2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS): SmartCity17: The 3rd IEEE INFOCOM Workshop on Smart Cities and Urban Computing

Characterizing and Classifying IoT Traffic in Smart Cities and Campuses

Arunan Sivanathan*, Daniel Sherratt*, Hassan Habibi Gharakheili*,
Adam Radford†, Chamith Wijenayake*, Arun Vishwanath‡ and Vijay Sivaraman*
*Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia.
†Cisco Systems, Sydney, Australia. ‡IBM Research, Melbourne, Australia.
Emails: {a.sivanathan, d.sherratt}@student.unsw.edu.au, h.habibi@unsw.edu.au, aradford@cisco.com,
c.wijenayake@unsw.edu.au, arvishwa@au1.ibm.com, vijay@unsw.edu.au

***Abstract*—Campuses and cities of the near future will be equipped with vast numbers of IoT devices. Operators of such environments may not even be fully aware of their IoT assets, let alone whether each IoT device is functioning properly safe from cyber-attacks. This paper proposes the use of network traffic analytics to characterize IoT devices, including their typical behaviour mode. We first collect and synthesize traffic traces from a smart-campus environment instrumented with a diversity of IoT devices including cameras, lights, appliances, and health-monitors; our traces, collected over a period of 3 weeks, are released as open data to the public. We then analyze the traffic traces to characterize statistical attributes such as data rates and burstiness, activity cycles, and signalling patterns, for over 20 IoT devices deployed in our environment. Finally, using these attributes, we develop a classification method that can not only distinguish IoT from non-IoT traffic, but also identify specific IoT devices with over 95% accuracy. Our study empowers operators of smart cities and campuses to discover and monitor their IoT assets based on their network behaviour.**

I. Introduction

The Internet of Things (IoT), comprising everyday objects such as lights, cameras, motion sensors, power switches and appliances, is heralded to bring the next wave of Internet growth. Cisco predicts that IoT connections will reach 12.2 billion by 2020 [1], representing nearly half of all connected devices. Homes, enterprises, campuses and cities are expected to be instrumented with thousands of "smart" IoT devices that can autonomously interact with each other and be remotely monitored/controlled.

This rapid growth in scale creates an operational challenge – knowing what IoT devices are connected and whether they are functioning normally can become difficult for the administrator. This might arise from different departments being involved in asset management. For example, sensors for lighting may be installed by the local council, sewage and garbage sensors by the sanitation department and cameras by the local police division. Consolidating IoT assets from these various departments may be onerous or error-prone, making it difficult to ascertain what IoT devices are operating on the network at any point in time. This lack of "visibility" into IoT devices can make it very complex for the administrator to trouble-shoot problems in their smart-city/campus infrastructure, and can become particularly disastrous when cyber-security attacks have breached this critical infrastructure.

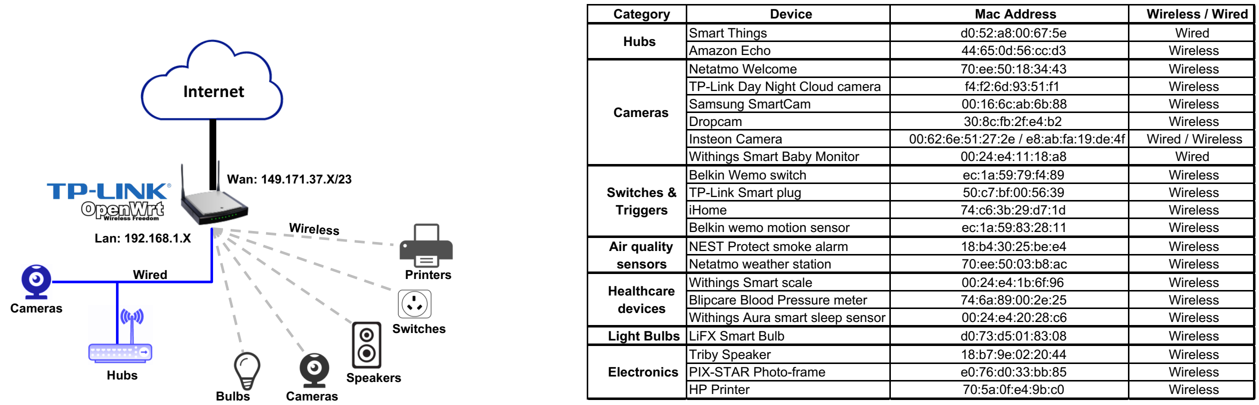
This paper addresses the above problem by characterizing IoT traffic at the network-level, and using this to identify and classify IoT devices, alongside detecting anomalous behaviour. Qualitatively, we expect most IoT devices to send short bursts of data sporadically. However, there has been no quantitative study in the literature to profile how much traffic they send in a burst, how long they idle between bursts, and whether these patterns are periodic or not. We also lack understanding on how much signalling they perform (e.g. DNS lookups or time synchronization using NTP) in comparison to the data traffic they generate, or how much multicast/broadcast traffic they generate, needed for service discovery. No study has identified how these aspects vary from one IoT device to another, performing different functions (e.g. camera compared to a smoke-alarm) or similar functions (one brand of camera versus another). To our knowledge there are no openly available IoT traffic traces that researchers can use to study these questions.

