STUDY PROJECT

CS F266

Submitted to: Dr. Vishal Gupta

Darknet Insights using R and Python

December 2019

PREPARED BY

SHRAMAY PALTA (2017A3TS0340P)

Project Problem

In layman's terms, Dark Net (or Darknet) is an umbrella term describing the portions of the Internet purposefully not open to public view or hidden networks whose architecture is superimposed on that of the Internet.

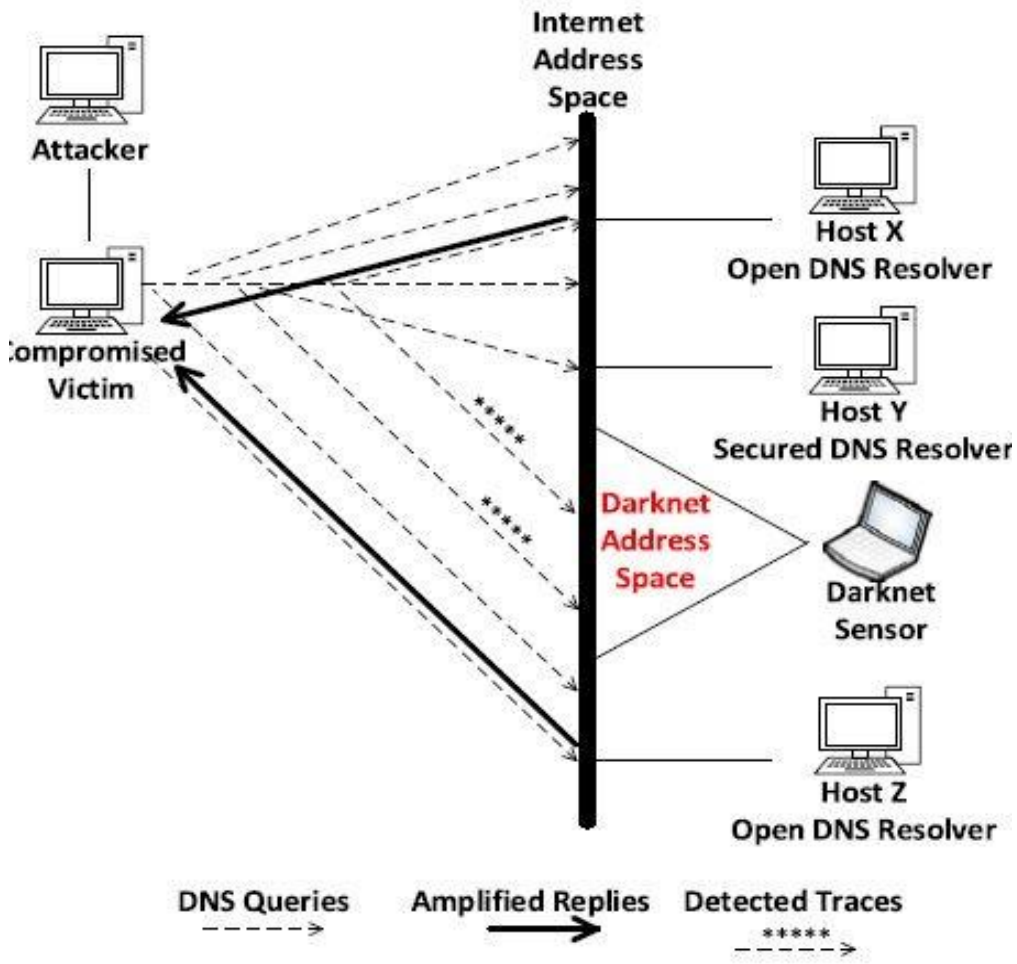
From the point of view of this project, we basically try to predict whether a Distributed Denial of Service (DDoS) Attack is happening or not based on some specific set of data sourced from the Center for Applied Internet Data Analysis(CAIDA) supercomputer servers of the University of California, San Diego (UCSD) and using their Corsaro tool.

