Multidimensional Behavioral Profiling of Internet-of-Things in Edge Networks

Kuai Xu, Yinxin Wan, Guoliang Xue, and Feng Wang
Arizona State University
{kuai.xu,ywan28,xue,fwang25}@asu.edu

ABSTRACT

The last decade has witnessed research advances and wide deployment of Internet-of-things (IoT) in smart homes and connected industry. However, the recent spate of cyber attacks exploiting the vulnerabilities and insufficient security management of IoT devices have created serious challenges for securing IoT devices and applications. As a first step towards understanding and mitigating diverse security threats of IoT devices, this paper develops a measurement framework to automatically collect network traffic of IoT devices in edge networks, and build multidimensional behavioral profiles of these devices which characterize who, when, what, and why on the behavioral patterns of IoT devices based on continuously collected traffic data. To the best of our knowledge, this paper is the first effort to shed light on the IP-spatial, temporal, and cloud service patterns of IoT devices in edge networks, and to explore these multidimensional behavioral fingerprints for IoT device classification, anomaly traffic detection, and network security monitoring for millions of vulnerable and resource-constrained IoT devices on the Internet.

ACM Reference Format:
Kuai Xu, Yinxin Wan, Guoliang Xue, and Feng Wang. 2019. Multidimensional Behavioral Profiling of Internet-of-Things in Edge Networks. In *IEEE/ACM International Symposium on Quality of Service (IWQoS '19), June 24–25, 2019, Phoenix, AZ, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3326285.3329072

1 INTRODUCTION

The last decade has witnessed the research advances and explosive deployment of Internet-of-things (IoT) and cyber-physical systems in smart homes, smart cities, and industry 4.0 for a wide spectrum of critical applications and services [23, 25, 29]. However, the recent spate of cyber attacks towards IoT devices in smart homes or small offices have created substantial challenges for Internet users without network and security expertises to manage and secure heterogeneous and poorly protected IoT devices [5, 6, 22] .

The burgeoning and insecure IoT devices in millions of edge networks call for effective techniques to detect, recognize, characterize, and address security threats towards these devices and applications. As a first step of securing IoT devices in edge networks, this paper develops a measurement framework to automatically collect, process, characterize, and *profile* communication patterns of IoT devices with a variety of traffic features from IP-spatial, temporal, and service dimensions. Specifically, we leverage intelligent and programmable edge routers with commodity hardware to continuously collect incoming and outgoing network flow traffic in real-time for connected IoT devices in distributed edge networks.

The availability of network traffic data makes it possible to develop multidimensional traffic profiles[1] of IoT devices for gaining an in-depth understanding of communication patterns and traffic behaviors of IoT devices, and more importantly, detecting and mitigating suspicious activities and cyber attacks towards vulnerable IoT devices. The additional benefit of measuring and monitoring network traffic of IoT devices is to have the full visibility of data communications and network configurations of IoT devices, e.g., Chromecast, a streaming media player developed by Google, configuring Google DNS servers as default rather than using the local ISP's DNS servers [3]. Such bogus behaviors are very hard to discovery if the measurement functions are not available on home routers for capturing and profiling traffic activities of IoT devices in edge networks.

In this study we build the behavioral profile of IoT devices from a wide spectrum of their traffic features based on three dimensions: *IP-spatial*, *temporal*, and *cloud*. The IP-spatial dimension is centered on the analysis of remote IP addresses of Internet end hosts such as domain name system (DNS) servers or network time protocol (NTP) servers which IoT devices have communicated with. In addition, aggregating these remote IP addresses into Border Gateway Protocol (BGP) network prefixes [32] and ASNs allows us to analyze IP-spatial correlations of Internet end hosts communicating with IoT devices. Our experimental results on IP-spatial behaviors of deployed IoT devices in the wild have discovered that most IoT devices engage with cloud servers from a small set of network prefixes and ASNs due to their single-purpose applications and specific functions. For example, our experiment study discovers Philips Hue smart light bulbs mostly communicate with cloud servers, which are owned by Philips and deployed on Google cloud platforms, via Philips Hue smart hub for sending *on* or *off* commands.

Our proposed measurement framework characterizes behavioral profiles of IoT devices from the temporal dimension through identifying three distinct temporal traffic patterns from connected objects in edge networks, and classify IoT devices into always-on and on-demand devices. For the analysis on the cloud dimension, our study shows that IoT devices typically only engage with a small and fixed

IWQoS '19, June 24–25, 2019, Phoenix, AZ, USA

ACM ISBN 978-1-4503-6778-3/19/06…$15.00
https://doi.org/10.1145/3326285.3329072

[1] Throughout the paper we will use the terms profile and fingerprint as well as profiling and fingerprinting interchangeably.

set of common applications such as Hyper Text Transfer Protocol (HTTP), DNS, and NTP due to their specific functionalities.

In light of the prevalent cybersecurity threats against IoT devices in edge networks, we explore the benefits of multidimensional behavioral profiles for a wide spectrum of applications including anomaly traffic detection, IoT device detection and classification, and network security monitoring. Specifically, we introduce a simple yet effective pattern-based anomaly detection approach for encoding common network traffic patterns with short encoded length, and encoding infrequent and unusual patterns with longer encoded length. The experimental evaluation shows that the approach is able to uncover suspicious traffic activities with high precisions. Moreover, we leverage multidimensional profiles of IoT devices for recognizing and detecting new and unknown IoT devices based on the profiles of existing and known IoT devices. Finally we outline how the behavioral profiles could facilitate network security monitoring via effectively capturing behavioral dynamics or deviations caused by cyber attacks such as port scanning activities and repeated failed login attempts.

The contributions of this paper are summarized as follows:

- This paper presents a measurement framework for collecting network traffic of IoT devices to characterize and model behavioral fingerprints of IoT devices in edge networks.
- This paper introduces a multi-dimensional approach to model the IP-spatial, temporal, and cloud behaviors of heterogeneous IoT devices, and presents the experiments results based on real world IoT devices.
- This paper explores multidimensional behavioral profiles of IoT devices for a spectrum of applications including IoT device classification, anomaly traffic detection, and network security monitoring.

The remainder of this paper is organized as follows. Section 2 introduces the measurement framework we have developed for multidimensional behavioral profiling of IoT devices and briefly describes the data-sets used in this study. Section 3 introduces how we profile behavioral patterns of IoT devices in edge networks with traffic features discovered from IP-spatial, temporal, and cloud dimensions. In Section 4, we explore behavioral profiles of IoT devices for a variety of critical applications such as IoT device classification, anomaly traffic detection, and network security monitoring. Section 5 discusses related work in this research area, while Section 6 concludes this paper and outlines our future work.