Understanding the nature of IoT traffic is important for several reasons: operators of smart-cities and campuses need to support appropriate performance levels of reliability, loss, and latency (needed by environmental, health, or safety applications), while also containing IoT multicast/broadcast discovery traffic that can impact other applications. The most compelling reason for profiling IoT traffic is to enhance cyber-security: it is well recognized that IoT devices are by their nature easier to infiltrate [2], and every month new stories emerge of how IoT devices have been compromised and used to launch large-scale attacks [3]. The large heterogeneity in IoT devices has led researchers to propose network-level security mechanisms that analyse traffic patterns to identify attacks (see [4] and our recent work [5]); success of these approaches relies on a good understanding of what "normal" IoT traffic profile looks like.

This paper fills the gap in the literature relating to availability of IoT traffic traces, characterization of IoT traffic profiles, and classification of IoT devices based on their profiles. We instrument a campus environment with over 20 IoT devices, comprising cameras, lights, activity sensors, health and well-being monitors, and consumer electronics. Our first contribution is to collect data traces from this environment

Funding for this project was provided by the Australian Research Council (ARC) Linkage Grant LP150100666.

978-1-5386-2784-6/17/$31.00 ©2017 IEEE



(a) Experimental setup depicting our smart environment.

Category	Device	Mac Address	Wireless / Wired
Hubs	Smart Things	d0:52:a8:00:67:5e	Wired
	Amazon Echo	44:65:0d:56:cc:d3	Wireless
Cameras	Netatmo Welcome	70:ee:50:18:34:43	Wireless
	TP-Link Day Night Cloud camera	f4:f2:6d:93:51:f1	Wireless
	Samsung SmartCam	00:16:6c:ab:6b:88	Wireless
	Dropcam	30:8c:fb:2f:e4:b2	Wireless
	Insteon Camera	00:62:6e:51:27:2e / e8:ab:fa:19:de:4f	Wired / Wireless
	Withings Smart Baby Monitor	00:24:e4:11:18:a8	Wired
Switches & Triggers	Belkin Wemo switch	ec:1a:59:79:f4:89	Wireless
	TP-Link Smart plug	50:c7:bf:00:56:39	Wireless
	iHome	74:c6:3b:29:d7:1d	Wireless
	Belkin wemo motion sensor	ec:1a:59:83:28:11	Wireless
Air quality sensors	NEST Protect smoke alarm	18:b4:30:25:be:e4	Wireless
	Netatmo weather station	70:ee:50:03:b8:ac	Wireless
Healthcare devices	Withings Smart scale	00:24:e4:1b:6f:96	Wireless
	Blipcare Blood Pressure meter	74:6a:89:00:2e:25	Wireless
	Withings Aura smart sleep sensor	00:24:e4:20:28:c6	Wireless
Light Bulbs	LiFX Smart Bulb	d0:73:d5:01:83:08	Wireless
Electronics	Triby Speaker	18:b7:9e:02:20:44	Wireless
	PIX-STAR Photo-frame	e0:76:d0:33:bb:85	Wireless
	HP Printer	70:5a:0f:e4:9b:c0	Wireless

(b) List of IoT devices in the smart environment.

Fig. 1. (a) Testbed showing the IoT devices and gateway, (b) Break down of IoT devices into different categories.

over a period of 3 weeks, and to make these openly available to the research community. These traces include raw packets (pcap) and flow information, annotated with specific device attributes, providing researchers a rich data-set to investigate many aspects of IoT. Our second contribution is to characterize the traffic corresponding to the various IoT devices, in terms of their activity pattern (traffic rate, burstiness, idle durations) and signalling overheads (broadcasts, DNS, NTP). Our last contribution is to develop a classification technique that learns the behaviour of an IoT device and is able to identify it based on its traffic profile. The data and characterization presented in this paper lay the foundation for enhancing visibility into IoT devices in a smart-city/campus network, upon which future works on IoT security and performance can be built.

The rest of this paper is organised as follows: §II describes relevant prior work. We present our IoT setup and data traces in §III. The traffic attributes are characterized in §IV, and in §IV-E we develop a classification method to identify IoT devices. The paper is concluded in §V.

II. Related Work

There is a large body of work characterizing general Internet traffic. However, studies focusing on characterizing IoT traffic (also referred to as machine-to-machine – M2M – traffic) are still in its infancy. The work in [6] is one of the first large-scale studies to delve into the nature of M2M traffic. It is motivated by the need to understand whether M2M traffic imposes new challenges for the design and management of cellular networks. The work uses a traffic trace spanning one week from a tier-1 cellular network operator and compares M2M traffic with traditional smartphone traffic from a number of different perspectives – temporal variations, mobility, network performance, and so on. The study informs network operators to be cognizant of these factors when managing their networks.