AIMS AND OBJECTIVES

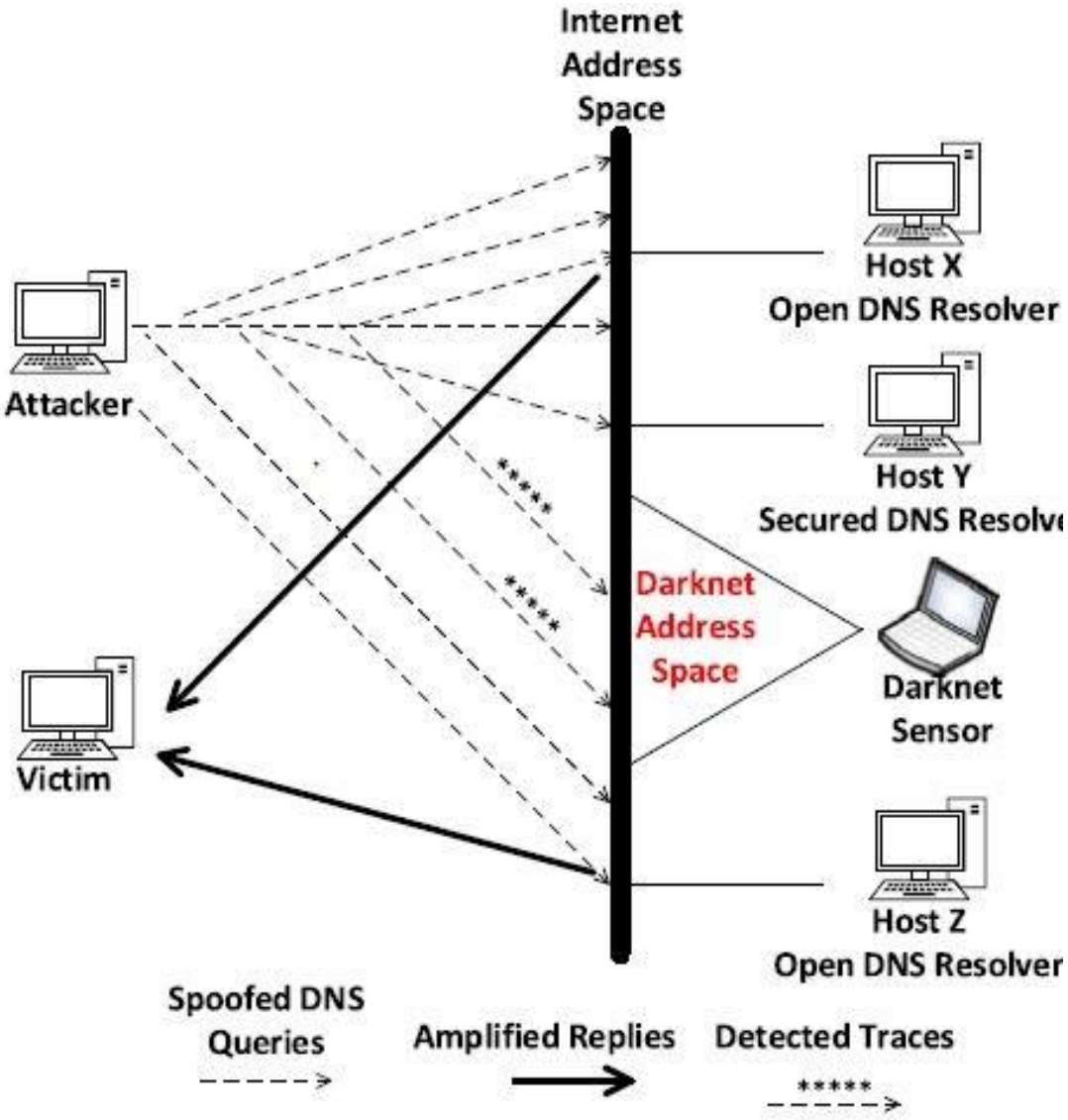
Throughout the duration of this project, our main aim is to try and determine parameters of importance that can be used to train an algorithm to determine whether a DDoS attack is happening or not.
We would then seek to draw out some useful information using some clustering algorithms like K-means and EM algorithm.

We have studied a variety of models from various sources to help understand what a DDoS attack is. Some of the Models that we have studied are:

1) Compromised Victim Scenario



2) IP Spoofing Scenario



In our pursuit relevant and important parameters for prediction, We consider a very specific scenario:

- We analyze only those packets whose either source or destination ports are equal to port number 53, which is mainly used by Domain Name System (DNS) as TCP.
- We try to determine the intensity of the DDoS attack by trying to incorporate the importance to some primary factors like IP length, number of packets and TTL (Time to Live).

A majority of the work of this project is based on the paper “Inferring Distributed Reflection Denial of Service Attacks from Darknet”, wherein the authors have first attempted to generate a flow and then used it as a detection parameter that is used in rate classification. Finally, clustering algorithms are used.

