22/02/2023

Towards a Better Understanding of IoT Domain Names

Workshop on DNS and Internet Naming Research (DINR2023)

Ibrahim Ayoub, Martine S. Lenders, Benoît Ampeau, Sandoche Balakrichenan, Kinda Khawam, Thomas C. Schmidt, Matthias Wählisch

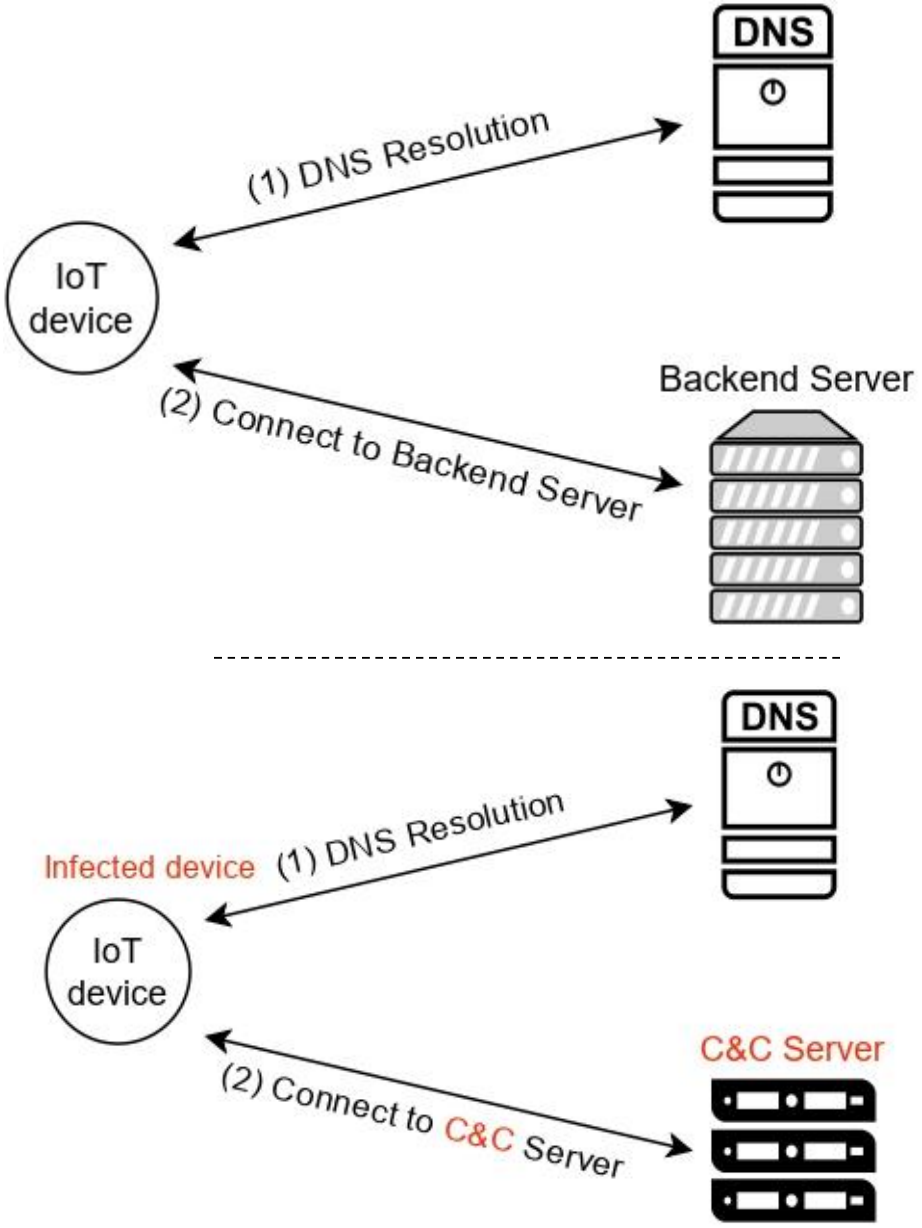
Motivation

1. **Security**

- IoT devices contact backend servers to receive commands & instructions, send and store data, receive software updates
- Malicious IoT devices contact C&C servers, which may exhibit different domain name properties