In [7], the authors note that the amount of traffic generated by a single M2M device is likely to be small, but the total traffic generated by hundreds or thousands of M2M devices would be substantial. These observations are to some extent corroborated by [8], [9], which note that a remote patient monitoring application is expected to generate about 0.35 MB per day and smart meters roughly 0.07 MB per day.

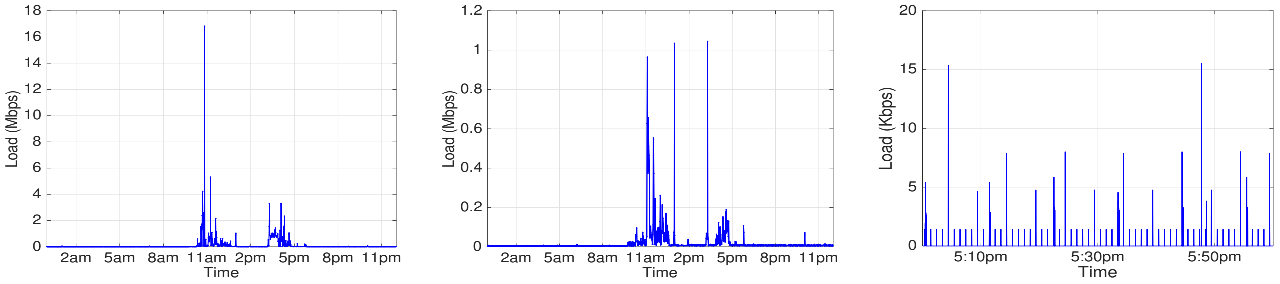
A Coupled Markov Modulated Poisson Processes framework to capture the behaviour of a single machine-type communication as well as the collective behaviour of tens of thousands of M2M devices is proposed in [10].

A simple model to estimate the volume of M2M traffic generated in a wireless sensor network enabled connected home is constructed in [11]. Since behaviour of sensors is very application specific, the work identifies certain common communication patterns that can be attributed to any sensor device. Using these attributes, four generalised equations are proposed to estimate the volume of traffic generated by a sensor network enabled connected apartment/home.

While all the above works make important contributions, they do not undertake fine-grained profiling and characterization of IoT traffic in a smart environment such as a city/campus. Furthermore, statistical models are not developed that enable IoT device classification based on their traffic profiles. Most importantly, prior works do not make any data set publicly available for the research community to use and build upon. Our work addresses these shortcomings.

III. IoT Data Collection

Our experimental setup – housed at our campus facility – comprises a wide range of IoT devices emulating a "smart environment", as depicted in Fig. 1(a). The TP Link Archer C7 v2, flashed with the OpenWrt firmware release Chaos Calmer (15.05.1, r48532), serves as the gateway to the public Internet. We also installed additional OpenWrt packages on the gateway, namely `tcpdump (4.5.1-4)` for capturing traffic, `bash (4.3.39-1)` for scripting, `block-mount` package for mounting external USB storage on the gateway, and `kmod-usb-core`, `kmod-usb-storage (3.18.23-1)` for storing the traffic data. As shown in Fig. 1(a), the WAN interface of the gateway is connected to the public Internet via the university network, while the IoT devices are connected to the LAN and WLAN interfaces respectively. Our smart environment has a total of 21 unique IoT devices representing different categories, see Fig. 1(b). These include cameras (Nest Dropcam, Samsung SmartCam, Netatmo Welcome, Insteon Camera, TP-Link Day Night Cloud Camera, Withings Smart Baby Monitor), switches and triggers (iHome, TP-Link Smart



(a) Load: IoT and non-IoT.

(b) Load: IoT only.

(c) Load: LiFX light bulb.

Fig. 2. Network load from IoT and non-IoT devices on a representative day.

Plug, Belkin Wemo Motion Sensor, Belkin Wemo Switch), hubs (Smart Things, Amazon Echo), air quality sensors (NEST Protect smoke alarm, Netatmo Weather station), electronics (Triby speaker, PIXSTAR Photoframe, HP Printer), healthcare devices (Withings Smart scale, Withings Aura smart sleep sensor, Blipcare blood pressure meter) and light bulbs (LiFX Smart Bulb). Several non-IoT devices were also connected to the testbed, such as laptops, mobile phones and an Android tablet. IoT devices were configured via apps, recommended by the device manufacturers, installed on the tablet.

All the traffic on the LAN side was collected using the `tcpdump` tool running on OpenWrt [12]. Capturing the pre-NAT traffic allowed us to map packets to specific devices directly; the MAC addresses in the packet headers reveal the identity of the devices (see column three in Fig. 1(b)). We developed a script to automate the process of data collection and storage. The resulting traces were stored as `pcap` files on an external hard drive of 1 TB storage attached to the gateway.

