UDC 004.7:004.8

Machine learning based network traffic classification approach for Internet of Things devices

Vadym Melnik[1, 3, a], Pavlo Haleta[2, 3, b], Nazar Golphamid[3, c]

[1] *National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Institute of Physics and Technology*

[2] *National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»*

[3] *Samsung R&D Institute Ukraine (SRK)*

Abstract

Due to design flaws, problems with implementations and improper network configuration, the Internet of Things devices become vulnerable in the network. They can be easily compromised and can also be attached to the Botnet network. IoT devices classification allows for strengthening of the overall network security through better VLAN planning and better firewall rule fine-tuning (e.g. per device class). In this paper only two classes of devices are considered: single-purpose devices (such as a bulb) and multi-purpose devices (such as mobile phone). Existing solutions do not provide the required accuracy within the given timeframe. We propose ML-based classification method based on supervised machine learning technology (Random Forest). With advanced packets flow analysis, our proposed approach demonstrates 94% of accuracy (7% better than the existing prior art). Additionally a very low False Positive rate is guaranteed for single-purpose IoT devices (e.g. a bulb must never be classified as a multi-purpose device).

Keywords: Internet of things, machine learning, local network, internet traffic, control packet

Introduction

Internet of things is gaining rapid popularity today, due to the fact that devices of this type can perform everyday tasks autonomously, without direct human intervention. More and more Internet of Things devices join the networks. In accordance with the statistical predictions [1], the number of IoT devices by 2020 will be increased to 25-30 billion which gives reason to talk about the beginning of the era of the Internet of things.

However, with the growing popularity of the Internet of Things, the danger of their use is also growing. Typically security level of mass-product devices is low [2] because of their low cost, lack of support and improper configuration. Therefore owners of smart environments (smart home, smart building, smart city) are faced with the big cyber security threat while dealing with an unmanageable amount of networking IoT agents. For example, in 2017 a university campus was attacked with the help of vending machines located on the territory of the university [3]. As a result, 5000 IoT devices were damaged. Therefore, according to a Cisco report [4], identifying and classifying each device is one of the ways to ensure that each device is in a secure network segment for it and receives the required quality of service configuration (setup). Automatic network micro-segmentation might give a baseline security to address the issue. To perform this segmentation as well as keeping it up-to-date an automatic IoT device classification technology is needed.

[a] v.melnik@samsung.com

[b] p.haleta@samsung.com

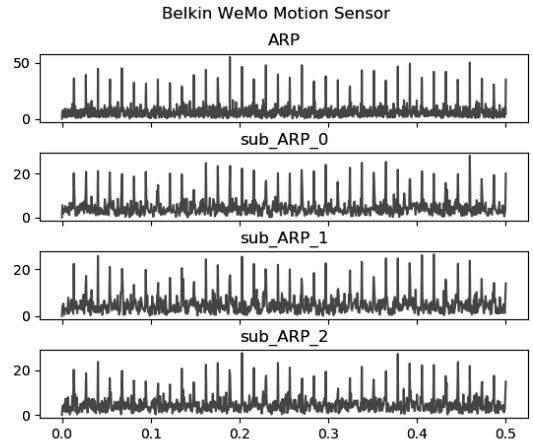
[c] n.golphamid@samsung.com

The main goal of this article is to suggest an effective mechanism for network devices classification using network packets flow analysis. Suggested technology allows for reliable differentiation between two classes of devices: single-purpose (resource-limited devices, such as sensors, cameras with limited and/or uncomplicated functionality) and multi-purpose (high-tech devices with better hardware resources) IoT devices. For many reasons (as an example, increasing the accuracy of determining the device type by their basic functionality) automatic network segmentation of just these two types of devices is preferable. Suggested technology has been tested in several SmartHome network environments consisting of 50 single and multi-purpose devices connected wirelessly, via Ethernet and using IoT Bridge (Bluetooth, ZigBee).

