Comprehending the IoT Cyber Threat Landscape: A Data Dimensionality Reduction Technique to Infer and Characterize Internet-scale IoT Probing Campaigns

Article *in* Digital Investigation · January 2019

DOI: 10.1016/j.diin.2019.01.014

Morteza Safaei Pour
University of Texas at San Antonio
8 PUBLICATIONS **30** CITATIONS

SEE PROFILE

Elias Bou-Harb
University of Texas at San Antonio
85 PUBLICATIONS **785** CITATIONS

SEE PROFILE

Kavita Varma
Florida Atlantic University
2 PUBLICATIONS **8** CITATIONS

SEE PROFILE

Nataliia Neshenko
Florida Atlantic University
12 PUBLICATIONS **84** CITATIONS

SEE PROFILE

Comprehending the IoT Cyber Threat Landscape: A Data Dimensionality Reduction Technique to Infer and Characterize Internet-scale IoT Probing Campaigns

Morteza Safaei Pour[a,*], Elias Bou-Harb[a], Kavita Varma[b], Nataliia Neshenko[a], Dimitris Pados[b], Kim-Kwang Raymond Choo[c]

[a] *Cyber Threat Intelligence Laboratory, College of Engineering & Computer Science, Florida Atlantic University, Florida, USA*
[b] *College of Engineering & Computer Science, Florida Atlantic University, Florida, USA*
[c] *Department of Information Systems and Cyber Security, University of Texas at San Antonio, Texas, USA*

Abstract

The resource-constrained and heterogeneous nature of Internet-of-Things (IoT) devices coupled with the placement of such devices in publicly accessible venues complicate efforts to secure these devices and the networks they are connected to. The Internet-wide deployment of IoT devices also makes it challenging to operate security solutions at strategic locations within the network or to identify orchestrated activities from seemingly independent malicious events from such devices. Therefore, in this paper, we initially seek to determine the magnitude of IoT exploitations by examining more than 1 TB of passive measurement data collected from a /8 network telescope and by correlating it with 400 GB of information from the Shodan service. In the second phase of the study, we conduct in-depth discussions with Internet Service Providers (ISPs) and backbone network operators, as well as leverage geolocation databases to not only attribute such exploitations to their hosting environment (ISPs, countries, etc.) but also to classify such inferred IoT devices based on their hosting sector type (financial, education, manufacturing, etc.) and most abused IoT manufacturers. In the third phase, we automate the task of alerting realms that are determined to be hosting exploited IoT devices. Additionally, to address the problem of inferring orchestrated IoT campaigns by solely observing their activities targeting the network telescope, we further introduce a theoretically sound technique based on L1-norm PCA, and validate the utility of the proposed data dimensionality reduction technique against the conventional L2-norm PCA. Specifically, we identify "in the wild" IoT coordinated probing campaigns that are targeting generic ports and campaigns specifically searching for open resolvers (for amplification purposes). The results reveal more than 120,000 Internet-scale exploited IoT devices, some of which are operating in critical infrastructure sectors such as health and manufacturing. We also infer 140 large-scale IoT-centric probing campaigns; a sample of which includes a worldwide distributed campaign where close to 40% of its population includes video surveillance cameras from Dahua, and another very large inferred coordinated campaign consisting of more than 50,000 IoT devices. The reported findings highlight the insecurity of the IoT paradigm at large and thus demonstrate the importance of understanding such evolving threat landscape.

Keywords: IoT forensics, Big data, Probing, Network Telescopes, Network forensics, L1-norm PCA

1. Introduction

The Internet of Things (IoT) paradigm is increasingly deployed in critical infrastructure sectors, including sectors that are typically not as technologically-advanced such as agriculture. Most IoT devices are low-cost/inexpensive with limited computational capabilities (i.e., battery life and processing or storage capability). The competitive landscape (i.e., to drive cost as low as possible) and technical constraints on IoT devices also mean that it will be challenging for IoT device manufacturers to design and incorporate sophisticated security features in these devices. Intuitively, these devices can be targeted by adversaries in order to gain access to the underpinning system or the data sensed, collected or disseminated. In other words, these devices can be both the target and the tool in a cyber attack. For example, vulnerabilities in IoT devices can be exploited to attack the system or their data, as well as allow the re-programming of a device to facilitate other cyber attacks (i.e., distributed denial of service attacks or false data injection attacks) or simply to fail (Rose et al. (2015)). Malfunctioning devices can also create a number of serious security vulnerabilities. When coupled with the highly interconnected nature of IoT devices, every poorly secured Internet-connected device could potentially have an impact on the security and the resiliency of the Internet at large. For example, an infected or compromised smart home device physically located in the United States can be (ab)used to send a large volume of malicious phishing emails to recipients worldwide or to facilitate other attacks using the owner's home Wi-Fi Internet connection (Gupta et al. (2017); Do et al. (2018); Bertino and Islam (2017)). Examples of noteworthy IoT-centric malware include Mirai (Antonakakis et al. (2017)), Hajime (Edwards and Profetis (2016)), Brickerbot (Cimpanu (2017)),

[*] Corresponding author. Tel.: +1 561 931 7531
Email addresses: msafaeipour2017@fau.edu (Morteza Safaei Pour), ebouharb@fau.edu (Elias Bou-Harb), kvarma2018@fau.edu (Kavita Varma), nneshenko2016@fau.edu (Nataliia Neshenko), dpados@fau.edu (Dimitris Pados), raymond.choo@fulbrightmail.org (Kim-Kwang Raymond Choo)

Reaper (Wired.com (2017)) and VPNFilter malware (Register (2018)). Further, Soltan et al. (2018) have lately highlighted on a new class of IoT potential attacks on power grids dubbed as MadIoT, which can leverage IoT-specific botnets of high wattage devices to cause local power outages and large-scale blackouts.

Existing mitigation strategies include IoT context-aware permission models (Yu et al. (2015); Jia et al. (2017)) and those focusing on identifying and addressing protocol weaknesses (Ur et al. (2013); Ronen and Shamir (2016)). However, mitigation strategies focusing on Internet-scale issues appear to be a topic that is understudied. This is indeed due to the challenge of acquiring and curating large volumes of IoT-relevant empirical data from a large range of (heterogeneous) IoT devices. In addition, acquiring IoT-centric malware and their signatures, for example to train machine learning algorithms, is also challenging (Azmoodeh et al. (2018a,b)).

In this paper, we seek to contribute to existing IoT security efforts by proposing a macroscopic approach to infer and characterize Internet-scale unsolicited/malicious IoT devices and campaigns. Specifically, the key contributions of this paper are as follows:

- We seek to contribute to a better understanding of the IoT-specific threat landscape by correlating more than 1 TB of passive measurements with the results of active measurements from the Shodan service. This allows us to investigate the magnitude of Internet-wide IoT exploitations. We also report on the hosting environment of such IoT devices, such as the type of the hosting sector (financial, education, critical infrastructure, etc.) and most exploited IoT manufacturers and device types. We also automate the proposed methodology in near real-time and distribute alerts to relevant stakeholders that are determined to be hosting exploited IoT devices to facilitate early remediation. Further, we make available the generated near real-time IoT threat information through an indexed database and a secure (authenticated) front-end service.

- We propose a novel data dimensionality reduction technique, based on L1-norm Principal Component Analysis (PCA), which is robust against outliers and noisy/corrupted data samples. To process the large volume of IoT-specific data and to mitigate the efficiency limitation of L1-norm PCA, we present a sub-optimal greedy algorithm. By relying on a ground truth related to a large-scale orchestrated probing campaign, we validate the utility of the proposed technique against the conventional L2-norm PCA in terms of clustering accuracy. We also make the proposed technique publicly available to the research community, which can be utilized for other data-intensive applications including network forensics.

- We apply the proposed technique on passive measurements data to infer, characterize and report on unsolicited and malicious "in the wild" orchestrated IoT probing campaigns. The results reveal 140 large-scale probing campaigns that have exploited more than 120,000 IoT devices, operating in a number of different countries and hosted by ISPs in numerous countries. This allows us to identify the sources of IoT insecurity and also pave the way for further research exploration relating to such orchestrated campaigns (malware attribution issues, microscopic remediation methodologies, etc.).