A. Trace Data

We started logging all network traffic in our smart environment from 23-Sep-2016. The process of data collection/storage begins at midnight local time each day using the `Cron` job on OpenWrt. We wrote a monitoring script on the OpenWrt to ensure that data collection/storage was proceeding smoothly. To make the trace data publicly available, we set up an Apache server on a virtual machine (VM) in our university data center and wrote a script to periodically transfer the trace data from the previous day, stored on the hard drive, onto the VM. The trace data is openly available for download at: http://149.171.189.1/. The size of the daily logs varies between 61 MB and 2 GB, with an average of 365 MB.

IV. Profiling and Characterizing the IoT Traffic

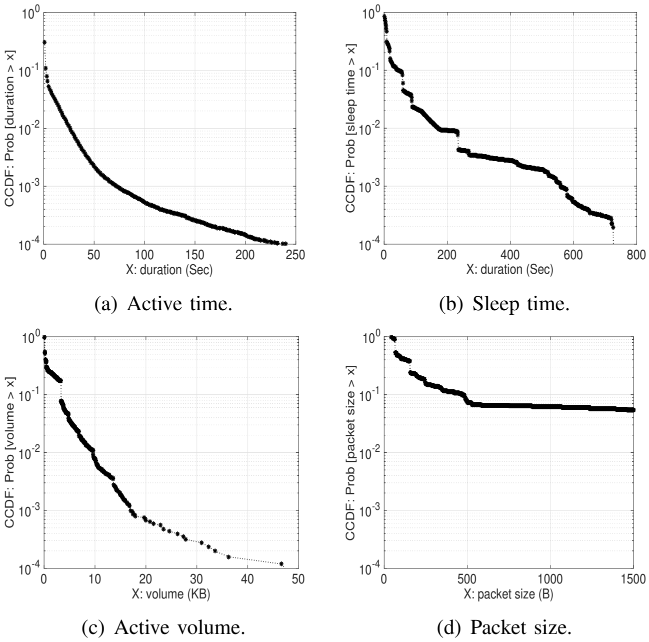
We now present our observations using passive packet-level analysis of traffic from 21 IoT devices over the course of two weeks (i.e. 23 Sep 2016 to 06 Oct 2016). We study a broad range of IoT traffic characteristics, including traffic load and signalling patterns, packet size distribution, dominant protocols used, and the distribution of active and sleep times.

A. IoT Activity

We first plot in Fig. 2(a) the total network load seen over a 24 hour period in our testbed on a representative day, 28 Sep 2016, chosen for illustrative purposes. The total load comprises traffic from all the non-IoT and IoT devices connected to the network. Non-IoT traffic comprises video and web traffic generated by the occupants of the environment on that day. IoT traffic constitute (i) traffic generated by the devices autonomously (e.g. DNS, NTP, etc. that are unaffected by human interaction), as well as (ii) traffic generated due to occupants interacting with the devices (e.g. Belkin Wemo sensor responding to detection of movement, Amazon Echo responding to voice commands issued by a user, LiFX light bulb changing colour and intensity upon user request, Netatmo Welcome camera detecting an occupant and instructing the LiFX light bulb to turn on with a specific colour, and so on).

We see from Fig. 2(a) that when there is activity from both IoT devices and non-IoT devices (e.g. between 10 am and 12 pm), the network load peaks at around 17 Mbps, while the average load is 400 Kbps. However, if we consider only the load imposed by the IoT devices, then there is a dramatic reduction in the peak load (1 Mbps) and average loads (66 Kbps), as depicted in Fig. 2(b), implying that traffic generated by IoT devices is small compared to traditional non-IoT traffic. If we zoom into the traffic pattern of one IoT device (for e.g. the LiFX light bulb) during a short interval, say one hour between 5 pm and 6 pm, a pattern of active/sleep communication emerges, as observed in Fig. 2(c). We therefore use the notion of active/sleep periods (defined as the duration over which an IoT device is generating traffic or remains idle, respectively) and the volume of traffic generated during active periods as attributes to capture the behaviour of IoT devices.

In Fig. 3(a) we plot the CCDF (Complementary Cumulative Distribution Function) of IoT active time and observe that it decays rapidly initially (only 5% of sessions last longer than 5 seconds), with the maximum active time being 250 seconds in our trace. This shows that IoT activities are short-lived in general. The CCDF in Fig. 3(b) shows that the IoT sleep duration (intervals during which no packet is exchanged) is less than 20 seconds 85% of the time, and only 4% of sleep times are longer than one minute, meaning IoT devices wake up very frequently and generate some network activity each time. The CCDF of IoT active volume is depicted in Fig. 3(c), and shows that more than 75% of IoT sessions transfer less than 1 KB, and only fewer than 1% of the sessions exchange more than 10 KB, suggesting that the majority of IoT devices generate only a small burst of traffic. Finally, we observe from Fig. 3(d) that IoT packet size decays slowly, with only about



(a) Active time.

