Characterizing and Mining Traffic Patterns of IoT Devices in Edge Networks

Yinxin Wan, *Student Member, IEEE*, Kuai Xu, *Senior Member, IEEE*, Feng Wang, *Member, IEEE*, and Guoliang Xue, *Fellow, IEEE*

***Abstract*—As connected Internet-of-things (IoT) devices in smart homes, smart cities, and smart industries continue to grow in size and complexity, managing and securing them in distributed edge networks have become daunting but crucial tasks. The recent spate of cyber attacks exploiting the vulnerabilities and insufficient security management of IoT devices have highlighted the urgency and challenges for securing billions of IoT devices and applications. As a first step towards understanding and mitigating diverse security threats of IoT devices, this paper develops an IoT traffic measurement framework on programmable and intelligent edge routers to automatically collect incoming, outgoing, and internal network traffic of IoT devices in edge networks, and to build multidimensional behavioral profiles 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, entropy, 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 vulnerable and resource-constrained IoT devices on the Internet.**

***Index Terms*—Internet-of-Things, measurement, smart home, network monitoring, anomaly traffic detection.**

I. Introduction

THE rapid development and deployment of IoT devices have introduced a wide spectrum of innovative applications and services such as industrial automation, smart homes, and remote healthcare monitoring. However, the burgeoning and insecure IoT devices in millions of edge networks such as smart home networks have left backdoors for Internet attackers to launch data theft, device hijacking, and distributed denial of service attacks (DDoS), e.g., the well-known Mirai botnet [4], [16]. Therefore, there is an urgent call for effective techniques to detect, recognize, characterize, and address security threats towards these devices and applications.

Manuscript received May 18, 2020; revised August 6, 2020; accepted August 22, 2020. Date of publication September 25, 2020; date of current version March 17, 2021. The information reported here does not reflect the position or the policy of the funding agency. This research was supported by NSF under Grants 1816995, 1717197, and 1704092. Recommended for acceptance by Dr. Shiwen Mao. This paper was presented in part at the IEEE/ACM IWQoS, 2019, Phoenix, AZ, USA, Jun. 2019. [34] *(Corresponding author: Guoliang Xue.)*

The authors are with the Arizona State University, Tempe, AZ 85281 USA (e-mail: ywan28@asu.edu; kuai.xu@asu.edu; fwang25@asu.edu; xue@asu.edu).

Digital Object Identifier 10.1109/TNSE.2020.3026961

In this paper, we focus on understanding and analyzing the network traffic patterns of IoT devices, which is a critical step to secure IoT devices in edge networks. Specifically, we propose an IoT traffic measurement framework to automatically collect, process, characterize, and profile communication patterns of IoT devices. The key component of our design is the programmable commercial edge router which continuously collects and fingerprints network flow data of IoT devices in real-time. We implement the measurement and monitoring functions at the edge routers because both incoming and outgoing traffic as well as the internal local area network (LAN) traffic are visible at such gateway locations.

Our proposed measurement framework enables us to have a delineated view of data communications and network configurations of IoT devices. For example, we discovered that Chromecast, a streaming media player developed by Google, configures Google domain name system (DNS) servers for DNS queries rather than adopting the default local DNS servers [6]. Such behaviors are very hard to discover if the measurement functions are not available on edge routers.

The availability of the large volume of real world network traffic makes it possible to develop multidimensional traffic profiles of IoT devices for gaining an in-depth understanding of communication patterns and traffic behaviors, and more importantly, detecting and mitigating suspicious activities and cyber attacks towards vulnerable IoT devices. Towards this end, we build the behavioral profile of IoT devices based on a wide spectrum of their traffic features from *IP-spatial*, *temporal*, *cloud*, and *internal* traffic dimensions. The IP-spatial dimension is centered on the analysis of remote IP addresses of Internet end hosts such as DNS servers or network time protocol (NTP) servers. In addition, aggregating these remote IP addresses into Border Gateway Protocol (BGP) network prefixes [24] and ASNs allows us to analyze IP-spatial correlations of Internet end hosts communicating with IoT devices. Our experimental results reveal that most IoT devices engage with cloud servers from a small set of network prefixes and ASNs due to their designs for single-purpose applications and specific functions.

Additionally, we explore the entropy concept to gain a deeper insight of the temporal dynamics and predictability of IoT devices. Our experimental results of measuring the sample entropy of different IoT devices reveal interesting observations on how IoT devices differ from each other in communication patterns and how entropy measures fluctuate over time and correlate with user-triggered activities. Through analysis on the cloud

dimension, our study discovers that IoT devices typically only engage with a small and fixed set of common applications such as hypertext transfer protocol (HTTP), HTTP secure (HTTPS), DNS, and NTP. At last, benefiting from the strategical location of the programmable edge router, we are able to further investigate the communication patterns of IoT traffic within the LAN by categorizing the traffic into communication between IoT devices and the router, communication among IoT devices, and communication between IoT and non-IoT devices.

In light of the prevalent cybersecurity threats, we explore the benefits of multidimensional behavioral profiles for a variety 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, which encodes common network traffic patterns with short encoded length and captures infrequent and unusual patterns with longer encoded length. Our experimental evaluation shows that our approach is able to uncover suspicious traffic activities with high precision. Moreover, we leverage multidimensional profiles of IoT devices for recognizing and detecting new and unknown IoT devices based on the knowledge 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:

- We present a measurement framework for capturing and collecting network traffic of IoT devices to characterize and model behavioral fingerprint of IoT devices in edge networks.
- We introduce a multidimensional approach to model the IP-spatial, temporal, entropy, and cloud behaviors of heterogeneous IoT devices, as well as the communication patterns of the internal LAN traffic.
- We explore 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 III gives a brief explanation of the research background and introduces the proposed measurement framework. Section IV presents the multidimensional behavioral profiles of IoT devices. In Section V, 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 II discusses related work in this research area, while Section VI concludes this paper and outlines our future work.

II. Related Work

The recent rapid development and deployment of IoT devices in smart homes, cities, and industry 4.0 have attracted significant interests from the system, networking, and security research communities in understanding their applications, security and privacy threats, vulnerabilities, and ecosystems [8], [9], [14], [27], [33], [35], [36], [38]. IoT behavioral profiling and fingerprinting is one of the crucial topics where we have witnessed a lot of recent research efforts. 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 [10], [13], [30] or characterizing packet headers and IP payload [5], [20].

Most of the existing studies on IoT behavioral fingerprinting are centered on the protocols of physical and link layers for the applications of device classification [10], [13], [15], [30]. For example, [30] introduces a real-time system that passively scans and analyzes the data communication over WiFi, Bluetooth, and Zigbee for classifying IoT devices and detecting privacy threats, while [13] 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, [15] 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 characterization [5], [20]. For example, [20] achieves IoT device fingerprints with 20 binary features of protocol fields extracted from packet 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. [5] characterizes the behavioral fingerprints of IoT devices with a subset of binary features identified in [20], and 3 payload-based features including the entropy of payload, TCP payload size, and TCP window size. Complement to these studies, our paper focuses on behavioral fingerprinting of IoT devices in edge networks based on network flow records rather than the raw IP data packets which raise privacy concerns of IoT users and incur expensive computational and storage cost on resource-constrained commodity edge routers such as off-the-shelf home routers.

A very recent paper studying the IoT devices on home networks [17] provides a large-scale empirical analysis with the ISP level network traffic data. Different from [17], our study explores programmable edge routers to build an IoT measurement framework from the perspective of edge networks and sheds light on multidimensional traffic patterns of IoT devices from incoming and outgoing network traffic as well as from the local LAN traffic within edge networks.

To summarize, our paper is different with most existing works in the way that it designs and implements an IoT traffic measurement framework based on programmable edge routers. To the best of our knowledge, our work is the first to build and study multidimensional behavioral profiles of heterogeneous IoT devices using network traffic data.

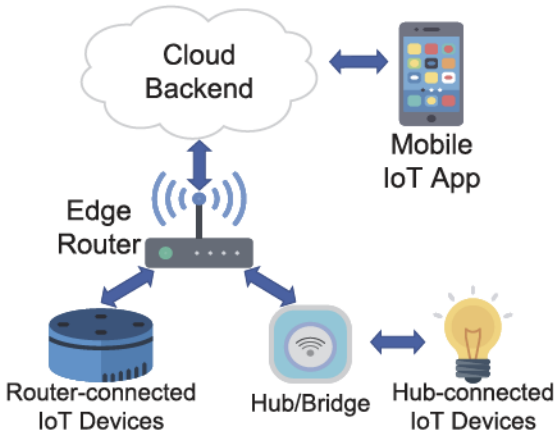


Fig. 1. An example communication flow of smart IoT platforms.

