Characterizing and Classifying IoT Traffic in Smart Cities and Campuses

Paper's analysis

ITPA 2019-2020

Andrea Graziani - 0273395

Università degli Studi di Roma "Tor Vergata"
FACOLTA' DI INGEGNERIA
Corso di Laurea Magistrale in Ingegneria Informatica

December 1, 2020

Research's goal

According to Sivanathan et al. [1], research's goal is to:

> *"[...] develop a classification method that can not only distinguish IoT from non-IoT traffic, but also identify specific IoT devices with over 95% accuracy."*

What is the reason according to which is important to profile IoT traffic?

1. To understand IoT devices "*normal*" **traffic pattern** in terms of their **activity pattern** (traffic rate, idle durations, etc.) and **signalling overheads** (DNS, NTP, etc.).
2. To enhance **cyber-security** involving IoT devices which administration belong to **different authorities**.
   According to Sivanathan et al. [1], is possible to improve security deploying a network-level security mechanisms which, analysing traffic patterns, is capable to **identify attacks knowing the normal traffic pattern of monitored IoT devices**.

Research's goal

In other words, research's goal is to build an classification model for IoT devices based on **machine learning** techniques, which building passes through following steps:

1. Collect data from an IoT environment.
2. Characterize traffic pattern corresponding to the various IoT devices.
3. Develop a classification technique that learns the behaviour of an IoT device and is able to identify it based on its traffic pattern.

Data-set building - Part 1

Since Sivanathan et al. [1] adopted a **supervised machine learning algorithms** to build their classification model, is necessary to generate a **data-set** in order to provide an appropriate input during the **learning phase**.

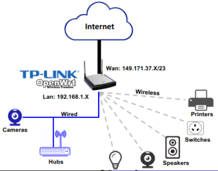
Data-set building - Part 2

- Sivanathan et al. [1] collected traffic over 3 weeks generated from a so-called "*Smart Environment*" built by them.
- Is very important to precise that collected data are **time series**, where each instance, indexed by time, contains several **attributes** (or **features**) including:
  - Sleep Time.
  - Active time.
  - Average packet size.
  - peak/mean rate.
  - Number of used protocols.
  - Unique `DNS` requests.
- Clearly, every instance contains a **label** identifying the IoT device, which is necessary during supervised learning.

The "Smart Environment" - Part 1

- In order to build the necessary dataset, Sivanathan et al. [1] built the aforementioned "*Smart Environment*" to simulate a real usage scenario, collecting required data.
- This environment is made up of:
  - 21 unique IoT devices representing different categories, like **cameras**, **healthcare devices**, **hub**, **air quality sensors** and so on.
  - A router, the TP LINK ARCHER C7[1]
  - Several non-IoT devices were also used, such as laptops, mobile phones and tablet.

Figure: "Smart environment"'s scheme



The "Smart Environment" - Part 2

Several IoT devices used by Sivanathan et al. [1] for their experiments **are battery operated**.
For instance:

- The *Withings Smart scale* device is powered by 4 1.5 V alkaline cells (AAA).[a]
- Similarly, the *Netatmo Weather station* device is powered by 2 1.5 V alkaline cells (AAA) with an **estimated autonomy of about 2 years**.[b]
- The *Blipcare blood pressure meter* device is powered by an internal battery.[c]

[a] https://www.withings.com/it/en/body
[b] https://www.netatmo.com/it-it/weather/weatherstation/specifications
[c] http://www.blipcare.com/

"Smart Environment" Architecture - Part 1

The architecture of the "Smart Environment" can be splitted into:

Front-end which contains the *router* and the *IoT/non-IoT devices*

Back-end represented by the *cloud* which is responsible for computations, storing received information, filtering duplicate packets and so on. through the gateway

Cloud resources are exploited through so-called **cyber-foraging techniques** in order to overcome the very strictly constrains of any IoT devices.

Proposed architecture is intended for **cloud-native** applications.

"Smart Environment" Architecture - Part 2

The front-end of the "Smart Environment", build by researchers, is characterized by a **star network topology**.