(b) Sleep time.

(c) Active volume.

(d) Packet size.

Fig. 3. CCDF of IoT active/sleep time, active volume and packet size.

10% of packets being larger than 500 Bytes.

B. IoT Application Layer Protocols

We now focus on the application layer protocol (inferred using the destination port numbers) that IoT devices use to communicate locally in the LAN and/or externally with servers on the public Internet. Fig. 4 shows the probability histogram of destination port numbers for all IoT packets. It can be seen that HTTPS (i.e. TCP port 443) is the dominant protocol used by the IoT devices. Nevertheless, about 45% of IoT traffic (by number of packets) is not sent over HTTPS to the servers on the public Internet indicating that a sizeable fraction of IoT traffic is not being securely transported over the Internet. This raises serious security concerns for the users of those IoT devices. We note that this observation is in contrast to a forecast made by Sandvine in 2016 [13], in notes that 70% of traffic on the Internet is encrypted. Not surprisingly, the second most dominant application layer protocol is HTTP (i.e. TCP port 80), which constitutes 11% of IoT Internet traffic.

IoT devices commonly advertise their presence in order to discover other devices in the network. This behaviour is visible in our trace data. It can be seen in the figure that UDP port 1900, indicative of the SSDP protocol, appears in 8% of IoT packets. We note that SSDP traffic is communicated to a multicast address 239.255.255.250 and is only visible within the internal network. Our analysis also shows that UDP ports 53 and 123, representing DNS and NTP respectively, are among the other frequently used protocols by the IoT devices. In the next subsection, we will study these two signalling protocols in more detail. Lastly, TCP port 1935 accounted for 7% of IoT packets that were sent to the Internet. This port number was only used by the Whitings baby monitor camera.

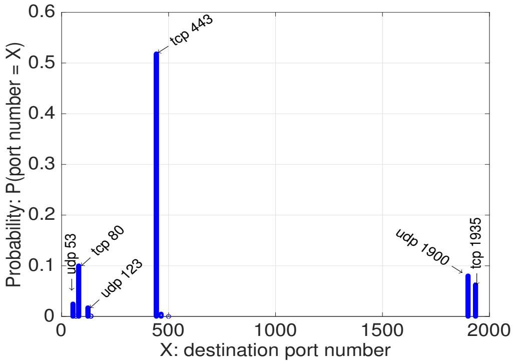


Fig. 4. Probability histogram of destination port numbers for IoT packets destined to both the local network and the Internet.

TABLE I
MOST FREQUENT DESTINATION PORT NUMBER.

Device	Belkin switch	Blipcare BP meter	HP printer	Insteon camera	LiFX bulb
port number	TCP 3478	TCP 8777	TCP 5222	UDP 10001	TCP 56700

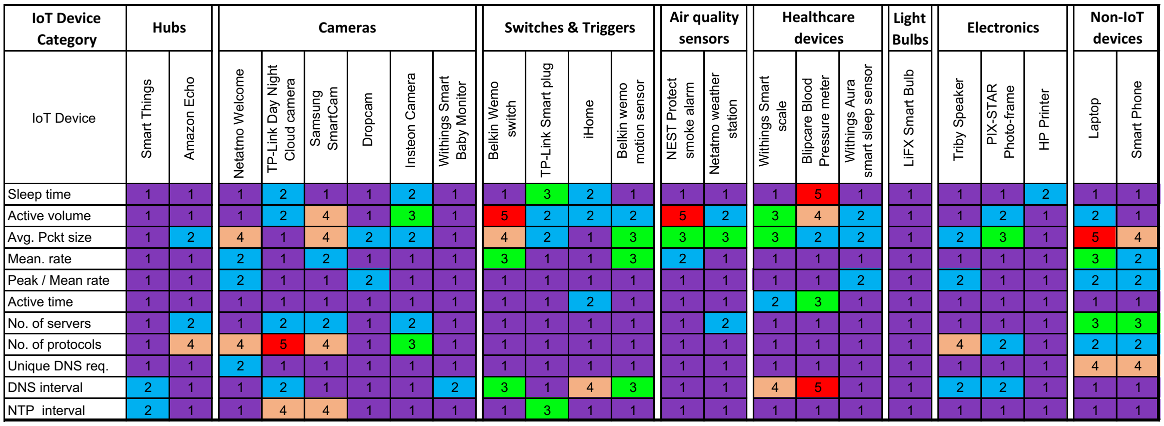
Device	NEST Protect	Netatmo weather	TPLink camera	Triby speaker	Withings camera
port number	TCP 11095	TCP 25050	TCP 50443	TCP 5228	TCP 1935

C. IoT Device Specific Attributes

Our aim is to examine characteristics of IoT devices from different viewpoints and highlight their dominant attributes, enabling us to (a) distinguish an IoT device from a non-IoT device such as a laptop or mobile phone, and (b) identify a certain IoT device or its category (as listed in Fig. 1(b)).