III. Background and Traffic Measurement

A. Background

Recent advances in embedded systems have enabled the wide deployment of IoT devices in edge networks. The major players in IoT domains have also developed their own smart IoT platforms such as Samsung's SmartThings [29], Google's Nest [11], and Amazon's Alexa [3] to support broad IoT compatibility and rapid application development. These IoT platforms adopt similar system architectures that consist of IoT devices, cloud-based servers, and IoT applications. For example, to remotely turn on a Philips Smart Hue bulb in the front yard of a house at night, the user could simply open the Philips Hue App installed on the smartphone and send the *turn-on-bulb* command to a Philips Lighting server located in Google cloud. The cloud server then communicates with the Philips Hue bridge in the user's home, which in turn forwards the *turn-on-bulb* command to the smart bulb in the front yard. This communication flow is illustrated in Fig. 1.

These existing IoT platforms are primarily function-driven and feature-driven, thus leaving security and privacy concerns as the secondary or optional goals. As a result, today's IoT devices are often vulnerable to a variety of security threats and many of them have already been compromised [17]. For example, the insecure configuration and design flaws of IoT devices have contributed to one of the largest botnets, the Mirai botnet [4], [16], which commands and controls over 600,000 IoT devices at its peak. In addition, the coarse access control policy, malicious applications, and exposures in open wireless channels have created broad attack vectors towards heterogeneous IoT devices [7], [9], [14], [27].

Protecting and securing millions of vulnerable IoT devices is a complicated and challenging task. As a recent security evaluation study [2] pointed out, IoT measurement and monitoring is an important early step towards this goal. Specifically, the first step of IoT security lies in the measurement, monitoring, and analysis of communication patterns and behavioral profiles of IoT devices. For example, what do remote *hosts* on the Internet talk with the smart speakers or thermostats, at what *time*, for what *reasons*? Answering these questions is of great importance to understanding if, when, and how the connected IoT devices in edge networks are targeted, compromised, and controlled by cyber attacks.

B. Traffic Measurement via Programmable Edge Routers

In this study we advocate an edge router-based IoT traffic measurement and monitoring platform for continuously monitoring *the incoming and outgoing traffic between edge networks and the Internet* as well as *the internal traffic within edge networks*. Compared with ISP-based solutions [4], [17], which are based on the network address translation (NAT) router translated traffic, the edge router-based platform has a more detailed and comprehensive view of IoT traffic. More specifically, the edge router can see the non-translated incoming and outgoing traffic to and from an IoT device. The internal LAN traffic in the edge network is also visible to the internal interface of the edge router. In addition, the edge router-based solutions are transparent to IoT devices and therefore there is no need for the users to install or update additional packages and applications on IoT devices. Furthermore, the edge router is a central security checkpoint at an ideal location for control and policy enforcement. These unique strengths motivate us to develop a router-based traffic measurement platforms to capture, store, characterize, and mine traffic patterns of IoT devices in edge networks.

Fig. 2 illustrates our proposed IoT traffic measurement framework via programmable routers at edge networks. In this framework, 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, we collect 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, duration, byte count, and packet count. We have deployed the prototype framework in 22 real-world home edge networks across the United States, Hong Kong SAR, and mainland China since August 2018. These smart homes house hundreds of IoT devices for a variety of purposes, among which we have observed 20 unique models of IoT devices, as summarized in Table I. We choose not to collect raw IP packets from IoT devices in this study since most data packets originating from or destined to IoT devices are encrypted, and applying MITM proxy to bypass the TLS/SSL encryption is impractical because it requires full root privileges of the IoT devices [21]. The storage of raw data packets of IoT devices such as smart TVs or IP cameras could also bring undesired system challenges for resource-constrained edge routers. In fact, network flow records are widely used for Internet traffic classification, network measurement and analysis [31], [37] thanks to their diverse and informative traffic features and marginal computational and storage resource overheads.

The availability of millions of network traffic flow data allows us to characterize and model the multidimensional behavioral profiles of heterogeneous IoT devices. Specifically, we explore the behavior in four dimensions: IP-spatial, temporal, cloud, and internal traffic. The study of *IP-spatial* behavior focuses on remote IP addresses engaging with IoT devices and aggregates these IP addresses into BGP network prefixes and ASNs for correlation analysis. The *temporal behavior* focuses on the temporal dynamics and predictability

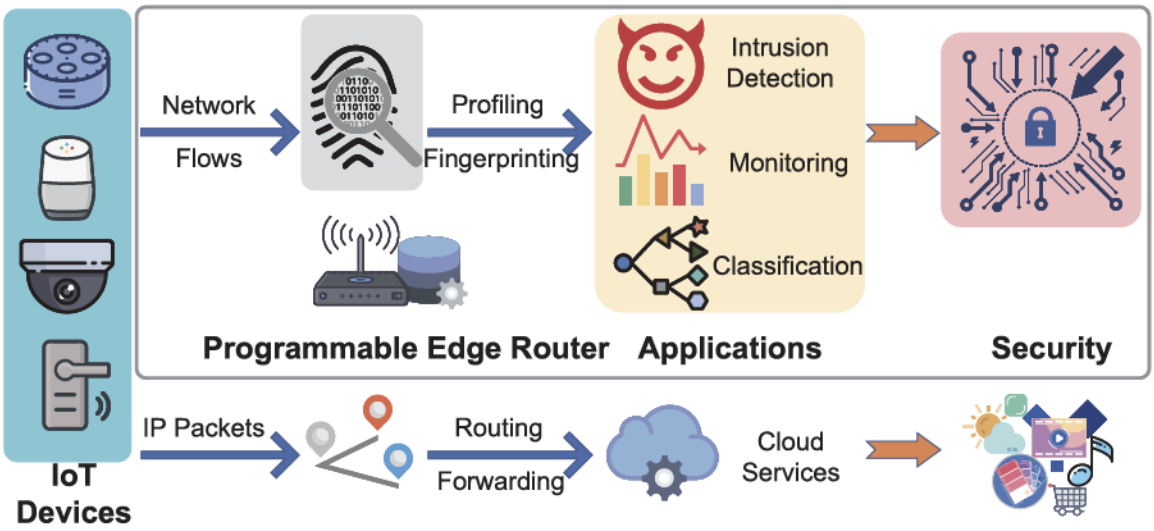


Fig. 2. An IoT traffic measurement framework via programmable routers at edge networks.

TABLE I
HETEROGENEOUS IOT DEVICES DEPLOYED IN 22 HOMES

Device Name	Device Type	Connectivity
Amazon Echo	Voice Assistant	WiFi
Amazon Echo Dot	Voice Assistant	WiFi
Amazon Echo Show	Voice Assistant and Camera	WiFi
Amcrest ProHD Camera	Cloud Camera	Ethernet & WiFi
Arlo Ultra Camera	Cloud Camera	WiFi
August Smart Lock Connect	Hub	WiFi
August Doorbell Cam Pro	Doorbell	WiFi
Google Home	Voice Assistant	WiFi
Google Nest Camera	Cloud Camera	WiFi
Gosund WiFi Smart Socket	Smart Plug	WiFi
LG Smart TV	Smart TV	Ethernet & WiFi
Philips Hue Smart Bridge	Hub	Ethernet & WiFi
Reolink Camera	Cloud Camera	Ethernet & WiFi
Ring Video Doorbell	Doorbell	WiFi
Samsung Smarthings Hub	Gateway Ethernet & WiFi	
Schlage Wireless Lock	Smart Lock	WiFi
TCL Smart TV	Smart TV	Ethernet & WiFi
TP-LINK Smart Bulb LB130	Light Bulb	WiFi
TP-LINK Wi-Fi Smart Plug	Smart Plug	WiFi
YI Home Camera	Cloud Camera	WiFi

of traffic features of IoT devices over time. Specifically, *temporal dynamics* studies how traffic features and behvaioral patterns of IoT devices change over time, while *temporal predictability* explores sample entropy concepts to understand if IoT traffic features and patterns are predictable based on prior observations.