The road-map of this paper is as follows. In the next section, we review related literature on IoT security, passive measurements and analysis, and PCA for dimensionality reduction. In Section 3, we present the proposed approach designed to infer unsolicited IoT network flows from darknet data, as well as introduce the L1-norm PCA technique and its corresponding validation. We will also explain how it can be applied to identify orchestrated IoT probing campaigns. Section 4 presents the findings of applying the proposed approach on empirical darknet data to infer and characterize Internet-scale unsolicited IoT devices and probing campaigns. A discussion is presented in Section 5 while Section 6 provides the concluding remarks.

2. Related Work

IoT context-aware permission models. One particularly popular line of IoT security research is IoT context-aware permission models, where collaborative models are designed to secure IoT environment from malicious actors. For instance, Yu et al. (2015) proposed a policy abstraction language that is capable of capturing relevant environmental IoT factors, security-relevant details, and cross-device interactions, to vet IoT-specific network activities. Further, the authors proposed a crowd-sourced repository where IoT operators can share derived attack signatures, which deviate from the captured benign policies. Along a similar research direction, Jia et al. (2017) proposed ContextIoT, a system that is capable of supporting complex IoT-relevant permission models by performing program-flow and runtime taint analysis. In another closely related work, Fernandes et al. (2016) proposed an approach to restrict generated traffic flows from an exploited IoT application. The approach is based on taint arithmetic, which initially tracks an application's program flow to subsequently flag policy violations. Further, He et al. (2018) also studied the problem of access control and authentication for residential IoT settings, in which multiple users with complex social relationships interact with a single device.

IoT protocol vulnerabilities. Researchers have also analyzed specific IoT security issues and protocol weaknesses. For instance, Ronen and Shamir (2016) demonstrated information leakage attacks using a set of IoT smart lights. The authors exploited protocol weaknesses to gain access to the connected local network and subsequently exfiltrated sensitive data from an air gapped office building. In a similar fashion, Ho et al. (2016) investigated state consistency and unlocking attacks by exploring protocol and system vulnerabilities in IoT smart locks. The authors demonstrated how trust models, network designs and replay activities can be instrumented to cause security issues related to the revocation procedures of such locks in addition to forcing secure locks to be accidentally unlocked. Other related work on IoT device insecurity and data exfiltration include those of Do et al. (2018) and D'Orazio et al. (2017). The related research discussed so far (i.e., IoT context-aware permission models and IoT protocol vulnerabilities) are at a *microscopic* level, focusing on specific devices or specific contexts. In contrast, this paper develops and investigates an Internet-scale, *macroscopic* perspective of IoT maliciousness by characterizing and investigating Internet-scale traffic.

IoT data capturing initiatives. Obtaining IoT-relevant empirical data is challenging, and so is sharing such datasets securely (Banerjee et al. (2018)); it is thus not surprising that there has been a few efforts to collect, curate and analyze such data. The first IoT tailored honeypot, namely, `IoTPOT`, was designed and deployed by Pa et al. (2016). IoTPOT emulates telnet services of various IoT devices running on different CPU architectures. The proposed honeypot demonstrated its capability to capture various types of malware samples which can then be used for subsequent in-depth analysis of IoT targeted attacks. Guarnizo et al. (2017) also presented the Scalable high-Interaction Honeypot (SIPHON) platform for IoT devices. The authors demonstrated how by leveraging worldwide wormholes and few physical devices, they were able to mimic various IoT devices on the Internet and to attract significant malicious traffic. The authors further characterized such traffic by elaborating on attackers' frequency and their employed protocols. Vervier and Shen (2018) elaborated on outcomes derived from low and high interaction IoT honeypots for a period of six months, to report on the IoT botnet ecosystem. Additionally, several attempts to fingerprint IoT devices were executed. For instance, recently, Meidan et al. (2017) leveraged network traffic analysis to classify IoT devices connected to an organization's network by applying techniques rooted in supervised data classification. However, unlike these approaches, we collect Internet-scale data generated from real-world devices, including those operating within orchestrated IoT campaigns.

Empirical measurements for Internet-scale characterization. In the context of empirical measurements for device characterization and vulnerability analysis, Heidemann et al. (2008); Cui and Stolfo (2010) presented empirical measurements obtained from wide-area scans. Costin et al. (2014) also statically analyzed more than 30,000 firmware images derived from embedded IoT devices to understand their insecurity, while Fachkha et al. (2017) conducted passive measurements to analyze attackers' intentions when targeting protocols of Internet-facing cyber-physical systems. The latter approach is quite similar to (Bodenheim et al. (2014)), where the authors evaluated the `Shodan` service (Materly (2009)), a search engine for Internet-connected devices, in terms of its capability in scanning and indexing online industrial control systems. Husák et al. (2018) empirically assessed Internet-wide malicious activities (e.g. DDoS attacks and cyber scanning) generated from and targeted towards business sectors and critical infrastructure. Different from these existing works, the proposed work intends to explore, develop and deploy non-intrusive passive methods and algorithms that aim at inferring and attributing Internet-wide compromised IoT devices.

Network Telescope: Measurements and Analysis. The idea of leveraging network telescopes (darknet) to monitor unused IP addresses for security purposes was first proposed in the early 1990's by Bellovin (1993) for AT&T's Bell Labs Internet-connected computers. Since then, the focus of network telescope studies has shifted; for example to facilitate discovery of the relationship between backscattered traffic and DDoS attacks (Moore et al. (2006)), worm propagation analysis (Bailey et al. (2005)), the use of time series and data mining techniques on telescope traffic (Limthong et al. (2008)), study of large-scale orchestrated probing activities (Bou-Harb et al. (2013, 2014a)), the monitoring of large-scale cyber events through telescopes (Dainotti et al. (2015); Bou-Harb et al. (2014c)), and more recently the study of amplification DDoS attacks using telescope sensors (Fachkha et al. (2014); Rossow (2014)). Additionally, Pour and Bou-Harb (2018) provided formal stochastic analysis to compare different detection systems employed on network telescopes based on their parameters such as darknet size, attacker behavior, minimum detection time and probability. This paper extends network telescope research to specially address the problem of IoT security, and at the time of this research this is a yet to be attempted approach. We will also formally analyze the orchestration behavior of compromised IoT devices by scrutinizing their network traffic extracted from passive measurements to infer and report on evolving IoT campaigns. Furthermore, the proposed and envisioned effort will be geared towards providing operational/actionable cyber security and forensic capabilities through the development of an IoT-centric cyberinfrastructure to facilitate IoT threat sharing.

Botnet detection systems. Different botnet detection systems have been proposed in the literature, such as those of Gu et al. (2007); Karasaridis et al. (2007); Gu et al. (2008); Zhao et al. (2013); Bou-Harb et al. (2016); Meidan et al. (2018). Some investigate specific protocol channels, others might require deep packet inspection or training periods, while the majority depends on malware infections and/or attack life-cycles. In this paper, we focus on inferring large-scale orchestrated IoT botnets, a yet to be investigated topic. Specifically, we analyze artifacts/features extracted from network telescope traffic, without requiring content analysis or training periods.

Applications of L1-PCA. The L1-PCA method has been utilized in a broad range of applications such as Direction of Arrival (DoA) estimation (Markopoulos et al. (2014a)), robust face recognition (Johnson and Savakis (2014); Maritato et al. (2016)), extraction of compressed-sensing surveillance of video sequences (Liu and Pados (2015); Pierantozzi et al. (2016)), indoor monitoring of patients and the elderly (Markopoulos and Ahmad (2017)), and radar-based indoor human motion classifier (Markopoulos and Ahmad (2018)). Khalid et al. (2015) also demonstrated the effectiveness of L1-PCA for dimensionality reduction in intrusion detection systems with the presence of outliers and compared it with the conventional L2-PCA. In this work, we address the efficiency issue related to the use of L1-PCA by introducing a sub-optimal, fast greedy algorithm. This allows us to evaluate the application of the L1-PCA on large network telescope datasets to infer orchestrated IoT probing campaigns. We also validate the effectiveness of L1-

PCA against the traditional L2-PCA on a (ground-truth) probing campaign, in the presence of other non-orchestrated events and outliers.

3. Proposed Approach

In this section, we detail the proposed approach to extract IoT-relevant unsolicited flows from network telescope data. Additionally, we introduce the theory behind L1-PCA including the proposed sub-optimal solution, and validate its effectiveness against the typical L2-PCA in the context of detecting coordinated probing campaigns.