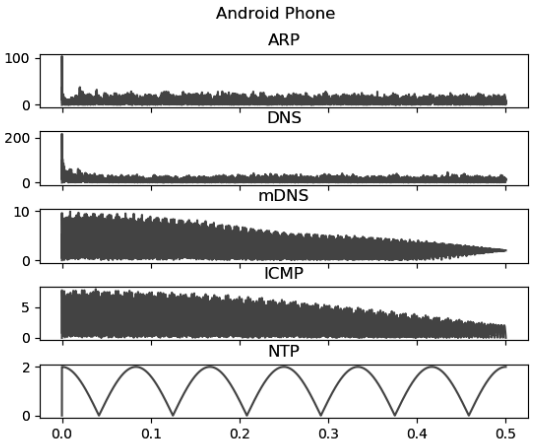
1. Related works

Nowadays, there are numerous statistical methods for packet-level device classification.

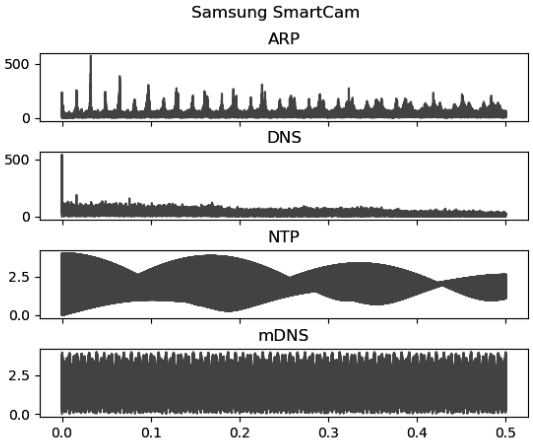
For example, in [5] spectrum features received by the discrete Fourier transform are applied to service protocols such as Address Resolution Protocol (ARP), Domain Name System (DNS), Network Time Protocol (NTP). The advantage of this method is that it provides hi-speed device classification. The paper claims that this method is able to classify devices in 90 minutes after connecting them to the network. However, this method has limited applicability because it assumes that IoT devices are utilizing service protocols frequently and periodicaly, which might not be the case (Fig. 1a, 1b, 1c).



(a) Active time of Belkin wemo motion sensor after DFT



(b) Active time of Android Phone after DFT



(c) Active time of Samsung SmartCam after DFT

Fig. 1. Active time of some devices in local network after DFT

In [6] a classification method that uses domain names which IoT devices use while they communicate was proposed. With filtering domain names that belongs to the device manufacturer, a sufficiently high classification accuracy is achieved. However, the main disadvantage of this method is that not all devices are associated with domains that contain key name of the device manufacturer. Another feature is that single-purpose and multi-purpose devices can belong to the same manufacturer, using a common domain name.

In [7], a classification method that analyzes the statistical characteristics of the packets flow of each device was proposed. The results show fairly accurate classification.

Presented approach in this paper improves [7] by adding custom features and by limiting the number of classes to single and multi-purpose IoT only.

The main task of this work is to analyze all possible characteristics of traffic to search for features that are clearly different for single-purpose and multi-purpose devices and that can increase classification accuracy.

2. Proposed concept and analysis

The proposed method can be applied to any local network in which all devices have an Internet connection. A device, which has network connection possibility and any kind of built-in sensors can be assigned to the Internet of Things. On this basis, the following types of devices can be distinguished:

- Single-purpose devices: resource-limited devices, such as sensors, cameras with limited and / or uncomplicated functionality. Also, these devices do not require human intervention.
- Multi-purpose devices: this group includes high-tech devices with better hardware resources, such as smartphones, personal computers and so on.

To carry out operations and functions IoT devices require network. The connection of these devices were divided into two categories:

- Device-to-Device (D2D) communications: this group includes communications between devices in the local network.
- Device-to-Infrastructure (D2I) communications: this group includes communications between devices of one network (local network) and devices or services of another network (remote network).

The proposed classification method, based on the behavioral profile, is depicted on Fig. 2. Behavioral profile is built using the statistical characteristics of the packets flow.