By analyzing how IoT devices interact with cloud servers, we build the profiles of their *cloud behaviors*. In addition to studying the IP-spatial, temporal, and cloud behaviors based on the incoming and outgoing network traffic between IoT devices and end systems on the Internet, our measurement framework also enables us to study the communication patterns of IoT devices within edge networks. These multidimensional behavioral profiles of IoT devices built by our proposed IoT measurement framework effectively capture who, when, what, and why on the behavioral patterns of IoT devices in edge networks, and ultimately lead to a variety of practical applications such as intrusion detection, IoT device detection and classification, and security monitoring.

TABLE II
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
Reolink IP Camera	Yes	2	2	1
Philips Hue	Yes	1	1	1
Samsung Smart Plug	Yes	3	2	1
LG Smart TV	Yes	4	3	2
Android Smartphone	No	37	24	13
Macbook Laptop	No	172	102	39

IV. MULTIDIMENSIONAL BEHAVIORAL PROFILING OF IOT DEVICES

In this section, we present a multidimensional approach to characterizing the behaviors of IoT devices based on a wide spectrum of traffic features.

A. IP-Spatial Behavior of IoT Devices

We first characterize the IP-spatial behaviors of IoT devices by analyzing whom the IoT devices talk to. We propose to aggregate remote IP addresses that IoT devices communicate with into *BGP network prefixes* and *ASNs* in order to gain an in-depth understanding of "clustered" IP-spatial behaviors for IoT devices and make sense of the remote IP addresses. 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 [32] and the official registry records from Internet assigned numbers authority (IANA). Our experiments following this strategy reveal an interesting observation. Even though most IoT devices communicate with a large number of remote hosts, they typically only engage with a very small subset of BGP network prefixes and ASNs, which are likely from the same server pool by the same service providers for efficient load balancing and content distributions. Table II illustrates the IP-spatial behavior

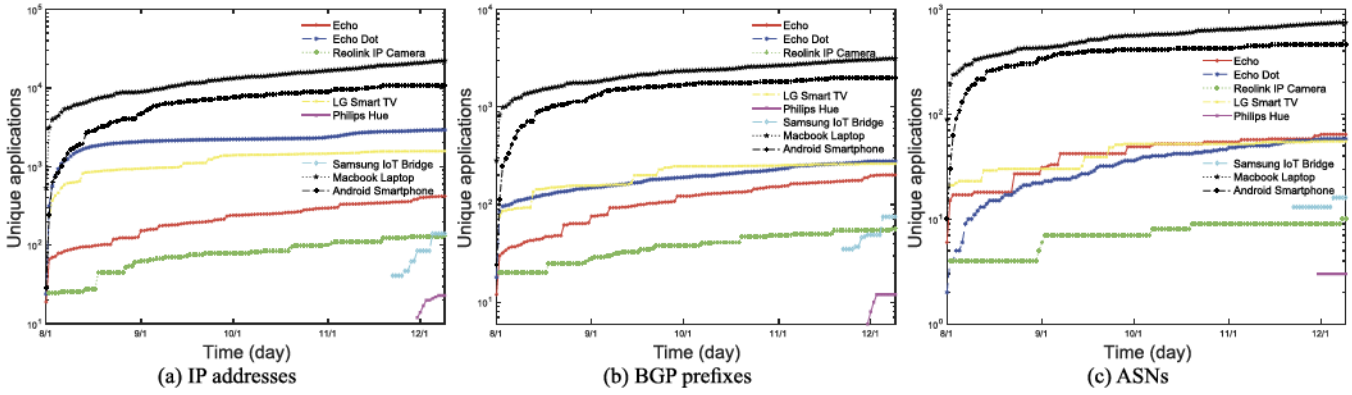


Fig. 3. The convergence of IP addresses, prefixes, and ASNs for IoT and non-IoT devices over the longitudinal measurement period (Samsung IoT Bridge and Philips Hue were added to the system late).

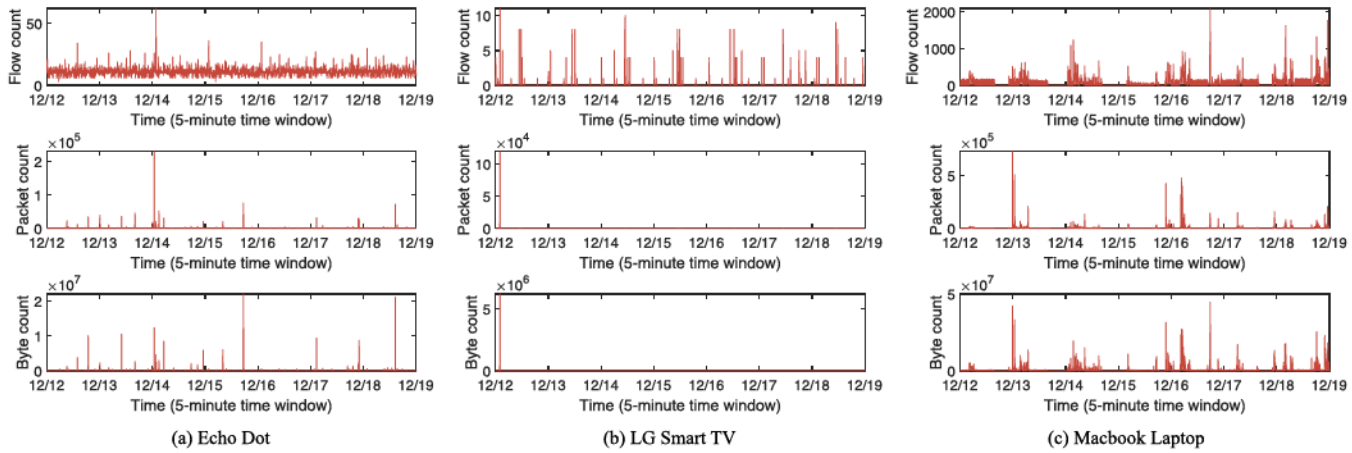


Fig. 4. Traffic characteristics of IoT devices and non-IoT devices over one-week time-span.

patterns of 6 IoT devices and 2 non-IoT devices in a same edge network during a 5-minute time window. As we can see from Table II, all 6 IoT devices only communicate with servers from a very limited number of *unique* ASNs, but the smartphone and laptop engage with remote end hosts from 13 and 39 *unique* ASNs, respectively.

Fig. 3[a-c] demonstrate the convergence of unique remote IP addresses, their network prefixes, and ASNs for a variety of IoT and non-IoT devices deployed in one edge network over a 4-month time span. This longitudinal measurement study for the IP-spatial behavior confirms that most IoT devices engage with a much smaller set of destination IP addresses, prefixes, and ASNs than smartphones and laptops.

B. Temporal Behavioral Dynamics of IoT Devices

We study the temporal behavior of IoT devices by measuring the number of distinct time slots in which IoT devices exhibit traffic activities. We select a 5-minute time window in the experiment in order to balance the computation overhead and monitor real-time traffic activities. Fig. 4[a-c] depict the flow, packet, and byte count of three different kinds of devices over a one-week time span. As shown in Fig. 4, the Echo Dot, LG Smart TV, and Macbook Laptop exhibit distinct traffic characteristics over time. These features reflect the activities of different devices. For example, LG smart TV has high peaks in both packet count and byte count at the very beginning, which corresponds to the activity that this device was turned on for network streaming at that time. The diversity of temporal patterns on flow, packet and byte count inspires us to measure and quantify the *variability* over the entire data collection period.

For each IoT device d in the edge network, let $W_{d,i}$ denote the number of time windows in which the device d is observed with network traffic on the i-th day. Considering that the connected devices are randomly added into the edge network, we use the average time window μ_d for each device rather than the total number of time windows during the entire measurement period, which is derived as $\mu_d = \frac{\sum_{i=1}^{N} W_{d,i}}{N}$, where N is the number of the days since device d is observed in the edge network and $1 \leq i \leq N$. So the temporal variability on time windows, measured by coefficient of variance, can be calculated as $CoV_d = \frac{\mu_d}{\sigma_d}$, where σ_d, the standard deviation, is derived as $\sigma_d = \sqrt{\frac{1}{N}\sum_{i=1}^{N} W_{d,i} - \mu_d}$.