3.1. Inferring and characterizing Internet-scale Exploited IoT Devices

Network telescopes (also referred to as darknets) are a collection of routable, allocated, yet unused IP addresses which operate no legitimate hosts (Fachkha and Debbabi (2016)). They are used solely to passively gather and amalgamate Internet-scale incoming traffic. Unsolicited/malicious IoT devices that attempt to probe the Internet space (searching for other devices or to exploit certain Internet-wide vulnerabilities) would inevitably target the network telescope space. Therefore, to have a broad vantage point into Internet-wide, IoT-specific probing activities, we draw upon near real-time data from a `/8` network telescope that is operated by the Center for Applied Internet Data Analysis (CAIDA (2018)); `/8` represents 1/256 of all the routable IPv4 address space. We operate a probing detection algorithm which generates darknet flows representing consecutive packets originating from each unique source IP address. The algorithm operates similar to the probing detection component embedded in `Bro` (Paxson (1999)); by receiving a packet from source *i*, the algorithm waits for the next packet. If the flow timeout expires before the arrival of the following packet from the same source *i*, the algorithm resets the threshold counter; otherwise, it increments it and compares it with a threshold to deem it as a probing event. The algorithm adopts typical parameters for scan detection, including the common probing threshold of 64 and minimum duration for a flow of 300 seconds (Rossow (2014)). Packets scanning open resolvers (for amplification purposes) which target the network telescope are analyzed with deep packet inspection to distinguish them from typical scans and attribute them to an amplification protocol (i.e., DNS, SSDP, etc.). From a performance perspective, the deployed algorithm can process close to 20,000 darknet flows per minute.

To infer unsolicited IoT devices, there is a need for identifying scans originating from IoT nodes versus those that are generated by typical hosts. Given that fingerprinting of IoT devices by solely observing network traffic is still at its infancy, we rely on the `Shodan` service (Materly (2009)) in this work, which indexes Internet-facing IoT devices. We retrieve entire IoT databases provided daily by `Shodan` and execute correlations on source IP addresses between the darknet probing sessions and `Shodan` indexed information. We automate such an approach to generate near real-time email alerts to realms hosting inferred unsolicited IoT nodes, empowered by an `ELK` backend stack and the `ELK Watcher` (Elastic (2018)).

For characterization purposes, we initially leverage a geolocation database provided by `Maxmind` to attribute the inferred unsolicited IoT sources to their hosted ISPs, ASN, cities, countries, etc. Additionally, for the past year, we have been involved in a collaborative large-scale effort, conducting discussions with numerous Internet entities across the globe to obtain rare and private information related to allocated IP blocks pertaining to certain sectors and critical infrastructure. To this end, we employ such information to further attribute such Internet-scale IoT nodes to such sectors and realms, in an attempt to provide an in-depth analysis of the global IoT cyber situational posture.

Further, given the lack of real, empirical IoT threat information, we believe that it is also very important to make the extracted threat intelligence publicly accessible to the wider research and operational communities. Along these lines, we are currently developing an IoT threat sharing facility to provide access to *(i)* raw IoT unsolicited traffic traces to support large-scale IoT data analytics, and *(ii)* generate signatures to allow further forensic investigations as well as to be employed at local realms for proactive IoT inference and mitigation. In particular, we make available the generated IoT threat insights via a front-end service at `http://faculty.eng.fau.edu/ebouharb/floridasoar/index.html`, which includes near real-time information related to Internet-scale compromised IoT devices (and geo-location and sector information) as well as basic `Snort` signatures (Roesch et al. (1999)) (automatically extracted from IoT-relevant darknet flows). As this setup requires authentication, interested parties are encouraged to contact the authors to gain free access to such information. Note that this database will be made publicly available after the conference notification date.

3.2. Feature engineering

The aim herein is to generate feature vectors related to the inferred IoT probing sources to facilitate the initial application of L1-PCA and subsequently the application of unsupervised data clustering on darknet data to infer orchestrated IoT campaigns. Clearly, the intuition here is that IoT bots operating within orchestrated campaigns share similar network traffic characteristics, which is a common "in the wild" observation (Heo and Shin (2018); Silva et al. (2013)). To this end, we select a set of features by extracting behavioral and statistical information from IoT probing events to capture the machinery of the IoT probing sources. The selected features are discussed next.

Probing Rate is the average number of received packets divided by the event duration. **Protocol** shows the Internet protocol of the IoT scan event and could take one of the following categorical values (TCP, UDP or ICMP). **Scan Type** determines whether the IoT source performs a horizontal scan, a vertical scan or a strobe scan. **Scan Trend** shows how the targets are being probed. There exists several probing strategies such as IP-sequential, reverse IP-sequential or non-sequential (Bou-Harb et al. (2014b)) (typically referred to as permutation probing). To this end, we apply the Man-Kendall statistical trend test (Kendall (1955)) on the sequence of targeted IP addresses for each IoT scanning event. Man-Kendall is a non-parametric hypothesis statistical test which checks for monotonicity in the sequence. By setting the significance level to 0.5%, we can avoid a high false positive rate. This can also distinguish between a positive and a negative slope of the trend which respectively reveals the IP-sequential and reverse IP-sequential strategies. Another behavioral property of an IoT scan event can be a metric to characterize how much the probing traffic is dispersed or focused towards the network telescope. We capture this with entropy and dispersion features. **Entropy** can be calculated by binning the set of darknet IP addresses, counting the number of target IP addresses in each bin in the intended scan event, and calculating the entropy of the resulting distribution. Therefore, an entropy which equals 0 indicates that the scan event is not targeted but in contrast hits all the bins with approximately the same frequency. Based on the minimum number of packets in the scan events, we chose the number of bins to have enough samples (Li et al. (2009)). **Dispersion** measures the level of dispersion of the target IP addresses in a scan. For this purpose, we calculate the number of non-constant least significant bits which are not mutual among all the destination IP addresses of a scan. For instance, in a given scan event, if all the destination IP addresses are of the form `108.32.x.x`, the dispersion would then be equal to 16. Hence, a dispersion which equals to 0 (minimum dispersion) refers to a scanner who is trying to scan just one target host (probably vertically), and a dispersion which equals to 32 (maximum dispersion) refers to a scanner that tries to scan all the IPv4 Internet space. This metric can aid in clustering orchestrated probing campaigns in scenarios when stealthy botnets assign a distinct, small sub-part of the cyber space to each of the bots to scan. It is worthy to note that this metric is very efficient in comparison with other typically employed statistical methods in terms of required memory and computation since it can be applied without the need to store all the packet information for every scan event. **Targeted Port** is the most probed port in the scan event. This value reports the targeted service that the scanner was most interested in.

We should state the feature engineering is highly dependent on the application domain and the expertise of the data analysis and thus we do not claim that the aforementioned set of features are comprehensive. Nevertheless, for the sake of the proposed work which employs darknet data for addressing the problem of inferring IoT probing campaigns, and as validated in Section 3.4, these set of features seem to be valid in practice.

3.3. Dimensionality reduction using L1-PCA

Principal component analysis (PCA) is a powerful dimensionality reduction tool for analyzing datasets which are formulated in the language of linear algebra and has been used for more than a century (Pearson (1901)). Broadly, PCA involves finding the orthonormal basis (the principal axes of the data) over which the variance of the projected data points is maximized. The basis is a low-dimensional subspace by which the original input space data structure can be effectively captured.

Despite the historical success of the traditional L2-norm PCA, it is well known that PCA is sensitive to outlier data values since the effect of outliers is exaggerated due to the square operation of the L2-norm (Ke and Kanade (2003)). On nominal, clean training data, L1-PCA is almost indistinguishable from L2-PCA. However, L1-PCA shows remarkable relative resistance to faulty data contamination in the dataset, due to the linear emphasis placed by the L1-norm optimization metric on each data point (Markopoulos et al. (2014b, 2017)). Hence, it is quite desirable as a prior methodology to data clustering.

In order to maximize the aggregate absolute magnitude of the projected data points, the works in (Markopoulos et al. (2013, 2014b)) proved that L1-PCA is not NP hard for a fixed data dimension D and offered two optimal algorithms for exact computation. The two methods presented in (Markopoulos et al. (2014b)) compute the L1-optimal principal components of a dataset of size N, $\mathbf{X} \in \mathbb{R}_{D\times N}$ with exact complexity of $O(2^{NK})$ or $O(N^{rank(X)K-K+1})$, where $K < rank(\mathbf{X})$ is the desired number of principal components. Nevertheless, such optimal algorithms are still of high complexity, especially given the problem at hand which deals with significant network data. To this end, we propose in this paper a fast greedy approximation algorithm with complexity $O(\min\{ND^2, N^2D\} + N^2(K+1) + ND)$, which is similar to the standard Singular Vector Decomposition (SVD) (Golub and Van Loan (2012)).