The first step of this algorithm is data collection in the local network. The main requirement is to obtain as few packets as possible to identify device as quickly as possible. The second step is to create model for supervised machine learning that consist of statistical characteristics of packets flow. Typical characteristics of the packets flow in the proposed model [7] are:

1) Sleep time.
2) Active volume.
3) Avg. Pckt size.

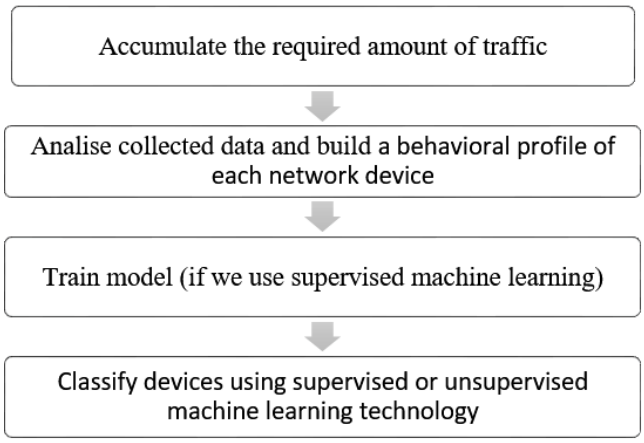


Fig. 2. Algorithm of the proposed classification method

4) Mean rate.
5) Peak/Mean rate.
6) Active time.
7) No. of servers.
8) No. of protocols.
9) Unique DNS req.
10) DNS interval.
11) NTP interval.

By analyzing each of the features it was established that not all features in this list are able to accurately distinguish the device into two groups – single-purpose and multi-purpose devices. In Fig. 3a, 3b it was shown that NTP and DNS intervals of each device from two groups are different and from this features we can't exactly say what device is it. So this features do not give enough information for classifier.

After analyzing the mean rate it can be seen the next fact.

$$Rate = Active\ volume \div Active\ time \qquad (1)$$

Active volume is a total sum of downloaded and uploaded bytes in a TCP session. Active time is a time between the first and the last packet in a TCP session.

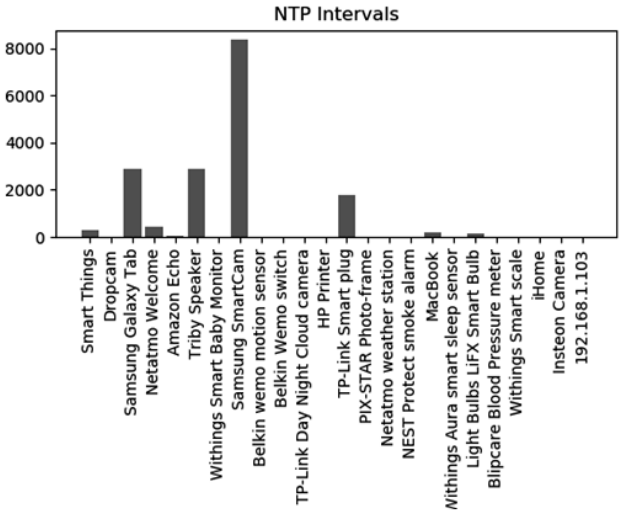
From (1) we can see that Rate depends on volume and time. As a result, Rate strongly correlates with this two features and can be discarded as redundant.

By analyzing the sleep time of each device, we can make the next observation. Many multi-purpose devices, when they are connected to the network and not affected by user intervention, can have the same behavior as single-purpose devices. For example, services of operation systems in mobile phone or laptop can make a connection for a system purpose such as: get updates, synchronize time and so on. Result of this observation is shown on Fig. 4.