2 AN IOT TRAFFIC MEASUREMENT FRAMEWORK VIA PROGRAMMABLE EDGE ROUTERS

Recent advances on embedded systems, sensors, robotics, and machine learning have enabled the wide deployment of IoT devices in edge networks. The first step of protecting and securing millions of IoT devices is to measure, monitor, and understand their normal communication patterns and behavioral profiles. For example, what remote hosts on the Internet are talking with the smart speakers or thermostats at home networks, at what time, for what reasons? A recent security evaluation study [20] on IoT deployment has also pointed out that measurement is a crucial step for protecting the security of IoT devices and the privacy of end users.

Answering these questions is very critical to understand if and when connected IoT devices in edge networks are compromised by cyber attacks such as Mirai botnet [16]. The Mirai botnet has successfully infected over 60,000 IoT devices including IP cameras and consumer-grade routers in the first 20 hours after being released to the Internet, and launched more than 15,000 cyber attacks towards game servers, telecoms, anti-DDoS providers, and other high-profile Web sites.

Towards profiling communication patterns of IoT devices, we leverage the computational resources on intelligent and programmable edge routers to develop a prototype measurement framework, which is able to capture network traffic flows of IoT devices for real-time traffic monitoring and behavioral profiling. As shown in Figure 1, the programmable edge router continuously captures, stores, and analyzes the incoming, outgoing, and internal network traffic flow records of all IoT devices in the edge network. For each flow record, our measurement framework collects the well-known 5-tuples of a network conversation or session, i.e., source IP address (`srcIP`), source port number (`srcPort`), destination IP address (`dstIP`), destination port number (`dstPort`), and protocol, as well as the start and end timestamps, byte count, and packet count.

Our measurement framework does not collect raw IP packets from IoT devices since most data packets originating from or destined to IoT devices are encrypted, and the storage of raw data packets of IoT devices such as smart TVs or IP cameras could bring undesired system challenges for resource-constrained edge routers. On the other hand, network flow records are widely used for Internet traffic classification, network measurement and analysis [10, 19] thanks to their diverse and informative traffic features and marginal computational and storage resource overheads.

In this study, we have collected network flow records of IoT devices from 22 home networks and small offices in the United States, Hong Kong, and China. The number of end systems including IoT devices and non-IoT devices connecting to each edge network ranges from 1 to 25. In total, these edge networks collectively connect over 50 IoT devices including Amazon Echo, Google Home, Philips Hue smart light bulbs, Samsung smart plug and motion sensor, YI home camera, August smart lock, LG smart TV, and a number of other IoT devices. To demonstrate the practical feasibility of the IoT traffic measurement framework, we deploy and evaluate the system with different brands of programmable routers including Linksys, Netgear, Buffalo, and CanaKit Raspberry Pi.

3 MULTIDIMENSIONAL BEHAVIORAL PROFILING OF IOT DEVICES

In this section we present a multidimensional behavioral profiling approach for fingerprinting the behaviors of IoT devices from a wide spectrum of traffic features based on network flow records collected from edge networks. First, we study the *IP-spatial* behavior of IoT devices via characterizing remote IP addresses engaging with IoT devices and aggregating these IP addresses into BGP networks prefixes and ASNs for correlation analysis. Subsequently, we study the *temporal* traffic patterns of IoT devices over our longitudinal measurement study, and profile the cloud behaviors of IoT devices via analyzing how they interact with cloud servers.

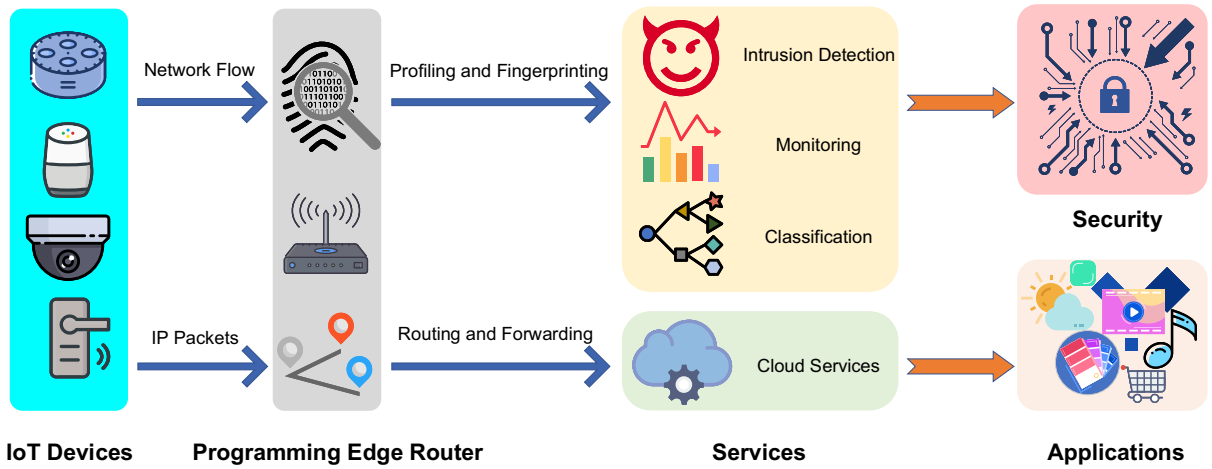


Figure 1: An IoT traffic measurement framework via programmable routers at edge networks.

3.1 IP-Spatial Behavior of IoT Devices

We characterize the IP-spatial behaviors of IoT devices by analyzing the remote IP addresses which communicate with these devices. More importantly, we aggregate and correlate these remote addresses into BGP network prefixes and ASNs for gaining an in-depth understanding of "clustered" IP-spatial behaviors for IoT devices. For example, the IP address of the DNS server for Google home smart voice assistant, `8.8.8.8`, is from the BGP prefix `8.0.0.0/9` and ASN 15169 owned by Google based on the latest snapshot of the BGP routing table [28] and the official registry records from Internet assigned numbers authority (IANA).

Aggregating and correlating remote individual IP addresses to network prefixes and ASNs reveal an interesting observation. IoT devices typically engage with a very small subset of BGP network prefixes and ASNs, even though they communicate with a large number of remote severs, which are likely from the same server pool by the same service providers for efficient load balancing and content distributions. Table 1 summarizes the clustered patterns of IP-spatial behavior of 6 IoT devices and 2 non-IoT devices in one edge network during a 5-minute time window. As shown in Table 1, each IoT device only engages with servers from one or two *unique* ASNs during the observation period, while the smartphone and laptop communicate with remote end hosts from 13 and 39 *unique* ASNs, respectively.