Fig. 5 is a scatter graph on the mean μ and coefficient of variance CoV of time slots observed with network activities for

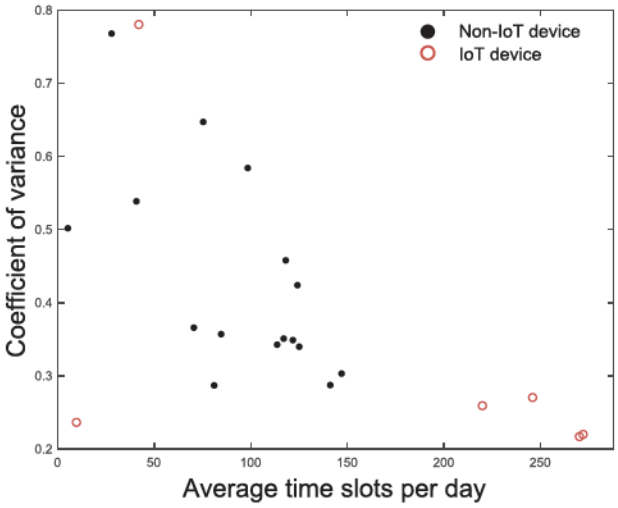


Fig. 5. The mean and coefficient of variance of time slots observed with traffic activities for IoT and non-IoT devices.

different IoT and non-IoT devices. In Fig. 5, 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 compared with non-IoT devices. One of the IoT devices, i.e., an IP camera, is only active for a small number of time slots per day, but exhibits low variability on the time window as well. The only IoT device exhibiting a high variability is a smart TV, and the main reason is that it is often turned on and off in an unpredictable fashion. Based on these observations, we can easily classify connected networked devices in edge networks into three categories: always-on IoT devices (e.g. Echo Dot), on-demand IoT devices (e.g. IP camera and smart TV), and non-IoT devices.

The self-similarity traffic patterns of IoT devices visualized in Fig. 4 also inspire us to analyze the autocorrelation on network traffic generated by all connected devices in edge networks. Autocorrelation is a metric that quantifies the correlation of the same variable across different and lagged periods of times, thus it is also often referred to as serial correlation and lagged correlation. The autocorrelation metric, $\rho_{d,k}$, for the IoT device d, between network traffic activity time series $X_{d,i}$ and a k-lagged copy of itself $X_{d,i+k}$ is captured by the autocorrelation function (ACF) as follows:

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

where μ and σ are the mean and standard deviation of network traffic activity time series $X_{d,1}, X_{d,2}, \ldots, X_{d,n}$ for the device 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 [19], [23].

Fig. 6 illustrates the autocorrelation plots, also referred to as correlograms, of network traffic time series for three selected IoT and non-IoT device. These plots reflect distinct repeating patterns of different devices. We can see noticeable peaks at the beginning where time lag is short for both Philips Hue and Amazon Echo. This indicates that communication patterns of IoT devices are typically stable and predictable. On the other hand, for Android smartphone, there is no significant peak in the autocorrelation plot, which corresponds to our intuition that non-IoT devices like smartphones often have messy and random network traffic.

C. Characterizing Traffic Predictability via Sample Entropy

To further study temporal dynamics and predictability of IoT network traffic in edge networks, we explore *sample entropy*, denoted as $\mathcal{SE}$, to quantify the randomness, uncertainty, or determinism of network traffic for IoT device over time due to the inherent ability of the sample entropy measure in capturing the complexity and predictability of time series data [26]. Given a traffic feature f, our continuous data collection generates a unique time series observation $f(t_1), f(t_2), \ldots, f(t_M)$ over M consecutive time windows. Let $Y(t_i)$ denote a vector of m continuous observations at time t_i, i.e., $\{f(t_i), f(t_{i+1}), \ldots, f(t_{i+m-1})\}$. For $1 \leq i \leq M - m + 1$, $B_i^m(r)$ represents the number of $Y(t_j)$ such that $D[Y(t_i), Y(t_j)] \leq r$ $(j \neq i)$, where $D[Y(t_i), Y(t_j)] = \max_k |f_i(t_k) - f_j(t_k)|$ where $f_i(t_k) \in Y(t_i), f_j(t_k) \in Y(t_j)$, and r specifies how much two sequences are expected to exhibit strong similarity, which is usually set as proportional to the standard deviation of the original time series.

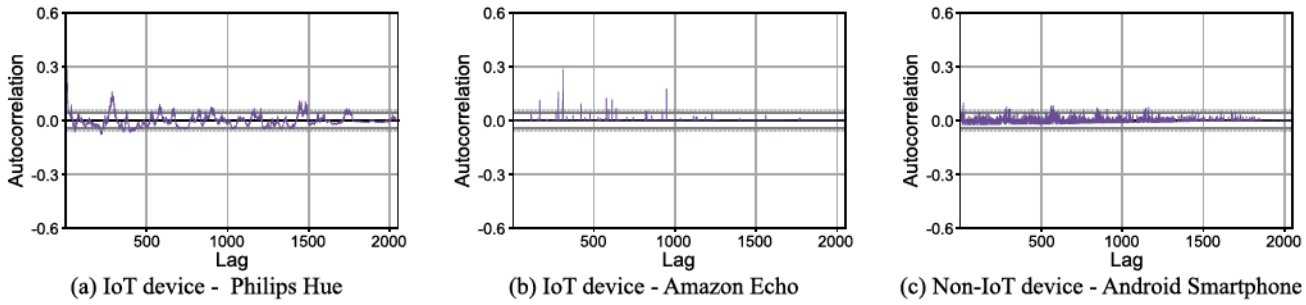
The sample entropy $\mathcal{SE}$ is defined as

$$\mathcal{SE} = -\ln(\Phi^{m+1}(r)/\Phi^m(r)), \tag{2}$$

where $\Phi^m(r)$ is the mean average value of $B_i^m(r)$, i.e., $\Phi^m(r) = (M - m + 1)^{-1} \sum_{i=1}^{M-m+1} B_i^m(r)$. In other words, the sample entropy reflects the conditional probability for two subsequences of $f(t_1), f(t_2), \ldots, f(t_M)$ that are similar along m consecutive observations continue to share similarity for $m + 1$ observations.

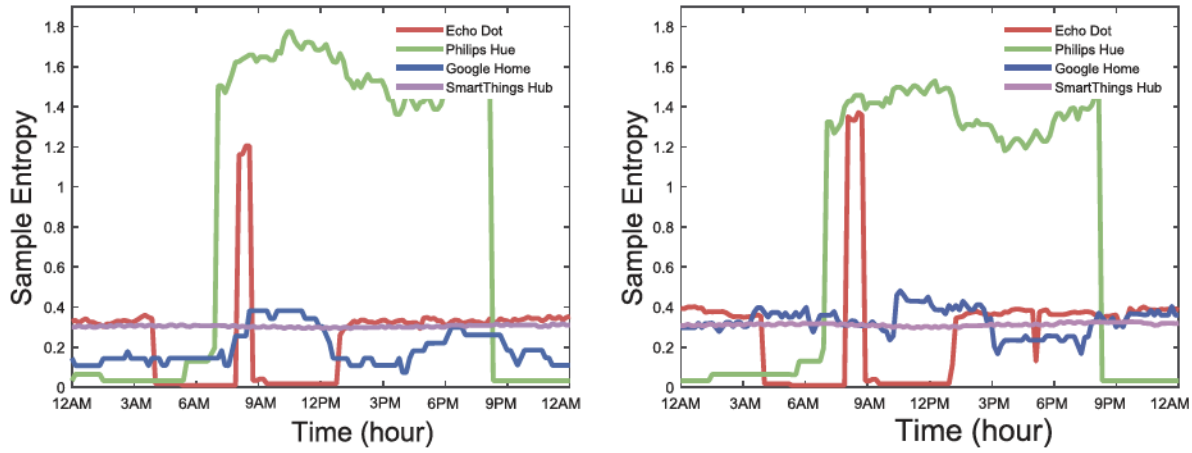
Applying the sliding window approach, we can estimate the entropy values for all traffic features of IoT devices in edge networks. In our experiments, we set each observation time window as 10 minutes and the overall time period as 4 hours, and then calculate the sample entropy of the time series traffic data collected in the past four hours to balance the computational overhead and real-time responses to traffic fluctuations.