This is a very important observation, because a star topology allows us to:

- **Preserve battery life** of IoT devices because they **do not have to forward** other nodes data; in other words, *any IoT device receives, or transmits, only its own data*.
- **Decrease the complexity** of the network.

The LPWAN example

The implementation of `LoRaWAN` network is based on the star network topology, and mostly, stars-of-stars network.
As known `LoRaWAN` network belongs to `LPWAN` category, *which are specifically designed to achieve the need for low power, long-range, low bit error rate, and low cost* needed in IoT context.

"Smart Environment"'s Wireless Networks Technology

According to vendor's specifications regarding the TP LINK ARCHER C7, is possible to know that the aforementioned router supports following protocols:

- `IEEE 802.11ac/n/a` at 5 GHz
- `IEEE 802.11n/b/g` at 2.4 GHz

Researchers use `IEEE 802.11` as **media access control** (`MAC`) and **physical layer** (`PHY`) protocol.

Researchers did *not* specify which **version** of IEEE 802.11 standard has been effectively used.
We don't know with which **frequencies** data has been transmitted.

"Smart Environment"'s Wireless Networks Technology

We believe that above protocols are **not** fully optimized for IoT business models and devices used in smart cities and campuses for following reasons:

- These technologies provide a **short/medium coverage** with **100-to-1000** meters range. Provided coverage range can be not enough to fulfil all use cases.
  - This is due to **mid/high frequencies** used by these protocol which are **vulnerable to several side effect during signal propagation** (*blocking*, *reflection*, *refraction* and so on)
- They are affected by **header-overhead** caused by **short packets transmission** which are very common is many IoT scenarios.

“Smart Environment”’s Wireless Networks Technology

Utilizing sub-1 GHz bands used by both `802.11ah` and `LoRaWAN`, is possible to provide **better propagation characteristics** in outdoor scenarios. **Low frequencies signal are less affected by obstacles presence.**

A combination of low-band, mid-band and high-band spectrum is desirable to manage all possible use cases.

	802.11ac	802.11n	802.11a	802.11ah	LoRaWAN
Frequency (GHz)	5	2.4,5	5	0.7/0.8/0.9	$\backsim 0.86(EU)$
Sensitivity (dbm)	-82	-82	-88	-98	$[-124, -137]$
Bit rate	6.5 (Mb/s)	6.5 (Mb/s)	1.5 (Mb/s)	0.15 (Mb/s)	5469 (Bit/s)
Max coverage range (km)	0.115	0.230	0.115	$\backsim 1$	$\backsim 15$

IoT Traffic - Part 1

According to their experimental results, Sivanathan et al. [1] stated that:

> *"[...] 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), [...], implying that traffic generated by IoT devices is small compared to traditional non-IoT traffic."[1, par. IV.A]*

> *"the traffic pattern of one IoT device [...] a pattern of active/sleep communication emerges. [...] IoT active time [...] 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."[1, par. IV.A]*

IoT Traffic - Part 2

Since many IoT devices are battery powered, **maximize energy efficiency**, in order to **preserve devices lifetime**, is critical.

According to Sivanathan et al. [1]'s results, the power management approach adopted by IoT devices is based on **periodic sleep**, during which radio transceiver are turned off.

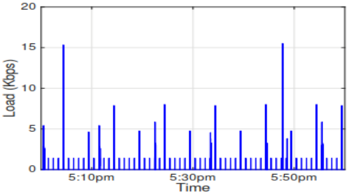
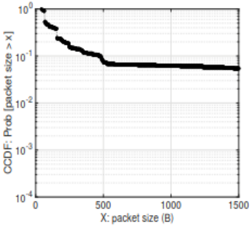


Figure: Load of LiFX light bulb device.

IoT Traffic - Part 3

- A very interesting observation by Sivanathan et al. [1] made by regard packet size, according to which only the 10% of packets are larger than 500 Bytes.
  - **header-overhead**, caused by **short packets transmission**, can occur frequently.