Dataset Used:

In our work, we have used the CAIDA Darknet Dataset which monitors traffic on the /8 part of the internet $(1/256)^{th}$. The dataset consists of flow tuple data of the form:

67.8.185.185|44.246.197.47|80|40123|6|45|0x12|44,1
67.8.185.185|44.153.198.47|80|50717|6|45|0x12|44,1
67.8.185.185|44.176.201.47|80|43155|6|45|0x12|44,1
67.8.185.185|44.190.202.47|80|56481|6|45|0x12|44,1

Here the first field represents the source IP, followed by the destination IP, source port, destination port, protocol number (for eg. 6 above represents TCP), flags, TTL, IP length and number of packets.

We have used the CAIDA designed tool CORSARO to extract the data. Corsaro helps in performing large scale analysis of trace data.

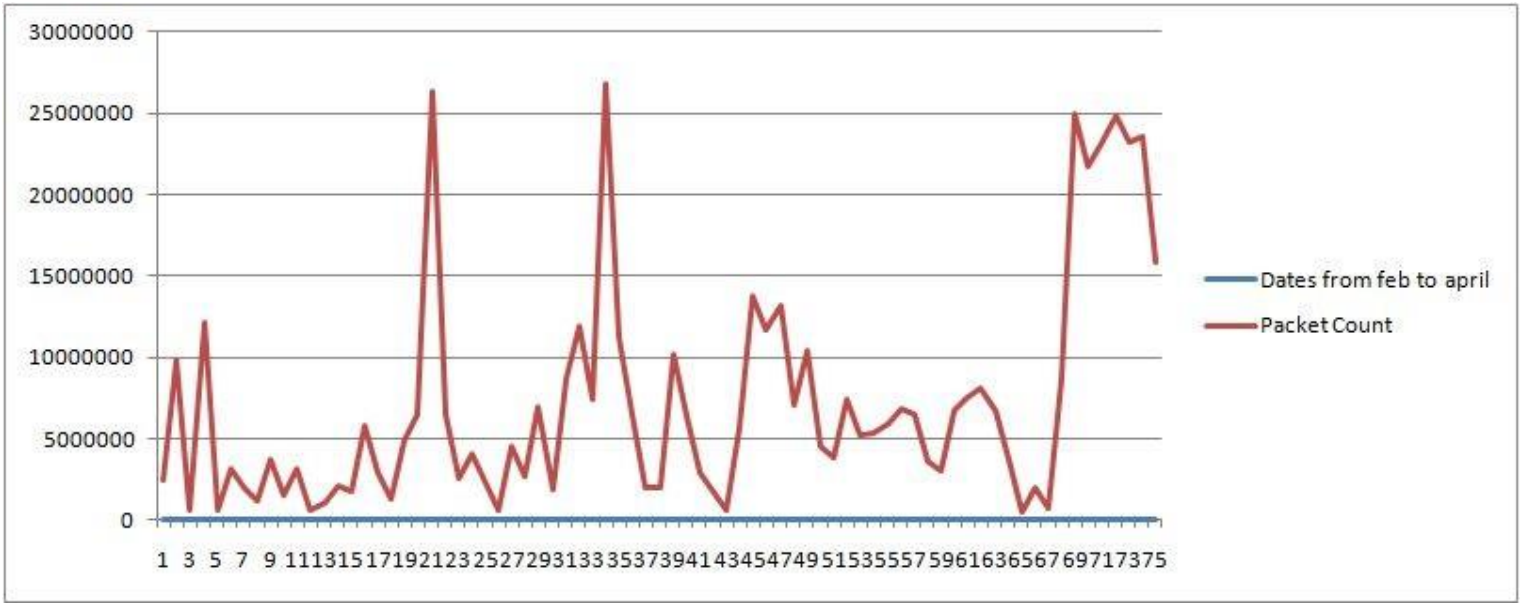
TASKS PERFORMED SO FAR:

- Studied research papers to try and grasp what a DDoS attack is.
- Became familiar with FreeBSD and the Corsaro tool.
- Used Python Scripts to extract data from a dataset spreading over three months which was further filtered on various attributes.
    - Source Port 53 and Packet Count
    - Destination Port 53 and Packet Count
    - Source Port 53 and IP Length and Number of Packets
    - Destination Port 53 and IP Length and Number of Packets
    - Source Port 53 and TTL and Number of Packets
    - Destination Port 53 and TTL and Number of Packets
- This data was further divided on the basis of protocol used i.e. TCP or UDP or IPV4.
- Although this data had been created on an hourly basis, we have implemented Python Scripts to aggregate this date over each data so as to better understand the attacks taking place.