We set the parameters m as 2 and r as 0.2 times the standard deviation of time series $f(t_1), f(t_2), \ldots, f(t_M)$ [25]. Based on our experimental results, such parameter settings will best estimate the time series entropy and depict the traffic and activity patterns of IoT devices, which confirms the findings in [25]. Fig. 7 illustrates the distinct sample entropy measures on packet count of four different IoT devices in the same edge network. Specifically, the entropy of Echo Dot exhibits a spike at 9AM due to the music playing on Spotify and a preconfigured weather forecast service during this time period, while Philips Hue communicates with the cloud server actively during the daytime and remains only the heart-beat communications with servers at night. Compared with Echo Dot and Philips Hue, Google Home and SmartThings Hub have more stable entropy over time, as our in-depth analysis



(a) IoT device - Philips Hue (b) IoT device - Amazon Echo (c) Non-IoT device - Android Smartphone

Fig. 6. The autocorrelation plots of network traffic time series for selected IoT and non-IoT devices.



(a) Incoming traffic of selected IoT devices (b) Outgoing traffic of selected IoT devices

Fig. 7. Distinct sample entropy measures of network traffic by different IoT devices over time.

TABLE III
THE DOMINANT APPLICATIONS USED BY IOT DEVICES IN EDGE NETWORKS

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					

reveals that most of their network traffic are predicable short-term connections with NTP, DNS, and cloud servers. In other words, the simple yet effective sample entropy measure is able to capture, characterize, and distinguish the temporal dynamics and predictability of network traffic for IoT devices, thus potentially could help develop new event detection and intrusion prevention algorithms for monitoring and securing IoT devices in edge networks.

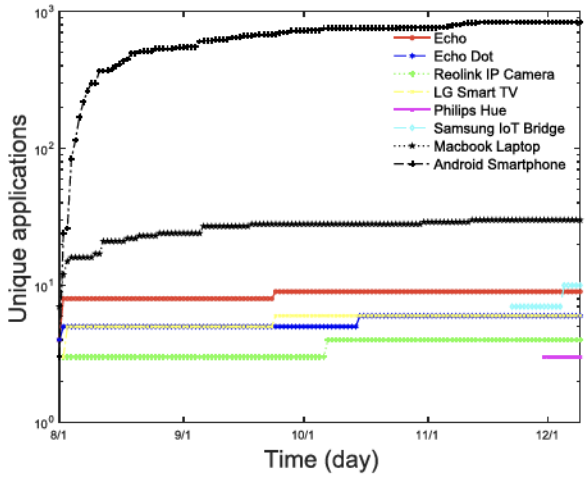


Fig. 8. The convergence of applications for IoT and non-IoT devices.

D. *Cloud Behavior of IoT Devices*

The objective of cloud behavior analysis is to understand why and how IoT devices communicate with remote cloud servers. Specifically, 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 III demonstrates 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, HTTPS, DNS, NTP, and Spotify music streaming. As a comparison, the Andriod Smartphone and the Macbook laptop in the same edge network engage with 11 and 15 distinct applications, respectively, during the same time period.

The limited and consistent set of common applications used by IoT devices again confirms that IoT devices are typically designed for very specific functions and dedicated utilities. Fig. 8 illustrates the convergence of cloud applications for IoT

TABLE IV
THE ENTROPY OF DESTINATION IP ADDRESSES, PREFIXES AND ASNS IOT DEVICES HAVE SENT HTTPS REQUESTS WITHIN A 24-HOUR TIME WINDOW

Device	Flows	Fanout			Normalized Entropy		
		IP	Prefix	ASN	IP	Prefix	ASN
Echo	148	20	6	1	0.5529	0.3158	0.0000
IP 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
LG 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

and non-IoT devices, where the number of applications for IoT devices converges rapidly.

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, for each application. In addition, we measure the distribution of network traffic across these remote servers, prefixes and ASNs by calculating the entropy and standardized entropy of these fanouts. For a given application a of an IoT device d, let F and R denote the number of network traffic flows and the *unique* numbers of the remote servers represented as $s_1, s_2, \ldots, s_R$. The probability of each remote server P_{s_i} is calculated as $P_{s_i} = \frac{C_{s_i}}{F}$, where C_{s_i} denotes the number of flows between d and s_i. Clearly $\sum_{i=1}^{R} C_{s_i} = F$. The normalized entropy on the remote servers for application a of device d is then derived as $\mathcal{NE}_{d,a} = -(\log R)^{-1} \sum_{i=1}^{R} P_{s_i} \times \log P_{s_i}$.

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, which means that this device only communicates with one or few servers on application a. While a $\mathcal{NE}_{d,a}$ value of 1 or near 1 means the high randomness on the remote servers. Following a similar process, we could also calculate the entropies and normalized entropies for their aggregated network prefixes or ASNs of remote servers. Table IV illustrates the entropy values of destination IP addresses, prefixes and ASNs of the hosts which IoT devices have sent HTTPS requests to within a 24-hour time window. As shown in Table IV, all IoT devices exhibit some 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 or classifying the newly added IoT devices in edge networks.

E. IoT Traffic Behaviors Within Edge Networks

To gain a comprehensive understanding of traffic behaviors of IoT devices, we further investigate the internal LAN traffic within edge networks which originates from IoT devices and destines to other IoT devices or non-IoT systems in the same edge network or vice versa. Based on the end systems involved in the traffic, we classify IoT traffic within edge networks into three categories: *1)* communication between IoT devices and edge routers, *2)* communication among IoT devices, and *3)* communication between IoT and Non-IoT devices.

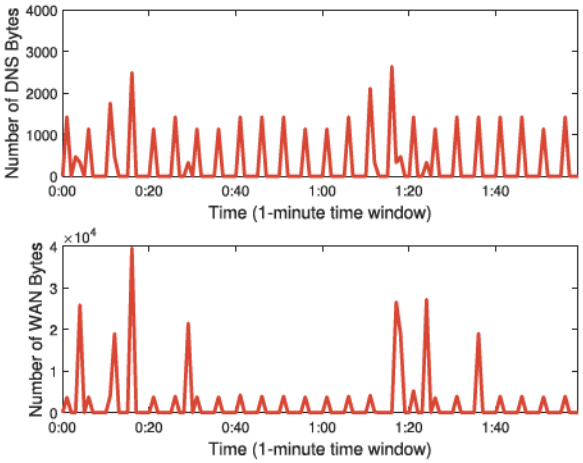


Fig. 9. The correlation of internal DNS traffic and incoming WAN data traffic for an Amazon Echo device.

1) Communication Between IoT Devices and Edge Routers: Based on our longitudinal measurement study of 22 edge networks, we have observed two dominant applications in this category - dynamic host configuration protocol (DHCP) and DNS. All of the IoT devices exchange the DHCP information with their respective edge routers periodically via UDP ports 67 and 68. This observation is consistent with the common DHCP configurations on today's home routers, which act as DHCP servers and automatically assign the IP address to all of the devices in edge networks. After the lease time is over, home edge routers will renew it if the IoT device is still active. It is interesting to note that all of the IoT devices in our study have their IP addresses renewed every 12 hours by exchanging DHCP packets with the edge routers, which indicates that the DHCP lease time is set as 12 hours on the router side.