Figure 2 shows the convergence of unique remote IP addresses, their network prefixes, and ASNs for a variety of IoT and non-IoT devices in the same edge network over a 4-month time span. As shown in the longitudinal measurement study for the IP-spatial behavior, it is very interesting to observe that all IoT devices have engaged with a much smaller set of destination IP addresses, prefixes, and ASNs than smartphones and laptops.

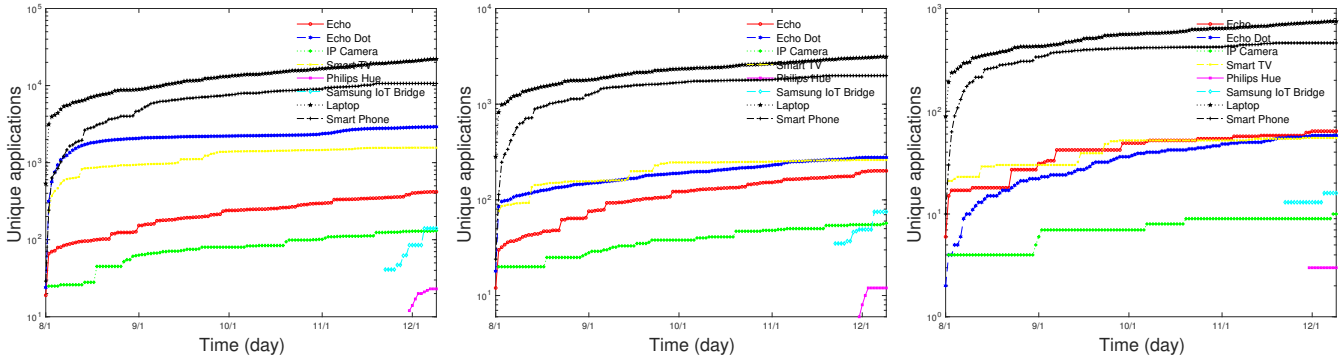
3.2 Temporal Behavior of IoT Devices

For the temporal behavior of IoT devices, we first measure the number of distinct time slots in which IoT devices exhibit traffic activities during the longitudinal measurement study. In this study, we select 5 minutes as the time unit for analysis to balance the computation overhead and monitoring real-time traffic activities, thus the maximum of time slots an IoT device is observed is 288 in one day. Figure 3 shows the flow, packet, and byte counts of three different connected devices in edge networks over one-week time span. As shown in Figure 3, the smart voice assistant, smart TV, and smartphone exhibit distinct traffic characteristics over time, and have very unique and diverse temporal patterns on flow, packet and byte counts over time, which leads us to measure and quantify the *variability* on the number of time windows for IoT devices over the entire data collection period.

For each IoT device d in the edge network, let $t_{d,i}$ represent the number of time windows the device d is observed with network traffic on the i-th day. Considering connected devices are randomly added into the edge network, we use the average time window for each device μ_d rather than the total number of time windows during the entire measurement period. The average of time windows μ_d is derived as $\mu_d = \frac{\sum_{i=1}^{N} t_{d,i}}{N}$, where N is the number of the days since the device d is observed in the edge network and $1 \leq i \leq N$.

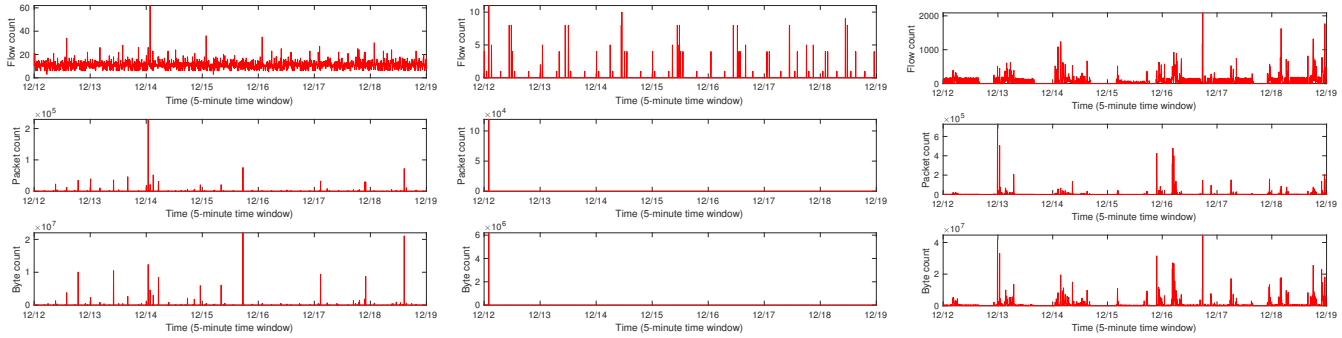
Table 1: The clustered patterns of IP-spatial behavior of IoT devices in the same edge network during a 5-minute time window.

Device	IoT	`dstIPs`	prefixes	ASNs
Amazon Echo	Yes	3	3	1
Echo Dot	Yes	5	4	1
IP Camera	Yes	2	2	1
Philips Hue	Yes	1	1	1
Samsung smart plug	Yes	3	2	1
Smart TV	Yes	4	3	2
Smart Phone	No	37	24	13
Laptop	No	172	102	39



(a) IP addresses (b) BGP prefixes (c) ASNs

Figure 2: The convergence of IP addresses, prefixes, and ASNs for IoT and non-IoT devices over the longitudinal measurement period.



(a) A smart voice assistant (b) A smart TV (c) A laptop

Figure 3: Traffic characteristics of IoT devices and non-IoT devices over 1-week time-span.

Finally, the actual temporal variability on time windows, measured by coefficient of variance, is calculated as $CoV_d = \frac{\mu_d}{\sigma_d}$, where σ_d, the standard deviation, is calculated as $\sigma_d = \sqrt{\frac{1}{N}\sum_{i=1}^{N} t_{d,i} - \mu_d}$.

Figure 4 illustrates a scatter graph on the mean μ and coefficient of variance CoV of time slots observed with traffic activities for different IoT and non-IoT devices deployed in the same edge network. As shown in Figure 4, four out of the six IoT devices exhibit traffic activities during the majority of time windows in each and every day, and their *variability* on the number of time windows is much smaller than that of non-IoT devices. One IoT device, i.e., an IP camera, is only active for a small number of time slots per day, but exhibits lows variability on the time window as well. The only IoT device showing a high variability on the number of time windows across different days is a smart TV, which is turned on and off in an unpredictable fashion. Based on these observations on the temporal patterns of IoT and non-IoT devices, we can classify connected devices in edge networks in three categories: always-on IoT devices, on-demand devices, and non-IoT devices.