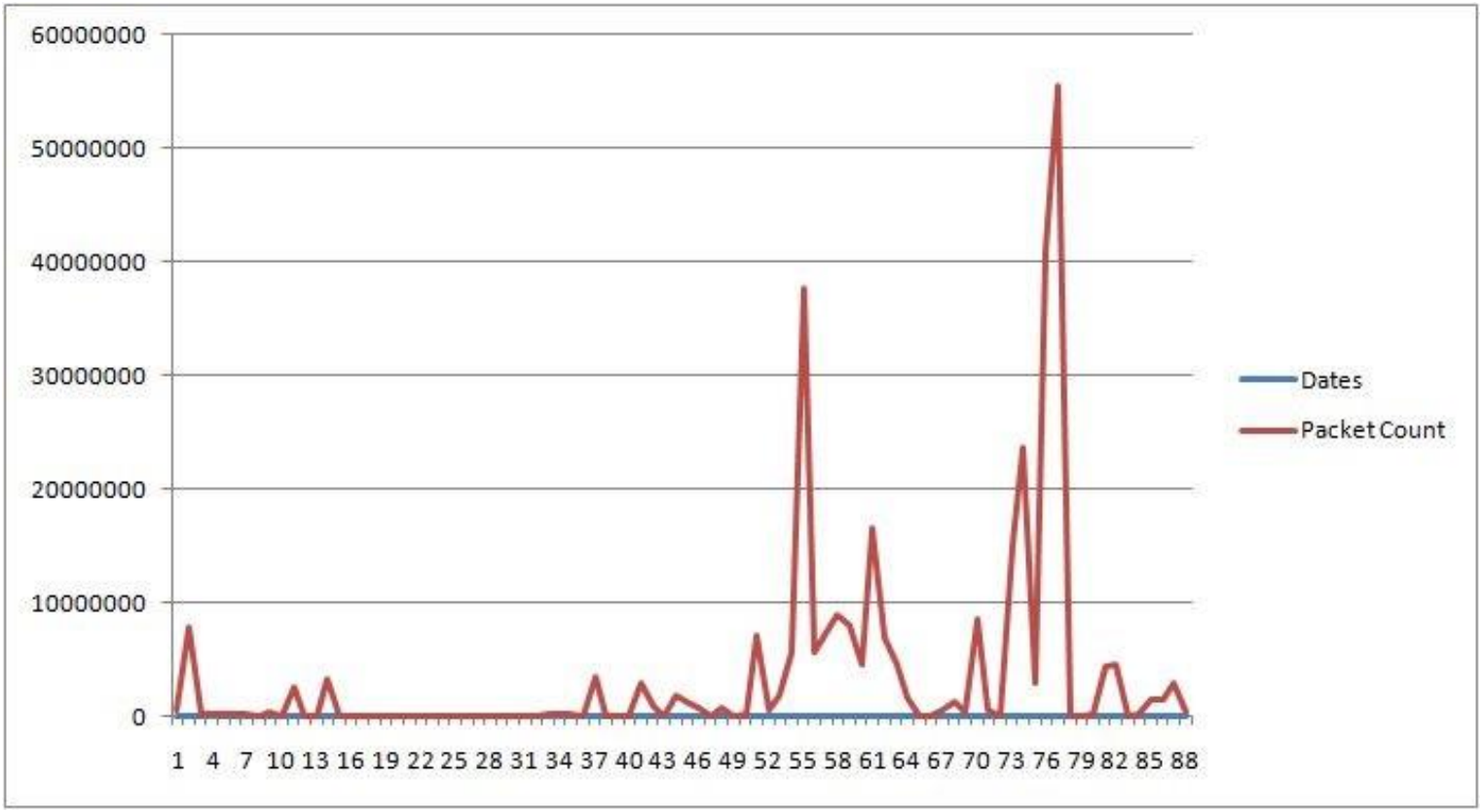
GRAPHS PLOTTED SO FAR

The dates represented here are from 1st February to 30th April and is represented as a continuous data.

