Progress Report

Current work: Using DNS data to find and separate IoT device streams [Nov 2016 onwards]

- **Outline:** http://www.cs.princeton.edu/~sgrover/reports/Outline.pdf
- Nov 2016: conceived the idea for "DNS data for real time IoT query system" and emailed folks at Comcast to discuss it in Dec 2016.
    - The response of the Comcast team was lukewarm as they were looking for a pure security solution while we presented the idea of a streaming platform. We decided to work on Princeton data from OIT to show our analysis before asking for private DNS data from Comcast again.
- Nov-Dec 2016: Analyzed Comcast data traces that were collected during the summer intern. Code: https://github.com/shahifaqeer/iot-dns
    - Concocted and tested three hypothesis to identify IoT devices in DNS traces from ISPs.
        - Hypothesis 1: There will not be many homes with IoT related DNS queries, i.e. spatial prevalence of IoT domains is lower in aggregate data and vice versa.
        - Hypothesis 2: Homes with IoTs will have similar query profiles, i.e. spatial correlation of IoT domains among filtered unpopular domains will be higher.
        - Hypothesis 3: Each IoT will perform at least 1 regular DNS lookup (heartbeat)
    - Results showed that our hypothesis worked but the traces were too short for proper testing of Hypothesis 3 (background data). Decided to collect data from OIT at Princeton university.
- Jan-Feb 2017: Collected and analyzed long traces (10 days) of DNS data from Princeton University.
    - Got IRB approval for the project and started collecting data at OIT
    - Created scripts to split and clean up the extremely large traces.
    - Gained access to ns1 cluster and set up an environment where large traces can be analyzed in parallel using spark.
    - Also met with Christine Murphy and started visiting CPS at this time.
- March 2017 onwards: Started analysis of short section of DNS traces. Code: https://github.com/shahifaqeer/oit-dns
    - Analyzed number of DNS domain queries and unique queries per IP address to filter to devices expected to be IoT devices (link to ipynb)
- March-April 2017: Analysis of 1 night of DNS data using pandas in ipython notebook
    - Analyze background DNS data pattern for a fitbit device
    - Extracted time-difference features to distinguish between background queries and bursty queries. However this technique did not work too well and we decided to come up with a better method to find regular DNS queries.
    - Algorithm tested on 1 night data and short summary of results

            - **Related doc:** **http://www.cs.princeton.edu/~sgrover/reports/Summaryof1nightanalysisforfitbitdevice-April2017.pdf**
- April 2017: Frequency and period for DNS domain queries and feature extraction
    - Previous approach based on time-difference was not applicable to general cases where we want to separate background data (with certain periods) from bursty DNS traffic
    - Proposed an approach to use autocorrelation to find periods, however this didn't work for normal DNS data. Some domains were too sparse, whereas in other cases queries that were a few seconds off could not be captured using autocorrelation. Edge detection to get the right autocorrelated period didn't work.
    - Extracted features for each device to run simple clustering algorithms https://github.com/shahifaqeer/oit-dns/blob/master/feature_extraction/pcap-feature-extractor.ipynb
        - **Related doc:** **http://www.cs.princeton.edu/~sgrover/reports/FeatureExtractionforML-April2017.pdf**
    - Arpit/Joon helped to extract hostdb data to combine with each device features. This gives us some ground truth regarding whether the device is a VM, a wireless device, or wired computer. https://github.com/shahifaqeer/oit-dns/tree/master/data_extraction
- May 2017: Solved frequency extraction for IoT devices in lab (link) and started working on Machine learning for a 1 hour sample of DNS data
    - As domain timing are not uniform, directly using FFT to find periods is not possible
    - Algorithm to uniformly sample DNS data using binning, and then extracting period using periodogram function
    - Applied algorithm to ACTIVE-DATASET collected earlier in the year.
        - **Related doc:** **http://www.cs.princeton.edu/~sgrover/reports/Active-dataset-DNS-analysis-May2017.pdf**
    - Read up and tried DBSCAN algorithm over extracted data (no results yet)
- **Next Steps to complete project:**
    - Refine frequency/period extraction algorithm to add frequency related features to OIT data [Expected time 1-2 days]
        - Currently can't deal with the case where a device uses phones to communicate with the server. Example: fitbit
    - Try DBSCAN to cluster devices using number of domains, timing, and frequency features. Use hostdb and DHCP data as ground truth [Expected time 1 week]
    - Apply above filtering and machine learning approach to complete database in Spark with the aim of identifying IoT related domains and their corresponding IP addresses. [Expected time 1 month]

