Internet-scale Insecurity of Consumer Internet of Things: An Empirical Measurements Perspective

Article *in* ACM Transactions on Management Information Systems · April 2020

DOI: 10.1145/3394504

CITATIONS
8

READS
349

3 authors:

Antonio Mangino
University of Texas at San Antonio
5 PUBLICATIONS **39** CITATIONS

SEE PROFILE

Morteza Safaei Pour
University of Texas at San Antonio
10 PUBLICATIONS **90** CITATIONS

SEE PROFILE

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

SEE PROFILE

Internet-scale Insecurity of Consumer Internet of Things: An Empirical Measurements Perspective

ANTONIO MANGINO, The Cyber Center for Security and Analytics, UT at San Antonio (UTSA)
MORTEZA SAFAEI POUR, The Cyber Center for Security and Analytics, UT at San Antonio (UTSA)
ELIAS BOU-HARB, The Cyber Center for Security and Analytics, UT at San Antonio (UTSA)

The number of Internet-of-Things (IoT) devices actively communicating across the Internet is continually increasing as these devices are deployed across a variety of sectors, constantly transferring private data across the Internet. Due to the extensive deployment of such devices, the continuous discovery and persistence of IoT-centric vulnerabilities in protocols, applications, hardware and the improper management of such IoT devices has resulted in the rampant, uncontrolled spread of malware threatening consumer IoT devices.
To this end, this work adopts a novel, macroscopic methodology for fingerprinting Internet-scale compromised IoT devices, revealing crucial cyber threat intelligence on the insecurity of consumer IoT devices. By developing data-driven techniques rooted in machine learning methods and analyzing 3.6 TB of network traffic data, we discover 855,916 compromised IP addresses, with 310,164 fingerprinted as IoT. Further analysis reveals China and Brazil to be hosting the most significant population of compromised IoT devices (100,000 and 55,000 respectively). Additionally, we provide a longitudinal analysis on data from one year ago against this work, revealing the evolving trends of IoT exploitation, such as the increased number of vendors targeted by malware, rising from 50 to 131. Moreover, countries such as China (420% increased infected IoT count) and Indonesia (177% increased infected IoT count) have seen notably high increases in infection rates. Lastly, we compare our geographic results against Global Cybersecurity Index ratings, verifying that countries with high GCI ratings such as the Netherlands and Germany had relatively low infection rates. However, upon further inspection, we find that the GCI rate does not accurately represent the consumer IoT market, with countries such as China and Russia being rated with 'high' CGI scores, yet hosting a large population of infected consumer IoT devices.

CCS Concepts: • **Security and privacy** → **Network security**; *Intrusion/anomaly detection and malware mitigation*; • **Computer systems organization** → **Embedded and cyber-physical systems**.

Additional Key Words and Phrases: Internet-of-Things, IoT security, Data Science, IoT forensics

ACM Reference Format:
Antonio Mangino, Morteza Safaei Pour, and Elias Bou-Harb. 2020. Internet-scale Insecurity of Consumer Internet of Things: An Empirical Measurements Perspective. 1, 1 (April 2020), 25 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The recent rise in embedded system technologies has instigated a significant increase in the development and deployment of Internet-of-Things (IoT) devices. The Internet-of-Things is generally

Authors' addresses: Antonio Mangino, The Cyber Center for Security and Analytics, UT at San Antonio (UTSA), antonio.mangino@my.utsa.edu; Morteza Safaei Pour, The Cyber Center for Security and Analytics, UT at San Antonio (UTSA), morteza.safaeipour@utsa.edu; Elias Bou-Harb, The Cyber Center for Security and Analytics, UT at San Antonio (UTSA), elias.bouharb@utsa.edu.

XXXX-XXXX/2020/4-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

defined to encompass any device capable of connecting to the Internet, interacting with other devices and exchanging data. Primarily adopted in consumer markets, government agencies and critical infrastructure [4], the mass manufacturing of such embedded devices has surpassed multiple benchmarks, each underestimating the impact and rapid increase of actively communicating IoT devices across the Internet [9]. The expansion of the IoT paradigm is especially evident within the consumer market as businesses and smart homes embrace these emerging technologies. In fact, over 70% of all North American households have at least one Internet-connected IoT device and the global median of 40% shows an upward trend [28]. Behind the enormous growth of the IoT paradigm resides consumers looking to make their daily lives more convenient through automated device functionality, including enhanced opportunities for communication made possible by Internet transmissions [33, 51]. Yet, despite the widespread deployment of consumer IoT devices, very little effort is taken towards securing such devices, leading to the uncontrolled spread of malware and development of large-scale, malicious IoT-centric botnets [34].

Botnets are a collection of devices infected by a common malware, in which a command-and-control infrastructure allows a bot master to launch malicious attacks [3, 27]. Once infected, bots will begin to scan the Internet space to find similarly vulnerable devices to compromise. Hackers specifically target IoT devices within the consumer market due to their abundant vulnerabilities and the lack of awareness towards device security. In late 2016, the most notorious IoT-specific botnet known as *Mirai* emerged. Consisting of over 400,000 compromised IoT devices, the Mirai botnet launched momentous Distributed Denial of Service (DDoS) attacks, reaching transmission rates of 1.1Tbps [2]. Targeting networked consumer-based IoT devices, the explosive growth and offensive capabilities of the Mirai botnet advertised the perfect breeding ground for similarly devastating botnets, with new strains of malware being identified daily [27]. The rampant spread and evolution of massive botnet campaigns emphasizes the necessity of researching Internet-scale mitigation and remediation techniques. Furthermore, consumer IoT devices are at the highest risk of exploitation by such malware due to the lack of information technology and security departments overseeing the security postures of such devices. As it stands, no trained or specialized department entity is responsible for monitoring, remediating and preventing IoT device exploitation in consumer realms.

Indeed, thoroughly securing IoT devices is a complex issue, with vulnerabilities existing in device hardware, firmware, protocols and cryptographic module implementations [40]. IoT devices are regularly configured with unused ports and services left open, virtually left unprotected with default configurations (e.g., credentials). Default credentials are extremely vulnerable to dictionary-based brute-force password attacks and similar intrusions on common services such as Telnet and FTP, allowing hackers to gain unauthorized access into devices [2, 28]. Moreover, IoT applications tend to request unnecessary administrative privileges, leading to an increased number of intrusion points. 55% of applications requested administrative rights for operations that they do not actively use and 42% were granted rights that were not explicitly requested [19]. Furthermore, vulnerabilities are actively exploited during the installation of patches and updating of firmware. Many updates are requested and installed through plain text network traffic, allowing for hackers to covertly capture network traffic originating from compromised devices and retrieve firmware binaries, which can then be reverse engineered to discover additional vulnerabilities and intrusion points [52].

Forced recruitment into botnet campaigns to launch malicious attacks is not the the only security risk threatening IoT devices. IoT devices within the consumer sector primarily include cameras, sensors, monitors and appliances that are networked together to remotely automate domestic household or business environments [51]. The inherent nature of such devices leads to an increased amount of private and public data that is actively recorded, cached and communicated across the Internet. Malicious entities may target consumer IoT devices to infiltrate and steal private data such

as financial records, personal identity characteristics and health-related information. Transmissions from compromised IoT devices can be used to profile the device owner, gathering evidence to reveal their identity and device usage statistics, including the times that the device owner is home or at work [46]. The threat to privacy is already affecting the consumer market, with 74% of smaller businesses reporting cyber-related attacks and 90% of large organizations experiencing similar breaches [29]. Consumer data is at risk of being exploited without proper measures for securing IoT devices, evident through examining the malicious entities which are actively developing and propagating malware to take advantage of poorly secured devices.

Research Problem Statement. The growth of the IoT paradigm has created an environment in which hundreds of thousands, if not millions, of unsecured IoT devices are currently deployed and in usage. Unprotected devices are at risk of being compromised, in which malicious entities threaten the security, privacy and safety of their operating environments, which may directly affect the device owner. Despite the threat to IoT devices and private information, public knowledge of device security is largely inadequate. It is common for device owners to be unaware of proper safety techniques and generally disregard implementing security measures. Moreover, cybersecurity efforts towards identifying, mitigating and remediating Internet-scale compromised IoT devices are largely impeded by multiple inherent IoT device characteristics. Specifically, the large number of vendors, each producing many variations of similar firmware and hardware, conforms a heterogeneous nature onto IoT devices. Therefore, the IoT paradigm is extremely diverse, a characteristic which hampers efforts to curate and analyze IoT-centric empirical data and artifacts. Furthermore, analyzing network traffic to discover specific identifiers of IoT devices, such as manufacturer, vendor, device type, model and operating system is extremely difficult. Only a small fraction (results from this work indicate less than 10%) of the Internet-wide IoT device population responds with relevant service banners when contacted, requiring the development of a comprehensive machine learning classifier to effectively learn from the smaller population to fingerprint the remaining compromised IoT devices. To this end, in order to provide IoT-specific cyber threat intelligence and artifacts, this work develops a novel, data-driven methodology for inferring compromised IoT devices and categorizing device-specific information. Moreover, this work focuses on device metrics derived from sorting infected IoT devices within the consumer market sector.

Contributions. Motivated by the aforementioned IoT-centric challenges and the deficit of Internet-scale measurements related to the IoT paradigm at large, this work offers a number of contributions towards the identification, characterization and indexing of Internet-scale compromised consumer IoT devices.

- **A learning-driven classification method for inferring infected consumer IoT devices.** Leveraging passively-collected network telescope data, the proposed methodology identifies compromised IP addresses, extracts features from packet headers and retrieves text input from service banners to create a valid training and test data set. The aggregated data is then leveraged for supervised training of shallow machine learning classifiers to accurately fingerprint Internet-scale compromised IoT devices. To expedite the reproducibility of our results while motivating additional studies and measurements of IoT security, we make our developed models and source code available via https://github.com/ccsa-rd/TMIS-IoT.
- **Empirical consumer IoT statistics.** This work analyzes 3.6 TB of recent network traffic over a 24-hour time interval to fingerprint compromised IoT devices. 855,916 infected IP addresses were discovered actively scanning the Internet space, with a total of 310,164 identified as compromised IoT devices. Further results on the geographic location, targeted ports and

IoT vendors reveal malware trends across the global IoT environment. Within the consumer IoT market, Internet Service Provider Rostelcom in Russia and Turk Telekom in Turkey show the highest infection rates with 3,589 and 3,559 compromised devices, respectively. Of the 20 reported ISPs, 7 originate from China, including China Telecom Hubei (3,238 devices) and China Telecom Hunan (2,987 devices).