2. **Protocol design**

- Constrained IoT introduces different requirements compared to common Internet (e.g., smaller MTU, less memory)
- DNS-related protocol (e.g., DNS over CoAP**[1]**) benefit from a better understanding of IoT domain names



[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/

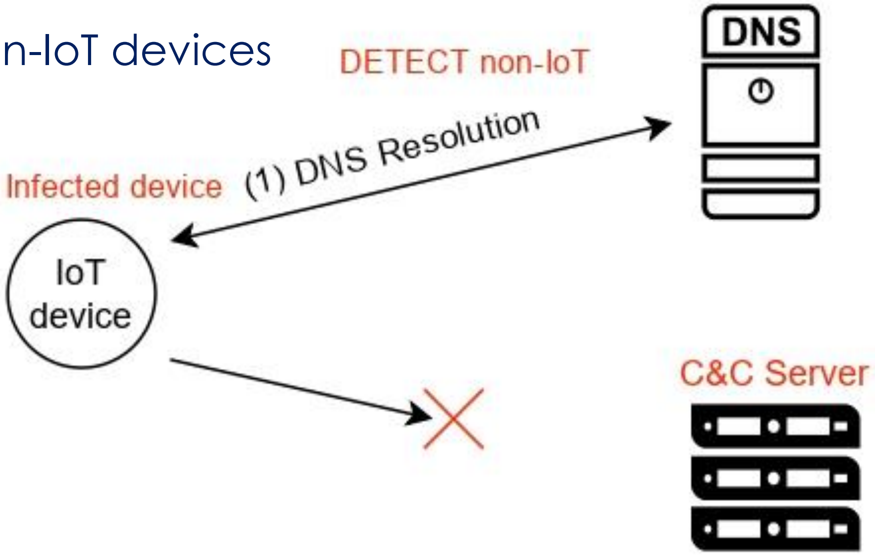
Objective and Approach

Objective

- Study statistical properties of domain names used by IoT devices to contact backend servers
- Compare different ML models

Approach

- Train Machine Learning models to classify between IoT and non-IoT devices
  - Detect non-IoT domain names during DNS resolution
  - Lightweight: raw domain names + label (IoT/non-IoT)



Analysis. Datasets.

IoT Dataset:

- Public datasets from *IoTFinder*, *Yourthings***[2]** & *IoTLS***[3]**
- Testbeds with real IoT devices
- Extract IoT domain names from DNS traffic
- Result: 7415 unique domain names

Non-IoT Dataset:

- Use several top-lists
  - The Cisco Umbrella 1 Million top domains
  - Majestic top 1 Million
  - Tranco top 1 Million

[2] "Yourthings data." [Online]. Available: https://yourthings.info/data/
[3] "Iotls: Understanding tls usage in consumer iot devices," in Proceedings of the 21st ACM Internet Measurement Conference. Available: https://doi.org/10.1145/3487552.3487830

Analysis. Processing Data.