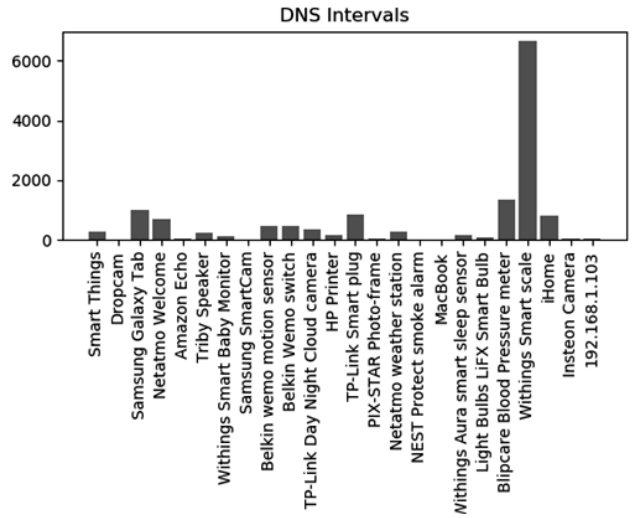
The device classification in the proposed method is based on three main assumptions:

***Assumption* 1.** The number of DNS queries from multi-purpose IoT devices significantly exceeds the number of DNS queries from single-purpose IoT devices.

DNS is one of the most popular protocols in the Internet of Things. By analyzing the packet flows over a long observation period, it was found that the number of DNS queries of the multi-purpose Internet of Things devices significantly exceeds the number of DNS queries



(a) NTP intervals



(b) DNS intervals

Fig. 3. Avg. NTP and DNS intervals of each device by one day

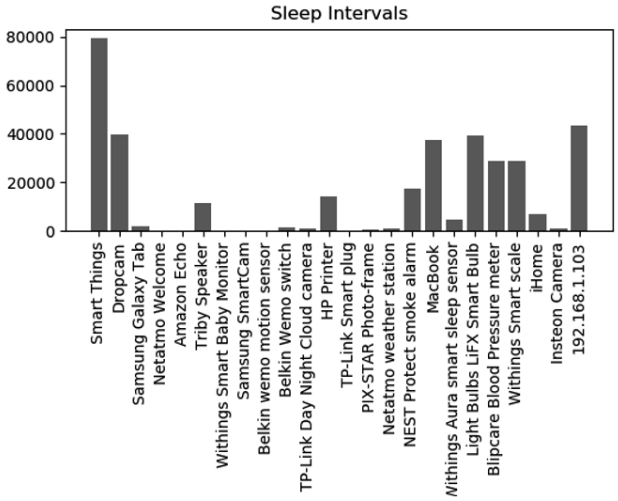


Fig. 4. Sleep intervals of each device by one day

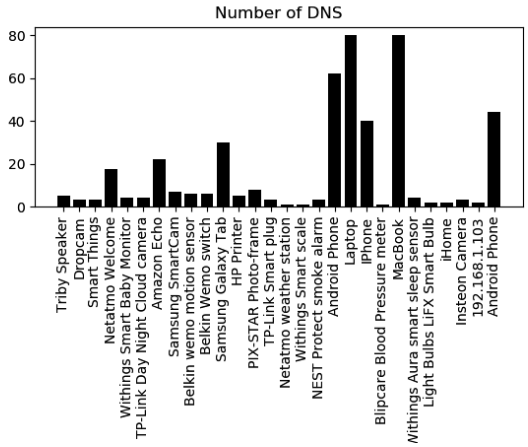


Fig. 5. The number of DNS queries of single-purpose and multi-purpose devices (for 1 week)

of the single-purpose Internet of Things devices. Also most of the DNS queries of the single-purpose Internet of Things contain the names of their manufacturers. For multi-purpose devices like a mobile phone or a laptop, this situation is not typical. For comparison, the analysis results are shown in Fig. 5.

Thus, the number of unique DNS queries and the number of unique domain names, with which the IoT devices are connected, are important indicators that can be used to classify the single-purpose Internet of Things devices and multi-purpose Internet of Things devices.

***Assumption* 2.** The connection strength of a multi-purpose device exceeds the connection strength of a single-purpose device due to the fact that the multi-purpose device must be able to establish connection with many single-purpose devices. The connection strength indicates how many devices are connected over a network. As an example: mobile phone can control many sensors, cameras, single-purpose speakers and so on.