Figure: “Smart environment”’s scheme



IoT Application Layer Protocol: Overview

- According to [**REALSMARTIOT**], smart city and campus services services are based on a centralized architecture where a dense and heterogeneous set of IoT devices generate differ- ent types of data that are then delivered through suitable com- munication technologies to a control center, where data storage and processing are performed
  peripheral devices deployed over the urban area generate differ- ent types of data that are then delivered through suitable com- munication technologies to a control center, where data storage and processing are performed
- A very important aspect is the necessity to make (part of) the data collected by the urban IoT easily accessible by authorities and citizens,

The Most Dominant Application Layer Protocols - Part 1

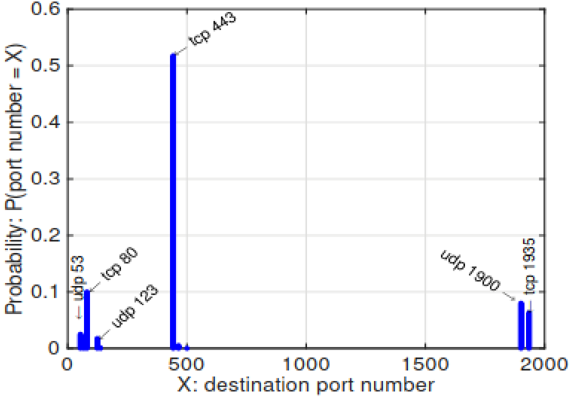


Figure: Probability histogram of destination port numbers for IoT packets destined to both the local network and the Internet.

The Most Dominant Application Layer Protocols - Part 2

HTTPS (TCP port 443) is the dominant protocol used by the IoT devices since it represents over the 55% of total IoT traffic.

HTTP (TCP port 80) represent the second most dominant application layer protocol constituting the 11% of total traffic.

SSDP (UDP port 1900) is the next most dominant application layer protocol representing the 8% of traffic.

- SSDP, which stands for **Simple Service Discovery Protocol**, is used to for *advertisement* and *discovery* purposes of network services without the assistance of server-based configuration mechanisms, such as DHCP or DNS.

RTMP (TCP port 1935) represent the fourth most dominant protocol representing the 7% of traffic

- RTMP, which stands for **Real-Time Messaging Protocol**, is a proprietary protocol used for streaming audio, video and data over the Internet, generally used by cameras. It is owned by Adobe.

The Most Dominant Application Layer Protocols - Part 3

DNS (UDP port 53) represents less than 5/4% of total traffic.

NTP (UDP port 123) constitutes less than 2/3% of IoT traffic.

Application specific Sivanathan et al. [1]'s results shows that, regarding remaining IoT traffic, each IoT device use an own **application-specific** protocol.
In table reported below, are reported most frequent *transportation protocol* and *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

The role of HTTP - Part 1

Why results that HTTPS and HTTP are the most used protocols by IoT devices?

- A very important aspect of an urban, or campus, IoT infrastructure is the **necessity to make data collected by the urban IoT devices easily accessible** by both authorities and citizens.
- In order to achieve this objective, IoT devices adopt a very well known web-based paradigm called **Representational State Transfer** (**ReST**), which plays a very important role into **Web of Things Architecture** (**WoT**).

The role of HTTP - Part 2

- Exploiting REST paradigm, HTTP and HTTPS are used very frequently because they facilitate both the **integration of IoT devices** with existing services currently available on the Web and the **Web applications development**.
- HTTP and HTTPS offer a **direct access** for users to IoT devices data and services, without the need for installing additional software.
  In fact, using a Web browser (or any HTTP library in the case of a software client) client are able to to directly extract, save and share smart things data and services.
  This ensures the usability of the architecture and minimizes the entry barriers for final users.

The disadvantages of HTTP

- The **verbosity** and **complexity** of native HTTPS/HTTP make them **unsuitable** for constrained IoT devices.
  - In fact, the **human-readable format** of HTTP, which has been one of the reasons of its success in traditional networks, turns out to be a limiting factor due to the large amount of heavily correlated (and, hence, **redundant**) data.
- HTTPS/HTTP rely upon the TCP transport protocol that, however, does not scale well on constrained devices, yielding poor performance for small data flows in lossy environments.