The self-similarity traffic patterns of IoT devices visualized on Figure 3 also inspire us to analyze the autocorrelation on network

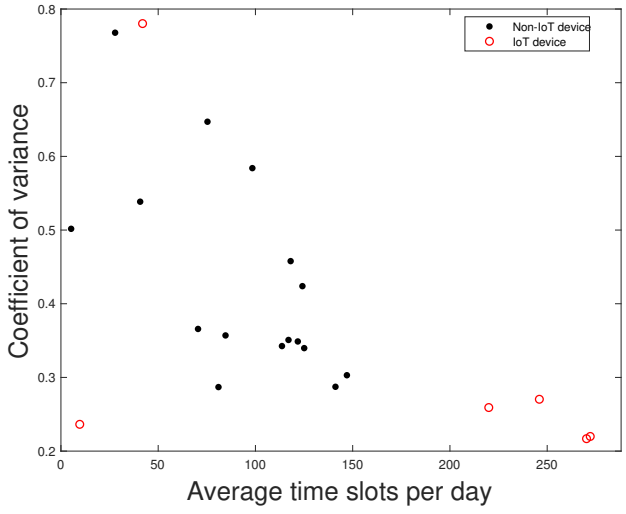


Figure 4: The mean and coefficient of variance of time slots observed with traffic activities for IoT and non-IoT devices.

traffic for all connected devices in edge networks. The autocorrelation metric quantifies the correlation of the same variable across different and lagged periods of times, thus the metric is also referred as to serial correlation and lagged correlation. The autocorrelation metric, $\rho_{d,k}$, for the IoT device d, between network traffic activity time series $X_{d,t}$ and a k-lagged copy of itself $X_{d,t+k}$ is captured by the autocorrelation function (ACF) as follows:

$$\rho_{d,k} = \frac{\sum_{t=k+1}^{n-k}(X_t - \mu)(X_{t+k} - \mu)}{\sigma^2}, \quad (1)$$

where μ and σ are the mean and standard deviation of network traffic activity time series X_d, respectively. An autocorrelation value of 0 suggests independent and random observations on the traffic time series of connected devices in edge networks, while a significant autocorrelation reveals substantial correlations among adjacent observations or determines predictable seasonality in the time series [12, 31].

Figure 5 illustrates the autocorrelation plots, also referred to as correlograms, of network traffic time series for three selected IoT and non-IoT device. As shown in Figure 5, the network traffic time series of IoT devices in edge networks indeed exhibit various extents of self similarity patterns.

3.3 Cloud Behavior of IoT Devices

The objective of characterizing cloud behavior of IoT devices is to understand why IoT devices communicate with remote servers in the cloud. In particular, we profile cloud behaviors of IoT devices based on the *dominant* applications or services observed from `dstPort` and protocol of their outgoing network traffic flows. Table 2 illustrates all the observed 5 applications for the 6 IoT devices deployed in one edge network during a 24-hour time window. These 5 applications are HTTP, Hyper Text Transfer Protocol Secure (HTTPS), DNS, NTP, and Spotify music streaming. As a comparison, one smartphone and one laptop in the same edge network engage with 11 and 15 distinct applications, respectively during the same time period.

Application	Service	Echo	Camera	Echo Dot	Philips Hue	Smart TV	IoT Hub
443/TCP	HTTPS	Y	Y	Y	Y	Y	Y
80/TCP	HTTP	Y		Y	Y	Y	Y
53/UDP	DNS	Y	Y	Y		Y	
123/UDP	NTP	Y	Y	Y	Y		
4070/TCP	Spotify	Y					

Table 2: The dominant applications used by IoT devices in edge networks.

The limited and consistent set of common applications used by IoT devices confirms that IoT devices are typically designed for very specific functions and dedicated utilities. Figure 6 illustrates the convergence of cloud applications for IoT and non-IoT devices. As shown in Figure 6, the number of applications for IoT devices converges in a very rapid fashion. It is very interesting to note that all IoT devices use HTTPS for secure and encrypted Web services, which shows the security awareness and investment of IoT manufactories and application developers. On the other hand, the non-encrypted HTTP service is still observed for five IoT devices.

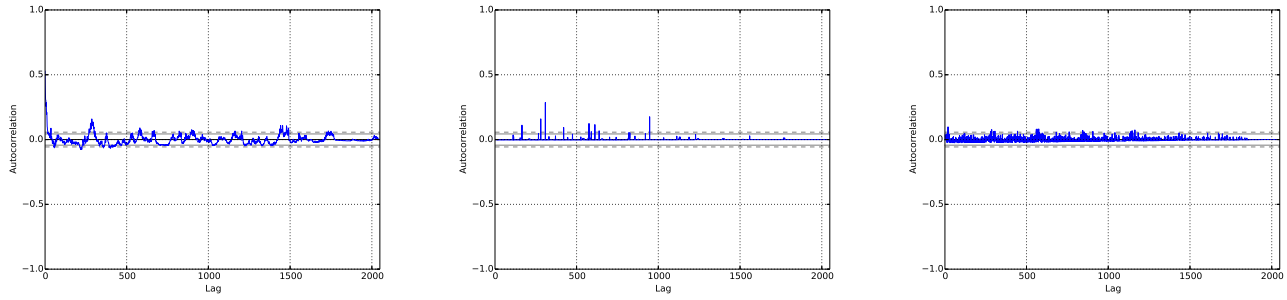
For each application, we continue to characterize the remote servers and their aggregated network prefixes or ASNs via analyzing the *fanouts*, i.e., unique numbers of destination IP address, BGP prefixes, and ASNs. In addition, we measure the distribution of network traffic across these remote servers, prefixes and ASNs via calculating the entropy and standardized entropy of these fanouts. For a given application a for an IoT device d, let N and m denote the number of network traffic flows and the *unique* number of the remote servers represented as $s_1, s_2, \ldots, s_m$. The probability of each remote server p_{s_i} is calculated as $p_{s_i} = \frac{f_{s_i}}{N}$, where f_{s_i} denotes the number of flows between d and s_i. Clearly $\sum f_{s_i} = N$. The entropy on the remote servers for the application a for the device d is then derived as $\mathcal{E}_{d,a} = -\sum_{i=1}^{m} p_{s_i} \log p_{s_i}$, while the normalized entropy is derived as $\mathcal{NE}_{d,a} = \frac{\mathcal{E}_{d,a}}{\log m}$.