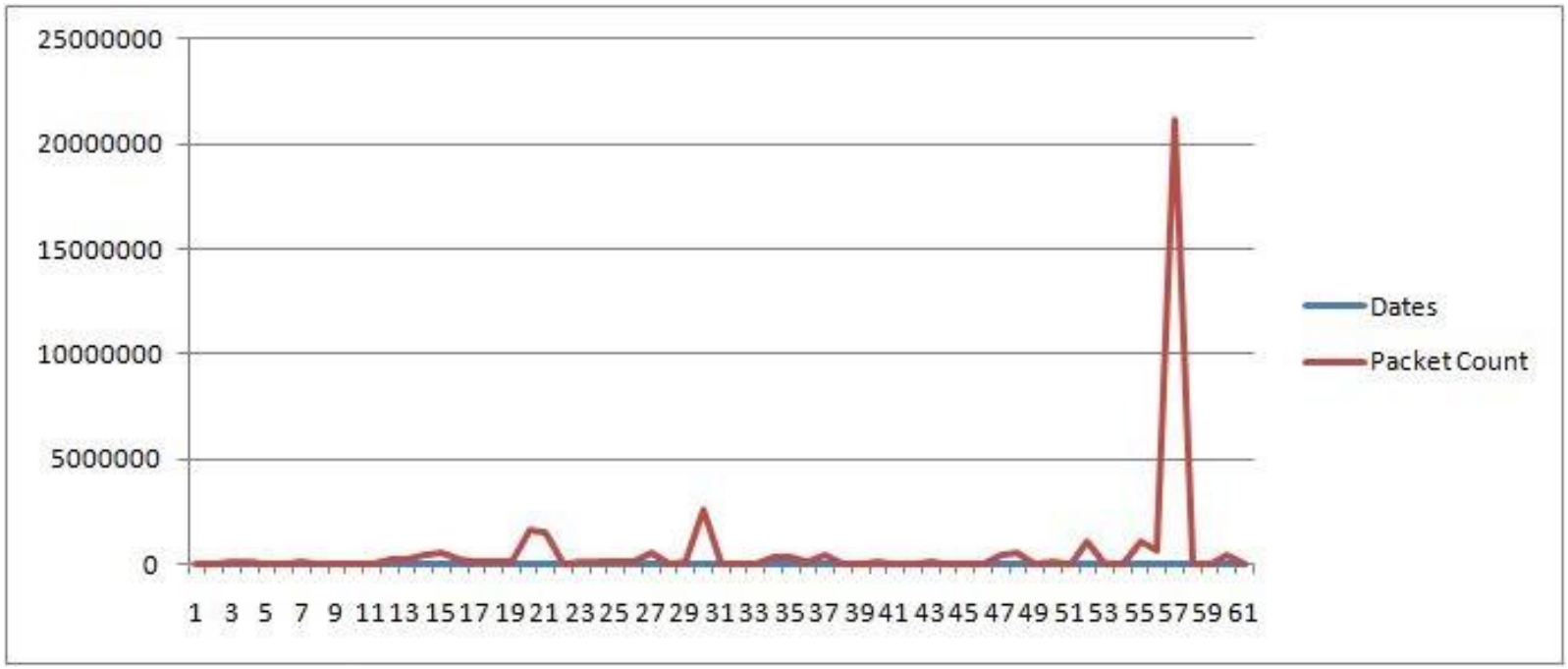
Source Port 53 and Number of Packets using TCP (Doesn't include 16th to 28thFebruary)



Source Port 53 and Number of Packets using UDP.

