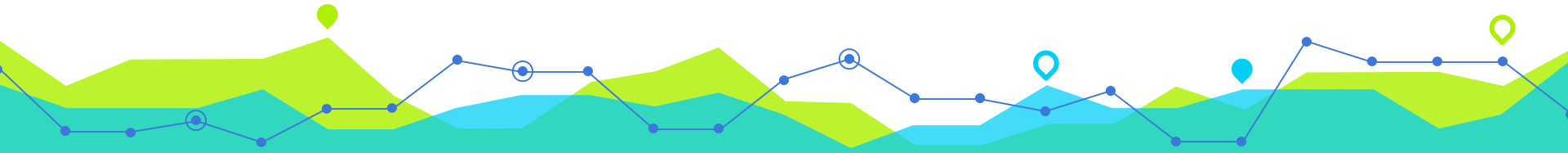


Team 05: Network Visualization of IoT Devices

Tracey Chen, Bernice Meja, Jessica Yang

Our Motivation

In recent years, campuses have seen a spike in IoT devices. IoT (Internet of Things) are “smart” devices that can connect to the internet but do not include laptops, smartphones, and smart watches. The influx of IoT devices poses new security threats. Duke OIT wishes to have a complete and accurate overview of the types of devices on the Duke network, to improve device management and security, by quickening the process of device identification.

Objective

Identify and label devices on a network

Features

avgSleep	avgActive	totalActive	totalByte	avgByte	avgFlow	uniqIP	uniqPort	DNS	NTP
82.94	118.71	12464.52	565840	2815.12	8695.79	53	4	264.44	21600
90.73	142.48	13107.74	176331	1160.07	2073.00	33	4	334.63	21600
46.63	69.91	12863.91	17317443	22173.4	1656469	150	4	34.70	21600
91.26	137.61	12797.36	295116	1844.48	5982.88	42	4	384.35	21600

Table 1: Feature information extracted from network data

Each line is the information over a six-hour period, for one device

avgSleep average duration of the device being inactive
avgActive average duration of the device being active
totalActive total duration of the device being active
totalByte size of all package delivered
avgByte average size of packages delivered
avgFlow average size of packages delivered and received
uniqIP number of unique destination IP’s from the device
uniqPort number of unique destination ports from the device
DNS average interval between each DNS query
NTP average interval between each NTP query

Our Approach

Data Collection and Visualization A lab at Duke OIT was set up, containing 10 IoT constantly turned-on devices, as well as 6 personal, non-IoT devices. The network data was captured using Argus Spark, and kept as csv files. Various features of the raw data were graphed.

Data Collection and Visualization

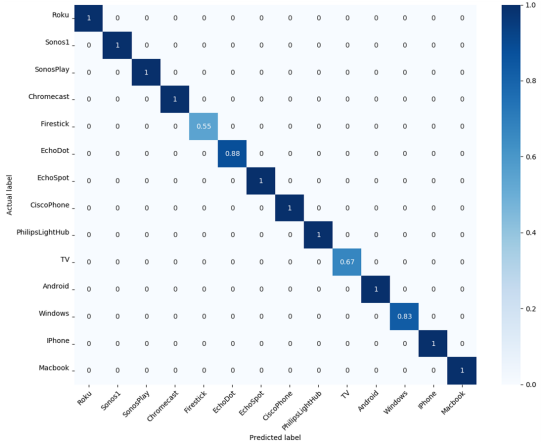
Features and Model Creation

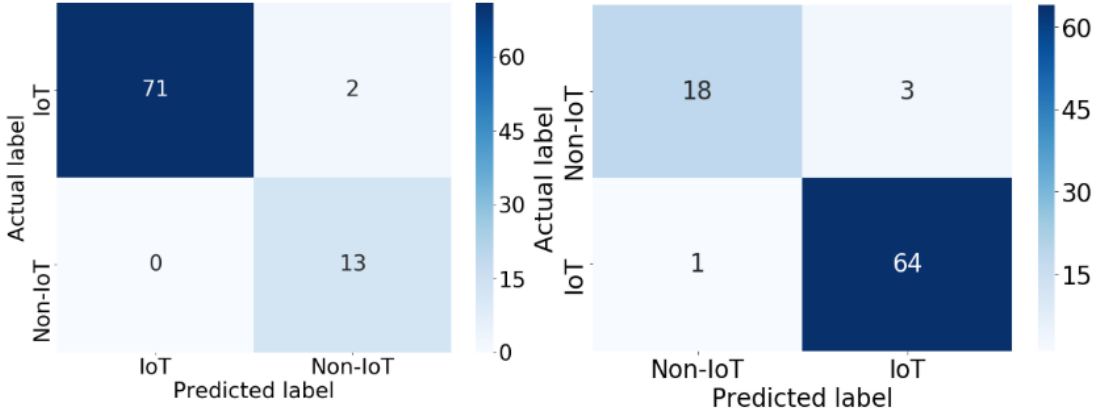
Feature Extraction An automated script was written to automatically extract the following features from the raw network data: average sleep time, total active time, average active time, average time between DNS queries, average time between NTP queries, average package size, total flow volume, number of unique destination ports, number of unique destination IP's.

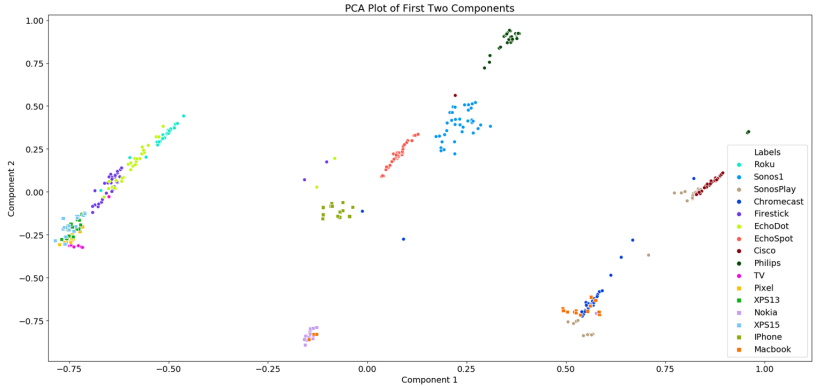
Modelling Unsupervised learning methods—PCA and k-means clustering—were performed to visualize and identify the devices' natural clusters. Supervised learning methods—logistic regression and random forest—were performed to determine if the devices could be correctly classified with binary labels (as IoT or non-IoT), and with multi-labels (as individual devices).

Mass Application

Our Models







Top Left: Random Forest Confusion matrix for individual devices

Top Right: Random Forest and Logistic Regression Confusion matrices for IoT vs Non-IoT

Bottom: PCA Plot for first two components

Conclusion

With our current data, our model can identify individual devices if it has already been encountered; otherwise, it can determine if a device is IoT or non-IoT.

Our major limitation in this project is the lack of data. We suffer from a small data problem, as we collected data using only 16 unique devices, merely 10 of which were IoT devices. In addition, the lab data is not representative of real-world data, as shown in the tabel to the right.
Learning to apply our model to the mass, real life network is an on-going challenge.

Future work for this project includes understanding why the models are as successful as they are, and determining how this model performs with real-world data.

Lab Data	Real World Data
Static device IP	*Dynamic* device IP
16 unique devices (10 IoT)	*Countless* unique devices
Artificial and systematic device usage and data collection	*Organic and random* usage

Duke OIT

We're in IT together