Short Abstract: Towards a Better Understanding of IoT Domain Names

Ibrahim Ayoub
Afnic - Université Paris-Saclay

Martine S. Lenders
Freie Universität Berlin

Benoît Ampeau
Afnic

Sandoche Balakrichenan
Afnic

Kinda Khawam
Université de Versailles St-Quentin

Thomas C. Schmidt
HAW Hamburg

Matthias Wählisch
Freie Universität Berlin

***Abstract*—In this short abstract, we present our work-in-progress on studying IoT-related names. Our work aims for a better understanding of the characteristics of domain names commonly used to contact IoT backend servers. To analyze the properties of IoT backend servers names, we train a machine learning model to detect whether a given domain name is IoT-related (i.e., is resolved normally by an IoT device).**

Overview

The Internet of Things connects many resource-constrained devices to the global Internet. These devices frequently contact backend servers (*e.g.,* located in a cloud) to receive commands, deliver information, offload computation tasks, or receive firmware updates. To ease deployment and maintenance, those backend servers are not configured based on IP addresses but on names. The domain names are resolved via DNS [1] allowing the devices to connect directly to their intended server. Our work studies the domain names used for IoT backend servers. We construct two datasets: an IoT dataset, a dataset of domain names used by IoT devices, and a non-IoT dataset, which contains generic domain names.

Dataset of IoT names. We use public datasets that have been gathered in related work [2]–[4]. These datasets are based on packet captures collected in a testbed that consists of real IoT devices. The collected traffic includes DNS messages sent from and received by these devices.

We extract domain names from data gathered by IoTFinder, YourThings [2] and IoTLS [4] which were collected over two months (August 1, 2019 till September 30, 2019), over 10 days (April 10-19, 2018), and over two years (January 2018 till March 2020), respectively. The result is a set of 7415 unique domain names.

Dataset of non-IoT names. To construct the non-IoT dataset, we explored using the publicly available lists of top-visited websites. Such lists include Alexa, Cisco Umbrella, Tranco, and Majestic.

We pre-processed the resulting datasets by first testing them with a public resolver and discarding unresolvable domain names. We then studied their conformity to the most recent rules of domain names syntax used by Zonemaster [5], a tool developed by Afnic and the Swedish Internet Foundation to test domain names. Domain names that did not follow the syntax rules were discarded as well. This ensured that the resulting datasets after pre-processing contained legitimate resolvable domain names that respect the domain names' syntax rules. Moreover, this gave us an insight into the domain names either used by IoT devices or found in top visited websites lists and what percentage of these domain names is resolvable via public DNS or follows the syntax rules of domain names.

Analyzing statistical properties of domain names. The next step was to study the statistical properties of the datasets. For each list, we study the average, maximum, and minimum domain length and the average number of subdomains and other statistical properties. This allows us to compare different lists and helps identify which lists are statistically similar. This appears to be useful in the next step.

The next step includes training a machine learning model to classify domain names as IoT or non-IoT. One application of such a model is detecting whether an IoT device connects to a server whose domain name is not usually contacted by IoT devices and blocking the connection. This approach allows systems to be proactive and stop attacks before they happen. If a device attempts to connect to a domain name that is not IoT-related, the domain name should be treated as a phishing domain name, and the connection should be refused. A system adopting this policy, accompanied by high accuracy in detecting IoT and non-IoT domain names, is more secure and lighter in terms of performance than other approaches. Given the objective, the problem is a supervised learning binary classification problem.

We used Word2Vec as a word embedding technique to prepare our data for training. The goal of word embedding is to obtain real vectors representing the text data and use the real vectors to train the machine learning model. We used several algorithms, including Linear Regression and variations of Decision Tree. The results show good performance of the models. The values we obtain depend on the data we use as non-IoT as the statistical properties of domain names affect the model's performance.

Acknowledgements. This work has been supported by ANR and BMBF within the PIVOT project [6] (Privacy-Integrated design and Validation in the constrained IoT).

REFERENCES

[1] M. S. Lenders, C. Amsüss, C. Gündoğan, T. C. Schmidt, and M. Wählisch, "DNS over CoAP (DoC)," Internet Engineering Task Force, Internet-Draft draft-lenders-dns-over-coap-04, Jul. 2022, work in Progress. [Online]. Available: https://datatracker.ietf.org/doc/draft-lenders-dns-over-coap/04/

[2] "Yourthings data." [Online]. Available: https://yourthings.info/data/

[3] R. Perdisci, T. Papastergiou, O. Alrawi, and M. Antonakakis, "Iotfinder: Efficient large-scale identification of iot devices via passive dns traffic analysis," in *2020 IEEE European Symposium on Security and Privacy (EuroSP)*, 2020, pp. 474–489.

[4] M. T. Paracha, D. J. Dubois, N. Vallina-Rodriguez, and D. Choffnes, "Iotls: Understanding tls usage in consumer iot devices," in *Proceedings of the 21st ACM Internet Measurement Conference*, ser. IMC '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 165–178. [Online]. Available: https://doi.org/10.1145/3487552.3487830

[5] "Zonemaster: Requirements and normalization of domain names in input." [Online]. Available: https://github.com/zonemaster/zonemaster/blob/develop/docs/specifications/tests/RequirementsAndNormalizationOfDomainNames.md

[6] "Pivot." [Online]. Available: https://pivot-project.info/