not represent the activities of all existing IoT malware families. Nevertheless, our results can indeed validate our methodology in terms of identifying malware-infected IoT devices and attributing their generated scanning campaigns to the overall behaviors of known malware families. More importantly, given the fact that IoT malware are rapidly evolving towards targeting new discovered vulnerabilities, our approach can be leveraged to infer the behaviors of emerging IoT malware through the detection of scanning campaigns that target new/uncommon ports. Finally, it is important to realize that despite the identified behavioral similarities among real IoT malware and the exploited devices involved in scanning the darknet, finding the exact malware variant/family that infected these devices requires further in-depth investigation and fingerprinting, which is considered for future work.

5 Campaign Persistence and Evolution

In order to investigate the persistence and evolution of IoT-generated scanning campaigns, we compared our findings from analyzing the initial data that was collected during April 2017, with newly collected data from the darknet. We followed the steps described in Section 3 to process over 3TB of newly collected IoT traffic from the darknet between May 21–25, 2018 (108 hours). The new data represents about 107M packets generated by 2,902 IoT devices towards the darknet, among which, about 99% (over 106M packets) were TCP-SYN packets. These TCP-SYN packets were generated by 1,647 compromised IoT devices, with an average of about 390 IoT devices that were generating approximately 988,000 TCP-SYN packets towards the darknet per hour (Figure 11). In what follows, we compare the IoT-generated scanning campaigns from the two collected data sets and investigate scanning activities, campaign persistence, and evolution.

5.1 Scanning Activities

The analysis of the newly collected data resulted in identifying a significantly less number of compromised IoT devices (1,647), as compared to the 6,797 devices that were discovered in our initial data collection. Nevertheless, these fewer IoT devices were found to be more active in scanning the darknet, sending significantly more TCP-SYN packets (106M packets) towards the darknet over a relatively shorter period of time (Figure 11), as compared to the 54.6M TCP-SYN packets that were generated by the IoT devices in our initial data (Figures 2). In fact, a Mann-Whitney U Test confirms that the number of generated scanning packets by compromised IoT devices was significantly greater ($p < 0.0001$) for the newly identified devices ($median = 993,931$ packets) than for the devices identified in the initial data ($median = 371,486$ packets).

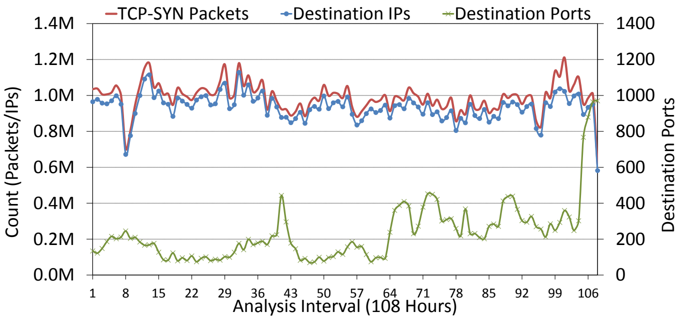


Fig. 11. The distribution of all TCP-SYN scanning packets generated by compromised IoT devices during the new analysis intervals (108 hours).

Moreover, the compromised IoT devices in the new data set scanned an average of 245 unique destination ports per hour, with Telnet/23 to be scanned by the highest number of TCP-SYN packets, followed by HTTP ports 80 and 8080 (Table 7). It is important to note that these ports have been continuously targeted by different variants of IoT malware/botnets (e.g., `Mirai`). Moreover, while these ports were scanned by less than 33% of all IoT devices, port 445, which is associated to the Server Message Block (SMB) protocol, was scanned by a relatively larger number of IoT device (44.5%), among which the majority (705 out of the 773) did not scan any other ports. Further investigation shows that the SMB protocol has been vulnerable to the `EternalBlue` exploit, which was leveraged by `WannaCry` ransomware to perform large-scale attacks towards computers running Windows OS in May 2017 (one month after our initial data collection). Interestingly, our findings indicate that compromised IoT devices have been used to perform reconnaissance activities to identify different types of vulnerable hosts, including non-IoT devices. Furthermore, we observe a considerable increase in the number of IoT devices that scanned port 445 in the new data set (Table 7), as compared to the initial data (Table 1). While the real reason behind the increased scanning activities towards port 445 is not known to us, we believe that our findings may provide an early indication of large-scale malware outbreaks, which target the vulnerable SMB protocol on port 445. Indeed, our findings have been corroborated by other reports, which highlight the growing number of scanning activities and malware-driven attacks towards port 445 in recent years [49].