Similar to many non-IoT devices such as laptops and smartphones, the majority of IoT devices leave the choice of DNS severs to the edge routers for the considerations of short DNS query and reply latency. Routers usually set themselves as the DNS server and adding their IP addresses in the "DNS servers" field of the DHCP Offer packets. Among the 20 types of IoT devices in our study, Google Home is the only device that prefers external DNS services, i.e., Google's own public IPv4 DNS servers `8.8.8.8` and `8.8.4.4` over the edge router-based DNS services. As DNS traffic is often triggered by many IoT applications such as HTTPS and NTP for retrieving IP addresses of the cloud servers, there tends to exist strong correlations between the internal DNS traffic of IoT devices and the actual incoming wide area networks (WAN) network traffic of these devices. Fig. 9 illustrates the byte count of DNS query/reply packets between Amazon Echo Dot and the home router (the top plot) as well as the total number of bytes exchanged between the Echo Dot and cloud servers (the bottom plot) over a 2-hour time window. As shown in Fig. 9, the two traffic measures exhibit strong temporal and structural correlations. This observation confirms the importance of hardening the security of edge routers that act as DNS servers as well as all other non-IoT devices, since the IoT

StartTime	Duration	SrcIP	DstIP	SrcPort	DstPort	Protocol	PacketSize
18:09:50.223	0.059s	192.168.1.216	192.168.1.195	59337	80	HTTP	678
18:09:50.223	0.059s	192.168.1.195	192.168.1.216	80	59337	HTTP	2634
18:09:50.226	0.075s	192.168.1.216	192.168.1.195	59338	80	HTTP	574
18:09:50.226	0.075s	192.168.1.195	192.168.1.216	80	59338	HTTP	2542
18:09:50.404	0.065s	192.168.1.216	192.168.1.195	59339	80	HTTP	750
18:09:50.404	0.065s	192.168.1.195	192.168.1.216	80	59339	HTTP	1650

Fig. 10. First 6 network flows captured when sending an "open" command from Amazon Echo Dot (192.168.1.216) to Philips Hue Bridge (192.168.1.195). The exact dates are removed for privacy concerns.

devices will be vulnerable to the DNS spoofing attack in edge networks [39] if these edge routers or non-IoT devices are compromised.

2) Communication Among IoT Devices: The data communication among IoT devices happens primarily during the *operation* stage between paired IoT devices in the same edge network such as an Amazon Echo and a Philips Hue bridge. Due to the improved network latency and simplified management and operations, different IoT devices in the same edge network can be "paired" with each other for better communication and cooperation. We notice that during the initial pairing stage, all IoT devices contact their respective vendors' cloud servers for authentication and registration. After a paired relationship is established, many of the paired devices continue to rely on the cloud servers as a proxy to communicate with each other for security and trust considerations.

We also noticed one pair of IoT devices, Amazon Echo Dot and Philips Hue bridge, directly talking with each other using the HTTP protocol, as illustrated in Fig. 10, after the Philips Hue bridge is added into the trusted device list on the Amazon Echo Dot. These packets are captured by the router when we press the "open" button in the Amazon Alexa App. The direct internal communication significantly improves the efficiency and latency of operating the Philips Hue bulbs via controlling Amazon Echo Dot in the same edge network, since the commands are not required to transfer through the long-latency path from mobile Apps on smartphones, cloud servers, the Philips Hue bridge, to the light bulbs. On the other hand, the direct communication using the insecure HTTP protocol could potentially leave both IoT devices vulnerable to attacks. Therefore, whether retaining the cloud servers as a communication proxy is a system design trade-off between security and efficiency.

3) Communication Between IoT and Non-IoT Devices: We discover three communication protocols, multicast DNS (mDNS), simple service discovery protocol (SSDP), and HTTP/HTTPS in this category. Many Apple devices, Linux-based networked systems, and Windows computers with Apple iTunes all periodically broadcast multicast DNS (mDNS) packets to identify and resolve the IP addresses of other devices in the same edge network. In our study, Philips Hue bridge is the only IoT device leveraging mDNS protocol to identify itself via replying mDNS queries but many IoT devices broadcast mDNS packets in order to find other devices.

The SSDP protocol is designed for the advertisement and discovery of network services and device existence. Many IoT devices adopt SSDP protocol to bootstrap the device discovery services. For example, Samsung SmartThings hub broadcasts SSDP messages whenever a user tries to pair a new IoT device

```
1 M-SEARCH * HTTP/1.1
2 HOST: 239.255.255.250:1900
3 MAN: "ssdp:discover"
4 MX: 4
5 ST: urn:schemas-upnp-org:device:basic:1
```

Fig. 11. An example of SSDP requests sent by SmartThings Hub.

to the hub using the SmartThings App on a smartphone. Fig. 11 shows an SSDP message sent from the Samsung SmartThings hub when the hub is requested to pair with a Philips Hue bridge. The Mandatory Extensions in HTTP (MAN) in Fig. 11 defines the scope of the extension and carries the value of "ssdp:discover" to indicate a device search request, and the maximum wait time in seconds (MX) is used for load balance when the hub processes the SSDP responses. Search target (ST) is in the format of `urn:schemas-upnp-org:device:DeviceType:version` in the case of searching for a particular device, specified by the device type and version.

The corresponding device type and version of Philips Hue bridge is `basic` and `1`,respectively, as included in the SSDP response messages. However we notice that the Samsung SmartThings hub is actually enumerating all the device types and versions by sending out different SSDP requests, which explain why our IoT measurement framework captures a large number of SSDP network flows during every device paring process. These SSDP requests also flood in the Wi-Fi networks even if the selected device pairs with the hub using other wireless communication protocols such as ZigBee and Z-Wave. In other words, the drive-by attackers, if receiving the broadcast SSDP packets sent by the hub, could pair with and potentially compromise the corresponding IoT devices. *These findings confirm the design flaws in the paring stage of these wireless protocols reported in the prior research in* [22], [27], [28], [40]. Our study also discovers an interesting behavior of Philips Hue bridge which proactively sends out the SSDP packets targeting a Windows PC in the same edge network every two minutes. Such unique traffic pattern could help effectively detect and distinguish this type of IoT devices.

The third type of communication protocol between IoT and non-IoT devices is HTTP/HTTPS, which is used by smartphones to directly communicate with a variety of IoT devices in the same edge network. Many IoT devices are controlled, configured and monitored by smartphone-based apps on Android or Apple iOS platforms, and these devices e.g. Philips Hue Bridge, Google Home and Reolink Camera, often allow the smartphones to directly communicate with them for reduced network latency using HTTP and HTTPS protocols if

and only if the device has been registered in the corresponding application on the smartphone and the IoT device and smartphone are in the same edge network. On the other hand, some IoT devices such as Echo Dot, SmartThings Hub, and Ring Video Doorbell strictly require that all data packets of command and control must first go through the trusted cloud servers and then be forwarded to the devices for security reasons.

V. Exploring the Applications of Multidimensional Behavioral Profiling

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

A. 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, and weak password management. As cyber attacks exploring the weakly protected IoT devices often leave substantial traffic footprints in edge networks, it is intuitive to explore multidimensional behavioral profiles to detect anomaly traffic and security threats.

In this study, we adopt an anomaly detection method based on minimum description length (MDL) principle because of its data-driven approach and parameter-free feature [1], [12], [18]. The intuition and novelty of the MDL principle lie in the pattern-based compression and encoding techniques which exploit coding tables to capture the underlying data distributions. In other words, this technique encodes a frequent and common pattern with a short encoded length, while a long encoded length reflects anomalies and irregularities in the original data [1]. The MDL principle essentially is a model selection framework for performing lossless compression and encoding on data with categorical features and attributes. The main process is to search and identify the best model e which minimizes the overall encoding size for the entire data, i.e.,

$$\arg\min_{e \in \mathcal{E}} \; L(e) + L(u \mid e), \tag{3}$$

where $\mathcal{E}$ is the model set and $L(e)$, $L(u \mid e)$ are the bit length describing the specific model e and the bit length of describing the data u with the model e, 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 l flow records, each of which has w categorical features, i.e., $\mathcal{F} = \{h_1, \ldots, h_w\}$. 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 $\mathcal{C}$. 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 the pattern p. Thus based on the entropy theory, the optimal coding length for the pattern p is

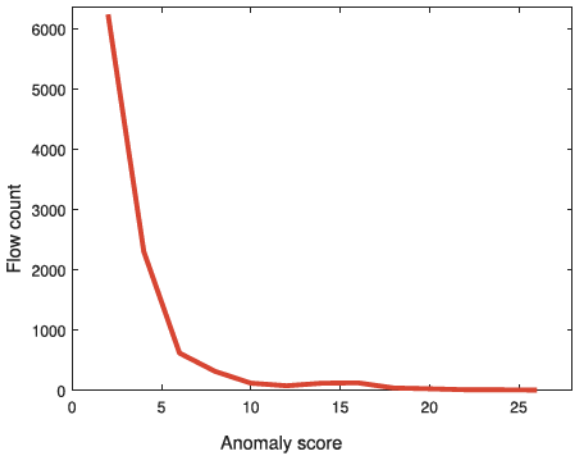


Fig. 12. The distribution of anomaly scores for all observed network traffic flows during a 24-hour time window for a smart voice assistant.

$$L(c(p) \mid CT) = -\log\left(\frac{\text{freq}(p)}{\sum_{p' \in CT} \text{freq}(p')}\right). \tag{4}$$

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

$$\begin{aligned} L(D \mid CT) &= \sum_{u \in D} L(u \mid CT) \\ &= \sum_{u \in D} \sum_{p \in \mathcal{N}(u)} L(c(p) \mid CT), \end{aligned} \tag{5}$$

where $\mathcal{N}(u)$ is the set of features which are used when encoding u. As shown in Eq. (6), the bit length of encoding the overall data is then calculated as:

$$L(CT) = \sum_{p \in CT} L(c(p) \mid CT) - \sum_{v \in \mathcal{V}} o_v \log(q_v), \tag{6}$$

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 $v \in \mathcal{V}$; q_v equals to o_v divided by the total length of all the patterns in the code table. Combining the entire feature set together, we can build multiple code tables to further reduce the overall encoding cost considering there may be correlation between different features.