The normalized entropy is in the range of [0, 1], revealing the degree of uncertainty, randomness, or variations on the remote servers which communicate with IoT devices in edge networks. Clearly, a $\mathcal{NE}_{d,a}$ value of 0 or near 0 indicates the uniformity on the remote servers, while a $\mathcal{NE}_{d,a}$ value of 1 or near 1 means the high randomness on the remote servers. The former scenario indicates the IoT device only communicates with one or a few servers on the application a, while the latter case reveals the device talking with a large number of random servers. Based on a similar process, we could calculate the entropies and normalized entropies for their aggregated network prefixes or ASNs of remote servers. Table 3 illustrates the entropy values of destination IP addresses, prefixes and ASNs IoT devices have sent HTTPS requests within a 24-hour time window. As shown in Table 3, all IoT devices exhibit how uncertainty on network prefixes and ASNs for their HTTPS traffic, while the laptop and smartphones exhibit much higher variations on the remote prefixes and ASNs for HTTPS traffic. These observations could potentially provide critical insights for detecting traffic anomalies of IoT devices or classifying newly added IoT devices to the edge network.

Device	Flows	Fanout			Normalized Entropy		
		IP	Prefix	ASN	IP	Prefix	ASN
Echo	148	20	6	1	0.5529	0.3158	0.0000
Camera	32	12	9	2	0.6023	0.5422	0.1792
Echo Dot	228	40	10	2	0.6197	0.3365	0.0051
Philips Hue	96	4	2	1	0.2163	0.0221	0.0000
Smart TV	429	109	39	7	0.6574	0.2968	0.1733
IoT Hub	258	3	2	1	0.1969	0.1115	0.0000
Laptop	3831	832	340	90	0.6782	0.5191	0.3064
Smartphone	1497	353	131	21	0.6274	0.4964	0.3077

Table 3: The entropy of destination IP addresses, prefixes and ASNs IoT devices have sent HTTPS requests within a 24-hour time window.

In summary, our multidimensional behavioral profiling of IoT devices have led to a number of discoveries. First, aggregating and correlating the remote IP addresses into BGP networks prefixes and ASNs reveal IoT devices typically engage with servers from a small number of networks and domains due to their specific and single-purchase functionalities. Second, the temporal traffic patterns could classify IoT devices into always-on devices such as smart voice assistants and on-demand devices such as smart TVs.



(a) IoT device A (b) IoT device B (c) Non-IoT device

Figure 5: The autocorrelation plots of network traffic time series for selected IoT and non-IoT devices.

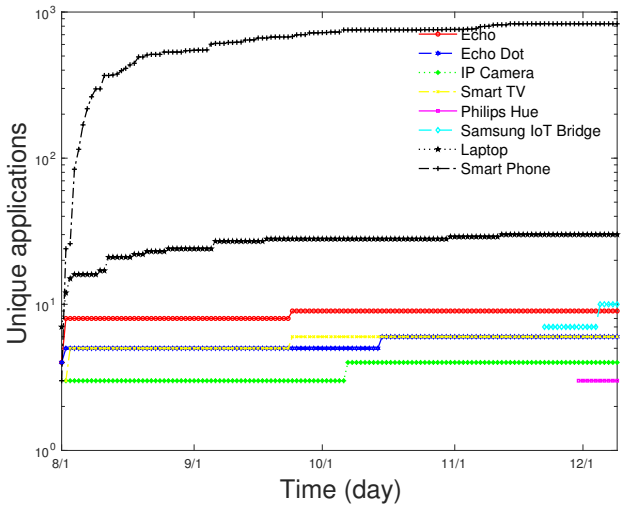


Figure 6: The convergence of applications for IoT and non-IoT devices.

Lastly, most IoT devices communicate with Internet servers for limited, fixed, and common applications such as HTTP, DNS, and NTP services. Profiling traffic behaviors of IoT devices not only uncover what, when and how IoT devices communicate with legitimate end hosts on the Internet, but also provide critical insights for detecting suspicious activities of IoT devices due to security threats and cyber attacks. Thus, the next section leverages IoT behavioral fingerprints for a wide variety of applications such as IoT device detection and classification, anomaly traffic detection, and cybersecurity monitoring.

4 EXPLORING THE APPLICATIONS OF MULTIDIMENSIONAL BEHAVIORAL PROFILING

In this section, we demonstrate the benefits of multidimensional behavioral profiles of IoT devices for a variety of applications including anomaly traffic detection, IoT device detection and classification, and network security monitoring.

4.1 Anomaly Traffic Detection for IoT Devices

Security and privacy are two key challenges faced by today's wide deployment of IoT devices in edge networks due to inadequate built-in security features, flawed authorization and authentication processes, weak password management, and other vulnerabilities. As cyber attacks exploring millions of weakly protected IoT devices often leave substantial traffic footprints in edge networks, we explore multidimensional behavioral profiles for detecting anomaly traffic and security threats.

In this study, we adopt an anomaly detection method based on minimum description length (MDL) principle due to its data-driven approach and parameter-free feature [15, 17, 21]. The intuition and novelty of the MDL principle lie in its pattern-based compression and encoding technique which exploit coding tables to capture the underlying data distributions. In other words, the technique encodes a frequent and common pattern with a short encoded length, and encodes a less frequent and unusual pattern with a long encoded length reflecting anomalies and irregularities in the original data [15].

The MDL principle essentially is a model selection framework for performing lossless compressions and encoding on data with categorical features and attributes. The main process is to search and identify the best model m which minimizes the overall encoding size for the entire data, i.e.,

$$\arg\min_{m \in \mathcal{M}} L(m) + L(d \mid m), \tag{2}$$

where $\mathcal{M}$, $L(m)$, $L(d \mid m)$ are the model set, the bit length describing the specific model m, and the bit length of describing the data d with the model m, respectively.

In the context of network flow traffic of IoT devices in edge networks, we consider all network flow data collected during a given time period as the data-set D consisting of n flow records, each of which has w categorical features, i.e., $\mathcal{F} = \{f_1, \ldots, f_w\}$. To encode the data with a code table, CT, we first extract all the patterns $\mathcal{P}$ in the data, and represent each pattern with a code c in the encoding set C. For a given pattern $p \in \mathcal{P}$ encoded as $c(p)$, we define its frequency, i.e., $freq(p)$ as the number of flow records in D containing p in their encoding. Thus based on the entropy theory, the optimal coding for the pattern p becomes

$$L(c(p) \mid CT) = -\log(\frac{freq(p)}{\sum_{q \in CT} freq(q)}).$$

In addition, the overall number of bits required to encode the entire data-set D is derived as:

$$\begin{aligned} L(D \mid CT) &= \sum_{r \in D} L(r \mid CT) \\ &= \sum_{r \in D} \sum_{p \in freq(r)} L(c(p) \mid CT). \end{aligned}$$

As shown in Eq. 2, the bit length of encoding the overall data is then calculated as:

$$L(CT) = \sum_{p \in CT} L(c(p) \mid CT) + \sum_{v \in \mathcal{V}} -o_v \log(p_v),$$

where $\mathcal{V}$ is the set of all unique categorical attributes appearing in the patterns of the code table, o_v is the the occurrence count of the category value $v \in \mathcal{V}$. p_i is calculated as $\frac{o_i}{\mathcal{L}}$ where $\mathcal{L}$ is the total length of all the patterns in the code table. Combining the entire feature set together, we can build multiple code tables for further reducing the overall encoding cost .