- **Evolution of compromised IoT devices over a one year time interval.** We analyze a data set previously collected in December, 2018 and compare the outcomes with our current results. Comparisons reveal critical insights on the evolution of vulnerabilities, revealing that China witnessed a 420% increase in infected IoT devices, while Brazil saw a 63.3% decrease. Further enumerating upon malware evolution, trends are detailed with their resulting implications on the consumer IoT landscape, such as the drastic increase of targeted vendors (50 in 2018, 131 in 2019) and the increased range of scans (i.e, the emergence of UDP scans on ports 16285, 8000 and 8080), illustrating new, evolved malware strains and a higher probability of future infections.

Organization. The rest of the paper is organized as follows. The following section introduces background terminology referenced later in this work. Next, Section 3 provides a comprehensive review on related literature that explore IoT protocol vulnerabilities, botnets and general IoT device security. Section 4 will detail the proposed methodology behind the classifier used to fingerprint Internet-scale infected consumer IoT devices. Section 5 executes and evaluates the proposed approach to reveal crucial IoT-specific cyber threat intelligence. Section 6 initiates a discussion and elaborates upon possible methodologies for mitigating and remediating compromised IoT devices. Lastly, Section 7 summarizes the contributions of this work while discussing potential improvements and several topics for future works.

2 BACKGROUND

Network telescopes, also known as darknets, are allocated sets of the IP address space that are routable, yet intentionally do not host any legitimate services. The lack of hosted services ensures that Internet traffic received by network telescopes is unsolicited and reliably contains discrete anomalies such as worms, DDoS backscatter and various other forms of malicious traffic [18, 24]. Network telescopes offer a large-scale, macroscopic vantage point for collecting Internet-wide traffic [25]. Similarly, honeypots are deployed systems that interactively act vulnerable to attract malicious activity and gather malware-specific information [16]. While honeypots do not offer extensively large vantage points, the collected data reveals crucial in-depth malware intelligence. Analyzing the network traffic captured by network telescopes, honeypots and similar architectures reveals, among other threat information, compromised IP addresses, which are a vital component for comprehensive methodologies that characterize exploited devices [6, 7].

The vast majority of packets collected by network telescopes are attributed to botnet campaigns actively scanning the Internet space for vulnerable devices in an attempt to propagate. The most common form of Internet scanning is performed through sending Transport Control Protocol (TCP) packets carrying a SYN flag without an attached payload. Such TCP SYN packets can be used to verify if a target is open for communication, which is why compromised devices will continually send these packets to multiple destination ports, scanning for open services to communicate with. Across a massively distributed Content Distribution Network's (CDN) firewall, the vast majority of TCP packets received were identified as scan-relate and subsequently dropped. In fact, 98% of all TCP packets received by the CDN's firewall had the SYN bit set and did not carry a payload,

indicating scanning activities [41]. Recent works leveraging passively collected scan packets are further detailed in the related works section.

Additionally, evaluations and results from this work are compared against geographical distributions developed by the Global Cybersecurity Index (GCI) [49]. GCI releases yearly, questionnaire-based measurements to identify the type, level and evolution over time of cybersecurity commitment in countries from a regional and global perspective, while comparing the level of engagement in cybersecurity programs and initiatives. GCI metrics shed light on societal awareness of cybersecurity based on five key ideologies: legal, technical, organizational, capacity building and cooperation. Because this work provides Internet-scale measurements of compromised IoT devices, as well as reveals geographically defined results, our comparisons and correlations with GCI measurements illustrate a detailed, global perspective of the current state of IoT security.

3 LITERATURE REVIEW

In this section, we review related works to provide insight on the foundations of this research. We begin by introducing the insecurities within IoT device management, then elaborate upon IoT-centric protocol vulnerabilities and finally review various methodologies and classifiers used to identify compromised IoT devices.

Vulnerable IoT device management. Consumer IoT devices are primarily deployed within households, which have recently adopted the label *smart homes*. Unlike IoT devices deployed in government, critical infrastructure and higher education, consumer IoT devices rarely have a designated authority for overseeing device health and mitigating the spread of malware [43]. Yet, even these information technology specialists face many challenges as they deviate from traditional roles. Security operations must continually introduce new policies to actively mitigate frequently evolving malware, phishing scams and related attacks against industry and infrastructure. Considering the gradually escalating roles of security analysts, the lack of a responsible party specialized in protecting against such threats within private households results in improper management and insufficient protection of consumer IoT devices. Facilitated by the lack of societal knowledge, device owners are generally left unaware of potential security risks and leave their devices unprotected [11].

Recent publications have identified common risks and vulnerabilities within the consumer IoT market. For instance, Ali et al. performed an information risk assessment on possible security threats against IoT devices [1]. Security risks were identified by the severity of what actions an attacker can accomplish after successfully exploiting a device, including the amount of data that can be stolen or lost. Of the identified vulnerabilities, user credentials posed the highest risk of being targeted by malicious attacks. If user credentials are hacked, the attacker will gain unauthorized access into the system, allowing for the execution of subsequent malicious operations and even preventing device owners from accessing their devices. Additional IoT risks include software exploitation, malicious code being injected within software applications and devices being compromised to spy on users, including location information services [1, 21].

Leveraging data compiled by Avast Software, Kumar et al. reported the current security posture of deployed consumer IoT devices [28]. It was discovered that many IoT devices had ports and protocols left open, even if they are not exclusively used by the device. These unused ports receive and communicate with incoming transmissions, resulting in unguarded intrusion points. Many devices were found to have left the File Transfer Protocol (FTP) and Telnet services open, and 17.4% of devices had weak FTP credentials while 2.1% had weak Telnet credentials. To validate the claim that credentials were weak, a list of 200 common username and password combinations was leveraged for a brute-force dictionary attack against the identified devices. The dictionary

attacks revealed *admin/admin* to be the most popular username/password combination, accounting for 88% of weak FTP credentials and 36% of weak Telnet credentials. Furthermore, these weak credentials are the default values installed by device manufacturers, with some defaults being less secure than others. 55.3% of TP-LINK routers with open FTP services were identified to have weak credentials, while 10.9% of D-LINK routers were equally vulnerable. The risk of exploitation through brute-force attacks on vulnerable credentials is extremely high and can be replicated across a myriad of consumer IoT devices [52]. Without device owners actively updating manufacturer defaults or similarly weak credentials, the threat of brute-force exploitation persists.

Furthermore, Fernandes et al. analyzed the Samsung SmartThings programming framework [19]. Consisting of over 500 IoT-oriented applications, known as SmartApps, the SmartThings framework is among the leading IoT development frameworks. With such a widespread distribution of developers and their applications, numerous security flaws leave consumer IoT devices vulnerable. 55% of SmartApps requested device operations that they never accessed and 42% of SmartApps were granted capabilities that they did not explicitly request, leading to the over-privilege of offending applications. These over-privileged applications are vulnerable to exploitation through design flaws, allowing hackers full access to the host device. Fernandes et al. exploited popular applications by launching a wide array of attacks against a variety of consumer IoT devices, including security locks and fire alarms. These preceding works highlight the dire state of consumer IoT security, vulnerable to exploitation through improper device administration and management.

Indeed, such works offer valuable insight on vulnerabilities within poorly managed devices. Improper management such as leaving unused ports and services open, failing to update default credentials and installing potentially unsafe applications leaves consumer IoT devices vulnerable to exploitation. Complementary to these findings, our methodology for retrieving device-specific information requires actively scanning and searching for open ports to recover service banners. By targeting a total of 45 unique ports that were selected based on the popularity of hosted services, as well as determined by previous artifacts of malware, we are able to communicate with improperly secured devices. Successfully retrieving service banners from open ports reveals text-based information detailing device specifications and assists with fingerprinting compromised devices.

IoT-centric protocol vulnerabilities. Poor device management leaves IoT devices vulnerable; however, it is not the only threat to consumer IoT device security. IoT-specific protocols are vulnerable to exploitation as well. The radio protocols Z-Wave and ZigBee are popular IoT protocols due to their reliability when communicating signals indoors, specifically through concrete walls. Fouladi and Ghanoun released a comprehensive work detailing the Z-Wave protocol stack layers, including the development of a device for intercepting transmissions [20]. The proposed device, labeled as *Z-Force*, decoded and disassembled Z-Wave transmissions, allowing for devices utilizing the Z-Wave protocol to be remotely exploited with the attacker remaining undetected. Fernandes et al. continued the analysis of the Z-Wave protocols, compromising encrypted keys used by automated IoT lock systems [19].

Similarly, Ur et al. [50] studied IoT access control within home automation devices. Multiple Internet-connected IoT devices were tested, with many utilizing the Z-Wave protocol, including a wireless LED lighting system, bathroom scale and an electronic door lock that are available in consumer markets. By investigating ownership processes, roles and monitoring device capabilities, the authors revealed that IoT devices allow for physical access control. Further, this work illustrates a major IoT security flaw, where many IoT devices do not properly log activities. Important information such as labeling which user account issued a specific operation, what time the operation was issued and the state of the device are not tracked. Improper logging prevents device owners

who regularly monitor their devices from accurately auditing device activity and user privileges, allowing attackers to remain stealthily hidden.

Alternatively, Ronen and Shamir demonstrated a number of vulnerabilities and attack vectors which can be used to remotely exploit IoT lighting systems, followed by launching several attacks using compromised lights [42]. The authors proved such devices were vulnerable to various misdemeanors and certain exploits threatened the entire host network; within a corporate setting, classified data was remotely ex-filtrated without a trace. Moreover, exploited lights can be used to harm device owners, revealed by testing light strobing at specific, seizure-inducing frequencies. Through similar vectors, Ho et al. investigated protocol and system vulnerabilities in IoT smart locks [23]. The authors demonstrated that exploited locks could be forcibly opened through remote commands and that safety procedures meant to alert device owners of exploitation can be revoked or disabled.

These works highlight the insecurity of IoT-specific protocols, which can be exploited for the remote compromising of consumer IoT devices. Similarly to the aforementioned improper management of IoT devices, the lack of secure IoT architectures continues to leave devices vulnerable. To quantify the number of infected IoT devices, this work provides Internet-scale empirical evidence of such remote exploitation of IoT devices, passively identifying the ports and services on vulnerable hosts that have been compromised.

Fingerprinting compromised IoT devices. Preceding works begin with aggregating network traffic data sets through network telescopes, honeypots and similar collection architectures. Passively collected network traffic is used to identify infected IP addresses, yet additional processing is required to identify IoT-specific characteristics. To this end, many works leverage Internet search engines such as Shodan [45] and Censys [47] to retrieve Internet-scale device information [37, 48]. Internet search engines actively scan the Internet space at fixed intervals to locate devices with open ports and services, retrieving available service banners. Banners contain text-based information regarding device specifications and scripts can be developed to search and retrieve specific regular expressions that correlate with device information [28]. Antonakakis et al. leveraged a network telescope and multiple honeypots to aggregate network traffic related to the *Mirai* malware [2]. Mirai-generated network traffic carries a unique identifier - the packet header's TCP sequence value is equal to the destination IP address. After identifying 1.2 million infected hosts utilizing their network telescope, Censys scans retrieved valuable device information and the work revealed the Mirai campaign's device composition and multiple attack vectors. Similarly, Cetin et al. deployed a network telescope consisting of 300,000 IP addresses to collect network generated by Mirai strains [11]. Using banners retrieved from Censys and *Nmap*, infected IoT devices were identified and notification-based remediation efforts were successfully tested. Further, Shaikh et al. combined Shodan and Censys database results to identify compromised IoT devices at large, reporting empirical statistics on infection rates [44].