In this article, Google's PageRank technology [8] is used to calculate the connection strength of each device. This technology analyzes an oriented graph of network connectivity, each edge of which has a weight equal to the number of packets generated by each device.

The connectivity strength is calculated by (2).

$$PR(u) = (1-d) + d \sum_{v \in M(v)} PR(v)/L(v), \qquad (2)$$

where u, v denotes devices, $PR(u)$ is a PageRank score, d is a dampening factor that is usually set to 0.85, $M(u)$ is the set of nodes that have links to the node u, and $L(v)$ is the number of outgoing links from node v.

As can be seen from Tab. 1, the connectivity index of multi-purpose devices exceeds the index of single-purpose devices.

But we can also notice that the index in the top of the table belongs to the hubs, since they are the central nodes in the network through which communication between devices occurs. The obtained result of the connectivity strength analysis proves that the connec-

Table 1. The connectivity strength of devices in the local network

Device	PageRank
Securify Almond	0.050131
Philips HUE Hub	0.041608
Samsung SmartThings Hub	0.030454
Android Tablet	0.026212
Apple HomePod	0.024202
Google Home	0.0221
Google Home mini	0.0221
Samsung SmartTV	0.020897
Sonos	0.018622
MiCasaVerde VeraLite	0.016152
Roku 4	0.016152
Roku TV	0.016152
iPad	0.016091
Wink 2 Hub	0.014857
Amazon Fire TV	0.014113
Apple TV	0.013064
D-Link DCS-5000L Camera	0.013052
Amazon Echo	0.012345
Belkin Netcam	0.12345
Logitech Hurmony Hub	0.011895
Bose SoundTouch 10	0.01888
iPhone	0.011069
August doorbell cam	0.008807
Belkin WeMo Link	0.008807
Belkin WeMo Motion Sensor	0.008807
Belkin WeMo Switch	0.008807
Canary	0.008807
Caseta Wireless Hub	0.008807
Chamberlain myQ garage opener	0.008807
Harmon Kardon Invoke	0.008807
Insteon Hub	0.008807
Koogeek Lightbulb	0.008807
LIFT Virtual Bulb	0.008807
Logitech Logi Circle	0.008807
Nest Camera	0.008807
Nest Cam IQ	0.008807
Nest Quard	0.008807
Netgear Arlo Camera	0.008807
nVidia Shield	0.008807

Table 2. Features importance of proposed model

Feature	Importance
Number of DNS queries	0.328768
Number of types of protocols	0.184130
Device type from the user-agent field	0.096616
Number of communications in the local network	0.090238
Connectivity strength of the devices	0.070912
Number of domain names with which device has TCP connections	0.065945
Number of unknown communications	0.047183
Average sessions time	0.044121
Average sessions volume	0.038351
Number of all communications	0.033735

tivity strength of the multi-purpose devices exceeds the connectivity strength of the single-purpose devices.

***Assumption* 3.** The User-Agent field in the HTTP header is inherent in most cases for multi-purpose devices. Thus, by analyzing this indicator, we can identify the type of device and make the necessary conclusion.

This characteristic is categorical and helps to increase accuracy in determining the devices class, but can only be used in the case of unencrypted traffic.

The random forest technology [9] was chosen to identify the devices. It is an ensemble method of machine learning for classification, which operates with the help of building numerous decision trees. The reason for choosing a random forest is its high resistance to reconfiguration in comparison with other decision tree classifiers.

The final list of features that is used in proposed method and their importance is shown in Tab. 2.

3. Evaluation results

To carry out the necessary experiments, two independent datasets [10], [11] were selected, each including both single-purpose and multi-purpose IoT devices.

The dataset [11] was selected for testing which includes 20 multi-purpose and 34 single-purpose devices. The network structure and device interconnections for the selected dataset are shown in Fig. 6.