The comparison of the total number of compromised devices hosted in different countries across the two analyzed data sets indicates a significant drop in the number of exploited devices hosted in Russia, followed by relatively smaller drops in the number of devices hosted in the U.S. and Thailand (Figures 6 and 12). In addition, while routers contributed to the largest portion of the IoT devices in

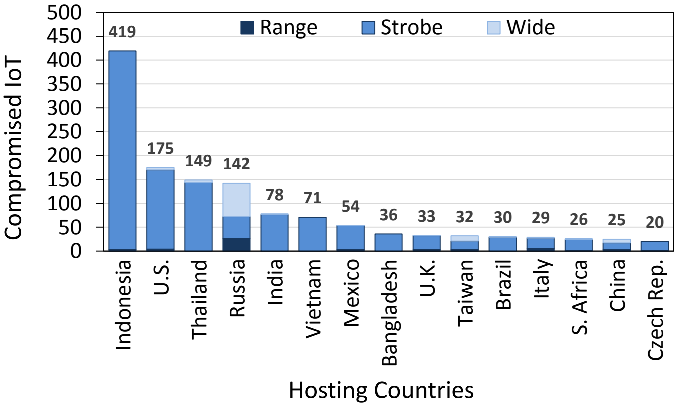


Fig. 12. Countries with the largest number of exploited IoT devices from each class (new data–May 2018).

TABLE 7
Top 18 scanned services/ports (CP=99%).

#	Service/Port	Packets		IoT Devices	
		(M)	%	Src. IP	%
1	*Telnet*/23	**14.32**	**13.42**	465	28.2
2	*HTTP*/80	8.14	7.64	556	33.8
3	*HTTP*/8080	8.06	7.56	508	30.8
4	*Unassigned*/81	6.14	5.76	126	7.7
5	*Kerberos*/88	5.96	5.59	119	7.2
6	*iRDMI*/8000	5.91	5.55	120	7.3
7	*Alt. Telnet*/2323	5.90	5.54	118	7.2
8	*XFER*/82	5.90	5.53	104	6.3
9	*MIT-ML-DEV*/85	5.90	5.53	102	6.2
10	*SUNPROXYADMIN*/8081	5.90	5.53	110	6.7
11	*DDI-TCP-1*/8888	5.90	5.53	104	6.3
12	*MIT-ML-DEV*/83	5.90	5.53	103	6.3
13	*RADAN-HTTP*/8088	5.90	5.53	108	6.6
14	*VCOM-TUNNEL*/8001	5.90	5.53	105	6.4
15	*CTF*/84	5.90	5.53	103	6.3
16	*SMB*/445	2.86	2.68	**773**	**44.5**
17	*RDP*/3389	0.66	0.62	20	1.2
18	*SSH*/22	0.45	0.42	15	0.9

our initial data (about 65%), IP cameras represented the largest population in the new data (about 50%). These changes can be justified by the significant decrease in the proportion of IoT devices within the range scanning class, which consist of mainly routers that were largely hosted in Russia. However, while the real reason behind the temporal change is unknown to us, we can only assume that these scanning campaigns have faded as a result of remediation and patching processes that took place after detecting the malware-infected devices and their malicious activities.

On the other hand, there is a noticeable increase in the total number of wide scanners hosted in Russia, as compared to other countries. Moreover, the number of devices that performed strobe scans almost doubled in Indonesia to reach slightly over 400 devices in the new data set. These temporal changes may in fact raise attention towards a number of points such as the weak security measures and/or remediation efforts put by consumers in those countries. Also, it may reflect the emergence of specific IoT malware variants, which target/exploit vulnerable devices that are widely deployed in those countries.