In contrast, a number of methodologies collect localized network traffic for IoT device classification. Meidan et al. generated network traffic using a collection of IoT devices, PCs and smartphones [31]. Specific packet features relating to TCP connections were extracted and used as a training data set for a machine learning classifier, which accurately differentiated IoT-generated traffic from non-IoT. Alternatively, Miettinen et al. captured IoT-specific network traffic generated during initial setup processes [32]. These signatures were mapped using random forest classification, achieving a relatively high accuracy when determining device characteristics. Guo et al. located manufacturer-specific servers used for device set up, downloading updates and running popular applications [22]. By using IoT devices within a controlled environment, the preconfigured IP routes used to automatically connect with manufacturer servers were revealed. Mapping the traffic sent

to each server's address, device-specific features were identified and successfully used to identify the IoT devices active within a college campus.

The aforementioned literature face certain limitations when translated to Internet-scale measurements. First, not every malware will carry such a profound signature as Mirai. Works that leverage malware-specific signatures are unable to identify newly emerged malware such as those exploiting zero-day vulnerabilities. Moreover, Internet search engines detect devices with open ports and services; however, there is no guarantee that they are already compromised. Further processing and correlation between data sets of known compromised addresses is required to accurately leverage Internet search engine data bases. Additionally, recent malware has been found to be closing open ports and services, preventing reinfection by competing malware and reducing the probability of Internet search engines detecting these compromised devices [2, 27]. Furthermore, testing within local networks does not offer an Internet-scale perspective, while these works additionally require physical IoT devices to generate specific, localized network traffic. In contrast, our work leverages a complex set of rules to aggregate a labeled data set from retrieved service banners. This data is then fed into machine learning classifiers for training to fingerprint unreachable compromised IoT devices and predict device characteristics from extracted packet features.

4 INFERRING INFECTED CONSUMER IOT DEVICES

The adopted classifier can be described by three sequential steps: (1) passive collection of network traffic by leveraging a network telescope (Step 1 of Figure 1); (2) Internet-wide active analysis through scanning, banner grabbing and characteristic labeling (Steps 2-4 of Figure 1); (3) and lastly, utilizing input data sets to develop machine learning classifiers to accurately fingerprint

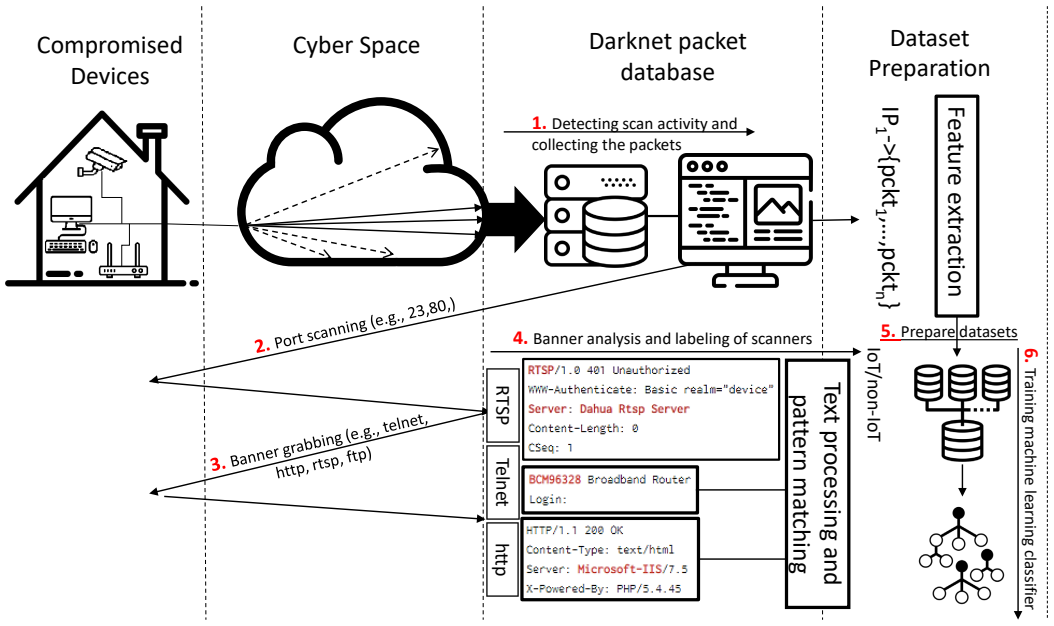


Fig. 1. The components of the proposed approach

4.1 Passive Measurements: Network telescope

In this section, we elaborate upon the methodology used to preprocess the network traffic captured at a network telescope before inferring Internet scanning activities.

Data collection and preparation. To begin aggregating the initial data set, we have utilized a /8 network telescope operated by the Center for Applied Internet Data Analysis (CAIDA) [10]. Encompassing over 16 million routable IP addresses, such a large-scale network telescope offers an enormous vantage point for collecting Internet-scale unsolicited traffic, with roughly 3.6 terabytes of network data collected per 24-hour time increment. While this specific dataset is subject to MOUs and cannot be shared, interested readers can request access to CAIDA's network telescope through their DHS IMPACT initiative. Despite the advantages of an extensive network telescope, not all of the collected traffic is relevant to our study, with unnecessary packets in the form of DDoS attack backscatter and misconfiguration traffic [18]. DDoS backscatter is generated as a result of malicious entities spoofing IP addresses located within the network telescope to launch attacks, followed by the targeted victim responding to the spoofed address. Misconfiguration traffic may be generated by hardware and software faults, or improperly configured network routing. To this end, this work employs, from a previous work [5], a darknet-specific, probabilistic sanitization model that identifies and filters out misconfiguration traffic, cleansing a network traffic data set to prepare it for further analysis and processing. Rather than relying on arbitrary thresholds for packet counts in specific time intervals, the probabilistic model calculates the likelihood that a traffic source intentionally scanned a specific destination. For space limitations and to keep the focus of this article on the presented work, we do not elaborate further on the inner details of the sanitization model, but kindly refer the reader to [5] for more details.

Inferring probing activities. Following the sanitization of misconfiguration traffic and related noise, compromised host addresses were identified through analyzing scan-like activities of scanner packets [35, 36]. A Threshold Random Walk (TRW) probing algorithm extracted packet flows corresponding to compromised hosts. The TRW algorithm searched for subsequently ordered packets attributed to the same source IP address within a 300 second time interval, employing a 64 packet threshold to determine if a source host was intentionally scanning the network telescope space. This metric is a typically sound threshold when attempting to extract scanners from darknet data sets [17]. In contrast, if the time interval expired without the threshold being met, the host was not labeled as a scanner and its associated records were removed from the data set. Flows that met the specified threshold within the time interval were extracted and labeled as scanning hosts.

Packet feature extraction for IoT classification. After grouping related packets into flows, packet features were extracted to further supplement the training data set. Because the majority of observed traffic is TCP SYN scans, the primarily extracted features reside within TCP and IP header fields. These features include values such as the application protocol, type of service, total length, time to live, source IP, destination IP, etc., for a total of 18 unique classifier fields. Additionally, previous results revealed the presence of assorted TCP OPTIONS being set, possibly to evade detection from firewalls or intrusion detection systems. To address our findings, 6 additional options were included as fields within the feature set: NOP, MSS, WSCALE, SACKOK, SACK and TIMESTAMP (displayed in Table 1). Furthermore, applicable features were extracted from packets utilizing UDP and ICMP, while a constant 0 value is used for unavailable fields. This constant is also used for TCP packet features that are not present or missing. Lastly, the minimum, first quantile,

median, third quantile and maximum values are computed for each feature, resulting in a total of 120 features comprising the classifier's training and test data sets, displayed in Figure 2.

The elements of our extensive feature set are not weighted equally, with a number of features offering distinct insights into malware scanning trends and IoT-centric bots. Primarily, the destination port and destination IP address of each packet was used to identify the services that a particular malware targeted. Moreover, other fields such as the TTL and TCP window size were used to investigate operating system firmware. However, while these features may require a higher preference within our feature set, malware is capable of altering packet header fields when scanning the Internet space. Commonly identified packet features and signatures are actively changed in

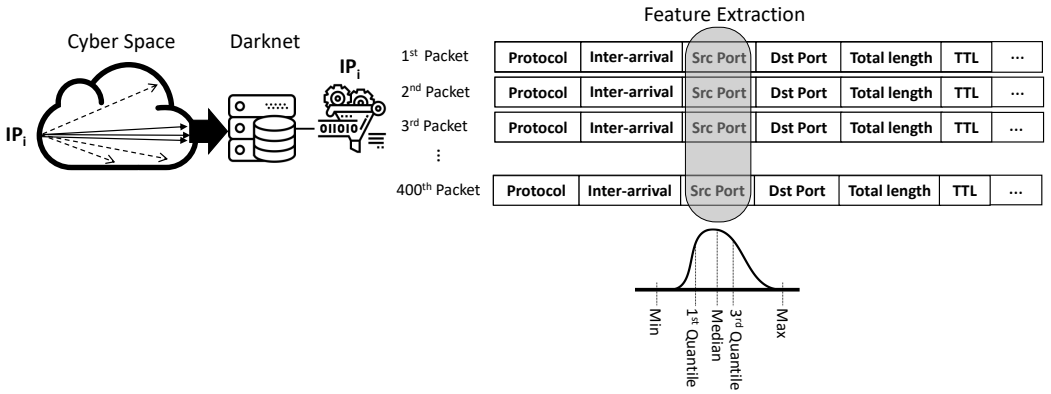


Fig. 2. Extracted components of each field

4.2 Active Measurements: Building the data set for labeling

This section elaborates upon our methodology for retrieving IoT-device characteristics that are necessary for creating an accurately labeled training data set.

Port scanning and banner grabbing. After identifying compromised hosts actively scanning the network telescope, we utilized the Internet scanning tool *ZMap* [14] and the application scanner *ZGrab* [13] to provide comprehensive analysis on device specifications. In order to verify the integrity of returned information, it was imperative to immediately reverse scan detected hosts to prevent any potential dynamic IP reallocation errors. Each identified IP address was scanned across the 45 most prevalent ports, selected based on the popularity of allocated default ports and services, typically left open in IoT devices. From these ports, we obtained service banner fields and application details from various protocols such as HTTP(s), TELNET, SMTP(s), IMAP(s), POP3(s), SSH and FTP, etc. Furthermore, additional scanning modules were developed to extract RTSP and SIP banners.