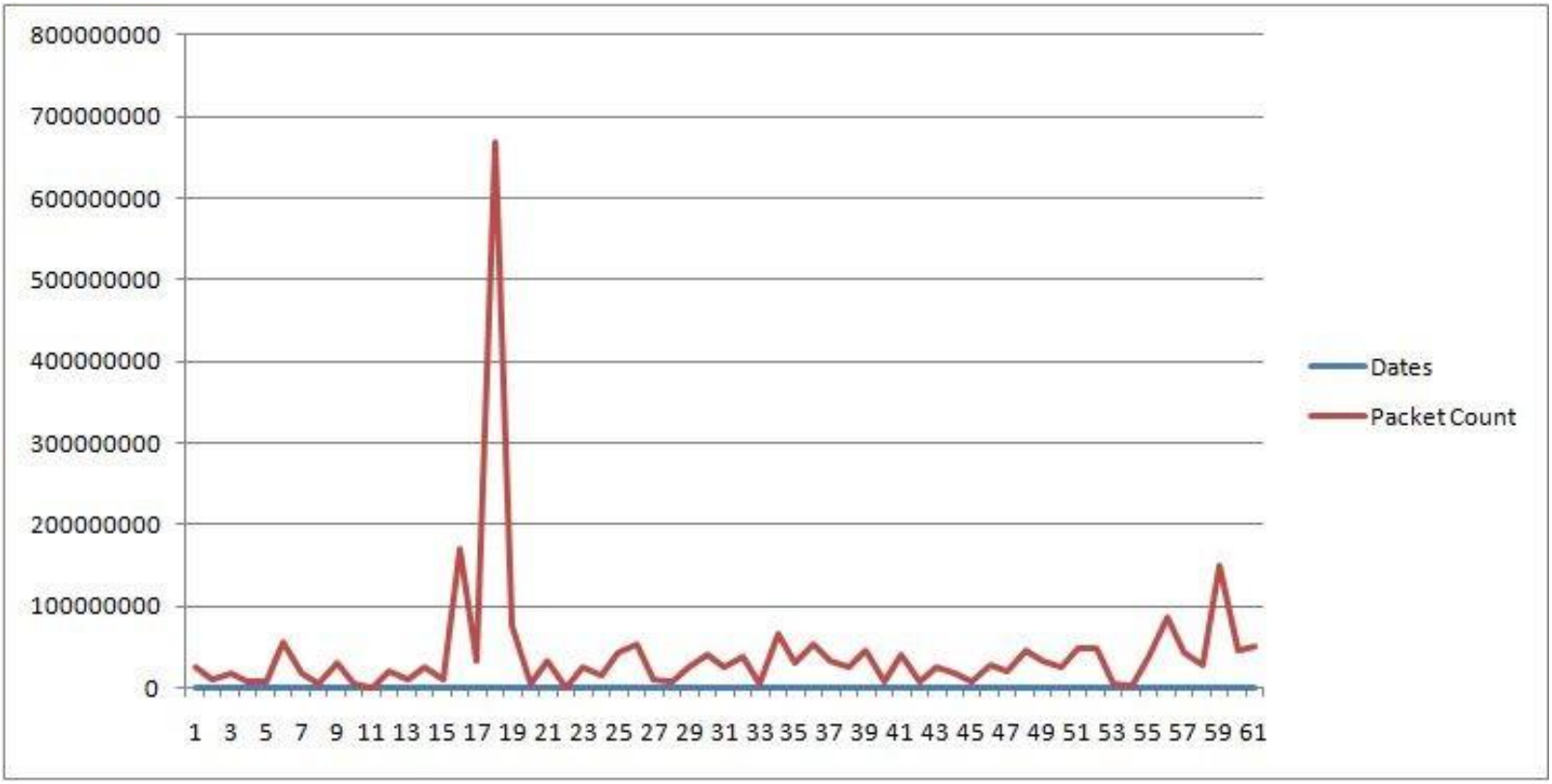


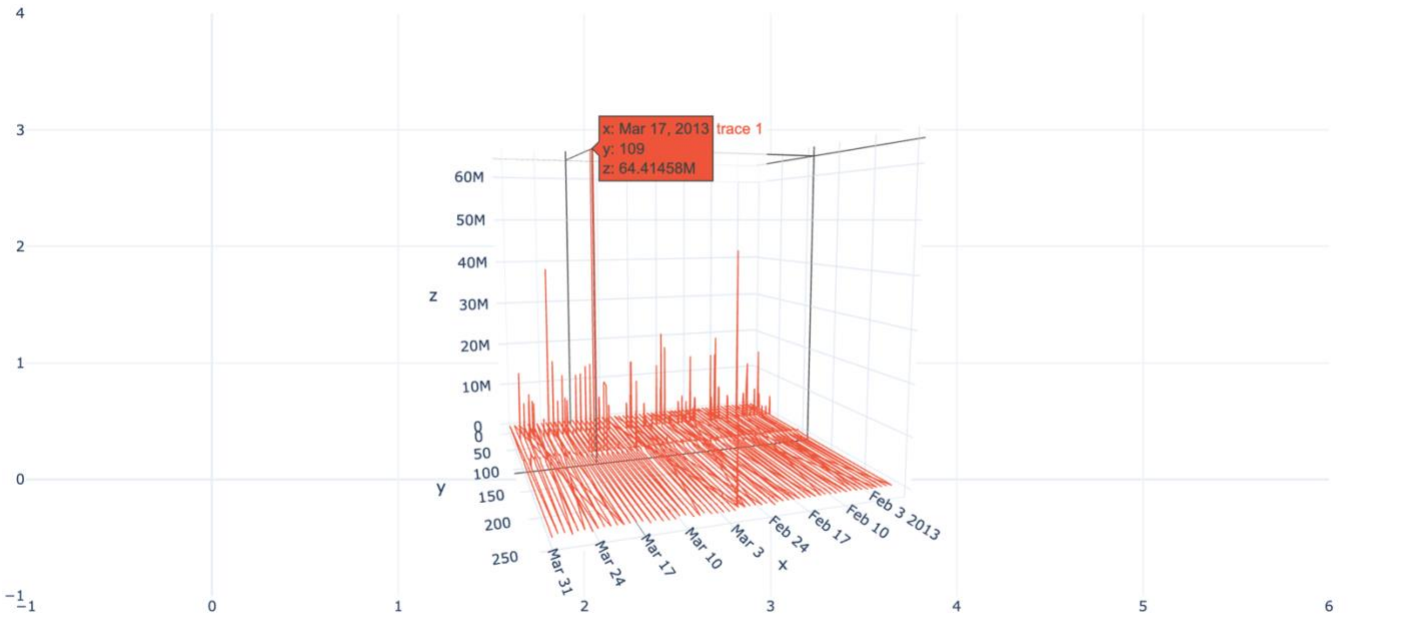
Destination Port 53 and number of packets using TCP (Doesn't include 16th to 28th February).