To get more accurate behavioral profile of each device, the observation period of 4 days was chosen.

The dataset [10] was selected for validation. The network structure and device interconnections for the selected dataset are shown in Fig. 7. Thus, 9 multi-purpose devices and 22 single-purpose devices were identified. It should be noted that during the observation a minimum of 21000 first obtained packets for each device was detected, with which it was possible to obtain 94% accuracy.

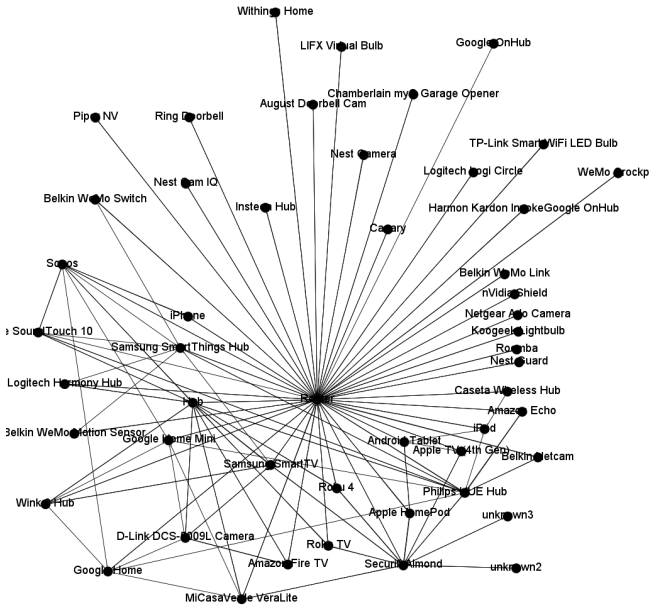


Fig. 6. The structure and communications in network that contains 20 multi-purpose and 34 single-purpose devices

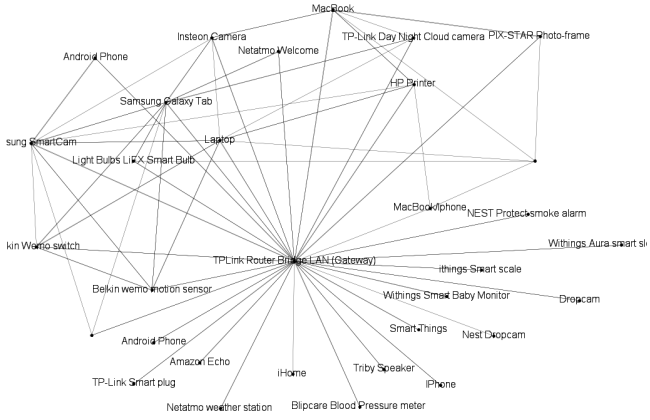


Fig. 7. The structure and communications in network that contains 9 multi-purpose and 22 single-purpose devices

Table 3. Result of proposed classification method with 21000 first obtained packets

Actual \ Predicted	Single-purpose device	Multi-purpose device
Single-purpose device	22	0
Multi-purpose device	2	7

Table 4. Result of prior-art classification method with 21000 first obtained packets

Actual \ Predicted	Single-purpose device	Multi-purpose device
Single-purpose device	22	0
Multi-purpose device	4	5

Tab. 3 shows that the Random Forest reaches 94% accuracy using different networks for testing and validation. The Random Forest classifier demonstrates better accuracy result among other classifiers (Tab. 6).

The result of the classification given in the Tab. 3 shows that the random forest technology reaches an accuracy of 94% in different local networks. For comparison, prior-art solution gives 87% (Tab. 4). Two solutions are false, that attributed acoustic systems with enhanced built-in functionality to single-purpose devices, due to passive (inactive) behavior in the network. However, when the number of packets was increased for classification to 560000, we get only one error and the accuracy was increased to 97% (Tab. 5). It is also important to note that in both results there are no errors related to the classification of single-purpose devices as multi-purpose.