Summer Intern at Comcast: Analyzing the Attack Surface of Consumer IoT Devices in the Comcast Network [June - Sept 2016]

- Final intern report link (do not share): http://www.cs.princeton.edu/~sgrover/reports/FinalInternReport-Grover.pdf
- June 2016 - obtain DNS traces from ISP edge router
- July 2016
  - Analyze individual devices and get a refined list of domains to extract DNS data.
  - Use DNS to filter homes with IoTs to collect and analyze only data from these devices
- August 2016
  - Final analysis to find insecure or malicious devices based on communication patterns

Active IoT Data Collection: Long term IoT device data collection [Jan - Feb 2016]

- Collected the following long term background traces

	Device	**Collection Time**	**Activity Mode**
D01	Nest Thermostat (10.0.0.7)	45 hrs	Normal Home Use
D02	Amazon Echo (10.0.0.4)	16 hrs	Background
D03a	Nest Dropcam (10.0.0.9)	24 hr	Background/No video
D03b	Nest Dropcam	13 hr	Active Video
D03b	Nest Dropcam	2 hr	Active Video + Viewstream
D04a	Sharx Security Camera (10.0.0.8)	24 hr	Background/No video
D04b	Sharx Security Camera	13 hr	Active Video
D04c	Sharx Security Camera	2 hr	Active Video + Viewstream
D05a	SmartThings (10.42.0.89)	4 hrs	Background
D05b	SmartThings + Door Sensor	14 hrs	Background

D05c	SmartThings + SmartSocket	10 hrs	Background
D05d	SmartThings + Door Sensor + SmartSocket	14 hrs	Background

- TA for Fog IoT coursera lectures with Prof. Mung Chiang

SpookyScan [Oct 2015 - June 2016]

- Jan 2016: Spooky scan new results and analysis - multiple meetings with Paul/Roya
- Feb 2016: worked on paper skeleton for spooky scan with Roya - responsible for certain analysis sections
- April 2016: worked on spooky scan for CCS deadline but had some family problems at this time and had to leave for India
- May 2016: hypothesis testing and stationarity analysis
- June 2016: Comcast Intern and progress on spooky scan stalled. Was informed later that the project has been dropped.
- Related code: https://github.com/shahifaqeer/spooky-analyzer

PrivacyCon: IoT device active and background state pcap analysis [Oct - Dec 2015/Summer 2015]

- Link to PrivacyCon 2016 presentation: http://www.cs.princeton.edu/~sgrover/reports/PrivacyCon.pptx
- Oct - Dec 2015: Security analysis of short term IoT traces collected in the lab
  - Security bug in nest thermostat updates that might show location data
  - User activity can be easily identified based on throughput even for secure devices
- Summer 2015: buying devices and setting up a collection infrastructure in the lab using a lenovo laptop as a router
  - Helped by Michele Mangili

MySpeedTest Data Analysis [Dec 2015 - Jan 2016]

- Analysis of MST data with Ava Chen

Comcast ISP bandwidth project: A Case Study of Traffic Demand Response to Broadband Service-Plan Upgrades [Jan - May 2015/Sept-Oct 2015]

- Apr 2015: Initial results http://www.cs.princeton.edu/~sgrover/reports/20150421-comcast-analysis-meeting.pdf
- Submitted to IMC and received a reject in Aug 2015
- Sept-Oct 2015: reanalyze data and fix paper for resubmission http://www.cs.princeton.edu/~sgrover/reports/PAM16_26-grover.pdf
- April 2016: presented at PAM http://www.cs.princeton.edu/~sgrover/reports/grover-Traffic_demand_response.pptx
- Oct 2016: presented at FCC
- Aug 2016: presented at CableLabs