3.3.1. Problem Formulation:

Consider our dataset
$\mathbf{X} = [\mathbf{x_1}, \ldots, \mathbf{x_N}] \in \mathbb{R}_{D\times N}$ with rank $d \leq \min\{D, N\}$, where

$$\mathbf{X}_c \triangleq \mathbf{X}_{D\times N}(\mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^T). \tag{1}$$

Here, $\mathbf{1}$ is the all ones' vector of dimension N and $\mathbf{X}_c$ is the mean centered of normalized data matrix $\mathbf{X}$ (features are normalized to the [0 1] range). Now, we are interested in calculating a low rank data subspace (K dimensions) in the form of the orthonormal basis $\mathbf{Q}_{L1} \in \mathbb{R}_{D\times K}$ which solves

$$\mathbf{Q}_{L1} = \underset{\substack{\mathbf{Q}=[\mathbf{q}_1,\ldots,\mathbf{q}_k]\in\mathbb{R}^{D\times K} \\ \mathbf{Q}^T\mathbf{Q}=\mathbf{I}_K}}{\arg\max} \sum_{k=1}^{K} \left\|\mathbf{X_c^T}\mathbf{q}_k\right\|_1 \tag{2}$$

where $\|.\|_1$ represents the L1-norm of the vector/matrix argument and returns the sum of absolute values of the individual entries.

3.3.2. Sub-optimal solution:

Assume $K = 1$; (2) reduces to

$$\mathbf{q}_{L1} = \underset{\substack{\mathbf{q}\in\mathbb{R}^D \\ \mathbf{q}^2=1}}{\arg\max} \left\|\mathbf{X_c^T}\mathbf{q}\right\|_1 \tag{3}$$

which can be rewritten as

$$\begin{aligned}\max_{\substack{\mathbf{q}\in\mathbb{R}^D\\ \mathbf{q}^2=1}} \max_{\mathbf{b}\in\{\pm1\}^N} \mathbf{b^T X_c^T q} &= \\ \max_{\mathbf{b}\in\{\pm1\}^N} \max_{\substack{\mathbf{q}\in\mathbb{R}^D\\ \mathbf{q^2=1}}} \mathbf{q^T X_c b} &= \\ \max_{\mathbf{b}\in\{\pm1\}^N} \|\mathbf{X_c b}\|_2 & \end{aligned} \tag{4}$$

The optimal solution can be achieved by exhaustive search in the space of the binary antipodal vector $\mathbf{b}$[1]. This approach can be generalized for $K > 1$. For a matrix $\mathbf{A} \in \mathbb{R}_{m\times n}$, considering the Singular Vector Decomposition (SVD), $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}_{n\times n}\mathbf{V}^T$, define $U(\mathbf{A}) \triangleq \mathbf{U}\mathbf{V}^T$. In Markopoulos et al. (2013, 2014b), it is shown that if

$$\mathbf{B}_{opt} = \arg\max_{\mathbf{B}\in\{\pm1\}^{N\times K}} \|\mathbf{XB}\|_* \tag{5}$$

then $\mathbf{Q}_{L1} = U(\mathbf{XB_{opt}})$ is a solution to (2). Additionally, $\left\|\mathbf{Q}_{L1}^T\mathbf{X}\right\|_1 = \left\|\mathbf{XB}_{opt}\right\|_*$ where $\|.\|_*$ represents the nuclear norm.

Consider the compact SVD of $\mathbf{X}$, $\mathbf{X} = \mathbf{U}_{D\times d}\mathbf{\Sigma}_{d\times d}\mathbf{V}_{N\times d}^T$ where $d = rank(\mathbf{X}) \leq \min\{D, N\}$. Define $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N] \triangleq \mathbf{\Sigma}\mathbf{V}^T \in \mathbb{R}_{d\times N}$, such that for any $\mathbf{B} \in \{\pm1\}^{N\times K}$, $\|\mathbf{XB}\|_* = \|\mathbf{YB}\|_*$.

The first step from here is to find $\mathbf{B}_{opt}$. To accomplish this, we introduce the L1-PCA Algorithm 1 (Markopoulos et al. (2017)), which begins from an initial matrix $\mathbf{B}$ and employs bit flipping iterations (Johnson and Savakis (2014)) (inspired from the image recognition literature) to reach an approximation to $\mathbf{B}_{opt}$, say $\mathbf{B}_{bf}$. The algorithm returns $\mathbf{Q}_{bf} = U(\mathbf{XB}_{bf})$ to be used as a low complexity, near optimal solution to the L1-PCA problem of Eq. (2). According to our problem dataset, the resulting reduced data matrix is given by

$$\mathbf{D} = \mathbf{Q}_{bf}^T\mathbf{X}_c \tag{6}$$

Algorithm 1: L1-BF for calculating K L1-norm principal components of $\mathbf{X}_c$

```
Input: $\mathbf{X_c} \in \mathbb{R}^{D\times N}$ of rank $d$, $k \leq d$
$(\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}) \leftarrow$ svd$(\mathbf{X}_c)$
$\mathbf{Y} \leftarrow \mathbf{\Sigma}\mathbf{V}^T, \mathbf{v} \leftarrow \mathbf{V}_{:,1}, \mathbf{B} = sgn(\mathbf{v}\mathbf{1_K^T})$
$\mathbf{B}_{bf} \leftarrow bf(\mathbf{Y}, \mathbf{B}, K)$
$(\hat{\mathbf{U}}_{D\times K}, \hat{\mathbf{\Sigma}}_{K\times K}, \hat{\mathbf{V}}_{K\times K}) \leftarrow svd(\mathbf{XB}_{bf})$
$\mathbf{Q}_{bf} \leftarrow \hat{\mathbf{U}}\hat{\mathbf{V}}^T$
Output: $\mathbf{Q}_{bf}$
```

In the sequel, we demonstrate that (1) the selected features (of Section 3.2) are effective for inferring probing campaigns and (2) that the proposed L1-PCA exceeds the typical L2-PCA when it comes to data clustering accuracy using big network telescope data and a known probing campaign (as ground truth). The latter point is quite important to make at this stage of the paper, before we proceed in applying the methodology on large-scale network telescope data for clustering "in the wild" IoT-centric probing campaigns, which are typically hard to validate, given the lack of their corresponding ground truth.

3.4. Accuracy validation using the SIP scan campaign

We aim herein to validate the proposed dimensionality-reduction technique on a probing campaign with a known ground truth in the presence of outliers. To this end, in the sequel, we first describe the creation of the test datasets, followed by demonstrating the results of applying L1-PCA and the typical L2-PCA using the simplistic k-means clustering method to compare and contrast the obtained results.

3.4.1. Creating the test dataset

We created 10 different test datasets based on merging empirical scan events extracted from (1) a known orchestrated probing campaign (i.e., the SIP/VoIP scan campaign (Dainotti et al. (2012))), (2) recent scan events targeting the same destination port as the SIP campaign and (3) other scan events. Note that all the scan events target the same network telescope.

Dainotti et al. (2012) investigated and presented, through the lens of CAIDA's network telescope, a horizontal scan of the entire IPv4 address space conducted by the `Sality` botnet in a heavily coordinated and covert manner to discover and compromise VoIP-related (SIP) infrastructure. CAIDA has published the dataset of this SIP scan campaign. We use such dataset to infer the scanning events of this SIP scan (targeting UDP port 5060) and then extract the features as detailed in Section 3.2. Further, we obtain scan events targeting the same port 5060 by analyzing one day (May 2nd, 2018) of packets arriving at CAIDA's darknet and extracting the same features. These scan events, may or may not be orchestrated, however, they are definitely not part of the SIP scan campaign which we consider as the ground truth campaign. Additionally, we randomly selected a different number of scan events targeting different ports from May 2nd, 2018. We proceeded by mixing 53 flows related to the orchestrated SIP scan events with 67 scan events targeting port 5060 and N random scan events (N ranging from 0 to 4,500 flows) to create 10 test datasets, in order to examine the proposed approach under different scenarios.