Device labeling. Internet search engines such as *Censys* release device information from their recurring Internet-wide scans. Statistics on regularly open protocols, manufacturers, vendors, device type, model and operating system are included. Furthermore, scanning software such as *Nmap* and *ZTag* release similarly orchestrated lists. While it is unrealistic to assume these lists

Categories	Description	Data Type	Range
Target-based			
Protocol	Targeted protocols that malware target	Categorical	(ICMP, TCP, UDP)
Destination Port	Targeted ports and associated services	Discrete	$[0, ..., 2^{16} - 1]$
Length-based			
Total Length	The length of an entire IP packet (in bytes)	Discrete	[20, ..., 52]
TCP offset	The size of the TCP header (in bytes)	Discrete	[0, ..., 60]
TCP Data Length	The size of the TCP data (in bytes)	Discrete	[0, ..., 60]
Time-based			
Inter-arrival time	The difference in time between the current and previous packet that are attributed to the same source address	Continuous	-
Other IP header fields			
Type of Service	Defines how the datagram should be used (e.g. delay, reliability, precedence, throughput, etc.)	Discrete	$[0, ..., 2^3 - 1]$
Identification	A unique number assigned to a datagram fragment, to be used during the reassembly of a fragmented datagram.	Discrete	$[0, ..., 2^{16} - 1]$
Time to Live	Counter used to determine the number of router hops before the packet is dropped. This field is OS-dependent and decrements with each hop in a route.	Discrete	$[0, ..., 2^8 - 1]$
Source IP	The IPv4 address of the sender's packet and can be leveraged for discovering the geolocation of the scanning host.	Discrete	$[0, ..., 2^{32} - 1]$
Destination IP	The IPv4 address of the receiver's packet, various malware use different patterns to scan the cyber space.	Discrete	$[0, ..., 2^{32} - 1]$
Other TCP header fields			
Source Port	Sending port	Discrete	$[0, ..., 2^{16} - 1]$
Sequence	Used to keep the packets of a segment in order. This field is manually changed by the IoT-centric Mirai malware (Seq.Number == Dest.Address).	Discrete	$[0, ..., 2^{32} - 1]$
ACK Sequence	Used to ensure that the packets of a segment are sent and received in the correct order.	Discrete	$[0, ..., 2^{32} - 1]$
Reserved	Aligns the total header size as a multiple of four bytes, to increase packet sending and retrieval processing.	Discrete	$[0, ..., 2^3 - 1]$
Flags	Eight control bits: CWR, ECE, URG, ACK, PSH, RST, SYN, FIN	Discrete	$[0, ..., 2^8 - 1]$
Window Size	The number of bytes the sender will buffer when receiving response packets. This field is often OS-dependent and can be used for OS fingerprinting.	Discrete	$[0, ..., 2^{16} - 1]$
Urgent Pointer	Pointer used to indicate the priority ranking of a packet and its related data.	Discrete	$[0, ..., 2^{16} - 1]$
TCP Options			
WSCALE	Window Scaling, determines the growth of the window size as packets are received.	Discrete	$[0, ..., 2^{16} - 1]$
MSS	Maximum Segment Size, defines the largest sized segment that will be used during a connection between two hosts.	Discrete	$[0, ..., 2^8 - 1]$
TIMESTAMP	Detailing the exact time the packet was sent.	Binary	exists/null
NOP	No Option, used as a buffer to separate different options within a packet.	Binary	exists/null
SACK-permitted	Selective Acknowledgment-permitted, identifies specific data that is allowed during a TCP connection.	Binary	exists/null
SACK	Selective Acknowledgment, used to convey extended acknowledgment for the number and specific sequence of packets received by the receiver to the sender.	Binary	exists/null

Table 1. Detailed list of extracted fields from scan packets. The minimum, first quantile, median, third quantile and maximum values are computed for each field to prepare the feature set.

include every IoT device being manufactured and sold within consumer markets, the combination of multiple publicly available lists results in an extensive, Internet-scale archive of text rules to determine device features.

We continue to expand our list of IoT device labels by creating a set of regular-expression (regex) keywords relating to IoT devices, manufacturers and vendors. These regex keywords leveraged common, recurring IoT naming conventions across multiple vendors, as well as more specific expressions to target and detail popular models. Beginning with a generalized expression developed by Yang et al. [53], we extracted IoT-specific device information from service banners. Next, we reduced false positive results with more advanced, strict regex filtering. Lastly, we extracted specific sequences with manually tailored regex. For instance, the regex used to identify all device models of the *HP Officejet Pro Printer* series is available in Table 2. Furthermore, devices found to be running multi-purpose operating systems were labeled as non-IoT, generally identified by keywords such as "Win64", "Ubuntu", "Microsoft", etc., while we labeled specialized devices as IoT when found to have operating systems such as "embedded", "RouterOS", "FritzOS" and similar systems.

Table 2. Example regular expressions used for feature extraction

	Regular Expression
General	`"[a-z]+[-]?[a-z!]*[0-9]+[-]?[-]?[a-z0-9]"`
Advanced	`"[A-Za-z0-9]+[-]?\s?[A-Za-z0-9!]*\d+[-]?\s?[A-Za-z0-9!]"`
Specific	`"Officejet\s[Pro]+?\s?[A-Z0-9]+\s?[A-Za-z0-9]+?"`

4.3 IoT-centric Machine learning classifier

This section details the machine learning classifier that was trained with our aggregated device labels. Once trained, the classifier was used to fingerprint compromised IoT devices by operating on newly received network telescope traffic.

Random Forest. Random forest classifiers are a combination of tree predictors assorted with each tree dependent on a random vector that is sampled independently, yet equally distributed across all trees in the forest. Random forest algorithms overcome several problems with decision trees, including a reduction in over-fitting and variance [8].

Support Vector Machine (SVM). Assuming the training data set of features and label pairs $(\mathbf{x_i}, y_i), i = 1, \ldots, l$ where $\mathbf{x_i} \in R^n$ and binary labels $y \in \{1, -1\}^l$, the support vector machine (SVM) [12] is based on the following optimization problem:

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \quad \text{subject to} \quad y_i(\mathbf{w}^T\phi(\mathbf{x_i}) + b) \geq 1 - \xi_i, \quad \text{and} \quad \xi_i \geq 0. \tag{1}$$

The features vector $\mathbf{x_i}$ is mapped into a higher dimensional space by the function ϕ. Next, SVM calculates the linear separating hyperplane containing the maximal margin value within this dimensional space. $C > 0$ is defined as the penalty parameter of the error term. Furthermore, $K(\mathbf{x_i}, \mathbf{x_j}) \equiv \phi(\mathbf{x_i})^T\phi(\mathbf{x_j})$ is called the kernel function. Although many kernels have been proposed, in this paper we leverage the radial basis function (RBF) which is defined as $K(\mathbf{x_i}, \mathbf{x_j}) = exp(-\gamma\|\mathbf{x_i} - \mathbf{x_j}\|^2), \gamma > 0$.

Gaussian Naive Bayes (GNB). The Naive Bayes Model is based on the Bayes Rule, which is defined as: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$. The supporting logic evaluated the posterior probability of output label (Y) given the input X. The posterior probability is derived from each of the individual probabilities from features(X_i) given the output label Y. The form of binary classification with two classes C_1 and C_2 is defined below:

$$P(C_i|x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n|C_i).P(C_i)}{P(x_1, x_2, \ldots, x_n)}, \qquad i \in 1, 2 \tag{2}$$

Additionally, Naive Bayes makes the assumption that all features, $x_1, \ldots, x_n$, are independent. Therefore we can derive the formula:

$$P(C_i|x_1, x_2, \ldots, x_n) = (\prod_{j=1}^{j=n} P(x_j|C_i)) \frac{P(C_i)}{P(x_1, x_2, \ldots, x_n)}, \qquad i \in 1, 2 \tag{3}$$

If a training data set contains multiple class labels, $P(C_i|X)$ is calculated for each class. The class with the maximum probability is chosen as the output. Lastly, the Gaussian Naive Bayes algorithm assumes that all features are following a Gaussian distribution.

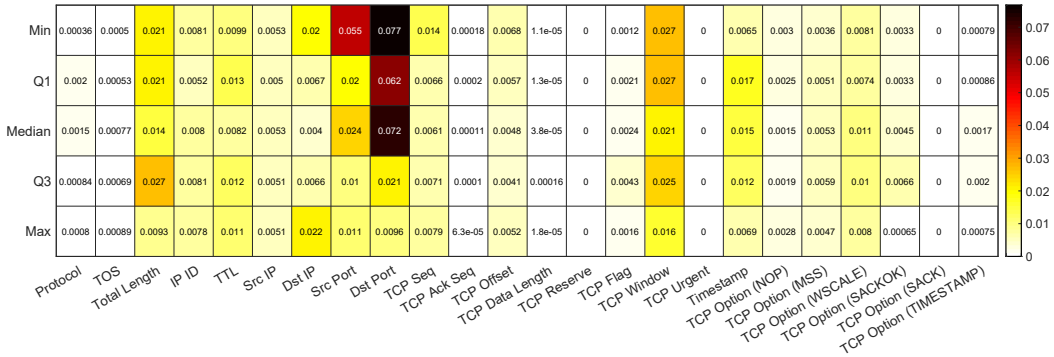
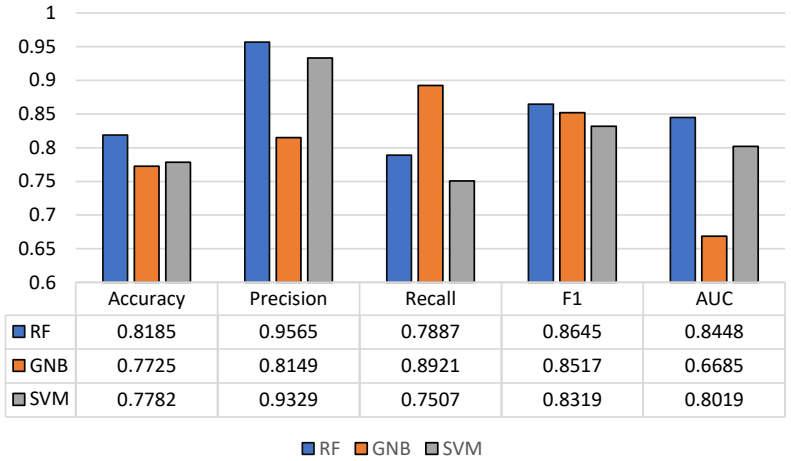


Fig. 3. Ranking of features' importance

Evaluation of machine learning classifiers. We begin by selecting three machine learning classifiers to be trained using the extracted packet features. Using the *sklearn* Python library, we investigated a Random Forest (RF), Gaussian Naive Bayes (GNB) and Support Vector Machine (SVM) classifier, evaluating their performance metrics and coming to the conclusion that the RF model was best suited for our data set, referenced in Figure 4. For the RF model, multiple hyperparameters were tested over 1000 model iterations to discover the most optimal classifier with respect to the AUC-ROC. We searched for a range of values (following Ebadi et al. [15]) on the number of estimators, maximum depth, minimum samples leaf, minimum samples split, bootstrap and criterion. The best RF model had the respective values: 70 estimators, a maximum depth of 16, 12 minimum samples leaf, 4 minimum samples split, bootstrap=false and criterion=entropy. For the SVM model, we tested different values for penalty and gamma. The best SVM model had the respective values: 1 and 0.1.