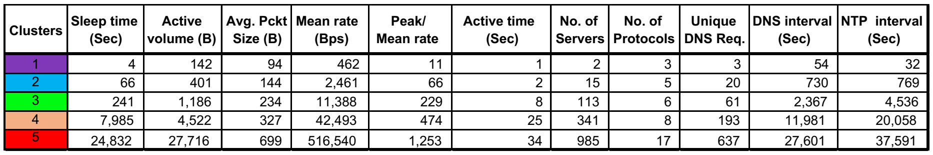
Data traffic pattern: In order to glean attributes from a data traffic and communication perspective, we first convert the raw pcap files into flows on a daily basis. Then for a given IoT device, we aggregate all the flows associated with that device. Given the resulting traffic profile, we study the probability histogram of the sleep time attribute and observe that there is a unique pattern for some IoT devices. For example, sleep times of 90, 60 and 20 seconds occur respectively for the HP Printer, iHome switch and Netatmo welcome camera with probability more than 70%. Another interesting attribute we find is that some devices exchange a unique volume of data (in bytes) for the most part during their active periods. For example, in the course of two weeks we observed that Samsung SmartThings, Samsung SmartCam and Netatmo weather station consistently exchanged 114, 3341 and 342 bytes during their active periods. Moreover, some devices such as Withings smart scale, Netatmo weather station and SmartThings exhibit signatures in terms of the average packet size; 225, 200 and 75 bytes respectively. Finally, as discussed in §IV-A, IoT devices generate short bursts of traffic frequently that result in fairly low bit-rate compared to non-IoT devices. We therefore measure, for each IoT device, the mean rate, ratio of peak to mean rate, active time and active volume from the daily traffic profiles. These attributes collectively help distinguish IoT devices from non-IoT devices.

Cloud servers: An important observation we make is that IoT devices differ from non-IoT devices based on the number



IoT Device Category	Hubs		Cameras						Switches & Triggers				Air quality sensors		Healthcare devices			Light Bulbs	Electronics			Non-IoT devices	
IoT Device	Smart Things	Amazon Echo	Netatmo Welcome	TP-Link Day Night Cloud camera	Samsung SmartCam	Dropcam	Insteon Camera	Withings Smart Baby Monitor	Belkin Wemo switch	TP-Link Smart plug	iHome	Belkin wemo motion sensor	NEST Protect smoke alarm	Netatmo weather station	Withings Smart scale	Blipcare Blood Pressure meter	Withings Aura smart sleep sensor	LiFX Smart Bulb	Triby Speaker	PIX-STAR Photo-frame	HP Printer	Laptop	Smart Phone
Sleep time	1	1	1	2	1	1	2	1	1	3	2	1	1	1	1	5	1	1	1	1	2	1	1
Active volume	1	1	1	2	4	1	3	1	5	2	2	2	5	2	3	4	2	1	1	2	1	2	1
Avg. Pckt size	1	2	4	1	4	2	2	1	4	2	1	3	3	3	3	2	2	1	2	3	1	5	4
Mean. rate	1	1	2	1	2	1	1	1	3	1	1	3	2	1	1	1	1	1	1	1	1	3	2
Peak / Mean rate	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1	2	1	1	2	2
Active time	1	1	1	1	1	1	1	1	1	1	2	1	1	1	2	3	1	1	1	1	1	1	1
No. of servers	1	2	1	2	2	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1	1	3	3
No. of protocols	1	4	4	5	4	1	3	1	1	1	1	1	1	1	1	1	1	1	4	2	1	2	2
Unique DNS req.	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	4	4
DNS interval	2	1	1	2	1	1	1	2	3	1	4	3	1	1	4	5	1	1	2	2	1	1	1
NTP interval	2	1	1	4	4	1	1	1	1	3	1	1	1	1	1	1	1	1	1	1	1	1	1

Fig. 5. Clustering of IoT attributes.



Clusters	Sleep time (Sec)	Active volume (B)	Avg. Pckt Size (B)	Mean rate (Bps)	Peak/ Mean rate	Active time (Sec)	No. of Servers	No. of Protocols	Unique DNS Req.	DNS interval (Sec)	NTP interval (Sec)
1	4	142	94	462	11	1	2	3	3	54	32
2	66	401	144	2,461	66	2	15	5	20	730	769
3	241	1,186	234	11,388	229	8	113	6	61	2,367	4,536
4	7,985	4,522	327	42,493	474	25	341	8	193	11,981	20,058
5	24,832	27,716	699	516,540	1,253	34	985	17	637	27,601	37,591

Fig. 6. Measure of attribute clusters.

of different Internet servers (excluding DNS and NTP servers) they communicate with over a 24-hour period. For example, a laptop in our testbed that was used for general Internet access contacted about 500 different servers, identified by unique IP addresses, within only two hours of activity. This is not surprising since a non-IoT device usually runs multiple applications and accesses numerous web sites. However, IoT devices are designed for specific purposes, and therefore communicate with their own server(s) (i.e. the device manufacturers') or selected cloud providers such as IFTTT servers. Our analysis shows that each IoT device communicates with less than 10 servers on average per day and the number of cloud servers contacted is fairly consistent across many IoT devices.

