Classification and Identification of IoT Devices

Abstract

In this day and age of technological advancement, not many days from now on, we will see our future to be fully equipped with IoT devices. The IoT devices will have major impact on our lives and become indispensable part of us. There lies the biggest threat. As number of devices will increase, the chances of their exposure to the attack will also be more. So, protecting IoT devices from the cyber attacks has become the major duty to be taken care of.

In this report we have done a detailed comprehensive assessment and summarisation of two papers [1] and [2] and as a part of our curriculum we have tried to implement few ideas from these papers. The paper [2] is a continuation of paper [1]. In this paper the authors work on the limitations of previous papers which is based on the fact that existing proposals may inspect the packet payload, and this creates risks to IoT users' privacy. This also increases computational complexity for classification and extraction of traffic characteristics.

Introduction

The first paper [1] proposes the use of network traffic analytics to characterize IoT devices, including their typical behaviour mode. The authors after collecting data across 20 diversified IoT devices over a period of 3 weeks, then analyze the traffic traces to characterize statistical attributes such as data rates and burstiness, activity cycles, and signalling patterns. Using these attributes, they develop a classification method that can first distinguish IoT from non-IoT traffic and then identify specific IoT devices with over 95 percent accuracy. Their study is helpful in empowering operators of smart cities and campuses to discover and monitor their IoT assets based on their network behaviour which is extreme helpful in campusses and cities.

The second paper[2] which has taken inspiration from [1] presents an approach which uses packet length statistics from encrypted traffic to characterize the behaviour of IoT devices and events. in a smart home scenario.The solution to characterize IoT devices and events is evaluated with traffic from two real-world testbeds and five classifiers. The evaluation included the algorithms k-Nearest Neighbors (k-NN), Decision Tree, Random Forest, Support Vector Machine (SVM) and Majority Voting. The results showed that the Random

Forest algorithm performed best among the five and could achieve up to 96 percent of accuracy in the identification of devices, 99 percent of precision in distinguishing between IoT and non-IoT devices and 99 percent of accuracy in the identification of IoT device events. It was seen that decision tree had lowest latency after k-NN, Random Forest, SVM and Majority Voting.They have used hypotheses testing to validate the results.

Their solution uses central tendencies such as statistical mean, the standard deviation and the number of bytes transmitted over a one-second window, which can be extracted from the encrypted traffic. This makes the use of TCP vectors redundant. This solution even identifies IoT devices and events, such as voice commands to smart assistants, and also distinguishes between IoT and non-IoT devices like the previous paper [1].

Important Terms

K nearest neighbour

This is a supervised machine learning algorithm that is used to derive nearest k neighbour for a particulat decision. This is also hepful in recognising patterns and estimating statistical features.

Logistic Regression

This is a model which takes the help of logistic function for modelling variables that are dependent.This also helps immensely in classification.

Naive Bayes

This classification approach is based on Bayes theorem where it assumes that the features or variable in a system are independent of each other.

Random Forest and Decision Trees

These two belong to supervised Machine learning algorithms that are non linear.During the time of training random forest constructs multitude of decision trees.Decision trees as the name suggests uses models similar to trees for showing outcomes.

Support Vector Machine

This is a set of supervised learning models which are effective in classifying in higher dimensional spaces. This classification approach is a takes the help of hyper plane to classify the data points.

Literature Survey

Characterizing and Classifying IoT Traffic in Smart Cities and Campuses

The authors work is the one of the first papers to profile, characterize and classify IoT devices in smart environments in an organised manner.There is a large body of work characterizing general Internet traffic. However, studies focusing on characterizing IoT traffic (also referred to as machine-to-machine – M2M – traffic) As most of the studies dealing with the characterizing IoT traffic are still in their infant stage., the authors motivated by the need to understand whether Machine to Machine(M2M) traffic imposes new challenges for the design and management of cellular networks.

They hosted their experimental at their campus facility which comprised of a large range of IoT devices that emulates a “smart environment”. They have taken into consideration, 21 unique IoT devices and extracted their traffic traces for over 21 days in a campus.

All the traffic on the LAN side was collected using the tcpdump tool was used to get all the traffic on the LAN side which runs on OpenWrt . Capturing the pre-NAT traffic helped them to map packets to specific devices directly. The MAC addresses in the packet headers shows the identity of every devices . Then they programmedd a script which was used to make the process of data collection and storage automatic. Finally the resulting traces were stored on an external hard drive which was attached to the gateway as pcap files .After getting the pcap files the researchers started logging all network traffic in our smart environment from 23-Sep-2016. The process of data collection and storage for each day began at midnight using Cron job on OprnWrt. They developed a monitoring script on the OpenWrt so that data collection and storage would proceed without any problems. By setting up an Apache server on a virtual machine (VM) in their university data center they made the data publically avaiable and progammed a script so that the trace data from the previous day wil be periodicaly transferred and stored on the hard drive, onto the VM.