An Unbalanced Data-Set - Part 1

The performance and the interpretation of a IoT device classification model **depend heavily on the data** on which it was **trained**.

- Scientific literature showed that classification model, which are trained on *imbalanced datasets*, are highly susceptible to producing inaccurate results.

The dataset produced by researchers can be unbalanced owing to several reasons including:

- Too few IoT device types.
- *Constrained* and *unconstrained* protocol stack are not equally represented into dataset.

An Unbalanced Data-Set - Part 2

- Generally smart city and campus services are based on a **very heterogeneous set of IoT devices**, generating **very different types of data** that have to be delivered through **suitable communication technologies**.
- For instance, possible applications can be:
  - Structural Health of Buildings.
  - Waste Management.
  - Air Quality.
  - Traffic Congestion.
  - Noise Monitoring.
  - City Energy Consumption.
  - Smart Parking.
  - Smart Lighting.

Proposed "*Smart environment*" seem to be more suitable for a smart home rather than a smart city or campus.

An Unbalanced Data-Set - Part 3

The "*Smart environment*" used by researchers *can* be *not* suitable for their purposes because is **too simple**.

- It includes **only** *unconstrained protocol stack* which include protocols that are currently the de-facto standards for Internet communications and are commonly used by regular Internet hosts (HTTP/TCP/IPv4).
  - In fact there is a prevalence of HTTPS/HTTP application layer protocol (66% of total IoT traffic according to Sivanathan et al. [1]) and of the TCP transport layer protocol (representing, more or less, the 85% of total transmitted packets according to Sivanathan et al. [1]'s results).
- It does **not** include any *constrained protocol stack*, the low-complexity counterparts of the de-facto standards for Internet, i.e., **Constrained Application Protocol** (CoAP), UDP, and 6LoWPAN, which are suitable even for very constrained devices.

Differences between IoT and Non-IoT Traffic

Experiment's results show following differences among IoT and Non-IoT traffic:

DNS traffic IoT devices initiate DNS queries for only a limited number of domains while non-IoT device, such as a laptop, looks for more than 300 domain names in a course of a few hours.

Number of Cloud servers IoT device communicates with less than 10 servers on average per day while non-IoT device contacts about 500 different servers

Security Problems Due To Unencrypted Traffic - Part 1

IoT devices communication is properly secured?

- According to Sivanathan et al. [1], about 45% of IoT traffic is not sent over HTTPS to the servers.
- Since the traffic transmitted using other protocols are typically not encrypted, Sivanathan et al. [1]'s results indicate that a sizeable fraction of IoT traffic is **not** being securely transported over the Internet.
  - The use of unencrypted protocols can leak sensitive information about users.

Security Problems Due To Unencrypted Traffic - Part 2

Why IoT devices transmit unencrypted data?

There may be various reasons according to which data are transmitted unencrypted:

- Due to **limitations** and **constrains** in the IoT device itself.
- As noted by Englehardt and Narayanan [25], IoT devices vendors may be hesitant to move to `HTTPS` if their products use any third-party resources that are `HTTP`-only. These resources are typically ads and trackers.
- Bad design.

available technologies. From the table, it clearly emerges that, in general, the practical realization of most of such services is not hindered by technical issues, but rather by the lack of a widely accepted communication and service architecture that can abstract from the specific features of the single technologies and provide harmonized access to the services.

- According to Sivanathan et al. [1], the set of IoT devices used for their experiments including a huge amount of **sensors**, including air quality sensors and health-care devices.
  As known, aforementioned kind of devices generate a huge amount of data modelled as **time series**, that is an array of values indexed by time.
  According to **TIMESERIES**, the stream of data generated by all these IoT sensors is generally interfaced with database, through a so-called *southbound* interface, using HTTP RESTful protocol.
  Similarly, all applications requiring access to the data stored in the database, using the same protocol, through a so-called northbound interface.

Some references

[1] A. Sivanathan et al. "Characterizing and classifying IoT traffic in smart cities and campuses". In: *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. 2017, pp. 559–564. DOI: 10.1109/INFCOMW.2017.8116438.