5.2 Persistence

The analysis of new IoT-generated data revealed scanning classes similar to those identified in Section 4.1.1. For instance, we found 44 IoT devices that were scanning the exact port ranges as specified in the range scans (e.g., 19328–19622), among which 19 devices were also common in both data sets. The persistence of such scanning activities after one year of initial observation might be justified in different ways: first, there are adversaries that are still interested in scanning possibly vulnerable hosts on these port ranges. Second, these compromised IoT devices were able to successfully evade detection and perform unsolicited scanning activities over a long period of time. Third, these exploited IoT devices were not updated/patched to receive the necessary remediation.

Furthermore, the majority of the compromised IoT devices in the new data set (1,421 out of 1,647) were performing strobe scans that are mainly targeting known services such as Telnet/23/2323, HTTP/80/8080, RDP/3389, SMB/445, and HTTPS/443. The scanned ports also include other less common/known services (i.e., ports 81–85, 8001, 8081, 8088, and 8888), which are thought to be used as alternative ports for HTTP by a number of online applications. More importantly, 8 out of the top 10 identified scanning objectives in the new data set (Table 8), which account for about 90% of all devices within strobe scanning class, also appeared among the top scanning objectives identified in the initial data set (Table 2). The similarities in terms of the identified scanning objectives and targeted ports demonstrate the persistence of IoT-generated campaigns over time, which tend to target a short list of vulnerable services using strobe scans.

In addition to range and strobe scanning classes, we identified 152 devices that performed wide scans, which consist of mainly routers (63%), followed by printers (19.7%), and IP cameras (15%). Furthermore, Russia hosted the largest number of these devices (about 46%), with significantly fewer number of devices distributed among other countries (Figure 12). Despite the fact that wide scans are less prevalent among compromised IoT devices in our data, the slight increase in the number of involved devices in the new data as compared to our initial data indicates the persistence of such campaigns. Confirming this however, requires further investigations that is beyond the scope of this paper and might be considered for future work.

5.3 Evolution

The analysis of the scanning objectives and classes within the newly analyzed data revealed 30 IoT devices (20 IP cameras and 10 routers) that were targeting a new range of destination ports (2–10000). These devices, which contributed to the high peaks in terms of the number of scanned destination ports throughout the analysis intervals (Figure 11), were performing distributed scans by targeting ports within the identified ranges on many destination addresses, resulting in a maximum rate of 5 packets per destination. Given the distinct scanned port ranges, we classify them as yet another variation of range scans, which reflect the behaviors of new or evolving IoT malware/botnets. It is also interesting to see that almost all of the devices scanned Telnet/23 and

TABLE 8
Frequent scanning objectives within strobe scanning class (CP=89.4%)

S_i	Frequency	%	scanning objective (ports)
1	**705**	**49.6**	445
2	228	16.0	23 80 8080
3	96	6.8	23 80 81 82 83 84 85 88 2323 8000 8001 8080 8081 8088 8888
4	79	5.6	80 443 8080
5	46	3.2	23
6	29	2.0	80 443
7	28	2.0	80 8080
8	26	1.8	80
9	19	1.3	3389
10	15	1.1	23 2323

HTTP/80/8080 ports, which is another sign of underlying correlation among these devices (i.e., scanning campaign).

Moreover, despite the similarities in the majority of the identified scanning objectives within the strobe scans when comparing both data sets (Tables 2 and 8), we observed the emergence of new scanning objectives that were in fact associated with recently discovered vulnerabilities. For instance, 12 compromised IoT devices were actively scanning port 5555, which is associated with ADB.Miner [3], the first Android worm to utilize port scanning code borrowed from Mirai. Similarly, we found traces of scans towards port 3333, which is associated with Fbot [50], a Satori variant that exploited various hosts on the Internet through their management port that runs the Claymore Miner software. Moreover, our results indicate possible traces of the Hajime botnet [41], which searches for vulnerable routers by scanning a list of ports including but not limited to 80–82, 8080, and 8081. Interestingly, while these ports appear in one of most frequent scanning objectives (S_3 from Table 8), they were also associated with other scanned ports (e.g., 8088 and 8888), which might reflect the behaviors of emerging IoT malware/botnets. In addition, we also noticed scans towards port 81, which is associated with a malware variant that extends Satori to exploit Goahead IP cameras [51]. Other newly scanned ports that were also related to a range of vulnerable services include: ports 83–85, 2000 (Cisco SSCP enabled phones [52] and Bashlite), 3389 (Mirai on RDP [53]), 8600, and 9000. It is important to understand that given the distinctive characteristics of the IoT devices in terms of the scanned ports, it is not anomalous to consider those devices to be correlated. In other words, they might be exploited by similar IoT malware, and therefore, involved in scanning campaigns as a part of a bigger botnet.