To validate the results of each classifier, we rely on standard machine learning metrics including precision, recall, F-measure and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for determining IoT devices. Precision is the ratio of correctly classified IoT devices over every



	Accuracy	Precision	Recall	F1	AUC
RF	0.8185	0.9565	0.7887	0.8645	0.8448
GNB	0.7725	0.8149	0.8921	0.8517	0.6685
SVM	0.7782	0.9329	0.7507	0.8319	0.8019

Fig. 4. Classifier metrics on training and test data set

For this work, the three aforementioned shallow machine learning algorithms were chosen after the deliberation and analysis of similar machine learning classifiers in our extended publications [38, 39]. The results of these works proved that with the correct feature set, both shallow and deep learning algorithms will have sufficiently acceptable results when fingerprinting compromised IoT devices. Utilizing the classifiers developed in the aforementioned works, this comprehensive analysis illustrates the weight of each packet header field that comprised the feature set. Visualized in Figure 3, the investigation revealed that correlating the distribution of a scanning hosts' destination ports was the highest weighted feature. The categorization of destination ports indicated the scanner's intentions, specifically hinting at malware-specific scanning trends. Similarly, the source port, total packet length, and window size features were weighted with high preference. While the remaining features did not reach the same level of weight as those mentioned, they were necessary for differentiating between hosts and IoT botnet-specific campaigns. To this end, for more in-depth analysis of our feature sets and comparisons between machine learning algorithms, indicating why we selected shallow learning classifiers, we kindly refer the reader to our extended publications [38, 39].

To create a training data set for our RF classifier, the previously extracted packet features are combined with the comprehensive list of IoT device specifications. The training data set contains device features corresponding to 40,140 identified devices, of which 28,457 are IoT and 11,683 are multi-purpose, non-IoT devices. The test data set is comprised of device features for 10,035 identified

devices, with 7,154 IoT devices and 2,881 multi-purpose, non-IoT devices. The aforementioned metrics are displayed in Figure 4.

Furthermore, despite the omission of the Mirai signature as a feature (TCP.Seq = Dest.Addr), the classifier successfully identified 58,675 Mirai-infected IoT devices, with only 259 IoT devices incorrectly classified (over 99.56% accuracy).

5 EMPIRICAL ANALYSIS

To shed light on the insecurity of consumer IoT devices, we leverage the proposed techniques to correlate the results with organizational databases on geolocation statistics and filtering devices by sector. Detailed analysis on the consumer IoT sector is reported below.

5.1 Empirical Results and Current State of Consumer IoT Insecurities

The CAIDA network telescope collected unsolicited scan traffic from 855,916 unique IP addresses throughout a 24-hour time frame on November 08, 2019. Of these identified hosts, 310,164 of them are identified as compromised IoT devices. After discovering the infected IoT devices, they were geographically identified using the Maxmind GeoLite2 database [30]. Leveraging additional business sector databases and the results from our Maxmind geolocation, we filtered IoT devices attributed to non-consumer sectors (e.g., health, financial, education, etc.).

Such databases are reliable for IP-geolocation due to the architecture implemented by the Internet Assigned Numbers Authority (IANA) for distributing the IP address space. Generally, entire IP address blocks are geographically assigned to the five major Regional Internet Registries (RIR). These RIR are then responsible for breaking their address blocks into smaller sizes and allocating them to nation-states within their jurisdiction (e.g., AFRINIC for Africa and APNIC for Asia/Pacific). Once allocated, the IP address block is then distributed to multiple ASNs, who proceed to sell or provide their services country-wide. While the IP addresses may shift within each nation-state (e.g., DHCP churn and IP reassignment), the national IP address block tends to remain constant - allowing for accurate measurements at a country-level [26].

Utilizing the MaxMind GeoLite2 database, identified devices were then organized by geographic location, with the results displayed in Figure 5 and Table 3. These statistics reveal China and Brazil to have an overwhelmingly higher number of compromised IoT devices than the rest of the world. After fingerprinting, the most significant number of infected addresses originated from China, with a total of 510,031 unique addresses (96,275 of which were IoT devices), while Brazil revealed a total of 60,181 (55,428 of which were IoT devices). The global population is not evenly distributed across countries, resulting in some having higher populations that others. This is directly correlated with the number of devices active within each country, so the aforementioned results may be skewed. Therefore, to overcome this variance, the percentages of compromised IoT-devices vs non-IoT devices is further explored.

Interestingly, only 18.87% of compromised Chinese devices were IoT, while 92.10% of Brazilian devices were IoT. Having such a high percentage of infected IoT devices indicates national emphasis on IoT security is extremely low, with little to no effort taken towards securing IoT devices in comparison towards securing traditional, multipurpose devices. Accordingly, Iran hosts the greatest number of compromised IoT devices in relation to total devices, with 94.62% of compromised devices classified as IoT. The third highest percentage was found to be Vietnam, with 71.67% addresses relating to IoT devices.

Figure 6 illustrates the top 20 Internet Service Providers (ISP) rated by the number of infected IoT devices originating from each company, with statistics spanning across multiple nations. The highest number of compromised IoT devices originated from Rostelecom in Russia, with 3,589 infected addresses. Closely following is Turkish company Turk Telekom, with 3,559 compromised

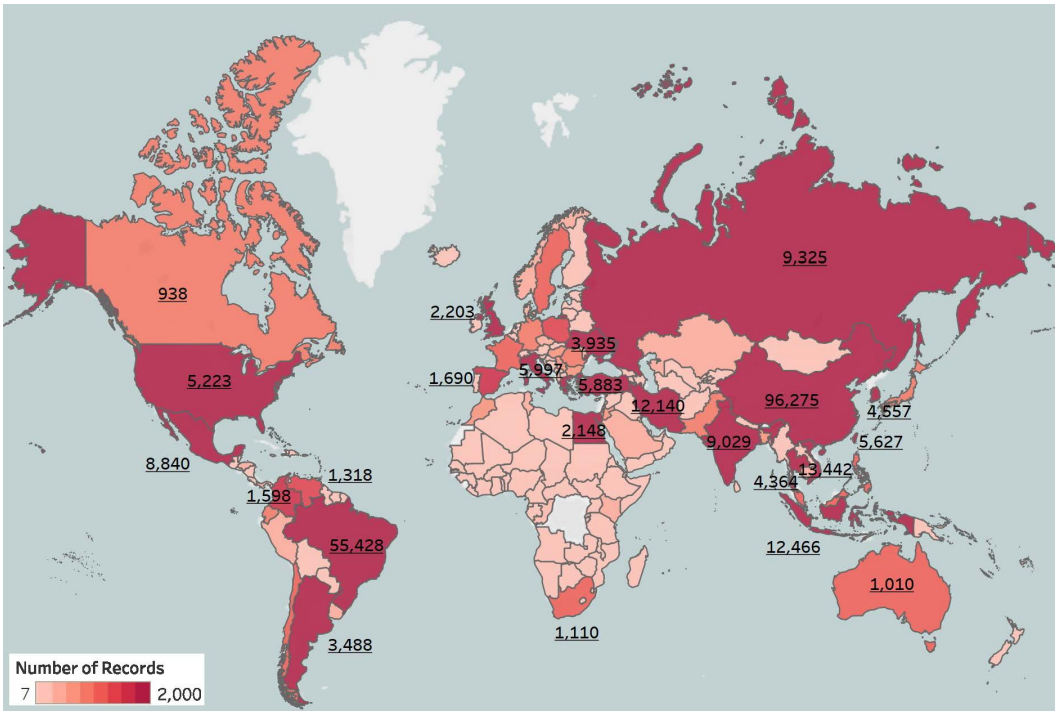


Fig. 5. Geographical location of infected IoT devices

Table 3. Relation of total compromised devices to IoT devices across the most populated regional countries.

Country	China	Brazil	Indonesia	India	Russia	USA	Vietnam	Mexico	Iran	Taiwan
Total	510,031	60,181	20,798	20,706	19,649	19,380	18,754	15,283	12,829	10,919
IoT	96,275	55,428	12,466	9,029	9,325	5,223	13,442	8,840	12,140	5,627
Percentage (%)	18.87	92.10	59.93	43.60	47.45	26.95	71.67	57.84	94.62	51.53

devices. Furthermore, 7 out of the 20 ISPs were located in China, including China Telecom Hubei (3,238) and China Telecom Hunan (2,987), ranking 3rd and 5th, respectively.

Next, we enumerate the port combinations most commonly targeted by unsolicited scan traffic. Over 1000 distinct combinations of port scanning were identified, with the top 17 most popular distributions displayed in 7. Roughly 91,000 IoT devices (29.34% of all identified IoT devices) were recorded scanning a three port distribution, 23 (TCP:Telnet), 80 (TCP:HTTP) and 8080 (TCP:HTTP/Alt). Nearly 11,500 devices (3.71% of all identified IoT devices) specifically targeted the Telnet protocol, solely scanning ports 23 and 2323. These results are consistent with historic measurements of IoT-centric malware targeting the numerous vulnerabilities in Telnet. Not all of the results follow typical trends. An unexpected 13,000 compromised IoT devices focused their scans on TCP port 5555, while 4,000 compromised IoT devices solely scanned TCP port 60001 (4.19% of all identified IoT devices). Similarly, the large number of IoT devices specifically focusing their scans against ports that do not host popular services (1024, 1588, 3128, 5984 and 9000), possibly indicates that malware campaigns are targeting a newly discovered vulnerability.