The simple yet effective pattern-based anomaly detection approach allows us to identify unusual or anomalous traffic flows from network traffic originating from or destined to IoT devices in edge networks. Our encoding process leverages the following multidimensional traffic features extracted from network flow records: flow duration, `srcIP`, `srcPort`, `dstIP`, `dstPort`, protocol, packet count, byte count, `dstIP`'s network prefix, and `dstIP`'s ASN. The MDL principle intends to encode unusual patterns with longer encoded lengths, thus we simply consider the encoding length $L(r \mid CT)$ for a network flow record r as the anomaly score.

Figure 7 illustrates the distribution of anomaly scores for all the observed network traffic flows originating from a Google Home smart voice assistant during a 24-hour time window. Based on the widely used elbow principle, we determine the anomaly score of 9 as the threshold for traffic anomalies for IoT devices in edge networks. To evaluate the quality of the anomaly detection, we manually validate all 526 network flows with an anomaly score of 9 or above.

Table 4 summarizes our in-depth analysis of all 526 network flows with high anomaly scores. As shown in Table 4, most of these network flows are long HTTPS connections between the smart voice assistant with Google cloud servers. In addition, a small number of network flows are related to ICMP, mDNS, and DHCP traffic. Thus the manual validation confirms the effectiveness of our proposed pattern-based anomaly detection for discovering unusual traffic activities from the multidimensional behavioral profiles of IoT devices.

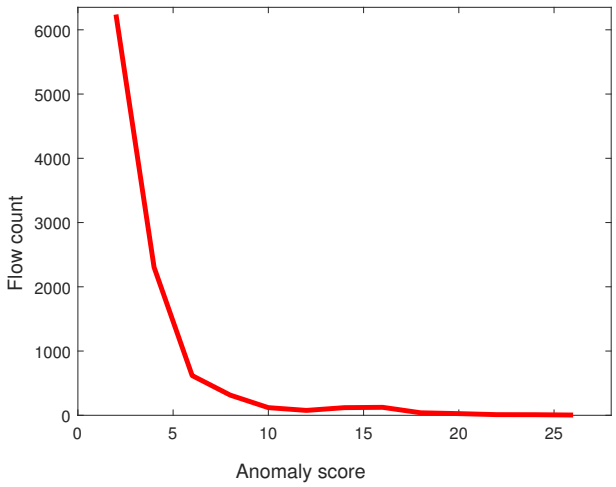


Figure 7: The distribution of anomaly scores for all observed network traffic flows during a 24-hour time window for a smart voice assistant.

Table 4: An in-depth analysis of network traffic flows high anomaly scores.

Protocols	Root cause analysis	Flows
HTTPS	long secure web sessions with cloud servers	489
ICMP	ping traffic	13
mDNS	multicast DNS query	3
DHCP	DHCP requests	9
DNS	Unusual number of Packets	2
8009/TCP	Optimized HTTP service running on the device.	2
5228/TCP	long TCP connections with Google Play services	8

4.2 IoT Device Detection and Classification

The multidimensional behavioral profiles of existing IoT devices in edge networks also provide unique and valualable features for detecting and classifying newly added devices to the network. Let i and j denote two IoT devices in the data-set. For each and every traffic feature in behavioral profiles over a given time window, we can quantify and measure the similarity and correlations of the feature between two devices i and j during the same time period. Assuming the feature b is the remote destination IP addresses (`dstIPs`) that communicate with IoT devices. Let $\mathcal{S}_{i,b}$ and $\mathcal{S}_{j,b}$ represent the unique sets of `dstIPs` observed for IoT devices i and j during the time window, respectively. The similarity on the `dstIP` feature, i.e., $s_{i,j,b}$, is calculated as

$$s_{i,j,b} = \frac{|\mathcal{S}_{i,b} \cap \mathcal{S}_{j,b}|}{|\mathcal{S}_{i,b} \cup \mathcal{S}_{i,b}|}. \quad (3)$$

Thus repeating the same process on the available features extracted from network flow data could lead to a *similarity vector* for any two IoT devices in the same or different edge networks. The similarity

matrix on traffic features among all IoT devices enables us to identify and cluster devices with similar behavioral fingerprints, and more importantly detect new suspicious IoT devices in the same edge network.

Figure 8 illustrates the distributions of similarity scores on three IP-spatial features including `dstIP`, destination prefixes and ASNs between IoT devices in two different edge networks. Each point represents one pair of IoT devices from two networks. As shown in Figure 8, most pairs of IoT devices exhibit low similarities, suggesting IoT devices communicating with diverse servers on the Internet. However, the high similarities between two pairs of IoT devices from two different edge networks are apparently worth in-depth investigations. Our further analysis discovers that two pairs of IoT devices are exactly the same IoT products, i.e., Amazon Echo Dot and Samsung SmartThings Hub, which happen to be deployed in both edge networks. In addition to the similarity scores on IP-spatial features, we also compare the scores on temporal and service dimensions. After ranking the average similarity score over all features, we find that the top pairs of IoT devices with the highest similarity scores, i.e., 0.65 and 0.47, are exactly the same two pairs of devices. We believe that the discovery of high similarity scores on behavioral features among similar IoT devices could help identify newly added or unknown IoT devices by monitoring and learning their behavioral fingerprints during the early phase after they join the edge networks.

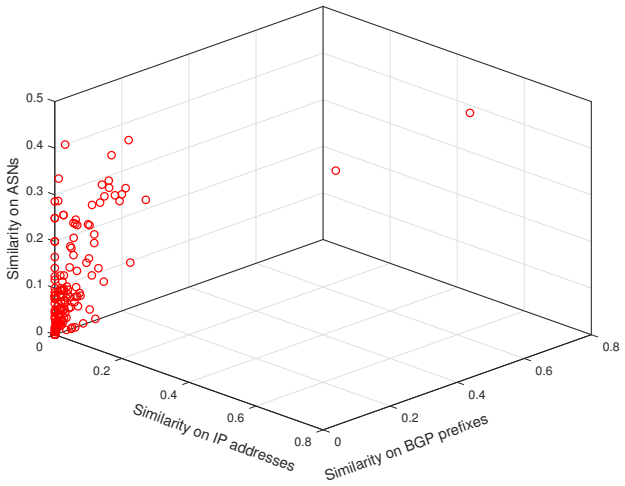


Figure 8: The scatter plot of similarity score on IP-spatial features.