6 Discussion

The rising number of IoT-driven cyber attacks in recent years have shed light on the activities of exploited IoT devices and the underlying IoT-tailored malware/botnets [4, 12]. Nevertheless, obtaining empirical IoT-related data for the purpose of generating cyber-threat intelligence has been shown to be a challenging task [23, 25]. In this paper, we extend our previous work [13] to provide a methodological approach for detecting and characterizing IoT-generating scanning campaigns, which reflect the unsolicited activities of exploited devices and the underlying malware/botnet.

In general, while adversaries implement various techniques for hiding traces of their malicious activities, they tend to utilize limited resources (compromised IoT devices) to perform distributed, Internet-wide scanning of vulnerable ports [54]. Despite that, our multi-stage investigation of compromised IoT devices and their scanning activities on the network telescope reveal their overall scanning objectives and behavioral characteristics over an aggregated period of time, which is used further to identify correlated devices participate in scanning campaigns. In addition, the in-depth analysis of the detected campaigns unveiled interesting characteristics of the underlying IoT malware/botnets. For instance, while the IoT-generated scanning campaigns were mainly targeting a short list of known and/or emerging vulnerabilities, a considerable portion of them were in fact targeting new, possibly unknown vulnerabilities over a range of uncommon ports (e.g., 19328–19622). These findings however, can motivate the security research community towards further investigation of new vulnerabilities (if any), while taking preemptive measures to protect existing assets against the newly identified exploitations.

Moreover, with support of data collected in the wild, our findings expose cyber-threat intelligence that contribute to the information infrastructure in the realm of IoT. These findings, which come in-line with the rapidly emerging IoT threat landscape [12], highlight the insecurity of current IoT devices, while providing empirical evidence that demonstrates the persistence of such threats, even long after being widely recognized by the security community [4]. Our findings also demonstrate the evolution of IoT-tailored malware/botnets, which reflect upon the growing efforts of adversaries towards exploiting more sophisticated vulnerabilities. For instance, we identified traces of recently discovered IoT malware such as ADB.Miner [3] and Fbot [50], to name a few, which represent the evolution of IoT exploitations towards well-crafted malware variants that target vulnerable devices on new/uncommon ports.

An important contribution of this paper is to present an approach for characterizing the behaviors of emerging IoT malware/botnets through the analysis of the detected scanning campaigns by leveraging publicly available data resources. Moreover, while we demonstrate the feasibility of our approach using data that was collected in the wild, our approach involves semi-automated processes that hampers the efficiency of the work. Indeed, we plan to address this limitation by developing a system that lays the foundation for future IoT-related cyber security research using passive network data and by leveraging the capabilities of big data analytics frameworks towards automating the data collection/analysis processes, while providing a powerful and scalable infrastructure for processing/analyzing data in near real-time. Finally, while our work contributes towards understanding the current state of the compromised IoT devices and the underlying malware-initiated scanning activities, we raise awareness towards exploring and delivering holistic counter measures for securing IoT devices and protecting future operations, especially, with the anticipated role of IoT devices as a fundamental driver of future generation of wireless networks (e.g., 5G networks) [55].

7 Conclusion

We introduced a practical approach for detecting and characterizing IoT-generated scanning campaigns. More specifically, by leveraging IoT device information and over 6TB of passive network traffic collected at a large-scale network telescope, we identified over 8,000 compromised IoT devices that were involved in a number of distinct scanning campaigns on the Internet. In fact, the multi-stage investigation of the devices and their generated scanning campaigns shed light on behavioral characteristics of the underlying IoT malware/botnets. Moreover, while our results corroborate findings with respect to known IoT malware/botnets, they extend our knowledge towards discovering emerging malware/botnets, which tend to target new vulnerabilities. In addition, we provide insights on the persistence and evolution of IoT-generated scanning campaigns over time. After all, while our findings shed light on the current state of exploited IoT devices, we also lay the foundation for future work towards building scalable, Internet-wide IoT threat detection systems that can help in building a better understanding of the threats landscape while developing proper countermeasures to limit their impact on future operations.