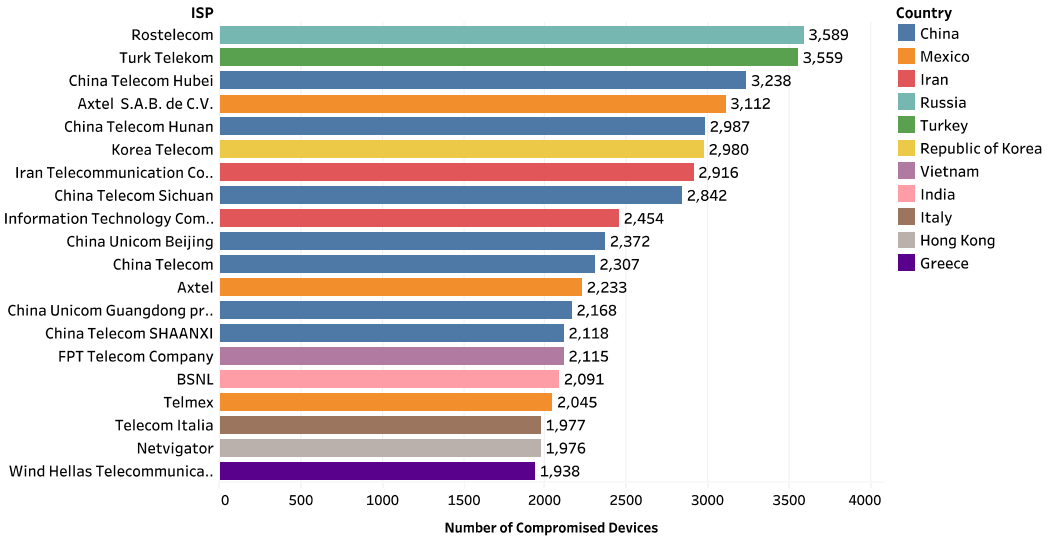


Fig. 6. Top ISPs ranked by number of compromised IoT devices

It is worth noting the centralization of compromised IoT scan patterns against the vast distribution of non-IoT patterns. From the top 17 port combinations, 216, 326 out of 310, 164 IoT devices are recorded (69.7% total). In contrast, only 123, 638 out of 545, 752 non-IoT devices are recorded (22.7% total). The stark contrast indicates that non-IoT malware target a wider distribution of ports and services while searching for vulnerabilities. While IoT scans are strictly TCP-based, non-IoT scans includes UDP and ICMP scans on ports 53, 16285, 8000 and 8080 (14.7% of top non-IoT scans).

Exploited device vendors were identified through extracted banner information, with the results displayed in Figure 8. Devices belonging to 131 total vendors were identified, with MikroTik devices demonstrating significantly higher exploitation rates than other vendors, accounting for 47.59% of all fingerprinted IoT devices. Aposonic and Hikvision displayed high exploitation rates with 11.71% and 6.57% devices carrying their brand name. Because we are unable to retrieve device information for every identified IoT device through retrieval, combined with the fact that some banners returned less information than others, the comparison of vendors against identified device types reveals interesting results. Consumer IoT devices primarily consisted of network routers (32.33% of IoT devices), cameras (28.17% of IoT devices) and DVRs (9.40% of IoT devices). Vendors such as MikroTik primarily manufacture network routers, while Aposonic and Hikvision emphasize cameras.

5.2 Evolution of Consumer IoT Infections

In order to facilitate a longitudinal analysis on the insecurity of consumer IoT devices, we compare and contrast our results with a data set from approximately one year ago, in December 2018. As mentioned in the previous section, the IANA's methodology for assigning IP address blocks results in each nation's IP address spaces remaining relatively constant. Therefore, while local IP addresses may alternate, at the national-level provided by this work, the IP addresses are considered relatively stable across the one year time frame.

Comparing the results of the two time frames indicates a 13.8% global decrease of compromised IoT devices, dropping from 359,826 (Dec. 2018) to 310,164 (Nov. 2019), as shown in Figure 9.

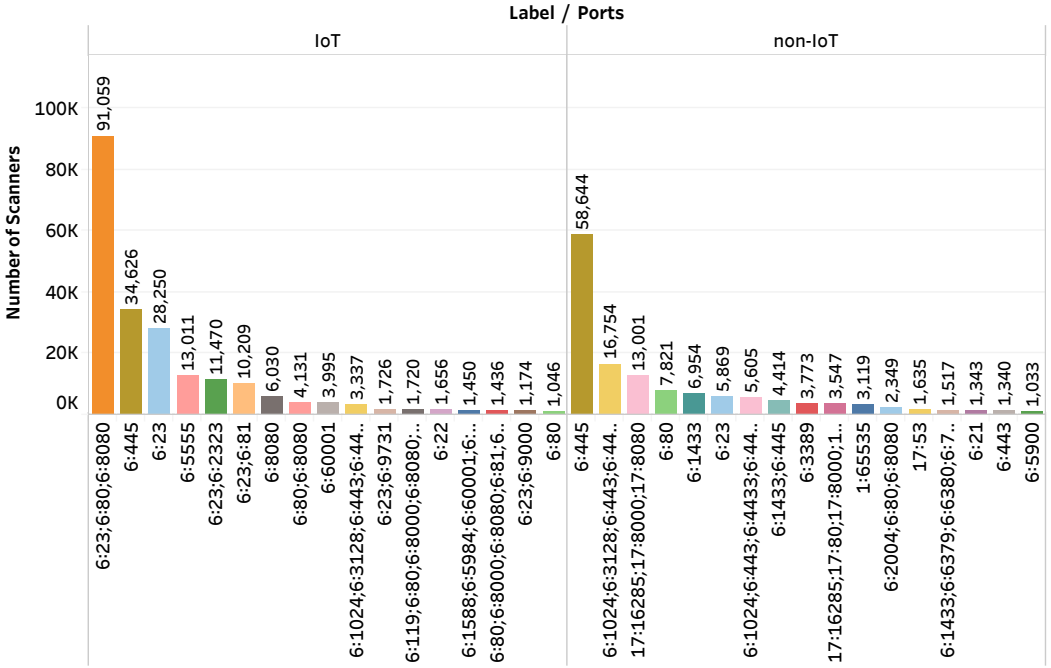


Fig. 7. Comparison of the most prevalent ports and protocols. The first number indicates protocol and the following numbers indicate the targeted ports (TCP=6, UDP=17, ICMP=1)

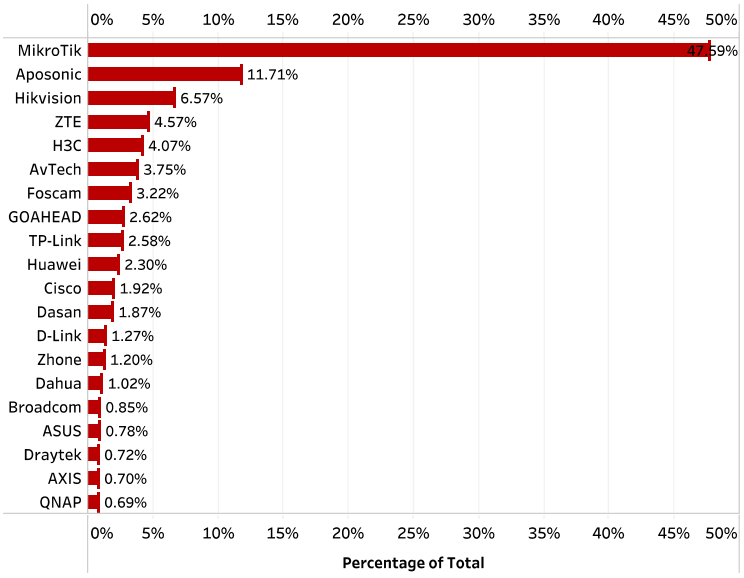


Fig. 8. Top vendors ranked by number of compromised IoT devices

Multiple conclusions can be derived from this decrease, from a global and nation-state perspective. Globally, a greater sense of cybersecurity awareness has led to the implementation of more secure

device management procedures and enhanced security practices. The employment of more rigorous security posture has resulted in lowered exploitation rates, coinciding with the visualized downward trend. Alternatively, after two years of rampantly spreading malware, specifically variants of Mirai, the vulnerable IoT ecosystem has become saturated with territorial malware. Saturation may be caused by a number of combined factors, namely, the peaked exploitation of unprotected legacy systems. At this point in time, the number of vulnerable legacy devices that have not been compromised has decreased to a point in which malware propagation is significantly slowed. This trend is further facilitated by the evolving territorial malware that prevent reinfection of compromised systems by competitors. Once saturation has occurred, combined with the gradual implementation of advanced security procedures for remediation and mitigation, the total number of infected devices will see a gradual decline.

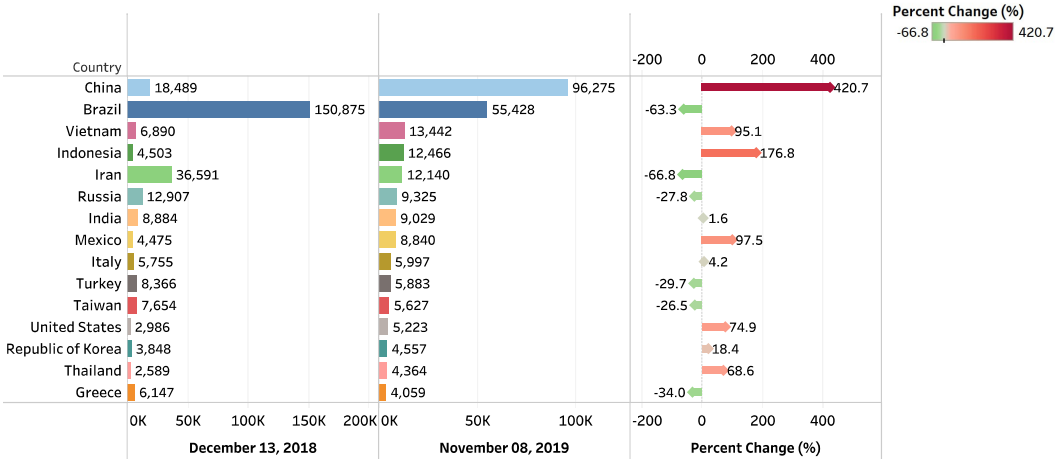


Fig. 9. December 2018 to November 2019 comparison of total number of compromised IoT devices within select countries

Regional trends relating to the increase of compromised IoT devices by country are contrasted against the effectiveness of additional mitigation and security techniques. Compared against results from last year, Brazil and Iran saw a significant decrease in compromised IoT devices. Last year, Brazil was recorded to have 150,875 compromised devices while Iran was recorded to have 36,591 devices. These figures have been reduced to 55,428 (66.3% decrease) and 12,140 (66.8% decrease), respectively. In contrast, China (420%), Indonesia (177%), Mexico (97.5%) and the United States (75%) faced momentous surges in infections. The results of this work infers that the gradual increase in the number of compromised IoT devices is attributed to evolving malware attack vectors, with malicious hackers no longer solely targeting legacy devices, but aiming to expand their exploitation across a growing number of modern manufacturer/firmware-specific vulnerabilities.

Moreover, a distinct decrease in the number of Mirai-infected devices was discovered. Last year, roughly 89,000 devices were fingerprinted with the Mirai malware, which has now dropped to 58,675 devices (roughly 34% decrease). Our results enumerates a significant decrease in port 23 and 2323 (Telnet) scans, yet an increase in previously unrecorded ports, such as 60001. This phenomenon illustrates the evolution of malware away from traditional Mirai-based attack vectors, supported by the increased number of ports targeted, specifically leveraging the UDP protocol. The downward trend of the Mirai botnet hints to a large-scale shift of targets for malware which are continually searching for and exploiting newly emerged vulnerabilities.