This 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 $L(u \mid CT)$ for a network flow record u as the anomaly score.

Fig. 12 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.

TABLE V
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

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 V summarizes our in-depth analysis of all 526 network flows with high anomaly scores. As shown in Table V, most of these network flows with high anomaly scores are long HTTPS connections between the smart voice assistant with Google cloud servers. Thousands of normal network traffic flows for the smart speaker are mostly periodical DNS queries and responses as well as short TCP/UDP data transfers. In addition, a small number of network flows are related to ICMP, mDNS, and DHCP protocols, which are corresponding to the network management and broadcast/multicast traffic. Although all of these network activities are benign in nature, our validation results confirm the ability and potential of our proposed pattern-based anomaly detection approach for discovering unusual and anomaly behaviors based on the multidimensional behavioral profiles of IoT devices.

B. IoT Device Detection and Classification

Our detailed analysis of IoT devices' behaviors also provides unique and valuable 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 traffic feature in behavioral profiles over a given time window, we can quantify and measure the similarities and correlations of this feature between i and j during the same time period. Assuming feature z is the remote destination IP addresses (`dstIPs`) that IoT devices communicate with. Let $\mathcal{A}_{i,z}$ and $\mathcal{A}_{j,z}$ 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., $S_{i,j,z}$, is calculated as

$$S_{i,j,z} = \frac{|\mathcal{A}_{i,z} \cap \mathcal{A}_{j,z}|}{|\mathcal{A}_{i,z} \cup \mathcal{A}_{j,z}|}. \quad (7)$$

Thus repeating the same process on all 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. This similarity matrix enables us to identify and cluster devices with similar behavioral fingerprints, and more importantly detect new suspicious IoT devices in the same edge network.

Fig. 13 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 the two networks. As shown in Fig. 13, most pairs of IoT devices exhibit low similarities, suggesting that IoT devices communicate 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 cloud 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 phrase after they join the edge networks.

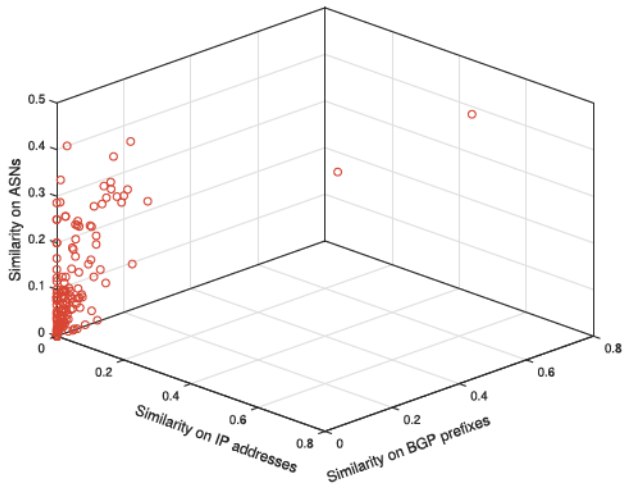


Fig. 13. The scatter plot of similarity score on IP-spatial features.

Several recent studies have explored machine learning techniques for IoT device detection and classification [13], [20]. But the multidimensional behavioral profiles we build from real world data could still provide additional features and unique insights for improving the quality and performance of these machine learning-based IoT device detection and classification.

C. Network Security Monitoring

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

To demonstrate the feasibility of our network security monitoring application, we simulate all the critical steps of Mirai botnet [4], [16] for infiltrating, infecting, and operating weakly protected IP cameras in a controlled edge network environment. We demonstrate that the behavioral fingerprints left by Mirai botnet 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. If successful, Mirai subsequently attempts to launch a dictionary attack to attempt the logins with 62 weak and widely used credentials (e.g. root:admin). Obviously the scanning activity and dictionary attack trigger substantial behavioral footprint deviations on the IP-spatial and application dimensions, since the IP address of the remote attacker is from an unusual network prefix and ASN, and the remote ports used in the scanning are rarely used in general. This infection stage also leaves unique behavioral fingerprints from IP-spatial, data volume, and application dimensions, as the loader, which could be different from the initial scanner, has to transfer the malware image to the compromised IP camera.

After being compromised, the IP camera now becomes a part of the IoT botnet and exhibits very unusual attacking behaviors. This compromised device starts to perform the aforementioned port scanning operations in order to infect more devices, and the device has to periodically communicate with control and command (C2) servers of the botnet. Eventually the device is directed to launch coordinated distributed denial-of-service attacks (DDoS) towards targets such as Dyn DNS infrastructure [4] with commands from C2. All of these malicious network activities generated by the IP camera leaves significant deviations on the behavioral fingerprints, thus our proposed multidimensional behavioral profiling framework for IoT devices could effectively detect, mitigate, and stop such attacks.

VI. Conclusions and Future Work

As the wide adoption of IoT devices continues 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 patterns 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, entropy, and cloud dimensions. We also leverage the benefits of our programmable router based scheme to take a deep look into the LAN network patterns of different IoT devices.

Based on real network traffic data collected from 22 edge networks over one-year time span, we have discovered a number of important and interesting findings. We notice that IoT devices typically communicate with cloud servers from a very small number of prefixes and ASNs, which belong to IoT manufactures, the cloud service providers, NTP service providers, and public DNS service providers. IoT devices also often exhibit repeated and predictable traffic activities over time due to heartbeat signals between IoT devices and cloud servers. Unlike laptops, desktops, or 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 characterize 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 will be 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 complement the current behavioral fingerprinting framework based on traffic features collected from network, transport, and application layers, which could collectively provide critical input for designing next-generation IoT security monitoring and threat prevention systems.

References

[1] L. Akoglu, H. Tong, J. Vreeken, and C. Faloutsos, "Fast and reliable anomaly detection in categorical data," in *Proc. ACM CIKM*, 2012, pp. 415–424

[2] O. Alrawi, C. Lever, M. Antonakakis, and F. Monrose, "SoK: Security evaluation of home-based IoT deployments," in *Proc. IEEE S&P*, 2019, pp. 1362–1380.

[3] Amazon, "Alexa Skills Kit," 2020. [Online]. Available: https://developer.amazon.com/alexa-skills-kit/

[4] M. Antonakakis *et al.*, "Understanding the Mirai Botnet," in *Proc. USENIX Secur.*, 2017, pp. 1093–1110.

[5] B. Bezawada, M. Bachani, J. Peterson, H. Shirazi, I. Ray, and I. Ray, "Behavioral fingerprinting of IoT devices," in *Proc. ACM ASHES*, 2018, pp. 41–50.

[6] Business Insider, "'Google, This is bogus as hell' - one of the fathers of the internet blasts google for how chromecast behaves on his home network," 2019. [Online]. Available: https://www.businessinsider.com/paul-vixie-blasts-google-chromecast-2019-2/

[7] Z.-B. Celik, L. Babun, A.-K. Sikder, H. Aksu, G. Tan, P. McDaniel, and A.-S. Uluagac, "Sensitive information tracking in commodity IoT," in *Proc. USENIX Secur.*, 2018, pp. 1687–1704.

[8] S. Feng, P. Setoodeh, and S. Haykin, "Smart home: Cognitive interactive people-centric internet of things," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 34–39, Feb. 2017.

[9] E. Fernandes, J. Jung, and A. Prakash, "Security analysis of emerging smart home applications," in *Proc. IEEE S&P*, 2016, pp. 636–654.