Acknowledgments

The work has been supported by Natural Sciences and Engineering Research Council of Canada (NSERC) and Concordia University. This work was also partially supported by a grant from the U.S. National Science Foundation (NSF), Office of Advanced Cyberinfrastructure (OAC), #1907821.

References

[1] W. Andy Greenberg, "The reaper IoT botnet has already infected a million networks," Online: https://www.wired.com/story/reaper-iot-botnet-infected-million-networks/, 2017.

[2] FBI, "Foreign Cyber Actors Target Home and Office Routers and Networked Devices Worldwide," https://www.ic3.gov/media/2018/180525.aspx, May 2018, public Service Announcement (Alert Number: I-052518-PSA).

[3] 360Netlab. (2018, February) ADB.Miner: More Information [Blog post]. Retrieved from https://blog.netlab.360.com/adb-miner-more-information-en/.

[4] M. Antonakakis *et al.*, "Understanding the Mirai Botnet," in *26th USENIX Security Symp.*, Vancouver, BC, 2017, pp. 1093–1110.

[5] T. Yu *et al.*, "Handling a trillion (unfixable) flaws on a billion devices: Rethinking network security for the Internet-of-Things," in *Proc. of the 14th ACM Workshop on Hot Topics in Networks*. ACM, 2015, p. 5.

[6] N. Neshenko *et al.*, "Demystifying IoT Security: An Exhaustive Survey on IoT Vulnerabilities and a First Empirical Look on Internet-Scale IoT Exploitations," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2702–2733, 2019.

[7] C. Labovitz, A. Ahuja, and M. Bailey, *Shining Light on Dark Address Space*. Arbor Networks Inc., 2001.

[8] C. Fachkha *et al.*, "Internet-scale Probing of CPS: Inference, Characterization and Orchestration Analysis," in *Proc. of the Network and Distributed Syst. Security Symp. (NDSS'17)*, San Diego, California, 2017.

[9] C. Fachkha and M. Debbabi, "Darknet as a Source of Cyber Intelligence: Survey, Taxonomy, and Characterization," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1197–1227, 2016.

[10] E. Bou-Harb, M. Debbabi, and C. Assi, "On Fingerprinting Probing Activities," *Computers & Security*, vol. 43, pp. 35–48, 2014.

[11] Z. Durumeric, M. Bailey, and J. A. Halderman, "An Internet-Wide View of Internet-Wide Scanning," in *Proc. of the 23rd USENIX Security Symp.*, San Diego, CA, 2014, pp. 65–78.

[12] P.-A. Vervier and Y. Shen, "Before Toasters Rise Up: A View into the Emerging IoT Threat Landscape," in *Int. Symp. on Research in Attacks, Intrusions, and Defenses*. Springer, 2018, pp. 556–576.

[13] S. Torabi *et al.*, "Inferring, Characterizing, and Investigating Internet-Scale Malicious IoT Device Activities: A Network Telescope Perspective," in *Proc. of the 48th Annual IEEE/IFIP Int. Conf. on Dependable Systems and Networks (DSN)*, June 2018, pp. 562–573.

[14] A. Cui and S. J. Stolfo, "A quantitative analysis of the insecurity of embedded network devices: results of a wide-area scan," in *Proc. of the 26th Annual Comput. Security Applicat. Conf.* ACM, 2010, pp. 97–106.

[15] V. Sachidananda *et al.*, "Let the Cat Out of the Bag: A Holistic Approach Towards Security Analysis of the Internet of Things," in *Proc. of the 3rd ACM Int. Workshop on IoT Privacy, Trust, and Security*, ser. IoTPTS '17, 2017, pp. 3–10.

[16] A. Costin *et al.*, "A large-scale analysis of the security of embedded firmwares," in *In 23rd USENIX Security Symp.*, 2014, pp. 95–110.

[17] D. D. Chen *et al.*, "Towards Automated Dynamic Analysis for Linux-based Embedded Firmware," in *Proc. of the Network and Distributed Syst. Security Symp. (NDSS)*, 2016.

[18] Y. J. Jia *et al.*, "ContexIoT: Towards Providing Contextual Integrity to Appified IoT Platforms," in *Proc. of the Network and Distributed Syst. Security Symp. (NDSS'17)*, 2017.