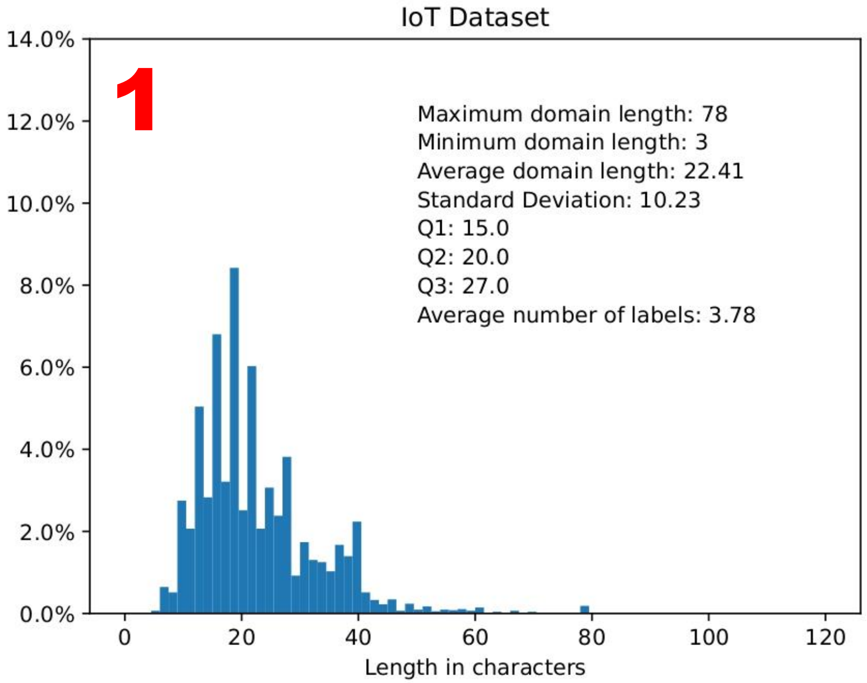
Cleaning Data:

- Resolving domain names and discarding the unresolvable
- Checking syntax (Zonemaster syntax rules**[4]**)

Statistical analysis:

- For each dataset we calculate:
  - Average, maximum, and minimum domain length
  - Average number of subdomains and other statistical properties

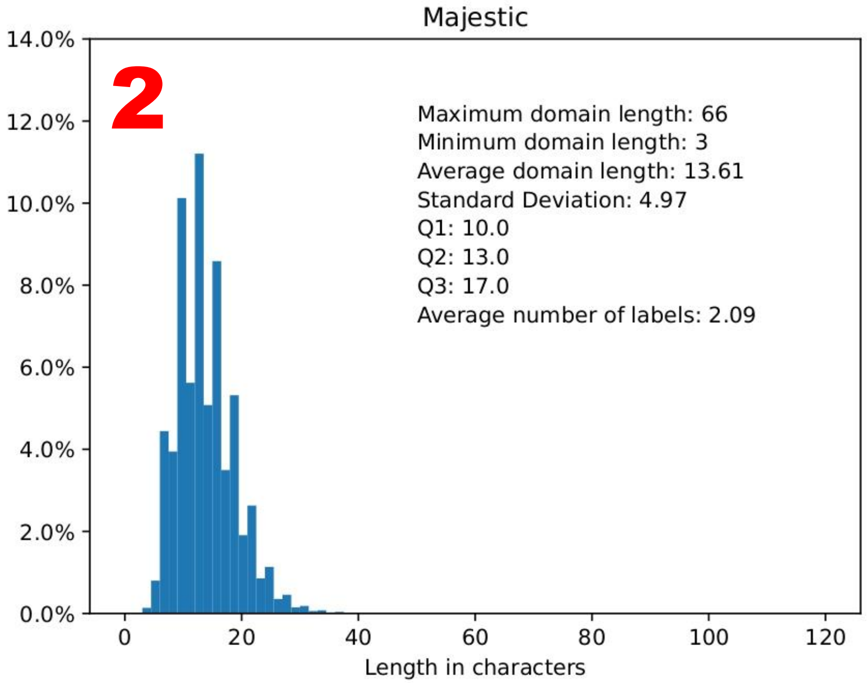
[4] “Zonemaster: Requirements and normalization of domain names in input.” [Online]. Available: https://github.com/zonemaster/zonemaster/blob/develop/docs/specifications/tests/RequirementsAndNormalizationOfDomainNames.md