Protocols: As noted in §IV-B, IoT devices predominantly tend to use a few specific application-layer protocols. If we examine the most frequently used destination port number for data exchange, then a unique device-specific signature emerges. Table I lists the dominant destination port numbers used by ten IoT devices used in our experiments.

DNS traffic: As mentioned in §IV-B, DNS is one the most popular protocols among IoT devices. In our dataset, we observed that IoT devices initiate DNS queries for only a limited number of domains (mostly domain name of their vendors or service providers) and repeat the queries in a consistent manner. For example, Amazon Echo, Samsung SmartThings and Belkin Wemo motion sensor issue periodic DNS requests every 5, 10 and 30 minutes respectively. Further, on average these three devices ask for 7, 3 and 5 unique domains within a day. However, a non-IoT device such as a laptop looks for more than 300 domain names in a course of a few hours. Therefore, we believe that the number of unique domains and the frequency of DNS queries are important attributes that characterize IoT devices.

NTP traffic: Precise and verifiable timing is crucial for IoT operations [14]. Our analysis indicates that UDP port 123 (NTP protocol) contributes to 2% of total packets sent from IoT devices to the Internet. We also find that the time synchronization occurs repeatedly in our testbed and many IoT devices exhibit a recognizable pattern in the use of NTP. For example, SmartThings, LiFX lightbulb and Amazon Echo send NTP requests every 600, 300 and 50 seconds respectively.

D. IoT Attributes Clustering

We have enumerated a multitude of attributes of interest for IoTs in the preceding subsection. To enable the reader to *visualize* the role played by each of these attributes, we apply the K-Means clustering algorithm – using the Weka tool [15] – to the attributes across all the IoT devices. We use five bins for each attribute. A smaller bin count was not effective in identifying the unique fingerprint that underpins each IoT device. Choosing a larger bin size renders the visualisation too onerous, while adding little additional insights.

The result of the clustering algorithm is shown in Fig. 5. Each row in the figure denotes an IoT traffic attribute. Each cell, corresponding to a specific device, has a colour code that represents the cluster bin (ranging from 1 to 5) assigned to the respective traffic attribute. It can be seen from the figure that the columns depict a *unique* colour map (i.e. a sequence of colours, corresponding to cluster bins), which denotes the signature/fingerprint of the traffic attributes that underpin each IoT device. In other words, no two columns share the same sequence of colours – equivalently, the cluster bins – permitting unique identification of the devices from the respective colour maps. For example, all the traffic attributes corresponding to the LiFX Smart Bulb belong to cluster 1 (i.e. coded purple), for the Withings Smart Baby Monitor the DNS interval attribute belongs to cluster 2 (i.e. blue) while the remaining attributes belong to cluster 1. Thus, given a colour sequence, the visualization aids in identifying the (unique) IoT device that matches the sequence. The K-Means algorithm also returns a single parameter for each cluster, called the "cluster center", which is shown in Fig. 6.

The following observations emerge from the above two figures. First, the attributes number of servers contacted and unique DNS requests can be used to distinguish if a device is non-IoT or IoT. For non-IoT devices, the number of servers is coded green (i.e. cluster 3), while for a vast majority of IoT devices it is coded purple (cluster 1), and only five are coded blue (cluster 2). Referring to Fig. 6 we note that, non-IoT devices communicate with a large number of servers (in excess of 100), owing to the diverse range of applications (e.g. video, web, etc.) that are executed on those devices. On the other hand, the number of servers contacted by IoT devices is substantially smaller, since IoT devices are custom-built for specific applications. Similarly, unique DNS requests issued by non-IoT devices is coded orange (cluster 4), while for all but one IoT device it is purple (cluster 1); only Netatmo Welcome camera is coded blue (cluster 2). We note from Fig. 6 that the number of DNS requests from non-IoT devices is nearly 200, while for IoT devices it is often less than 5. The analysis reveals that the number of servers and unique DNS requests can be used to infer whether a device is non-IoT or IoT.