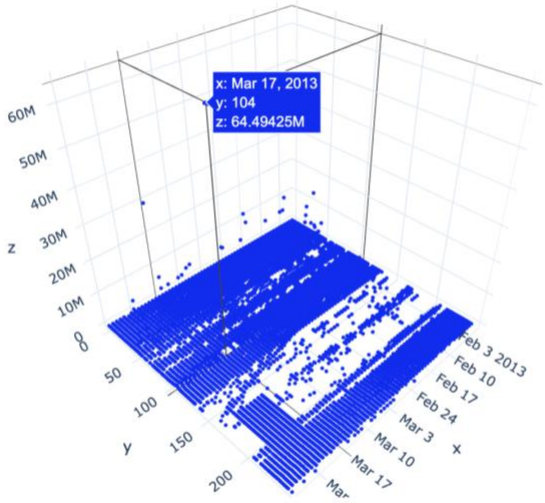
.

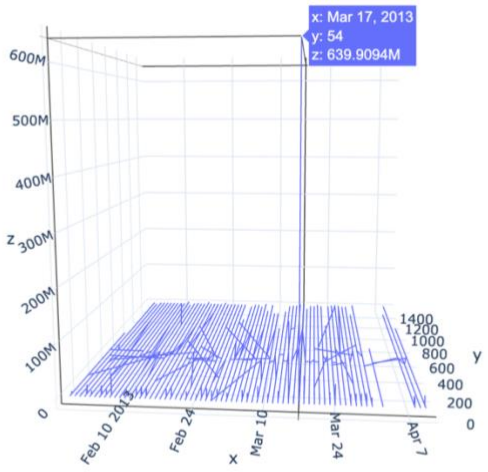
Destination Port 53 and number of packets using UDP





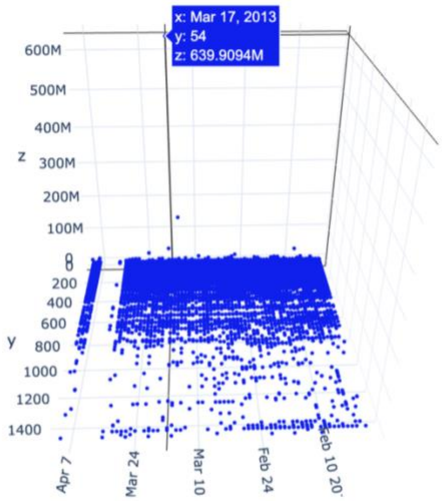
*LINE AND SCATTER GRAPHS FOR **DESTINATION PORT 53** SHOWING **TTL, PACKET COUNT AND DATES** ON THE 3 AXES. Click on the link to see the actual graphs.*