3.4.2. Executing the test dataset for validation purposes

For each dataset, we apply L1-PCA and L2-PCA to reduce the dimension of the feature space by projecting them on 3 main principal components. By leveraging the Silhouette Coefficient on the dimensionality-reduced data, we compute the optimal number of clusters to use in k-means clustering. Further, we pinpoint the cluster which contains the orchestrated SIP scan events and label all of its members accordingly. We also label those events that are not related to the SIP scan campaign. This allows us to compare the results with the true labels and calculate the confusion matrix of this binary classification. For each test dataset, we repeat this procedure 100 times and calculate the average confusion matrix to remove the effect of random centroid selection in k-means.

[1] If $\mathbf{b} \in \{\pm1\}^N$ is an optimal solution, $-\mathbf{b}$ is another optimal solution. By leveraging this, the complexity can be reduced.

Consequently, to compare the effect of L1-PCA and L2-PCA on the detection of the SIP scan campaign, we consider 3 typical metrics, namely, precision, recall and F-measure. Precision is the ratio of correctly clustered SIP scan events over all the scan events with true SIP scan label (53 in the analyzed dataset). Recall is the ratio of correctly clustered SIP scan events over all the events predicted as members of the SIP scan campaign. Therefore, values closer to one for precision and recall are more desirable. F-measure combines these two metrics as defined by $F\text{-}measure = 2 \times \frac{precision \times recall}{precision + recall}$ which is the square of the geometric mean divided by the arithmetic mean. These metrics are calculated for all the 10 test datasets and the results are reported in Table 1.

Considering the F-measure value in Table 1 for all the test datasets, we can infer that the introduced L1-PCA with its sub-optimal algorithm significantly outperforms the conventional L2-PCA in terms of clustering the orchestrated SIP scan campaign. This indicates that the proposed L1-PCA is not only significantly less computationally intensive than the typical L2-PCA, but is also more accurate in distinguishing orchestrated events, and quite robust against other non-orchestrated events and outliers; thus, highly applicable to the problem at hand.

Table 1: The results for L1-PCA and L2-PCA to compare their effectiveness on network data dimensionality reduction for the application of probing campaign detection using different datasets

	L1-PCA + k-means			L2-PCA + k-means		
Dataset	Precision	Recall	F-measure	Precision	Recall	F-measure
$N = 0$	0.8649	0.9724	0.9155	0.8649	0.9724	0.9155
$N = 500$	0.9434	0.8758	0.9083	0.9460	0.3206	0.4790
$N = 1000$	0.9811	0.5798	0.7289	0.9811	0.2798	0.4354
$N = 1500$	0.9811	0.8769	0.9261	0.9728	0.3413	0.5053
$N = 2000$	0.9777	0.6055	0.7479	0.9811	0.2870	0.4441
$N = 2500$	0.9811	0.3104	0.4716	0.9811	0.1609	0.2764
$N = 3000$	0.9766	0.7478	0.8470	0.9811	0.6122	0.7540
$N = 3500$	0.9811	0.7715	0.8638	0.9719	0.5844	0.7299
$N = 4000$	0.9811	0.8773	0.9263	0.9794	0.7919	0.8757
$N = 4500$	0.9811	0.8745	0.9248	0.9811	0.4898	0.6534

4. Empirical Evaluation

The aim of this section is to highlight the severity of the insecurity of the IoT paradigm by reporting on the exploitations of Internet-scale IoT devices. Further, we report on the existence of "in the wild" coordinated IoT-specific probing campaigns by applying the proposed L1-PCA technique. To achieve this, we analyze more than 1 TB of network telescope data captured on May 5th, 2018 from CAIDA's network telescope. We also leverage daily, entire IoT databases from the Shodan service. In terms of implementation details, the probing inference component is implemented in C using the libpcap library in a multi-threaded fashion, the correlation between darknet IP header information and Shodan data is invoked using python scripts, while the proposed L1-PCA in conjunction with its sub-optimal algorithm are both implemented in Matlab and are executed on a cluster of 3 nodes consisting of 20 cores each with a total available memory of 128 GB.

4.1. Inferring and characterizing unsolicited IoT devices

By analyzing 10 hours of CAIDA's darknet traffic on May 5th, 2018, we were able to infer unsolicited probing activities from 129,713 unique IoT devices, distributed in 199 countries, hosted by 43 various sectors, and hosted/operated by 8,540 ISPs. The top countries hosting such compromised devices were found to be Mexico (14%), Brazil (12%), China (9%), Indonesia (5%), Russia (4%), United States (4%), and Vietnam (4%). These countries hosted 52% of the inferred devices. Figure 1 illustrates the most affected sectors and their corresponding number of generated IoT-specific probes.

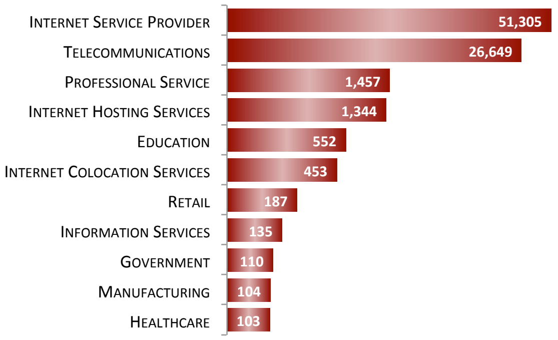

Figure 1: Top sectors generating IoT-specific probes

Note the existence of compromised IoT devices in critical environments such as the medical sector (101 devices, mostly generated by AVTECH sensors that are typically used for monitoring environmental factors such as temperature and humidity, and located in Iran, China and the U.S.), manufacturing and factory automation facilities (104 devices, mostly generated from IP cameras and routers, and located in China and the U.S.) and various governmental entities (110 devices, mostly generated from MikroTik routers and located in China, U.S. and the UK).

For the sake of further characterization, Table 2 summarizes the leading ISPs hosting unsolicited IoT devices which were inferred to be generating IoT-specific probes. We also characterize the entire set of inferred IoT probing events; 95% of them employed TCP, 91.3% adopted permutation probing, and 8.7% are scanning in a sequential manner (6.5% IP-sequential and 2.2% reverse IP-sequential). In addition, only 1.5% of the events are scanning small blocks of IP addresses, while others are not limited to some IP block. We also detected around 500 benign IoT-related scanning events from known entities such as Shadowserver, Team Cymru, Rapid7, and the University of Michigan (mostly generated from edge routers, flagged by Shodan as IoT and D-Link routers).

By investigating additional information returned from Shodan, and by contacting a few IoT operators, we gather some interesting information (depicted in Figure 2) related to well-known IoT manufacturers, in which their devices were deemed to be exploited.

We also made an auxiliary effort to contact some U.S. IoT operators, in which we observed probes from their devices. We generated automated emails to 250 realms by using their listed emails in WHOIS (Documented in RFC 3912). 169 did

Table 2: Top ISPs hosting unsolicited devices and generating IoT-specific probes

ISP	Country	Number of probes
Telmex	Mexico	17,622
Vivo	Brazil	6,130
PT Telkom Indonesia	Indonesia	3,468
VDC	Vietnam	2,453
Telefonica de Argentina	Argentina	2,243
HiNet	Taiwan	2,069
Turk Telekom	Turkey	2,043
Korea Telecom	Korea	1,777
China Unicom Liaoning	China	1,406
Viettel Corporation	Vietnam	1,284

not reply; our emails bounced to 27 recipients; 32 replied using automated emails that the issue will be investigated but we never heard from them again; and 22 acknowledged that their Internet-facing IoT device(s) might be compromised.