IoT Dataset

1

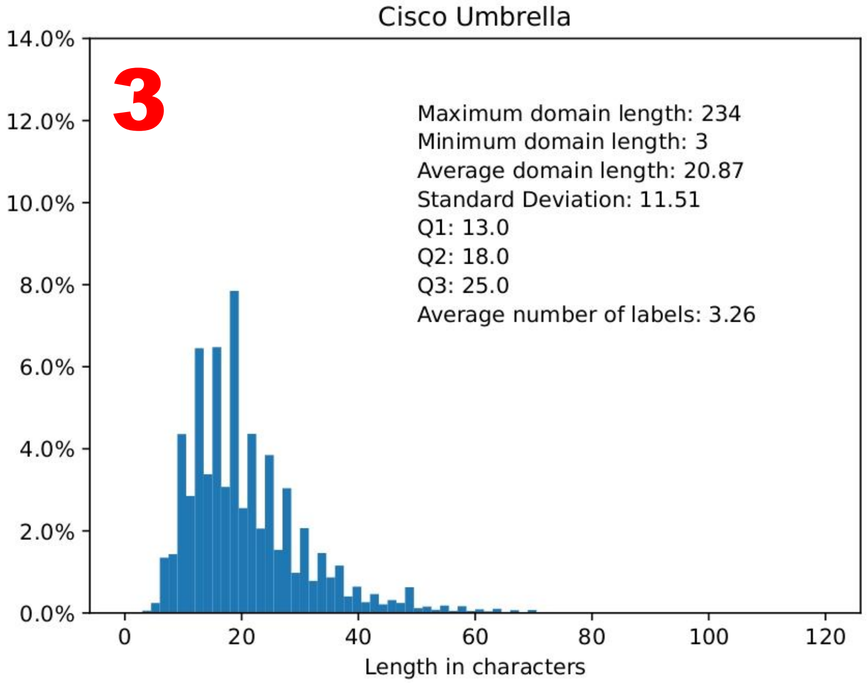
Maximum domain length: 78
Minimum domain length: 3
Average domain length: 22.41
Standard Deviation: 10.23
Q1: 15.0
Q2: 20.0
Q3: 27.0
Average number of labels: 3.78



Majestic

2

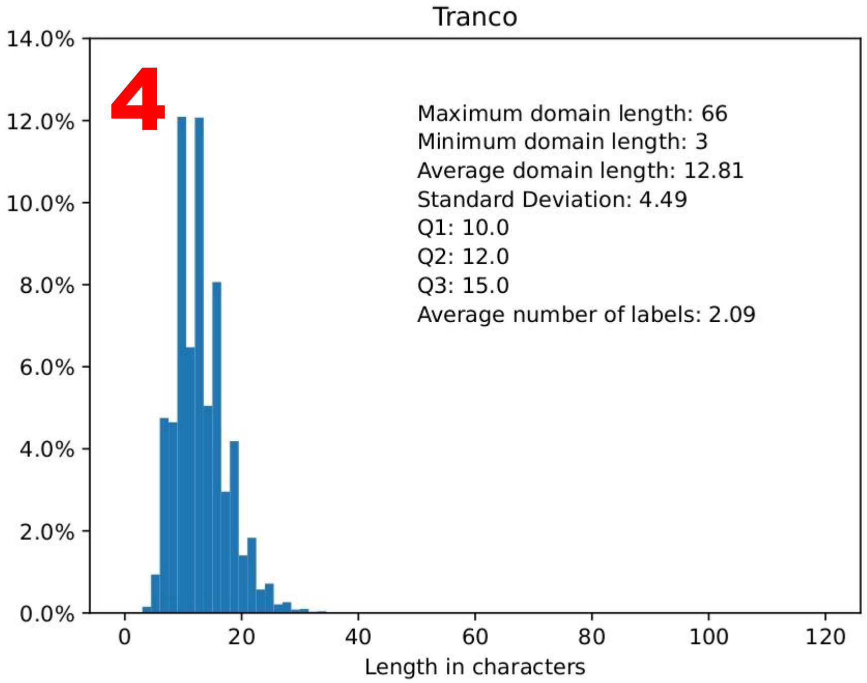
Maximum domain length: 66
Minimum domain length: 3
Average domain length: 13.61
Standard Deviation: 4.97
Q1: 10.0
Q2: 13.0
Q3: 17.0
Average number of labels: 2.09