Second, as described earlier, all the traffic attributes combined help distinguish one IoT device from another. To draw some insights on what the most dominant attributes are, we observe from Fig. 5 that active volume and DNS intervals taken together uniquely identify several IoT devices such as DropCam, Nest Protect, Insteon Camera, Blipcare Blood Pressure monitor and Belkin WeMo switch. We employ the *InfoGainAttributeEval* tool of the Weka to evaluate the relative importance of each attribute; the output confirms this intuitive observation – the ordering of attributes (ranked from the most dominant) that help distinguish one IoT device from another is active volume, DNS interval, average packet size, mean rate, sleep time, number of servers, number of protocols, unique DNS requests, NTP interval, peak/mean rate and active time.

E. IoT Device Classification

While the preceding subsection only examined the importance of different attributes via visualization, we now develop a classification technique, driven by supervised machine learning algorithms, to help identify IoT devices with high accuracy. To do so, we rely on numerous algorithms available in the Weka tool for classification (since we have labelled data sets), and present results from the Random Forest algorithm (for brevity, we have omitted results from the other algorithms).

We train the classifiers with dataset from two weeks (i.e. 23-Sep to 6-Oct). Each training instance of our dataset contains the following attributes – i.e. sleep time, active volume, average packet size, mean rate, peak to mean ratio, active time, number of servers, number of protocols, unique DNS requests, DNS interval, NTP interval, most frequent port number and a label identifying the IoT device. We then evaluate the efficacy of our classifiers using (i) 10-fold cross-validation method, and (ii) by applying it to an independent test dataset.

Our cross-validation method randomly splits the dataset into training (90% of total instances) and validation (10% of total instances) sets. This cross-validation is repeated 10 times. The results are then averaged to produce a single performance metric. For independent test data, we collect a new dataset spanning one week (7-13 Oct), that has not been seen before. The earlier two-week dataset is used for training and the newly collected one-week dataset is used for validation purposes.

The classification results indicate that the Random Forest algorithm reaches a high accuracy of over 97% in the 10-fold cross-validation test and over 95% in the independent test analysis, meaning that this algorithm is able to uniquely identify an IoT device with a very high probability. These results demonstrate the viability of the proposed traffic attributes in uniquely classifying and identifying IoT devices.

V. Conclusions

Despite the proliferation of smart IoT devices in cities and campuses around the world, operators of such environments lack an understanding of what IoT devices are connected to their networks, what their traffic profiles look like and whether the devices are functioning normally without their security being compromised. This work is the first to systematically profile, characterize and classify IoT devices in smart environments. We instrumented a campus facility with 21 unique IoT devices and collected traffic traces over 3 weeks, which we release to the public. We then statistically characterized the traffic in terms of activity patterns, signalling, protocols, etc. Finally, we developed a classification technique that not only distinguishes between IoT and non-IoT devices, but also uniquely identifies IoT devices with over 95% accuracy. This paper sets the stage for future work in performance and security in IoT-enabled smart cities and campus environments.

References

[1] Cisco Systems. (2016) Visual networking index (VNI).

[2] S. Notra et al., "An Experimental Study of Security and Privacy Risks with Emerging Household Appliances," in *Proc. M2MSec*, Oct 2014.

[3] The guardian. (2016) Why the internet of things is the new magic ingredient for cyber criminals. https://goo.gl/MuH8XS.

[4] T. Yu, V. Sekar, S. Sheshan, Y. Agarwal, and C. Xu, "Handling a Trillion (Unfixable) Flaws on a Billion Devices: Rethinking Network Security for the Internet-of-Things," in *Proc. ACM HotNets*, Nov 2015.

[5] A. Sivanathan *et al.*, "Low-Cost Flow-Based Security Solutions for Smart-Home IoT Devices," in *Proc. IEEE ANTS*, Nov 2016.

[6] M. Z. Shafiq *et al.*, "A First Look at Cellular Machine-to-Machine Traffic: Large Scale Measurement and Characterization," in *Proc. ACM Sigmetrics*, England, Jun 2012.

[7] N. Nikaein *et al.*, "Simple Traffic Modeling Framework for Machine Type Communication," in *Proc. ISWCS*, Germany, Aug 2013.

[8] M. Jadoul, Nokia. The IoT: The Network Can Make It or Break It. https://insight.nokia.com/iot-network-can-make-it-or-break-it.

[9] M. Simon, Alcatel-Lucent. Architecting Networks: Supporting IoT.

[10] M. Laner, P. Svoboda, N. Nikaein, and M. Rupp, "Traffic Models for Machine Type Communications," in *Proc. ISWCS*, Germany, Aug 2013.

[11] A. Orrevad, MS Thesis, KTH University, Sweden. M2M Traffic Characteristics: When Machines Participate in Communication.

[12] (2016) OpenWrt. https://openwrt.org/.

[13] Sandvine. (2016) Internet Traffic Encryption. https://www.sandvine.com/trends/encryption.html.

[14] M. Weiss et al. (2015) Time-Aware Applications, Computers, and Communication Systems. http://dx.doi.org/10.6028/NIST.TN.1867.

[15] (2016) Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/.