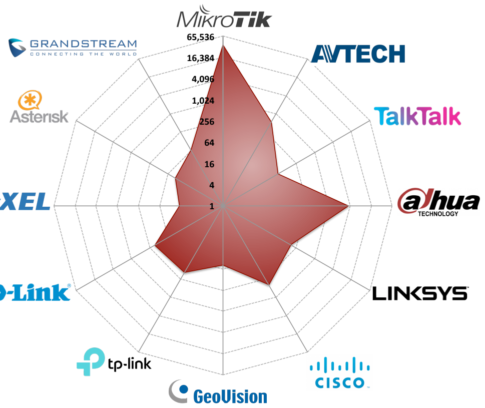

Figure 2: IoT probes generated from well-known manufacturers

4.2. Inferring and reporting on orchestrated IoT Probing Campaigns

We applied the proposed L1-PCA on the detected IoT-centric probing events and leveraged the Silhouette coefficient to estimate the optimal number of clusters to use with k-means. The output revealed 921 orchestrated, IoT-centric campaigns, 142 of which posses more than 50 IoT bots. Table 3 summarizes the top 15 largest campaigns, shows their widespread distribution related to the involved number of ISPs, countries, and sectors, in addition to pinpointing some insights related to their employed protocols and strategies, and most probed target port. We found some interesting characteristics by investigating these campaigns. For instance, campaign 2 seems to be quite distributed worldwide, involving 114 countries and 1,168 ISPs, where further analysis revealed that close to 40% of its IoT bots are related to video surveillance cameras from Dahua. Campaign 8 is also noteworthy, given that it had adopted the rare reverse IP-sequential probing strategy with the lowest probing rate relative to others. Most of the members of this campaign were exclusively operating from manufacturing sectors. Interesting also, we inferred a very large IoT probing campaign (C15 of Table 3) consisting of more than 50,000 IoT bots, distributed over 172 countries and 5,006 ISPs. We summarize some of the insights related to this campaign, which targeted the Telnet port in Figure 3.

IoT probing campaign targeting open resolvers. While initially analyzing the probing events, we noticed scans targeting the network telescope searching for open resolvers that have been specifically generated from IoT devices. Motivated by this phenomena, we applied our proposed clustering methodology on such inferred scans. To this end, we were able to infer 11 IoT coordinated probing campaigns searching for amplifiers as summarized in Table 4. Interestingly, we observe scans for Memcached servers from IoT cameras in campaign 3, high rate probing for DNS resolvers in campaign 8 by MikroTik routers, and co-occurring probes towards Chargen and QoTD from AvTech sensors. Future work will further explore such intriguing events.

5. Discussion

In this section, we discuss the following topics of interest.

Comprehensively inferring Internet-scale unsolicited IoT devices. While this work leveraged the Shodan service to gather a large dataset of IP information related to deployed IoT devices in order to facilitate their correlation with passive measurements, identifying technical information for Internet-wide IoT devices remains challenging. In addition, IoT malware often disable common outward facing services upon infection (Antonakakis et al. (2017)). Consequently, this makes indexing the infected IoT devices even more challenging for Internet scanning services such as Shodan and Censys (Team (2017)). Indeed, without addressing this limitation, approaches similar to the one presented in this paper would remain partially effective (at least operationally). In this context, we posit the following two potential solutions. The first is of a technical nature, rendered by exploring fuzzy matching algorithms, fuzzy hashes/signatures and machine learning techniques to extend the set of IoT devices (previously not indexed by Shodan) as perceived by the network telescope, by leveraging IoT-relevant darknet traffic (from previously inferred IoT devices). The second is a non-technical approach, requiring ISPs, local IoT operators and industry to collaborate to make such IoT information available. The sharing of this information can be performed securely, for example using permissioned blockchains (Banerjee et al. (2018)). We are also currently in touch with Cisco Systems to have access to Jasper, their IoT platform, to obtain access to a larger corpus of IoT device information.

Long-term analysis challenges. For a long-term investigation of this topic, challenges such as the effect of dynamic behaviors of IoT botnets and DHCP IP churn (Vu et al. (2014)) need to be taken into account. This will allow us to have a more sound estimation of compromised IoT devices within each inferred probing campaign.

Malware attribution for tailored remediation. With the continuous rise of new malware variants which specifically target IoT devices in consumer and critical sectors, the objective to

Table 3: Inferred “in the wild” IoT Probing Campaigns

Campaign	# IoT bots	Target Port	Rate (pps)	Scan Type	Protocol	Strategy	# Countries	# ISPs	# Sectors
C1	250	2323	272	Strobe	TCP	Rev. IP Seq.	40	132	9
C2	5,223	2323	591	Horizontal	TCP	Permutation	114	1168	16
C3	646	1433	106	Horizontal	TCP	Permutation	41	169	17
C4	555	-	461	Horizontal	ICMP	Permutation	53	220	11
C5	187	3879	57	Strobe	TCP	Permutation	33	97	12
C6	581	3333	174	Horizontal	TCP	Permutation	47	110	7
C7	2,532	5555	980	Horizontal	TCP	Permutation	69	389	13
C8	197	5555	33	Strobe	TCP	Rev. IP Seq.	30	92	9
C9	241	18183	70	Strobe	UDP	Permutation	8	43	6
C10	459	8080	227	Horizontal	TCP	Permutation	51	140	11
C11	1,493	3389	64	Strobe	TCP	Permutation	87	438	25
C12	1,763	80	62	Horizontal	TCP	Permutation	103	730	14
C13	2886	445	1,684	Horizontal	TCP	Permutation	91	576	22
C14	134	3389	83	Strobe	TCP	Rev. IP Seq.	35	98	10
C15	50,141	23	316	Horizontal	TCP	Permutation	172	5006	38

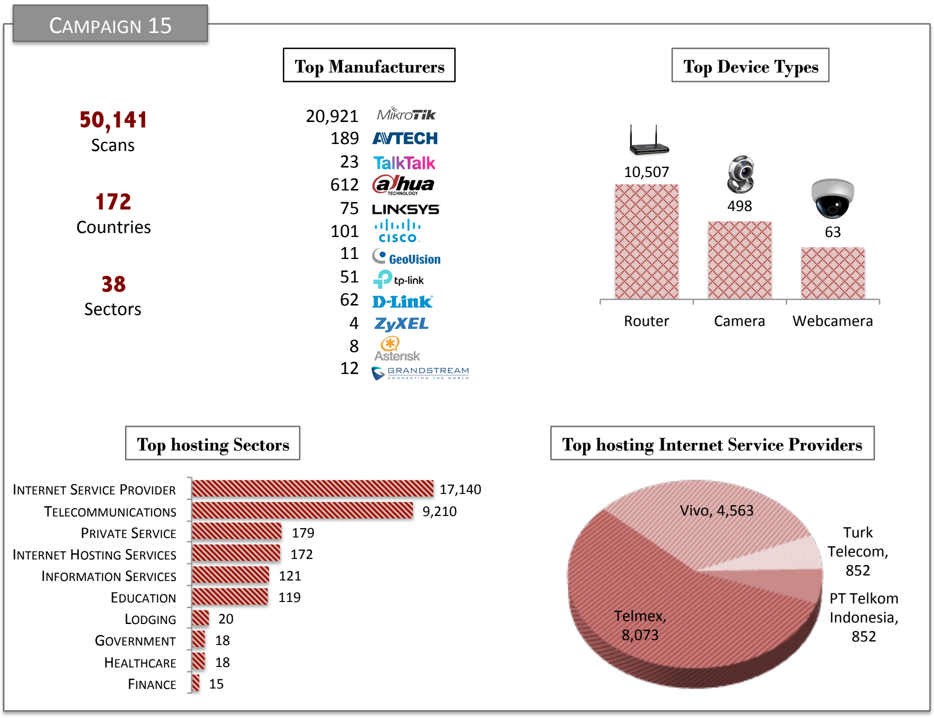

Figure 3: Characterization of a very large IoT probing campaign (**C15** of Table 3)

attribute such exploitations to certain malware variants is crucial. Thus, we are currently exploring formal correlation approaches between passive measurements and malware network traffic samples to provide an attribution evidence.

Feature extraction. Intuitively, as noted earlier in this paper, it is possible to extract other features besides the ones employed in Section 3.2. Nevertheless, given that the validation of the L1-PCA methodology was conducted using a ground truth data set provided by CAIDA, which did not provide complete packet details but rather limited their information to certain data (i.e., timestamp, source port, etc.), the mentioned features were thus solely utilized.