[19] E. Fernandes *et al.*, "FlowFence: Practical Data Protection for Emerging IoT Application Frameworks," in *25th USENIX Security Symp.*, 2016.

[20] B. Ur, J. Jung, and S. Schechter, "The Current State of Access Control for Smart Devices in Homes," in *Workshop on Home Usable Privacy and Security (HUPS)*, 2013.

[21] E. Ronen and A. Shamir, "Extended Functionality Attacks on IoT Devices: The Case of Smart Lights," in *IEEE European Symp. on Security and Privacy (EuroS&P)*. IEEE, 2016, pp. 3–12.

[22] S. Torabi *et al.*, "Detecting Internet Abuse by Analyzing Passive DNS Traffic: A Survey of Implemented Systems," *IEEE Commun. Surveys & Tutorials*, 2018.

[23] Y. M. P. Pa *et al.*, "IoTPOT: A Novel Honeypot for Revealing Current IoT Threats," *J. of Inform. Process.*, vol. 24, no. 3, pp. 522–533, 2016.

[24] J. D. Guarnizo *et al.*, "Siphon: Towards Scalable High-Interaction Physical Honeypots," in *Proc. of the 3rd ACM Workshop on Cyber-Physical System Security*. ACM, 2017, pp. 57–68.

[25] T. Luo *et al.*, "IoTCandyJar: Towards an Intelligent-Interaction Honeypot for IoT Devices," in *Blackhat*, 2017.

[26] M. A. Hakim *et al.*, "U-PoT: A Honeypot Framework for UPnP-Based IoT Devices," in *37th IEEE Int. Performance, Computing and Communications Conf. (IPCCC)*. IEEE, 2018, pp. 1–8.

[27] S. Bellovin, "There Be Dragons," in *USENIX Summer*, 1992.

[28] S. M. Bellovin, "Packets Found on an Internet," *ACM SIGCOMM Computer Communication Review*, vol. 23, no. 3, pp. 26–31, 1993.

[29] "Shodan," Retrieved from https://www.shodan.io/, 2019.

[30] E. Glatz and X. Dimitropoulos, "Classifying Internet One-way Traffic," in *Proceedings of the 2012 Internet Measurement Conference*, ser. IMC '12, Boston, MA, USA, 2012, pp. 37–50.

[31] E. Bou-Harb, C. Assi, and M. Debbabi, "CSC-Detector: A System to Infer Large-Scale Probing Campaigns," *IEEE Transactions on Dependable and Secure Computing*, 2016.

[32] N. Furutani *et al.*, "Detection of DDoS Backscatter Based on Traffic Features of Darknet TCP Packets," in *Ninth Asia Joint Conference on Information Security (ASIA JCIS)*. IEEE, 2014, pp. 39–43.

[33] "The CAIDA UCSD Real-Time Network Telescope Data," UCSD - Center for Applied Internet Data Analysis. Retrieved from http://www.caida.org/data/passive/telescope-near-real-time_dataset.xml, 2019.

[34] "Corsaro," Center for Applied Internet Data Analysis (CAIDA), https://www.caida.org/tools/measurement/corsaro/.

[35] "The ELK Stack," Retrieved from https://www.elastic.co/elk-stack.

[36] M. H. Bhuyan, D. Bhattacharyya, and J. Kalita, "Surveying Port Scans and Their Detection Methodologies," *The Computer Journal*, vol. 54, no. 10, pp. 1565–1581, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1093/comjnl/bxr035

[37] K. Thomas *et al.*, "The abuse sharing economy: Understanding the limits of threat exchanges," in *Int. Symp. on Research in Attacks, Intrusions, and Defenses*. Springer, 2016, pp. 143–164.

[38] A. Dainotti *et al.*, "Analysis of a/0 Stealth Scan From a Botnet," *IEEE/ACM Transactions on Networking (TON)*, vol. 23, no. 2, pp. 341–354, 2015.

[39] I. A. N. A. (IANA), "Service Name and Transport Protocol Port Number Registry," Retrieved March 1, 2019 from https://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml, 2019.

[40] "SX-Virtual Link Software Frequently Asked Questions," Retrieved from https://www.silextechnology.com/sx-virtual-link-faq, 2019.