Cisco Umbrella

3

Maximum domain length: 234
Minimum domain length: 3
Average domain length: 20.87
Standard Deviation: 11.51
Q1: 13.0
Q2: 18.0
Q3: 25.0
Average number of labels: 3.26



Tranco

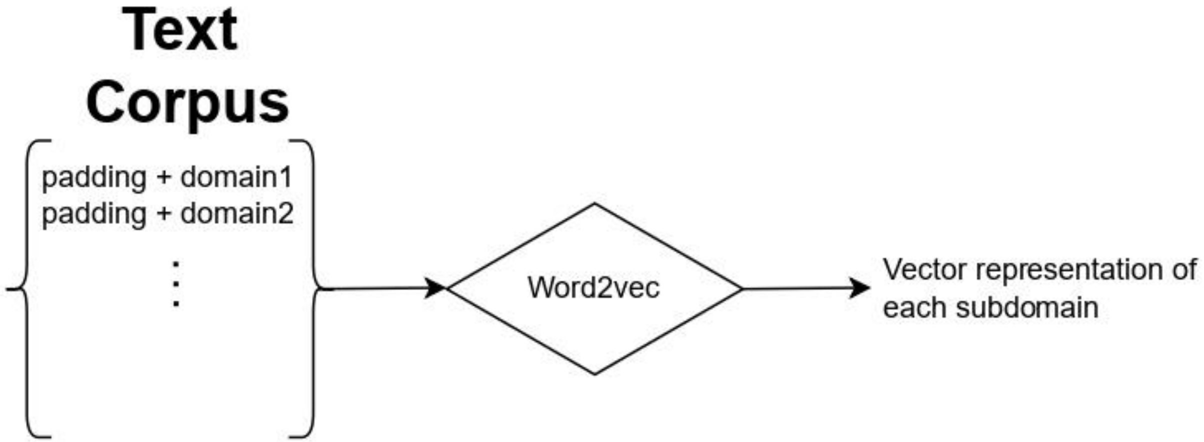
4

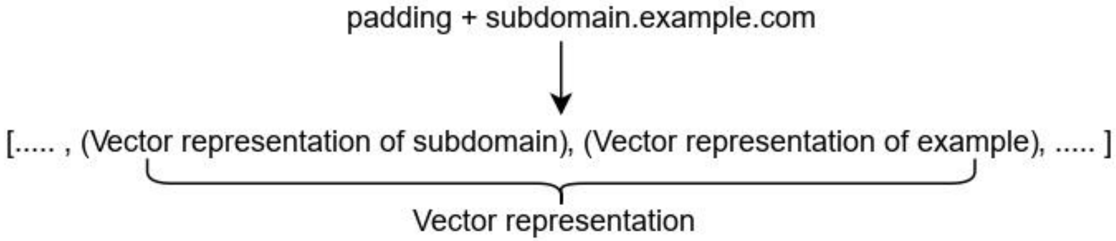
Maximum domain length: 66
Minimum domain length: 3
Average domain length: 12.81
Standard Deviation: 4.49
Q1: 10.0
Q2: 12.0
Q3: 15.0
Average number of labels: 2.09

Classifying Domain Names

- **Word Embedding:**
  - Word2Vec
  - Subdomains as words
  - Subdomain → Vector of size 32

- **Train several machine learning models:**
  - Linear Regression
  - Random Forest




padding + subdomain.example.com
[..... , (Vector representation of subdomain), (Vector representation of example), ...]
Vector representation

22/02/2023

Thank you!

ibrahim.ayoub@afnic.fr

afnic
Internet
made in France

Freie Universität Berlin

HAW
HAMBURG

UVSQ

université PARIS-SACLAY