Several recent studies have explored machine learning techniques for IoT device detection and classification [18, 27]. For examples, [18] presents a Random Forest classifier to automatically identify device types of the new IoT devices that are connected to a network for the enforcement of security polices and traffic rules, and [27] leverages the widely used supervised classification algorithm, i.e., Random Forest, for classifying authorized and unauthorized IoT devices based on the features extracted from the link and service layers of BLE protocol stacks. The multidimensional behavioral profiles of IoT devices we have developed in this study will provide additional features and unique insights for improving the quality and performance of these machine learning-based IoT device detection and classification.

4.3 Network Security Monitoring

In light of prevalent cyber attacks and exploits towards vulnerable IoT devices, it is crucial to develop effectively techniques for monitoring traffic activities of IoT devices for network security monitoring. Similar to a network telescope, our proposed measurement framework on programmable edge networks can build the fine-grained and multi-dimensional behavioral profiles of IoT devices, and provide critical insights for discovering the potential exploits and attacks towards IoT devices in real-time.

To demonstrate the feasibility of our proposed IoT measurement framework for network security monitoring, we simulate all the critical steps of Mirai botnet [8, 16] for infiltrating, infecting, and operateing weakly protected IP cameras in a controlled edge network environment. For each of the infiltration, infection, and operation steps, we demonstrate that the behavioral fingerprints left by Mirai botnet traffic reveals many unusual traffic patterns or substantial behavioral deviations that could raise anomalous alerts and security alarms.

During the infiltration step, Mirai first employs a port scan strategy for identifying open ports such as 22, 23, and 2323, and if successful, subsequently attempts to launch a dictionary attack to attempt the logins with 62 default credentials. Clearly the scanning activity and brute-force login process trigger substantial behavioral footprint deviations on the IP-spatial and application dimensions, since the IP address of the remote attacker is from a different network prefix and ASN, and the remote ports used in the scanning are very different from the limited set of applications used by IP cameras. The infection stage also leaves unique behavioral fingerprints on IP-spatial, data volumes, and applications, as the loader, which could be different from the initial scanner, has to transfer the malware image to the compromised IP camera.

During the operation stage, the compromised IP camera, as part of Mirai botnet now, exhibits very unusual attacking behaviors since the device starts to i) perform port scanning activities, 2) communicate with control and command (C2) servers of Mirai botnet, and eventually 3) launch coordinated distributed denial service attacks (DDoS) towards C2-specified targets such as Dyn DNS infrastructure [16]. All of these malicious traffic activities by the IP camera, a new Mirai bot, leaves significant deviations on the behavioral fingerprint on IP camera, thus our proposed multidimensional behavioral profiling framework for IoT devices could effectively detect, mitigate and stop such malicious activities.

5 RELATED WORK

The recent rapid development and deployment of IoT devices in smart homes, cities, and industry 4.0 have attracted signifiant interests from the research community in understanding the applications, security, threats, vulnerability, and the ecosystems [1, 4, 7, 13, 14, 24, 30]. IoT behavioral profiling and fingerprinting have recently attracted wide attention from the system, networking and security research communities. The fingerprinting techniques cover nearly all protocol layers of TCP/IP stacks such as applying wavelet

transform on the sequence of packet inter-arrival time (IAT) of wireless access points for device profiling [11, 26, 27] or characterizing packet headers and IP payload for device fingerprinting [2, 18].

Most of the existing studies on IoT behavioral fingerprinting are centered on the protocols of the physical and link layers for the applications of device classifications [9, 11, 26, 27]. For example, [26] introduces a real-time system that passively scans and analyze the data communication over WiFi, Bluetooth, and Zigbee for classifying IoT devices and detecting privacy threats, while [27] proposes to extract the unique features from the link and service layers of Bluetooth low energy (BLE) protocol stack for generating the IoT fingerprint for authenticating devices and defensing against spoofing attacks. In addition, [9] proposes a wireless device identification platform for distinguishing legitimate and adversarial IoT devices based on radio frequency (RF) fingerprinting over different ranges of signal-to-noise ratio (SNR) levels.

A few recent studies have shifted traffic data collection and analysis to the network, transport and application layers for device behavioral modeling and characterizations [2, 18]. For example [18] establishes IoT device fingerprints with 20 binary features of protocol fields extracted from packets headers collected from link, network, transport and application layers to reflect the protocol engagement of IoT devices headers such as ARP, IP, ICMP, TCP, UDP, NTP, DNS, DHCP, HTTP and HTPPS, and 3 numerical features including packet size, destination IP counter, source and destination port numbers, while [2] characterizes the behavioral fingerprints of IoT devices with a subset of binary features identified in [18], and 3 payload-based features including the entropy of payload, TCP payload size, and TCP window size. Compliment to these studies, our paper focuses on the behavioral fingerprinting of IoT devices in edge networks based network flow records, rather than the raw IP data packets which raise on privacy concerns of IoT users and computational resources on edge routers, for detecting new devices and traffic anomalies.

6 CONCLUSIONS AND FUTURE WORK

As the rapid and wide adoption of IoT devices continue to accelerate in smart homes, cities, and industries, it becomes increasingly urgent to design and implement Internet traffic measurement platforms to effectively monitor, characterize, and profile communications pattens of IoT devices with remote end hosts on the Internet and local systems on the same edge networks. Towards this end, this paper develops a systematic measurement framework for establishing multidimensional behavioral profiles of connected IoT devices based on a wide spectrum of traffic features from IP-spatial, temporal, and cloud dimensions. Based on real network traffic data collected from 22 edge networks over 4-month time span, we have discovered a number of important findings on behavioral fingerprints of IoT devices. First, IoT devices typically communicate with cloud servers from a very small number of prefixes and ASNs, which belong to IoT manufactories, the cloud service providers, NTP service providers, public DNS service providers. Second, IoT devices often exhibit repeated and predictable traffic activities over time due to heart-beat signals between IoT devices and cloud servers. Lastly, unlike laptops, desktops, and smartphones, IoT devices often engage with a limited and common number of applications such as DNS, HTTPS, HTTP, and NTP. These behavioral fingerprints not only summarize communication patterns of IoT devices with end systems on the Internet, but also benefit a range of security applications for IoT devices such as anomaly traffic detection, IoT detection and classification, and network security monitoring. Our future work is centered on exploring the traffic fingerprints at the link layer, i.e., studying wireless communications between IoT hubs and IoT sensors via Bluetooth, ZigBee, Z-Wave, and Wi-Fi. The link layer fingerprint could compliment the current behavioral fingerprinting framework based on traffic features collected from network, transport, and application layers.