[10] K. Gao, C. Corbett, and R. Beyah, "A passive approach to wireless device fingerprinting," in *Proc. IEEE/IFIP DSN*, 2010, pp. 383–392.

[11] Google, "Nest, create a connected home," 2020. [Online]. Available: https://nest.com/

[12] P. Grünwald, *The Minimum Description Length Principle*. Cambridge, MA, USA: MIT press, 2007.

[13] T. Gu and P. Mohapatra, "BF-IoT: Securing the IoT networks via fingerprinting-based device authentication," in *Proc. IEEE MASS*, 2018, pp. 254–262.

[14] G. Ho, D. Leung, P. Mishra, A. Hosseini, D. Song, and D. Wagner, "Smart locks: Lessons for securing commodity internet of things devices," in *Proc. ACM ASIACCS*, 2016, pp. 461–472.

[15] H. Jafari, O. Omotere, D. Adesina, H.-H. Wu, and L. Qian, "IoT devices fingerprinting using deep learning," in *Proc. IEEE MILCOM*, 2018, pp. 1–9.

[16] G. Kambourakis, C. Kolias, and A. Stavrou, "The mirai botnet and the iot zombie armies," in *Proc. IEEE MILCOM*, 2017, pp. 267–272.

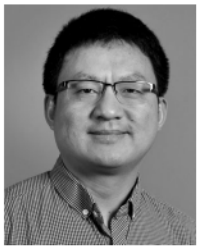
[17] D. Kumar *et al.*, "All things considered: An analysis of IoT devices on home networks," in *Proc. USENIX Secur.*, 2019, pp. 1169–1185.

[18] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*. Berlin, Germany: Springer, 1993.

[19] Y. Meidan *et al.*, "ProfilIoT: A machine learning approach for IoT device identification based on network traffic analysis," in *Proc. ACM Symp. Appl. Comput.*, Apr. 2017, pp. 506–509.

[20] M. Miettinen, S. Marchal, I. Hafeez, N. Asokan, A.-R. Sadeghi, and S. Tarkoma, "IoT SENTINEL: Automated device-type identification for security enforcement in IoT," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst.*, 2017, pp. 2177–2184.
[21] H.-M. Moghaddam *et al.*, "Watching you watch: The tracking ecosystem of over-the-top tv streaming devices," in *Proc. ACM CCS*, 2019, pp. 131–147.
[22] P. Morgner, S. Mattejat, Z. Benenson, C. Muller, and F. Armknecht, "Insecure to the touch: Attacking zigbee 3.0 via touchlink commissioning," in *Proc. ACM WiSec*, 2017, pp. 230–240.
[23] K. Park and W. Willinger, *Self-Similar Network Traffic and Performance Evaluation*. Hoboken, NJ, USA: Wiley, 2002.
[24] Y. Rekhter, S. Hares and T. Li, "A border gateway protocol 4 (BGP-4)," Internet Soc., Reston, VA, USA, RFC 4271, Jan. 2006. [Online]. Available: https://rfc-editor.org/rfc/rfc4271.txt
[25] J.-S. Richman, and J.-R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Amer. J. Physiology-Heart Circulatory Physiol.*, vol. 278, no. 6, pp. H2039–H2049, 2000.
[26] J. Riihijarvi, M. Wellens, and P. Mahonen, "Measuring complexity and predictability in networks with multiscale entropy analysis," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 1107–1115.
[27] E. Ronen, C. O'Flynn, A. Shamir, and A.-O. Weingarten, "IoT goes nuclear: Creating a zigbee chain reaction," in *Proc. IEEE S&P*, 2017, pp. 195–212.
[28] M. Ryan, "Bluetooth smart: The good, the bad, the ugly, and the fix," 2013. [Online]. Available: https://lacklustre.net/bluetooth/bluetooth_smart_good_bad_ugly_fix-mikeryan-blackhat_2013.pdf
[29] Samsung, "Smartthings, add a little smartness to your things," 2020. [Online]. Available: https://www.smartthings.com/
[30] S. Siby, R. Maiti, and N. Tippenhauer, "IoTScanner: detecting privacy threats in IoT neighborhoods," in *Proc. ACM IoTPTS*, 2017, pp. 23–30.
[31] M. Trevisan, D. Giordano, I. Drago, M. Mellia, and M. Munafo, "Five years at the edge: Watching internet from the ISP network," in *Proc. CoNEXT*, 2018, pp. 561–574.
[32] University of Oregon, "Route views project," 2020. [Online]. Available: http://www.routeviews.org/
[33] K. Xu, Y. Wan, and G. Xue, "Powering smart homes with information-centric networking," *IEEE Commun. Mag.*, vol. 57, no. 6, pp. 40–46, Jun. 2019.
[34] K. Xu, Y. Wan, G. Xue, and F. Wang, "Multidimensional behavioral profiling of internet-of-things in edge networks," in *Proc. IEEE/ACM IWQoS*, 2019, pp. 1–10.
[35] K. Xu, F. Wang, and X. Jia, "Secure the internet, one home at a time," *Secur. Commun. Netw.*, vol. 9, no. 16, pp. 3821–3832, Nov. 2016.
[36] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet Things J.*, vol. 1, no. 1, pp. 22–32, Feb. 2014.
[37] J. Zhang, X. Chen, Y. Xiang, W. Zhou, and J. Wu, "Robust network traffic classification," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1257–1270, Aug. 2015.
[38] W. Zhang, Y. Meng, Y. Liu, X. Zhang, Y. Zhang, and H. Zhu, "HoMonit: Monitoring smart home apps from encrypted traffic," in *Proc. ACM CCS*, 2018, pp. 1074–1088.
[39] L. Zhu, Z. Hu, J. Heidemann, D. Wessels, A. Mankin, and N. Somaiya, "Connection-oriented DNS to improve privacy and security," in *Proc. IEEE S&P*, 2015, pp. 171–186.
[40] T. Zillner, "Zigbee exploited: The good, the bad and the ugly," in *Proc. DeepSec Conf. Depth Secur.*, 2017, pp. 251–260.

Yinxin Wan (Student Member, IEEE) received the B.E degree in information security from the University of Science and Technology of China, in 2018. He is currently the Ph.D. student of Computer Science with Arizona State University. His research interests include cyber security, the Internet of Things, and data-driven networked system.



Kuai Xu (Senior Member, IEEE) received the B.S. and M.S. degrees in computer science from Peking University, China, in 1998 and 2001, respectively and received the Ph.D. degree in computer science from the University of Minnesota, in 2006. He is an Associate Professor with Arizona State University. His research interests include network security, Internet measurement, big data, data mining, and machine learning. He is a member of ACM.

Feng Wang (Member, IEEE) received the B.S. degree from Wuhan University in 1996, M.S. degree from Peking University in 1999, and the Ph.D. degree from the University of Minnesota, Twin Cities, in 2005, all in computer science. She is currently a Professor with School of Mathematical and Natural Sciences, Arizona State University. She has authored or coauthored more than 60 journal and conference papers and book chapters. Her research interests focus on network science, social media analysis, network optimization, network security, and wireless sensor networks. Her research is supported by National Science Foundation.

Guoliang Xue (Fellow, IEEE) received the Ph.D. degree in computer science from the University of Minnesota, in 1991. He is a Professor of Computer Science and Engineering with Arizona State University. He has authored or coauthored more than 300 papers in these areas, many of which in top conferences such as INFOCOM, MOBICOM, NDSS and top journals such as IEEE/ACM ToN, IEEE JSAC, IEEE TDSC, and IEEE TMC. His research interests span the areas of QoS provisioning, machine learning, wireless networking, network security and privacy, crowdsourcing and network economics, Internet of Things, smart city and smart grids. He has received the IEEE Communications Society William R. Bennett Prize in 2019 (Best Paper Award for IEEE/ACM TON and IEEE TNSM in the previous three years). He was a keynote speaker at IEEE LCN'2011 and ICNC'2014. He was a TPC Co-Chair of IEEE INFOCOM'2010 and a General Co-Chair of IEEE CNS'2014. He has served on the TPC of many conferences, including ACM CCS, ACM MOBIHOC, IEEE ICNP, and IEEE INFOCOM. He served on the editorial board of IEEE/ACM TRANSACTIONS ON NETWORKING and the Area Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, overseeing 13 editors in the Wireless Networking area. He is the Steering Committee Chair of IEEE INFOCOM.