[41] C. Cimpanu, "Hajime Botnet Makes a Comeback With Massive Scan for MikroTik Routers," Retrieved from https://www.bleepingcomputer.com/news/security/hajime-botnet-makes-a-comeback-with-massive-scan-for-mikrotik-routers/, March 2018.

[42] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *ACM SIGMOD record*, vol. 22, no. 2. ACM, 1993, pp. 207–216.

[43] M. Ester *et al.*, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in *Proc. of KDD*, vol. 96, no. 34, 1996, pp. 226–231.

[44] G. H. Shah, C. Bhensdadia, and A. P. Ganatra, "An Empirical Evaluation of Density-Based Clustering Techniques," *Int. J. of Soft Comput. and Eng. (IJSCE)*, vol. 22312307, pp. 216–223, 2012.

[45] J. Mazel *et al.*, "Hunting attacks in the dark: clustering and correlation analysis for unsupervised anomaly detection," *Int. J. of Network Management*, vol. 25, no. 5, pp. 283–305, 2015.

[46] G. Gu *et al.*, "BotMiner: Clustering Analysis of Network Traffic for Protocol-and Structure-Independent Botnet Detection," in *Proc. of the 17th USENIX Security Symp.*, 2008, pp. 139–154.

[47] "Qemu," Retrieved from https://qemu.org/, 2019.

[48] "Rasberry Pi," Retrieved from https://www.raspberrypi.org/, 2019.

[49] O. Kubovic. (2019, May) EternalBlue reaching new heights since WannaCryptor outbreak. [Blog post]. Retreived from https://www.welivesecurity.com/2019/05/17/eternalblue-new-heights-wannacryptor/. WeLiveSecurity.

[50] SISSDEN. (2018, February) Darknet - Satori strikes again. [Blog post]. Retreived from https://sissden.eu/blog/darknet-satori-dasan.

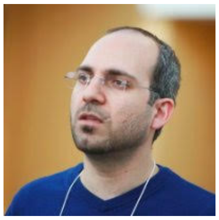
[51] L. Fengpei. (2017, April) New Threat Report: A new IoT Botnet is Spreading over HTTP 81 on a Large Scale. Retrieved from http://blog.netlab.360.com/a-new-threat-an-iot-botnet-scanning-internet-on-port-81-en/.

[52] (2011, December) Identifying and Mitigating Exploitation of the Cisco Unified Communications Manager Express and Cisco IOS Software H.323 and SIP DoS Vulnerabilities. Retreived from https://www.cisco.com/c/en/us/support/docs/cmb/cisco-amb-20100324-voice.html. Cisco.

[53] G. Escueta. (2017, February) Mirai Widens Distribution with New Trojan that Scans More Ports. Retreived from https://blog.trendmicro.com/trendlabs-security-intelligence/mirai-widens-distribution-new-trojan-scans-ports/. Trend Micro.

[54] J. Mazel, R. Fontugne, and K. Fukuda, "Identifying Coordination of Network Scans Using Probed Address Structure," in *8th Int. Workshop on Traffic Monitoring and Analysis*, 2016.

[55] "Internet of Things - Number of Connected Devices Worldwide from 2015 to 2025," In Statista - The Statistics Portal. Retrieved June 10, 2018 from https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/, 2016.



Sadegh Torabi received the B.Sc. and M.Sc. degrees (with Distinction) from the Computer Engineering Department, Kuwait University (KU), Kuwait, in 2005 and 2009, respectively, and the M.Sc. degree from the Electrical and Computer Engineering Department, University of British Columbia (UBC), Vancouver, BC, Canada, in 2016. He is currently a Ph.D. candidate at Concordia Institute for Information System Engineering, Montreal, QC, Canada. His current research interests are in the areas of cyber security including Internet of Things, cyber-physical systems, and usable security and privacy. He was a Research Assistant with UBC from 2011 to 2016. He was a recipient of the Concordia University Graduate Fellowship Award in 2016–2019.

