Who are You Talking to? Machine Learning Based IoT Traffic Fingerprinting

Alice Martynova¹ – Zubair Shafiq²

Los Gatos High School¹ - Department of Computer Science, The University of Iowa²

Background & Motivation

What is an IoT Device?
A device that connects to the Internet to complete a certain task such as Fitbits, Nest security cameras, and Amazon Echos.

Security Breaching of IoT Devices
Billions of IoT Devices are in use, yet their security measures are lackluster, and hackers can view a user’s activity and deduce private information about the user.

Research Objectives & Methods

Research Objectives
1. Investigate effectiveness of Machine Learning to fingerprint devices
2. Filter through spatial data of IoT devices to find features that optimize results
3. Train and Test a Random Forest Machine Learning Algorithm on Weka Platform

Methods
1. Convert JSON file to ARFF, a Weka acceptable file format
2. Features are number of bytes sent to hostnames or domain
   a) Hostname popular -> use hostname
   b) Hostname rare -> use domain
3. Conduct supervised training/testing using ablation on Weka
4. Analyze results to see whether machine learning was able to classify by device type

Data/Features Used
- Minim provided data, device activity in real homes for 2 weeks
- Data included: IP addresses, ports, start time and end time, hostnames, bytes, etc.
  - Used spatial data
  - bytes sent to hostname, device ID, and device type
  - 69 device types
  - Number of Domains and Hostnames
  - 27k domains
  - 57k attributes for hostnames of top 100 domains and the rest of the domains

Key Terms
- Non-IoT Devices vs IoT Devices: Non-IoT perform lots of tasks, IoT perform one task
- Non-Normalized is total number of bytes sent to device vs Normalized is percentage of bytes
- Accuracies
  - Precision, positive predicted value
  - Recall, sensitivity

Results

1. Continue improving accuracy of Machine Learning
   a) Obtain more data(including temporal data) for non-IoT devices
   b) Use combination of first, last, and middle features
   c) Explore effects of number of hostnames
   d) Test other algorithms
   e) Obtain data for longer period of time

2. Combine other methods with Machine Learning to fingerprint 100% of devices
3. Apply software to Minim routers so users can gain security and insight on device behavior

Future Studies

- Select IoT Device Recall And Precision
  - Google Home/Plume had 100% accuracy
  - Sonos Play/Amazon Echo peaked at 1500 features
  - Recall/Precision are consistent for Plume and Echo
  - Precision lower for Sonos Play and Google Home
    - Sometimes confused with other devices
  - Recall For All Devices
    - Peak for normalized and non-normalized at 1500
    - No huge disparity between normalized and non-normalized
  - Recall for IoT Devices
    - Peaked at 1500, consistent with IoT Device Recall and Recall for All Devices
    - Higher accuracy than recall for all devices
  - Non-IoT Devices
    - No consistent peak
    - Much lower accuracy than IoT Devices
    - Apple iPhone very inconsistent
    - Too much data from apple phone

References


Acknowledgements

Zubair Shafiq, Department of Computer Science, The University of Iowa, USA
Basil Chalita, Department of Computer Science, The University of Iowa, USA
Ms. Hamrad Manal, Department of Computer Science, The University of Iowa, USA