This means that this approach is able to detect single-purpose and multi-purpose devices in different networks using behavioral profiles from the packets flow of each device in the network.

Conclusion

In this paper the classification method for single-purpose and multi-purpose IoT devices is proposed. Comparing with the prior art on the mentioned datasets the classification accuracy has been increased by 7% (94% vs 87%). Validation on different datasets shows

Table 5. Result of proposed classification method with 560 000 first obtained packets

Actual \ Predicted	Single-purpose device	Multi-purpose device
Single-purpose device	22	0
Multi-purpose device	1	8

Table 6. Comparison of classifier productivity

ML technology	Accuracy
Random Forest	94%
Decision Tree	94%
kNN	84%
LOF	75%
NB	71%
K-Means	71%
SVM	71%
DBSCAN	71%

that the model is applicable for varying network environments. The result can be used for further device differentiation to create more flexible security policies and VLANs. It can also be used as a baseline for better QoS (Quality of Service) configuration.

This article gives impulse to future research in the field of network security to protect from unauthorized exposure of devices and in providing the necessary performance and quality of service in the Internet of Things environment.

References

[1] D. Knake, "IoT numbers vary drastically: devices and spending in 2020." https://www.wespeakiot.com/iot-numbers-devices-spendings-2020, accessed 2017-10-06.

[2] A. D. Rayome, "Security flaw made 175,000 IoT cameras vulnerable to becoming spy cams for hackers." https://www.techrepublic.com/article/security-flaw-made-175000-iot-cameras-vulnerable-to-becoming-spy-cams-for-hackers, accessed 2017-08-01.

[3] G. Mezzofiore, "A university was attacked by its light-bulbs, vending machines and lamp posts." https://mashable.com/2017/02/13/internet-of-things-university-network-/#9RqajU3A5Oqu.3, accessed 2017-02-13.

[4] Cisco, "Cisco 2017 Midyear Cybersecurity Report." https://www.cisco.com/c/dam/global/es_mx/solutions/security/pdf/cisco-2017-midyear-cybersecurity-report.pdf, accessed 2017-07-24.

[5] T. D. Nguyen, S. Marchal, M. Miettinen, H. Fereidooni, N. Asokan, and A.-R. Sadeghi, "DÏot: A federated self-learning anomaly detection system for iot," *Cryptography and Security*, 2019. https://export.arxiv.org/pdf/1804.07474.

[6] H. Guo and J. Heidemann, "Ip-based iot device detection," *IoT S&P '18 Proceedings of the 2018 Workshop on IoT Security and Privacy*, pp. 36–42, 8 2018. https://www.isi.edu/~johnh/PAPERS/Guo18b.pdf.

[7] A. Sivanathan, D. Sherratt, H. H. Gharakheili, A. Radford, C. Wijenayake, A. Vishwanath, and V. Sivaraman, "Characterizing and classifying iot traffic in smart cities and campuses," *IEEE Conference on Computer Com-*

munications Workshops (INFOCOM WKSHPS), 5 2017. http://www2.ee.unsw.edu.au/~vijay/pubs/conf/17infocom.pdf.

[8] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, 1998. http://snap.stanford.edu/class/cs224w-readings/Brin98Anatomy.pdf.

[9] T. K. Ho, "Random decision forests," *ICDAR '95 Proceedings of the Third International Conference on Document Analysis and Recognition*, 1995. https://ieeexplore.ieee.org/document/598994.

[10] A. Sivanathan, H. H. Gharakheili, F. Loi, A. Radford, C. Wijenayake, A. Vishwanath, and V. Sivaraman, "Classifying iot devices in smart environments using network traffic characteristics," *IEEE Transactions on Mobile Computing*, 8 2018. https://iotanalytics.unsw.edu.au/iottraces.

[11] O. Alrawi, C. Lever, M. Antonakakis, and F. Monrose, "Sok: Security evaluation of home-based iot deployments," *IEEE Symposium on Security and Privacy (SP)*, pp. 208–226, 2019. https://yourthings.info/data.