6. Concluding Remarks

In this paper, we contributed towards the IoT security literature by proposing a macroscopic, data-driven methodology to shed light on the large-scale IoT threat landscape. We correlated large volumes of network telescope data with IoT-specific information to infer and characterize Internet-scale IoT exploitations. We attributed such exploitations to their hosting realms, including sectors and manufacturers. Further, motivated by the potential application of big data in network forensics, we proposed the L1-PCA technique in conjunction with a sub-optimal algorithm to significantly reduce its complexity

Table 4: Inferred “in the wild” IoT Probing Campaigns searching for Amplifiers

Campaign	# IoT bots	Target Service	Rate (pps)	Scan Type	Protocol	Strategy	# Countries	# ISP	# Sectors
C1	151	NTP	4147	Horizontal	UDP	Permutation	8	31	5
C2	137	SSDP	5,741	Horizontal	UDP	IP Seq.	15	52	7
C3	75	MEMCACHED	3,188	Horizontal	UDP	Permutation	7	30	7
C4	70	QOTD	2,617	Strobe	UDP	Permutation	10	42	5
C5	58	SSDP	2,713	Horizontal	UDP	Permutation	9	32	4
C6	34	SSDP	2,689	Horizontal	UDP	IP Seq.	6	20	3
C7	31	QOTD	2,852	Strobe	UDP	Rev. IP Seq.	2	19	2
C8	24	DNS	3,320	Strobe	UDP	Permutation	6	20	5
C9	22	CHARGEN, QOTD	260	Strobe	UDP	IP Seq.	4	18	3
C10	11	CHARGEN, QOTD	488	Strobe	UDP	Permutation	4	10	2
C11	11	MEMCACHED	303	Horizontal	UDP	Rev. IP Seq.	4	8	4

while maintaining its superior clustering capabilities. We inferred a large number of exploited IoT devices and more than 140 coordinated IoT probing events, where a large inferred campaign consisted of more than 50,000 IoT devices. Interestingly, we also identified IoT orchestrated campaigns searching for open resolvers that can be abused for performing amplification attacks. Future work will include comprehensively fingerprinting IoT devices to automate the proposed clustering approaches and exploring malware forensics to strengthen the attribution evidence.

Acknowledgments

This work was supported by two grants from the U.S. National Science Foundation (NSF) (Office of Advanced Cyberinfrastructure (OAC) #1755179 and #1829698).

Antonakakis, M., April, T., Bailey, M., Bernhard, M., Bursztein, E., Cochran, J., Durumeric, Z., Halderman, J. A., Invernizzi, L., Kallitsis, M., et al., 2017. Understanding the mirai botnet. In: USENIX Security Symposium.

Azmoodeh, A., Dehghantanha, A., Choo, K.-K. R., 2018a. Robust malware detection for internet of (battlefield) things devices using deep eigenspace learning. IEEE Transactions on Sustainable Computing.

Azmoodeh, A., Dehghantanha, A., Conti, M., Choo, K.-K. R., 2018b. Detecting crypto-ransomware in iot networks based on energy consumption footprint. Journal of Ambient Intelligence and Humanized Computing 9 (4), 1141–1152.

Bailey, M., Cooke, E., Jahanian, F., Watson, D., Nazario, J., Jul. 2005. The blaster worm: Then and now. IEEE Security and Privacy 3 (4), 26–31.

Banerjee, M., Lee, J., Choo, K.-K. R., 2018. A blockchain future for internet of things security: a position paper. Digital Communications and Networks 4 (3), 149–160.

Bellovin, S. M., 1993. Packets found on an internet. ACM SIGCOMM Computer Communication Review 23 (3), 26–31.

Bertino, E., Islam, N., 2017. Botnets and internet of things security. Computer 50 (2), 76–79.

Bodenheim, R., Butts, J., Dunlap, S., Mullins, B., 2014. Evaluation of the ability of the shodan search engine to identify internet-facing industrial control devices. International Journal of Critical Infrastructure Protection 7 (2), 114–123.

Bou-Harb, E., Debbabi, M., Assi, C., 2013. A systematic approach for detecting and clustering distributed cyber scanning. Computer Networks 57 (18), 3826–3839.

Bou-Harb, E., Debbabi, M., Assi, C., 2014a. Behavioral analytics for inferring large-scale orchestrated probing events. In: Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on. IEEE, pp. 506–511.

Bou-Harb, E., Debbabi, M., Assi, C., Third 2014b. Cyber Scanning: A Comprehensive Survey. IEEE Communications Surveys Tutorials 16 (3), 1496–1519.

Bou-Harb, E., Debbabi, M., Assi, C., 2016. A novel cyber security capability: Inferring internet-scale infections by correlating malware and probing activities. Computer Networks 94, 327–343.

Bou-Harb, E., Lakhdari, N.-E., Binsalleeh, H., Debbabi, M., 2014c. Multidimensional investigation of source port 0 probing. Digital Investigation 11, S114–S123.

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 19-July-2008].

Cimpanu, C., 2017. Brickerbot author claims he bricked two million devices. Bleeping Computer, April.

Costin, A., Zaddach, J., Francillon, A., Balzarotti, D., Antipolis, S., 2014. A large-scale analysis of the security of embedded firmwares. In: USENIX Security. pp. 95–110.

Cui, A., Stolfo, S. J., 2010. A quantitative analysis of the insecurity of embedded network devices: results of a wide-area scan. In: Proceedings of the 26th Annual Computer Security Applications Conference. ACM, pp. 97–106.

Dainotti, A., King, A., Claffy, K., Papale, F., Pescapè, A., Nov 2012. Analysis of a "/0" Stealth Scan from a Botnet. In: Internet Measurement Conference (IMC). pp. 1–14.

Dainotti, A., King, A., Claffy, K., Papale, F., PescapÃĺ, A., Apr 2015. Analysis of a "/0" Stealth Scan from a Botnet. IEEE/ACM Transactions on Networking 23 (2), 341–354.

Do, Q., Martini, B., Choo, K.-K. R., 2018. Cyber-physical systems information gathering: A smart home case study. Computer Networks 138, 1–12.

D’Orazio, C., Choo, K.-K. R., Yang, L. T., 2017. Data exfiltration from internet of things devices: ios devices as case studies. IEEE Internet of Things Journal 4 (2), 524–535.

Edwards, S., Profetis, I., 2016. Hajime: Analysis of a decentralized internet worm for IoT devices. Rapidity Networks 16.

Elastic, 2018. What is the ELK Stack? `https://www.elastic.co/elk-stack`, accessed 2018-10-02.

Fachkha, C., Bou-Harb, E., Debbabi, M., 2014. Fingerprinting internet dns amplification ddos activities. In: New Technologies, Mobility and Security (NTMS), 2014 6th International Conference on. IEEE, pp. 1–5.

Fachkha, C., Bou-Harb, E., Keliris, A., Memon, N., Ahamad, M., 2017. Internet-scale probing of cps: Inference, characterization and orchestration analysis. In: Proceedings of the Network and Distributed System Security Symposium. Vol. 17.

Fachkha, C., Debbabi, M., 2016. Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization. IEEE Communications Surveys and Tutorials 18 (2), 1197–1227.

Fernandes, E., Paupore, J., Rahmati, A., Simionato, D., Conti, M., Prakash, A., 2016. Flowfence: Practical data protection for emerging iot application frameworks. In: USENIX Security Symposium.

Golub, G. H., Van Loan, C. F., 2012. Matrix computations. Vol. 3. JHU Press.

Gu, G., Perdisci, R., Zhang, J., Lee, W., et al., 2008. BotMiner: Clustering

Analysis of Network Traffic for Protocol-and Structure-Independent Botnet Detection. In: USENIX security symposium. Vol. 5. pp. 139–154.

Gu, G., Porras, P. A., Yegneswaran, V., Fong, M. W., Lee, W., 2007. BotHunter: Detecting Malware Infection Through IDS-Driven Dialog Correlation. In: USENIX Security Symposium. Vol. 7. pp. 1–16.

Guarnizo, J., Tambe, A., Bunia, S. S., Ochoa, M., Tippenhauer, N., Shabtai, A., Elovici, Y., 2017. Siphon: Towards scalable high-interaction physical honeypots. arXiv preprint arXiv:1701.02446.

Gupta, B., Tewari, A., Jain, A. K., Agrawal, D. P., 2017. Fighting against phishing attacks: state of the art and future challenges. Neural Computing and Applications 28 (12), 3629–3654.

He, W., Golla, M., Padhi, R., Ofek, J., Dürmuth, M., Fernandes, E., Ur, B., 2018. Rethinking access control and authentication for the home internet of things (iot). In: Proceedings of the 27th USENIX Conference on Security Symposium. USENIX Association, pp. 255–272.

Heidemann et al., J., 2008. Census and survey of the visible internet. In: Proceedings of the 8th ACM SIGCOMM conference on Internet measurement. ACM, pp. 169–182.