Elias Bou-Harb received the Ph.D. degree in computer science from Concordia University, Montreal, Canada. He was a visiting research scientist at Carnegie Mellon University (CMU), Pittsburgh, PA, USA, in 2015–2016. He joined the Department of Computer Science at Florida Atlantic University (FAU) as an Assistant Professor of cyber security and data analytics in 2016. He is currently an Associate Professor at the Cyber Center For Security and Analytics at the Department of Information Systems and Cyber Security at the University of Texas at San Antonio (UTSA). He is also a research scientist at the National Cyber Forensic and Training Alliance (NCFTA) of Canada. His current research interests are in the areas of operational cyber security, attacks detection and characterization, Internet measurement, cyber security for critical infrastructure, and mobile network security.

Chadi Assi received the Ph.D. degree from the City University of New York. He was a Visiting Researcher with Nokia Research Center, Boston, MA, USA, where he worked on quality of service in passive optical access networks. In 2003, he joined the Concordia Institute for Information Systems Engineering, Concordia University, as an Assistant Professor, where he is currently a Full Professor. His current research interests are in the areas of network design and optimization, network modeling and network reliability, and smart grids. He was a recipient of the prestigious Mina Rees Dissertation Award for his research on wavelength-division multiplexing optical networks. He is on the Editorial Board of the IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, the IEEE TRANSACTIONS ON COMMUNICATIONS, and the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.

ElMouatez Billah Karbab is a PhD Candidate at Concordia University. His research focuses on malware fingerprinting using machine learning techniques, cyber security, and embedded systems. He is a research scientist at the National Cyber Forensic and Training Alliance (NCFTA) of Canada. He is also serving as a data scientist and cyber-security specialist at NCFTA Canada. He served as an associate researcher at Research Center for Scientific and Technical Information (CERIST), Algeria, where he worked on international projects in collaboration with Univesity of Cape Town, South Africa, and Heudiasyc Lab, France.

Amine Boukhtouta received the computer science engineering degree from USTHB University, Algiers, Algeria, in 2005 and the M.A.Sc. degree in information systems security degree and the Ph.D. degree in electrical and computer engineering from Concordia University, Montreal, Canada, in 2009 and 2016, respectively. He has been a Cyber-Threat Researcher with National Cyber-Forensics Training Alliance Canada. He also completed a Postdoctoral industrial program in 2016, where he worked on security of evolving delivery network, big data analytics, and machine learning. He is currently working as an Experienced Researcher with ERICSSON Research Group, Montreal, QC, Canada. His current research interests include prevention, detection of cyber-threats, machine learning, and artificial intelligence. He was a recipient of OCTAS Prize in 2009 University Competition, the FQRNT Doctoral Scholarship in 2010–2011, the MITACS as well as PROMPT Postdoctoral Fellowships in 2016–2017.

Mourad Debbabi is a Full Professor at the Concordia Institute for Information Systems Engineering and Associate Dean Research and Graduate Studies at the Gina Cody School of Engineering and Computer Science. He holds the NSERC/Hydro-Quebec Thales Senior Industrial Research Chair in Smart Grid Security and the Concordia Research Chair Tier I in Information Systems Security. He is also the President of the National Cyber Forensics and Training Alliance (NCFTA) Canada. He is a member of CATAAlliance's Cybercrime Advisory Council. He serves/served on the boards of Canadian Police College, PROMPT Québec and Calcul Québec. He is the founder and one of the leaders of the Security Research Centre at Concordia University. In the past, he was the Specification Lead of four Standard JAIN (Java Intelligent Networks) Java Specification Requests dedicated to the elaboration of standard specifications for presence and instant messaging. Dr. Debbabi holds Ph.D. and M.Sc. degrees in computer science from Paris-XI Orsay, University, France. He published 5 books and more than 300 peer-reviewed research articles in international journals and conferences on cyber security, cyber forensics, smart grid, privacy, cryptographic protocols, threat intelligence generation, malware analysis, reverse engineering, specification and verification of safety-critical systems, programming languages and type theory. He supervised to successful completion 30 Ph.D. students, 72 Master students and 14 Postdoctoral Fellows. He served as a Senior Scientist at the Panasonic Information and Network Technologies Laboratory, Princeton, New Jersey, USA; Associate Professor at the Computer Science Department of Laval University, Canada; Senior Scientist at General Electric Research Center, New York, USA; Research Associate at the Computer Science Department of Stanford University, California, USA; and Permanent Researcher at the Bull Corporate Research Center, Paris, France.

View publication stats