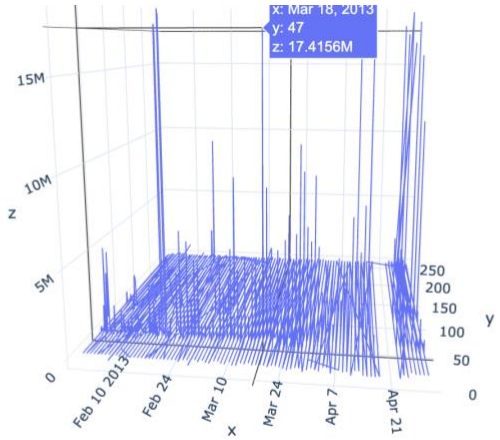


Export to plot.ly »

*LINE AND SCATTER GRAPHS FOR **DESTINATION PORT 53** SHOWING **IP LENGTH, PACKET COUNT AND DATES** ON THE 3 AXES. Click on the link to see the actual graphs.*

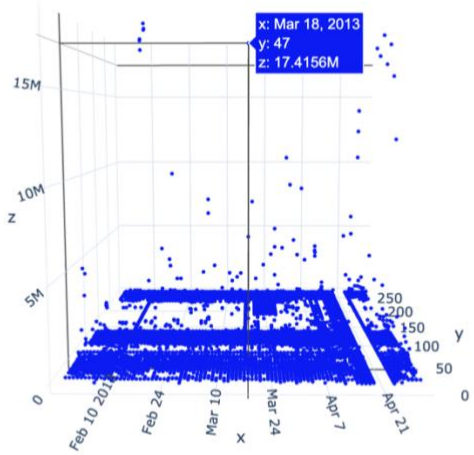


Export to plot.ly »

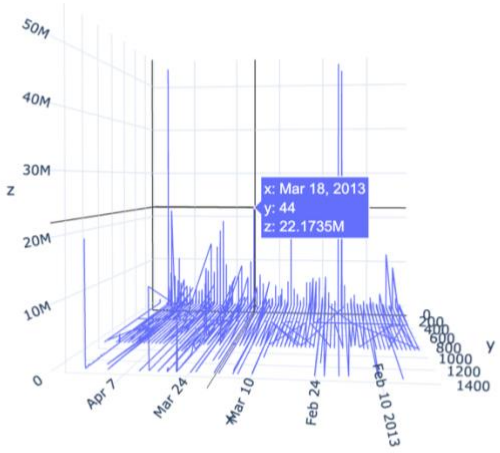


Export to plot.ly »

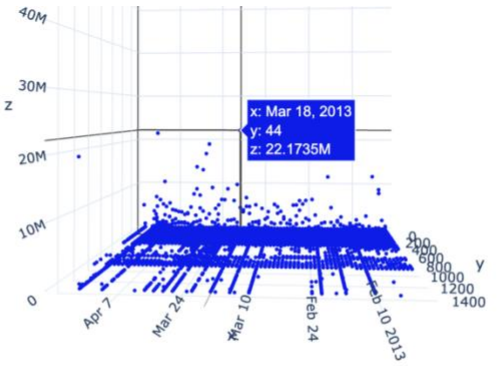
*LINE AND SCATTER GRAPHS FOR **SOURCE PORT 53** SHOWING **TTL, PACKET COUNT AND DATES** ON THE 3 AXES. Click on link to see the actual graphs.*



Export to plot.ly »



*LINE AND SCATTER GRAPHS FOR **SOURCE PORT 53** SHOWING **IP LENGTH, PACKET COUNT AND DATES** ON THE 3 AXES. Click on the link to see the actual graphs.*



Analysis of What We Have Achieved So Far.

From the above plotted graphs, we can conclude that attacks have taken place, which is evident from the peaks in the graphs.

The packet count for a particular set of dates has been enormous, reaching to the order of tens of millions of packets on a particular day. This clearly has to be classified as an attack.

The packet count for ipv4/6 was found to significantly smaller as compared to TCP and UDP and has thus been neglected.

We will further aggregate all the data categorically and define a threshold value, above which we will be safely able to identify an attack.

FUTURE ASPIRATIONS

The Future goals for this project involve the following:

- Try and establish a relation between the number of packets, IP Length and TTL, so as to be able to effectively determine the attack intensity.
- Find out if there is any relation between an attack and the values of IP length and TTL at that particular instance.
- Flow Generation for entire three-month dataset after we have been allotted more memory.
- Plot the data from sources other than the Darknet sources using Python Scripts and methods that have already been developed.
- Identify relevant parameters of importance that can be used to train clustering algorithms like K-Means and EM algorithm.
- Using the trained models to predict whether a DDoS attack is happening or not from a given DNS query.