Heo, H., Shin, S., 2018. Who is knocking on the telnet port: A large-scale empirical study of network scanning. In: Proceedings of the 2018 on Asia Conference on Computer and Communications Security. ACM, pp. 625–636.

Ho, G., Leung, D., Mishra, P., Hosseini, A., Song, D., Wagner, D., 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, pp. 461–472.

Husák, M., Neshenko, N., Pour, M. S., Bou-Harb, E., Čeleda, P., 2018. Assessing internet-wide cyber situational awareness of critical sectors. In: Proceedings of the 13th International Conference on Availability, Reliability and Security. ARES 2018. ACM, New York, NY, USA, pp. 29:1–29:6.

Jia, Y. J., Chen, Q. A., Wang, S., Rahmati, A., Fernandes, E., Mao, Z. M., Prakash, A., Unviersity, S. J., 2017. Contexiot: Towards providing contextual integrity to appified iot platforms. In: Proceedings of the 21st Network and Distributed System Security Symposium (NDSS'17).

Johnson, M., Savakis, A., 2014. Fast l 1-eigenfaces for robust face recogntion. In: Image and Signal Processing Workshop (WNYISPW), 2014 IEEE Western New York. IEEE, pp. 1–5.

Karasaridis, A., Rexroad, B., Hoeflin, D. A., et al., 2007. Wide-scale botnet detection and characterization. HotBots 7, 7–7.

Ke, Q., Kanade, T., 2003. Robust subspace computation using l1 norm. Tech. rep., Carnegie Mellon University.

Kendall, M. G., 1955. Rank correlation methods. Hafner Publishing Co.

Khalid, C., Zyad, E., Mohammed, B., 2015. Network intrusion detection system using l1-norm pca. In: Information Assurance and Security (IAS), 2015 11th International Conference on. IEEE, pp. 118–122.

Li, Z., Goyal, A., Chen, Y., Paxson, V., 2009. Automating analysis of large-scale botnet probing events. In: Proceedings of the 4th International Symposium on Information, Computer, and Communications Security. ACM, pp. 11–22.

Limthong, K., Kensuke, F., Watanapongse, P., 2008. Wavelet-based unwanted traffic time series analysis. In: International Conference on Computer and Electrical Engineering (ICCEE). IEEE, pp. 445–449.

Liu, Y., Pados, D. A., 2015. Compressed-sensed-domain l1-pca video surveillance. In: Compressive Sensing IV. Vol. 9484. International Society for Optics and Photonics, p. 94840B.

Maritato, F., Liu, Y., Colonnese, S., Pados, D. A., 2016. Face recognition with l1-norm subspaces. In: Compressive Sensing V: From Diverse Modalities to Big Data Analytics. Vol. 9857. International Society for Optics and Photonics, p. 98570L.

Markopoulos, P., Tsagkarakis, N., Pados, D., Karystinos, G., 2014a. Direction finding with l1-norm subspaces. In: Compressive Sensing III. Vol. 9109. International Society for Optics and Photonics, p. 91090J.

Markopoulos, P. P., Ahmad, F., 2017. Indoor human motion classification by l1-norm subspaces of micro-doppler signatures. In: Radar Conference (RadarConf), 2017 IEEE. IEEE, pp. 1807–1810.

Markopoulos, P. P., Ahmad, F., 2018. Robust radar-based human motion recognition with l1-norm linear discriminant analysis. In: 2018 IEEE International Microwave Biomedical Conference (IMBioC). IEEE, pp. 145–147.

Markopoulos, P. P., Karystinos, G. N., Pados, D. A., 2013. Some options for l1-subspace signal processing. In: Wireless Communication Systems (ISWCS 2013), Proceedings of the Tenth International Symposium on. VDE, pp. 1–5.

Markopoulos, P. P., Karystinos, G. N., Pados, D. A., 2014b. Optimal algorithms for *l*_{1}-subspace signal processing. IEEE Transactions on Signal Processing 62 (19), 5046–5058.

Markopoulos, P. P., Kundu, S., Chamadia, S., Pados, D. A., 2017. Efficient l1-norm principal-component analysis via bit flipping. IEEE Transactions on Signal Processing 65 (16), 4252–4264.

Materly, J., 2009. Shodan. https://shodan.io, accessed 2018-06-12.

Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Breitenbacher, D., Shabtai, A., Elovici, Y., 2018. N-baiot: Network-based detection of iot botnet attacks using deep autoencoders. arXiv preprint arXiv:1805.03409.

Meidan, Y., Bohadana, M., Shabtai, A., Guarnizo, J. D., Ochoa, M., Tippenhauer, N. O., Elovici, Y., 2017. Profiliot: a machine learning approach for iot device identification based on network traffic analysis. In: Proceedings of the Symposium on Applied Computing. ACM, pp. 506–509.

Moore, D., Shannon, C., Brown, D. J., Voelker, G. M., Savage, S., 2006. Inferring Internet denial-of-service activity. ACM Transactions on Computer Systems (TOCS) 24 (2), 115–139.

Pa, Y. M. P., Suzuki, S., Yoshioka, K., Matsumoto, T., Kasama, T., Rossow, C., 2016. Iotpot: A novel honeypot for revealing current iot threats. Journal of Information Processing 24 (3), 522–533.

Paxson, V., 1999. Bro: a system for detecting network intruders in real-time. Computer networks 31 (23-24), 2435–2463.

Pearson, K., 1901. Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2 (11), 559–572.

Pierantozzi, M., Liu, Y., Pados, D. A., Colonnese, S., 2016. Video background tracking and foreground extraction via l1-subspace updates. In: Compressive Sensing V: From Diverse Modalities to Big Data Analytics. Vol. 9857. International Society for Optics and Photonics, p. 985708.

Pour, M. S., Bou-Harb, E., 2018. Implications of theoretic derivations on empirical passive measurements for effective cyber threat intelligence generation. In: 2018 IEEE International Conference on Communications (ICC). IEEE, pp. 1–7.

Register, T., 2018. VPNFilter router malware is a lot worse than everyone thought. https://www.theregister.co.uk/2018/06/07/vpnfilter_is_much_worse_than_everyone_thought/, accessed 2018-06-14.

Roesch, M., et al., 1999. Snort: Lightweight intrusion detection for networks. In: Lisa. Vol. 99. pp. 229–238.

Ronen, E., Shamir, A., 2016. Extended functionality attacks on iot devices: The case of smart lights. In: Security and Privacy (EuroS&P), 2016 IEEE European Symposium on. IEEE, pp. 3–12.

Rose, K., Eldridge, S., Chapin, L., 2015. The internet of things: An overview. The Internet Society (ISOC), 1–50.

Rossow, C., 2014. Amplification Hell: Revisiting Network Protocols for DDoS Abuse. In: NDSS.

Silva, S. S., Silva, R. M., Pinto, R. C., Salles, R. M., 2013. Botnets: A survey. Computer Networks 57 (2), 378–403.

Soltan, S., Mittal, P., Poor, H. V., 2018. Blackiot: Iot botnet of high wattage devices can disrupt the power grid. In: Proc. USENIX Security. Vol. 18.

Team, C., 2017. Internet-wide scan data repository. Retrieved 22, 2017.

Ur, B., Jung, J., Schechter, S., 2013. The current state of access control for smart devices in homes. In: Workshop on Home Usable Privacy and Security (HUPS). HUPS 2014.

Vervier, P.-A., Shen, Y., 2018. Before toasters rise up: A view into the emerging iot threat landscape. In: International Symposium on Research in Attacks, Intrusions, and Defenses. Springer, pp. 556–576.

Vu, L., Turaga, D., Parthasarathy, S., 2014. Impact of dhcp churn on network characterization. In: ACM SIGMETRICS Performance Evaluation Review. Vol. 42. ACM, pp. 587–588.

Wired.com, 2017. The Reaper IoT botnet has already infected a million networks. https://www.wired.com/story/reaper-iot-botnet-infected-million-networks/, accessed 2018-06-14.

Yu, T., Sekar, V., Seshan, S., Agarwal, Y., Xu, C., 2015. Handling a trillion (unfixable) flaws on a billion devices: Rethinking network security for the internet-of-things. In: Proceedings of the 14th ACM Workshop on Hot Topics in Networks. ACM, p. 5.

Zhao, D., Traore, I., Sayed, B., Lu, W., Saad, S., Ghorbani, A., Garant, D., 2013. Botnet detection based on traffic behavior analysis and flow intervals. Computers & Security 39, 2–16.