In order to clean the attributes from a traffic they converted the raw pcap files into flows reguarly. For a specific IoT device, all the flows associated with that device was aggregated, which helped them to study the probability histogram of the sleep time attribute and observe that there is a unique pattern for some IoT devices. For example, sleep times of 90, 60 and 20 seconds occur respectively for the HP Printer, iHome switch and Netatmo welcome camera with probability more than 70 percent. They also observed that some devices exchange a unique volume of data for the most part during their active periods. For example, during those two weeks Samsung SmartThings, Samsung SmartCam and Netatmo weather station consistently exchanged 114, 3341 and 342 bytes during their active periods. Moreover, some devices such as Withings smart scale, Netatmo weather station and SmartThings exhibit signatures in terms of the average packet size; 225, 200 and 75 bytes respectively.

To make the readers visualize the role played by each of these attributes, the auhors have applied the K-Means clustering algorithm to the attributes across all the IoT devices using the Weka tool. They have used binning i.e five bins for each attribute. A smaller bin count was not effective in identifying the unique fingerprint that underpins each IoT device. Choosing a larger bin size renders the visualisation too onerous, while adding little additional insights.

Finally they observed the activity patterns,protocols and signaling patterns and characterised accordingly. After which they devised a classification technique which differentiated between IoT and non IoT devices in addition to identifying it over 95 percent accuracy.

Identifying IoT devices and events based on packet length from encrypted traffic

These two papers propose a solution for classifying encrypted IoT traffic based on packet length statistics. The solution takes three network traffic statistics into consideration to identify the devices and events: the mean and standard deviation of the packet length, and the num- ber of bytes transmitted by each device. The statistics enabled the classification of encrypted traffic.

The dataset taken by them was same as the [1], where there was extracted data of 21 IoT devices for a period of 3 weeks. In addition to this the researchers also took three consumer IoT devices and developed a testbed. Their proposed solution better performance than the [1] in distinguishing IoT and non-IoT devices, obtaining a minimum precision of 96 percent on random partitioning where was ti was 95 percent in earlier case. They have also achieved accuracy of 94 percent in identifying IoT devices using random partitioning. Finally, realizing up to 99 percent of accuracy on both random and chronological partitioning, the proposed solution accurately identified the device events. Using decision tree classifier they could achieve latency of 0.034 ms which is considered very fast generally. So, It was observed that while random forest was the best algorithm for identification, the decision tree out performs it in terms of speed. The Random Forest was the third fastest algorithm. But as it is an ensemble algorithm, the authors identified that there is a trade of accuracy and latency in IoT classification.

Implementation

We have tried to implement the above papers partially on the part of our curriculum. We have taken the datasets from the site iotanalytics.unsw.edu.au/iottraces which is the original datset that was created by the authors of [1]. First we have used Wireshark to access the pcap files. We converted the pcap files into csv using scapy which is one of the powerful packet manipulation program in python. There were 10 attributes such as PacketId, Time, Size, SourceMac, DestMac, SourceIP, DestIP,protocol, sourcePort, DestPort. After getting the csv files. The first step was explaratory data analysis of the data. We have extracted the data and then data cleaning is done to remove the attributes which are not necesssary.

The result was stored in output.csv file. In the cleaned file there were 18 entries and 4 attributes such as MacId, mean, median, mode of the packet length. This csv file made it clear that there were 18 devices active on that particular date. We have calculated central tendencies of the packet length as it would help us in distinguishing between the IoT devices. We plotted graphs such as 2D scatterplot,histogram,bargraph using matplotlib which helped us visulising data in better way. In the paper[2],it was mentioned As mentioned in the paper [2] where the authors have tried 5 different classifiers we have also implemented random forest classifier technique which is one of the 5 classification technique used by the paper[2]. We have considered 3 cases taking 10, 100, 1000 uncorrelated trees and the accuracy of them came out to be 0.9999003214632809,0.9999873214890180, 0.9999875401829101. This also underlines the observation of paper[2] which said random forest are the performs best with above 99 percent accuracy.

Then we have developed a graphic user interface(GUI) which combines all the above modules and represents them in a better way.The tkinter package (”Tk interface”) is the standard Python interface to the Tk GUI toolkit. It provides a powerful object-oriented interface to the Tk GUI toolkit fast makes it easy to create GUI applications. We have integrated GUI to the graphs where we can choose what features we want as x axis and y axis after choosing the dataset.

Conclusion

In this report we have done a compact and systematic analysis of research papers which identifies and classifies the IoT devices based on their packet length.Though there has been tremendous increase in the sage of IoT devices in the recent times in the campus and cities, the operators generally doesnot posses the good understanding of how IoT devices are connected to their networks, what the traffic profiles look like and if there has been breach in the securities.

The results of both the papers have revealed some of the important and intresting opporutnities for future research purposes. These include isolation of specific devices for security purposes,prioritization of IoT traffic, identification of abnormal events etc. These field can be explored in future works.

References

[1]. A. Sivanathan, H.H. Gharakheili, F. Loi, A. Radford, C. Wijenayake, A. Vish- wanath, V. Sivaraman, Classifying iot devices in smart environments using network traffic characteristics, IEEE Trans. Mob. Comput. (2018) 1, http://dx.doi.org/10.1109/TMC.2018.2866249.

[2]. Antônio J. Pinheiro, Jeandro de M. Bezerra, Caio A.P. Burgardt, Divanilson R. Campelo, Identifying IoT devices and events based on packet length from encrypted traffic, Computer Communications, Volume 144, 2019, Pages 8-17, ISSN 0140-3664, https://doi.org/10.1016/j.comcom.2019.05.012.