5.3 Comparative Analysis of Consumer IoT Infections and Countries' Security Posture

This section elaborates upon the classifier's results in comparison with the Global Cybersecurity Index (GCI). As noted in Section 2, the GCI is a yearly rating of the cybersecurity commitment in countries from a regional and global perspective, ranking the level of engagement of cybersecurity programs and initiatives.

The International Telecom Union offers three GCI classifications for countries, defining their security postures, ranging from high (1.000-0.670), medium (0.669-0.340) and low (0.339-0.000). Countries with a high GCI emphasize national cybersecurity efforts with laws and funded programs, while countries with a low GCI show little to no government-sponsored cybersecurity efforts. Leveraging geographical information reported by Yang et al. [53], the number of deployed Internet-facing IoT devices per country are revealed. For each reported country, we calculate the cybersecurity performance of IoT device protection using the metric: $(H/C[CountryX]) = \frac{\text{Estimated Number of Healthy IoT Devices in Country X}}{\text{Number of Compromised IoT Devices in Country X}}$.

Detailed results are displayed in Table 4, including the reported number of deployed IoT devices, the number of identified compromised IoT devices discovered by this work and the related CGI score for top hosting countries. The Netherlands (1465 H/C and 0.885 GCI), Germany (768 H/C and 0.849 GCI) and the United States (745.56 H/C and 0.926 GCI), Netherlands) earned 'high' GCI ratings. Validating the assumption that countries with a higher GCI typically have lower percentages of compromised IoT devices, all three nations were recorded to have no greater than 0.00125% infection rates. We further evaluate the consistency of GCI ratings by reporting that Brazil was rated to have a 'medium' GCI (6.92 H/C and 0.577 GCI), with a 12.6% IoT infection rate. Brazil's lower GCI rating directly correlates with a higher infection rate when compared with countries that earned 'high' GCI ratings.

Table 4. Detailed report relating to top hosting countries of Internet-facing IoT devices. The total number of deployed devices is retrieved from [53].

Country	USA	China	UK	Germany	Russia	Rep. of Korea	Brazil	Australia
Total Deployed	3,899,306	1,852,239	1,024,317	753,771	578,704	557,697	439,219	432,273
Compromised IoT	5,223	96,275	2,203	979	9,325	4,557	55,428	1,010
H/C	745.56	18.23	463.96	768.93	61.05	121.38	6.92	426.99
GCI	0.926	0.828	0.931	0.849	0.836	0.873	0.577	0.89
Country	France	Japan	Italy	Mexico	Canada	Argentina	Netherlands	India
Total Deployed	399,487	393,748	295,424	295,424	261,446	254,841	233,178	203,878
Compromised IoT	1,088	957	5,997	8,840	938	3,488	159	9,029
H/C	366.17	410.43	48.26	32.41	277.72	72.06	1,465.52	21.58
GCI	0.918	0.88	0.837	0.629	0.892	0.407	0.885	0.719

Figure 10 illustrates the GCI score of countries based on their corresponding H/C score. Although it is expected to see countries with a low CGI to have poor cybersecurity practices and a smaller H/C metric, countries such as China, Russia, Italy and India are categorized in the high CGI range, with very low H/C scores. Despite their emphasis on secure practices and device management, their high ratio of infected devices reveals that they are having difficulties securing IoT devices en masse. Further, this indicates that many cybersecurity practices are deployed at government and large corporate levels, yet consumer IoT devices are left unprotected. The variance of the H/C metric against the actual number of compromised devices emphasizes the disadvantages of the consumer market having no specialized oversight. Without a department or agency to enforce cybersecurity standards, consumer device owners do not effectively protect their devices, resulting in the discovery of infected IoT devices at alarming rates.

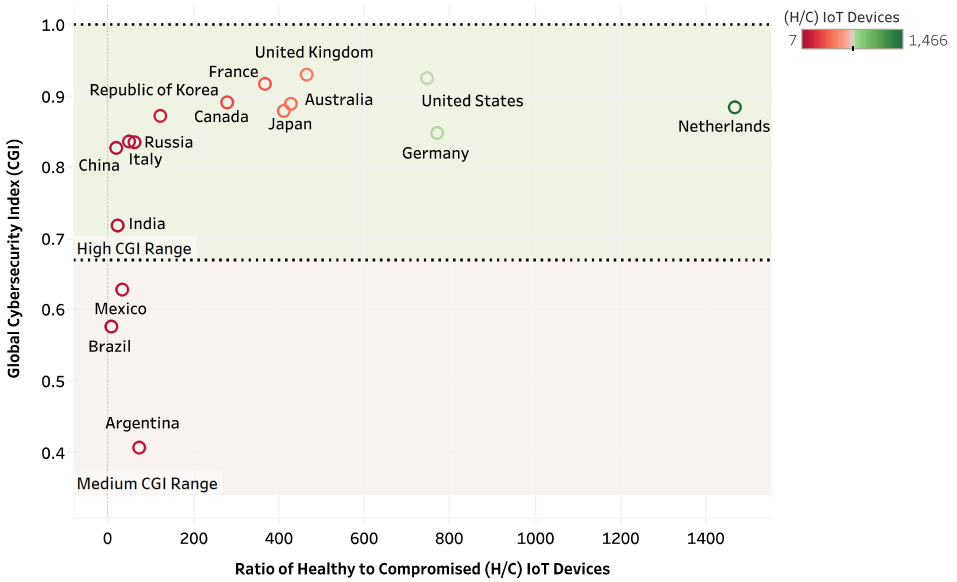


Fig. 10. Comparing Global Cybersecurity Index with the percentage of compromised IoT devices (the total number of deployed IoT devices in each country) for the top countries by number of deployed IoT devices

6 DISCUSSION

This section discusses global trends discovered during the evaluation of the proposed classifier and offers probable causes of such trends.

The results produced by this work illustrate the severity of IoT device security. In a single day, over 300,000 infected IoT devices were discovered to be actively scanning the Internet space to propagate and continue spreading their malware. Geolocation trends pinpoint the severity of malware threats within specific countries, specifically China and Brazil, with tens of thousands of compromised IoT devices. Additionally, this work's results reveal that IoT cameras were the single most exploited device, validating the claim that the consumer IoT sector is being explicitly targeted.

Furthermore, the results of this work were compared against data from last year, which indicated a global decrease in compromised IoT devices. One possible cause is the saturation of the vulnerable IoT environment. Following the emergence of Mirai and its variants, a monumental increase in exploited IoT devices was recorded. However, this trend has begun slowing down due to a variety of factors. First, the previous study revealed compromised devices across 50 vendors, many of which are industry leaders in their respective market (i.e, MikroTik routers or Aposonic cameras). However, this work revealed compromised devices manufactured by 131 different vendors, including brands that were unaffected or not recorded last year. The increased distribution of malware across multiple vendors indicates that the vast majority of legacy systems have already been compromised. Because embedded legacy systems are largely unprotected and do not receive regular patches or updates, they were the first to fall victim to Mirai-derived malware. As the number of vulnerable, yet non-exploited devices dwindled, malware aggressively competed with one another, actively closing vulnerable ports and services to prevent reinfection from other sources. Therefore, the number of legacy devices that can be infected has dramatically been reduced and malware authors have needed to branch out and target new vendors. Further, the distribution of targeted ports has changed in a similar manner, with previous Mirai-targeted ports seeing a decrease in traffic,

while newly recorded UDP ports have seen a significant increase of unsolicited scans. Secondly, the newly emerging trend of smaller vendors being exploited may result in a future catastrophic boom in malware. With hackers beginning to eye additional devices, the number of vulnerabilities and zero-day exploits being targeted across a larger number of systems may result in exponential malware propagation and an increased number of compromised devices.

Lastly, when compared against the GCI and related metrics, the results of this work illustrate the vulnerability of the consumer market. Despite many countries having a high GCI rating, strong cybersecurity programs and active laws or processes in place to secure their cyberspace, these countries still show a large number of infected IoT devices. The current global metrics for grading cybersecurity efforts is heavily geared towards government and institutional efforts, yet does not properly categorize the consumer market. Many device owners are largely unaware of device vulnerabilities and security remains an anomaly. To this end, it is paramount that works such as this shed light on the current state of insecurity within the consumer IoT market. Device owners need to be informed of secure device management procedures and take an active lead towards protecting their devices.

7 CONCLUDING REMARKS AND FUTURE WORKS

As the IoT paradigm continues to expand into critical infrastructure, governments and consumer sectors, the amount of malicious malware and entities attempting to exploit these IoT devices increases. The consumer sector is particularly at risk, as it lacks a single entity overseeing device management and its security posture. Moreover, global metrics such as GCI attempt to measure a country's emphasis on safe and secure cybersecurity practices yet do not successfully transition into the consumer sector, demonstrated by the results of this work revealing an enormous number of compromised consumer IoT devices.

This work expands upon related IoT-centric research by introducing a generalized, large-scale macroscopic methodology to infer Internet-scale compromised IoT devices and report crucial empirical metrics within the consumer IoT sector. Over a course of 24 hours, 855,916 compromised IP addresses were identified, with 310,164 being attributed to infected IoT devices. Country and vendor-specific results were compared against data from one year ago, indicating a global decrease in the number of infected IoT devices. While a number of countries such as Brazil and Iran saw a decrease in compromised IoT devices, countries such as China, Indonesia, Mexico and the United States of America showed significant increases. This work sheds light on the necessity of further mitigation and remediation efforts, specifically tailored towards consumer IoT devices.

Future work will expand upon the current methodology to overcome a number of limitations. This work leveraged shallow machine learning classifiers for compromised consumer IoT device classification; however, future studies can develop deep learning models, comparing the effectiveness against the provided shallow models. Furthermore, IoT-specific malware samples can be extracted and analyzed to create advanced feature sets and further improve the classifier's functionality. Lastly, remediation techniques can be explored to rectify compromised IoT devices, at a localized or global scale.

ACKNOWLEDGMENTS

The authors would like to acknowledge their gratitude in advance towards the anonymous reviewers and editors for their constructive assessment and feedback. This work was supported by the U.S. National Science Foundation (NSF) (Office of Advanced Cyberinfrastructure (OAC) #1907821).

REFERENCES

[1] Bako Ali and Ali Awad. 2018. Cyber and physical security vulnerability assessment for IoT-based smart homes. *Sensors* 18 (2018), 817.

[2] Manos Antonakakis, Tim April, Michael Bailey, Matt Bernhard, Elie Bursztein, Jaime Cochran, Zakir Durumeric, J Alex Halderman, Luca Invernizzi, Michalis Kallitsis, et al. 2017. Understanding the mirai botnet. In *26th {USENIX} Security Symposium ({USENIX} Security 17)*. 1093–1110.