ACKNOWLEDGMENTS

This research was supported in part by NSF grants 1816995, 1717197, and 1704092. The information reported here does not reflect the position or the policy of the funding agency.

REFERENCES

[1] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi. 2014. Internet of Things for Smart Cities. *IEEE Internet of Things Journal* 1, 1 (February 2014), 22 – 32.

[2] B. Bezawada, M. Bachani, J. Peterson, H. Shirazi, I. Ray, and I. Ray. 2018. Behavioral Fingerprinting of IoT Devices. In *Proceedings of ACM CCS Workshop on Attacks and Solutions in Hardware Security (ASHES)*.

[3] Business Insider. 2019. Google, This is Bogus as Hell - One of the Fathers of the Internet Blasts Google for how Chromecast Behaves on His Home Network. https://www.businessinsider.com/paul-vixie-blasts-google-chromecast-2019-2/.

[4] E. Fernandes, J. Jung, and A. Prakash. 2016. Security Analysis of Emerging Smart Home Applications. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*.

[5] E. Fernandes, J. Paupore, A. Rahmati, D. Simionato, M. Conti, and A. Prakash. 2016. FlowFence: Practical Data Protection for Emerging IoT Application Frameworks. In *Proceedings of USENIX Conference on Security Symposium*.

[6] E. Ronen, A. Shamir, A.-O. Weingarten, and C. O'Flynn. 2017. IoT Goes Nuclear: Creating a ZigBee Chain Reaction. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*.

[7] G. Ho, D. Leung, P. Mishra, A. Hosseini, D. Song, and D. Wagner. 2016. Smart Locks: Lessons for Securing Commodity Internet of Things Devices. In *Proceedings of ACM on Asia Conference on Computer and Communications Security (ASIACCS)*.

[8] G. Kambourakis, C. Kolias, and A. Stavrou. 2017. The Mirai botnet and the IoT Zombie Armies. In *Proceedings of EEE Military Communications Conference (MILCOM)*.

[9] H. Jafari, O. Omotere, D. Adesina, H.-H. Wu, and L. Qian. 2018. IoT Devices Fingerprinting Using Deep Learning. In *Proceedings of IEEE Military Communications Conference (MILCOM)*.

[10] J. Zhang, X. Chen, Y. Xiang, W. Zhou, and J. Wu. 2015. Robust Network Traffic Classification. *IEEE/ACM Transactions on Networking* 23, 4 (August 2015), 1257 – 1270.

[11] K. Gao, C. Corbett, and R. Beyah. 2010. A Passive Approach to Wireless Device Fingerprinting. In *Proceedings of IEEE/IFIP International Conference on Dependable Systems & Networks (DSN)*.

[12] K. Park and W. Willinger. 2002. *Self-Similar Network Traffic and Performance Evaluation*. John Wiley & Sons.

[13] K. Xu, F. Wang, and X. Jia. 2016. Secure the Internet, One Home at a Time. *Security and Communication Networks* 9, 16 (November 2016), 3821 – 3832.

[14] K. Xu, Y. Wan, and G. Xue. 2019. Powering Smart Homes with Information-Centric Networking. *IEEE Communication Magazine* (2019).

[15] L. Akoglu, H. Tong, J. Vreeken, and C. Faloutsos. 2012. Fast and Reliable Anomaly Detection in Categorical Data. In *Proceedings of ACM international conference on Information and knowledge management (CIKM)*.

[16] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou. 2017. Understanding the Mirai Botnet. In *Proceedings of USENIX Security Symposium*.

[17] M. Li and P. Vitányi. 1993. *An Introduction to Kolmogorov Complexity and its Applications*. Springer.

[18] M. Miettinen, S. Marchal, I. Hafeez, N. Asokan, A.-R. Sadeghi, and S. Tarkoma. 2017. IoT SENTINEL: Automated Device-Type Identification for Security Enforcement in IoT. In *Proceedings of IEEE International Conference on Distributed Computing Systems (ICDCS)*.

[19] M. Trevisan, D. Giordano, I. Drago, M. Mellia, and M. Munafo. 2018. Five Years at the Edge: Watching Internet from the ISP Network. In *Proceedings of International Conference on emerging Networking EXperiments and Technologies (CoNEXT)*.

[20] O. Alrawi, C. Lever, M. Antonakakis, and F. Monrose. 2019. SoK: Security Evaluation of Home-Based IoT Deployments. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*.

[21] P. Grünwald. 2007. *The Minimum Description Length Principle.* MIT press.

[22] Q. Wang, W. Hassan, A. Bates, and C. Gunter. 2018. Fear and Logging in the Internet of Things. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*.

[23] R. Want, B. Schilit, and S. Jenson. 2015. Enabling the Internet of Things. *Computer* 48, 1 (February 2015), 28 – 35.

[24] S. Feng, P. Setoodeh, and S. Haykin. 2017. Smart Home: Cognitive Interactive People-Centric Internet of Things. *IEEE Communications Magazine* 55, 2 (February 2017), 34 – 39.

[25] S. Shams, S. Goswami, K. Lee, S. Yang, and S.-J. Park. 2018. Towards Distributed Cyberinfrastructure for Smart Cities Using Big Data and Deep Learning Technologies. In *Proceedings of IEEE International Conference on Distributed Computing Systems (ICDCS)*.

[26] S. Siby, R. Maiti, and N. Tippenhauer. 2017. IoTScanner: Detecting Privacy Threats in IoT Neighborhoods. In *Proceedings of ACM International Workshop on IoT Privacy, Trust, and Security (IoTPTS)*.

[27] T. Gu and P. Mohapatra. 2018. BF-IoT: Securing the IoT Networks via Fingerprinting-Based Device Authentication. In *Proceedings of IEEE International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*.

[28] University of Oregon. [n.d.]. Route Views Project. http://www.routeviews.org/.

[29] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu. 2016. Edge Computing: Vision and Challenges. *IEEE Internet of Things Journal* 3, 5 (October 2016), 637–646.

[30] W. Zhang, Y. Meng, Y. Liu, X. Zhang, Y. Zhang, and H. Zhu. 2018. HoMonit: Monitoring Smart Home Apps from Encrypted Traffic. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

[31] Y. Meidan, M. Bohadana, A. Shabtai, J. Guarnizo, M. Ochoa, N. Tippenhauer, and Y. Elovici. 2017. ProfilIoT: a Machine Learning Approach for IoT Device Identification Based on Network Traffic Analysis. In *Proceedings of ACM Symposium on Applied Computing*.

[32] Y. Rekhter and T. Li. 1995. A Border Gateway Protocol 4 (BGP-4).