[3] Elisa Bertino and Nayeem Islam. 2017. Botnets and internet of things security. *Computer* (2017), 76–79. https://doi.org/10.1109/MC.2017.62

[4] Elias Bou-Harb. 2016. A brief survey of security approaches for cyber-physical systems. In *2016 8th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*. IEEE, 1–5.

[5] Elias Bou-Harb. 2016. A probabilistic model to preprocess darknet data for cyber threat intelligence generation. In *2016 IEEE International Conference on Communications (ICC)*. IEEE, 1–6.

[6] Elias Bou-Harb, Mourad Debbabi, and Chadi Assi. 2016. A novel cyber security capability: Inferring Internet-scale infections by correlating malware and probing activities. *Computer Networks* 94 (2016), 327–343.

[7] Elias Bou-Harb, Claude Fachkha, Mourad Debbabi, and Chadi Assi. 2014. Inferring internet-scale infections by correlating malware and probing activities. In *2014 IEEE International Conference on Communications (ICC)*. IEEE, 640–646.

[8] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[9] Joseph Bugeja, Andreas Jacobsson, and Paul Davidsson. 2016. On privacy and security challenges in smart connected homes. In *2016 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 172–175.

[10] CAIDA. 2018. UCSD Network Telescope – Near-Real-Time Network Telescope Dataset. http://www.caida.org/data/passive/telescope-near-real-time_dataset.xml. [Online; accessed 10-Feb-2019].

[11] Orçun Çetin, Carlos Gañán, Lisette Altena, Takahiro Kasama, Daisuke Inoue, Kazuki Tamiya, Ying Tie, Katsunari Yoshioka, and Michel van Eeten. 2019. Cleaning Up the Internet of Evil Things: Real-World Evidence on ISP and Consumer Efforts to Remove Mirai.. In *NDSS*.

[12] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[13] Zakir Durumeric, David Adrian, Ariana Mirian, Michael Bailey, and J Alex Halderman. 2015. A search engine backed by Internet-wide scanning. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 542–553.

[14] Zakir Durumeric, Eric Wustrow, and J Alex Halderman. 2013. ZMap: Fast Internet-wide Scanning and Its Security Applications.. In *USENIX Security Symposium*, Vol. 8. 47–53.

[15] Nima Ebadi, Brandon Lwowski, Mehrad Jaloli, and Paul Rad. 2019. Implicit Life Event Discovery From Call Transcripts Using Temporal Input Transformation Network. *IEEE Access* 7 (2019), 172178–172189.

[16] Sam Edwards and Ioannis Profetis. 2016. Hajime: Analysis of a decentralized internet worm for IoT devices. *Rapidity Networks* 16 (2016).

[17] Claude Fachkha, Elias Bou-Harb, Anastasis Keliris, Nasir D Memon, and Mustaque Ahamad. 2017. Internet-scale Probing of CPS: Inference, Characterization and Orchestration Analysis.. In *NDSS*.

[18] Claude Fachkha and Mourad Debbabi. 2016. Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization. *IEEE Communications Surveys & Tutorials* 18 (2016), 1197–1227. https://doi.org/10.1109/COMST.2015.2497690

[19] Earlence Fernandes, Jaeyeon Jung, and Atul Prakash. 2016. Security analysis of emerging smart home applications. In *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 636–654.

[20] Behrang Fouladi and Sahand Ghanoun. 2013. Security evaluation of the Z-Wave wireless protocol. *Black hat USA* 24 (2013), 1–2.

[21] Dimitris Geneiatakis, Ioannis Kounelis, Ricardo Neisse, Igor Nai-Fovino, Gary Steri, and Gianmarco Baldini. 2017. Security and privacy issues for an IoT based smart home. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 1292–1297.

[22] Hang Guo and John Heidemann. 2018. IP-based IoT device detection. In *Proceedings of the 2018 Workshop on IoT Security and Privacy*. ACM, 36–42.

[23] Grant Ho, Derek Leung, Pratyush Mishra, Ashkan Hosseini, Dawn Song, and David Wagner. 2016. Smart locks: Lessons for securing commodity internet of things devices. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*. ACM, 461–472.

[24] Martin Husák, Jana Komárková, Elias Bou-Harb, and Pavel Čeleda. 2018. Survey of attack projection, prediction, and forecasting in cyber security. *IEEE Communications Surveys & Tutorials* 21, 1 (2018), 640–660.

[25] Martin Husák, Nataliia Neshenko, Morteza Safaei Pour, Elias Bou-Harb, and Pavel Čeleda. 2018. Assessing internet-wide cyber situational awareness of critical sectors. In *Proceedings of the 13th International Conference on Availability, Reliability and Security*. 1–6.

[26] IANA. 2020. https://www.iana.org/.
[27] Constantinos Kolias, Georgios Kambourakis, Angelos Stavrou, and Jeffrey Voas. 2017. DDoS in the IoT: Mirai and other botnets. *Computer* 50 (2017), 80–84. https://doi.org/10.1109/MC.2017.201
[28] Deepak Kumar, Kelly Shen, Benton Case, Deepali Garg, Galina Alperovich, Dmitry Kuznetsov, Rajarshi Gupta, and Zakir Durumeric. 2019. All things considered: an analysis of IoT devices on home networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 1169–1185.
[29] Huichen Lin and Neil Bergmann. 2016. IoT privacy and security challenges for smart home environments. *Information* 7 (2016), 44.
[30] MaxMind. 2019. https://dev.maxmind.com/geoip/geoip2/geolite2/.
[31] Yair Meidan, Michael Bohadana, Asaf Shabtai, Juan David Guarnizo, Martín Ochoa, Nils Ole Tippenhauer, and Yuval Elovici. 2017. ProfilIoT: a machine learning approach for IoT device identification based on network traffic analysis. In *Proceedings of the symposium on applied computing*. ACM, 506–509.
[32] Markus Miettinen, Samuel Marchal, Ibbad Hafeez, N Asokan, Ahmad-Reza Sadeghi, and Sasu Tarkoma. 2017. IoT Sentinel: Automated device-type identification for security enforcement in IoT. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2177–2184.
[33] Hamidreza Moradi, Wei Wang, and Dakai Zhu. 2019. Adaptive Performance Modeling and Prediction of Applications in Multi-Tenant Clouds. In *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*. IEEE, 638–645.
[34] Joern Ploennigs, John Cohn, and Andy Stanford-Clark. 2018. The Future of IoT. *IEEE Internet of Things Magazine* 1 (2018), 28–33. https://doi.org/10.1109/IOTM.2018.1700021
[35] Morteza Safaei Pour and Elias Bou-Harb. 2018. Implications of theoretic derivations on empirical passive measurements for effective cyber threat intelligence generation. In *2018 IEEE International Conference on Communications (ICC)*. IEEE, 1–7.
[36] Morteza Safaei Pour and Elias Bou-Harb. 2019. Theoretic derivations of scan detection operating on darknet traffic. *Computer Communications* 147 (2019), 111–121.
[37] Morteza Safaei Pour, Elias Bou-Harb, Kavita Varma, Nataliia Neshenko, Dimitris A Pados, and Kim-Kwang Raymond Choo. 2019. Comprehending the IoT cyber threat landscape: A data dimensionality reduction technique to infer and characterize Internet-scale IoT probing campaigns. *Digital Investigation* 28 (2019), S40–S49.
[38] Morteza Safaei Pour, Antonio Mangino, Kurt Friday, Matthias Rathbun, Elias Bou-Harb, Farkhund Iqbal, Sagar Samtani, Jorge Crichigno, and Nasir Ghani. 2019. On Data-driven Curation, Learning, and Analysis for Inferring Evolving Internet-of-Things (IoT) Botnets in the Wild. *Computers & Security* (2019), 101707.
[39] Morteza Safaei Pour, Antonio Mangino, Kurt Friday, Matthias Rathbun, Elias Bou-Harb, Farkhund Iqbal, Khaled Shaban, and Abdelkarim Erradi. 2019. Data-driven Curation, Learning and Analysis for Inferring Evolving IoT Botnets in the Wild. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*. ACM, 6. https://doi.org/3339252.3339272
[40] Morteza Safaei Pour and Mahmoud Salmasizadeh. 2017. A New CPA Resistant Software Implementation for Symmetric Ciphers with Smoothed Power Consumption: SIMON Case Study. *ISeCure* 9, 2 (2017).
[41] Philipp Richter and Arthur Berger. 2019. Scanning the Scanners: Sensing the Internet from a Massively Distributed Network Telescope. In *Proceedings of the Internet Measurement Conference*. ACM, 144–157.
[42] Eyal Ronen and Adi Shamir. 2016. Extended functionality attacks on IoT devices: The case of smart lights. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*. IEEE, 3–12.
[43] Ahmad-Reza Sadeghi, Christian Wachsmann, and Michael Waidner. 2015. Security and privacy challenges in industrial internet of things. In *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*. IEEE, 1–6.
[44] Farooq Shaikh, Elias Bou-Harb, Nataliia Neshenko, Andrea P Wright, and Nasir Ghani. 2018. Internet of malicious things: correlating active and passive measurements for inferring and characterizing internet-scale unsolicited IoT devices. *IEEE Communications Magazine* 56, 9 (2018), 170–177.
[45] Shodan. 2019. The search engine for Internet of things. http://shodan.io.
[46] Tianyi Song, Ruinian Li, Bo Mei, Jiguo Yu, Xiaoshuang Xing, and Xiuzhen Cheng. 2017. A privacy preserving communication protocol for IoT applications in smart homes. *IEEE Internet of Things Journal* 4 (2017), 1844–1852.
[47] Censys Team. 2017. Internet-Wide Scan Data Repository. *Retrieved* (2017), 2017.
[48] Sadegh Torabi, Elias Bou-Harb, Chadi Assi, Mario Galluscio, Amine Boukhtouta, and Mourad Debbabi. 2018. Inferring, characterizing, and investigating Internet-scale malicious IoT device activities: A network telescope perspective. In *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, 562–573.
[49] International Telecommunication Union. 2018. https://www.itu.int/en/ITU-T/.
[50] Blase Ur, Jaeyeon Jung, and Stuart Schechter. 2013. The current state of access control for smart devices in homes. In *Workshop on Home Usable Privacy and Security (HUPS)*. HUPS 2014.

[51] Charlie Wilson, Tom Hargreaves, and Richard Hauxwell-Baldwin. 2017. Benefits and risks of smart home technologies. *Energy Policy* 103 (2017), 72–83.
[52] Jacob Wurm, Khoa Hoang, Orlando Arias, Ahmad-Reza Sadeghi, and Yier Jin. 2016. Security analysis on consumer and industrial IoT devices. In *2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 519–524.
[53] Kai Yang, Qiang Li, and Limin Sun. 2019. Towards automatic fingerprinting of IoT devices in the cyberspace. *Computer Networks* 148 (2019), 318–327.