Analysis of Network Flow Traffic Data

Name – Anshul Roonwal
SBU ID – 110554783

Data Set Analyzed - network_data1

Introduction

1. The data set – "network_data1" consists of data packets being transferred among several nodes in a network over the period of 9 continues days starting from midnight and incrementing by real-valued incremental seconds to 6 decimal places. There are multiple 'Source' and 'Destination'. Though the actual data bit stream is missing in the data set, information about every transfer has been captured in the 'info' column of the data set.

2. The two main csv (comma separated files) are –
   A) AllNetworkData.csv
   B) TestNetworkData.csv

   While TestNetworkData.csv is a small file and can be considered as training data, AllNetworkData.csv is the bigger file containing data over the span of more than 9 days. In order to not miss any data while analysis, I have performed all the operations on the bigger file ie; AllNetworkData.csv

Problem Statements

1. Cleaning the data in hand. This involves (and is not limited to) handling missing values, removing junk characters, discarding the information that may not be useful for analysis.
2. Finding proper tools to load and visualize the entire data since the data size is very huge.
3. Analyzing the network activity at Source and Destination nodes individually. Finding out which nodes in the network are most active over the span of 9 days.
4. Analyzing the protocols which were being used to send the data units.
5. Analyzing data transmission with respect to time.
6. Analyzing the Network in terms of connectivity among the nodes.

Observations and Results

1. Cleaning the data

A. Junk Characters

While parsing the data, I encountered error saying –
Error - "Expected Integer, got String at index 2579614" for the 1st column 'No.'

At index 2579614, 2579616, 2579618 – A String of Junk Characters was present like the one below –
"**Fûà'"”!]ý…ÛH2ØÐðŒEi^Å1ßÚ{ÿóòÔ„d¼¤àÃ<□~xG×”0)ßökv¡ñ½VÚ'5<;lŠ□V²X)H%2&a>&Ãï´i×ÅØ RgV**"
So, I removed those three rows.

B. Missing Values

Observation –

As there were Junk Characters present in 3 rows for only the first column – 'No.', the rest of the columns for those 3 indices had missing values.

Conclusion -

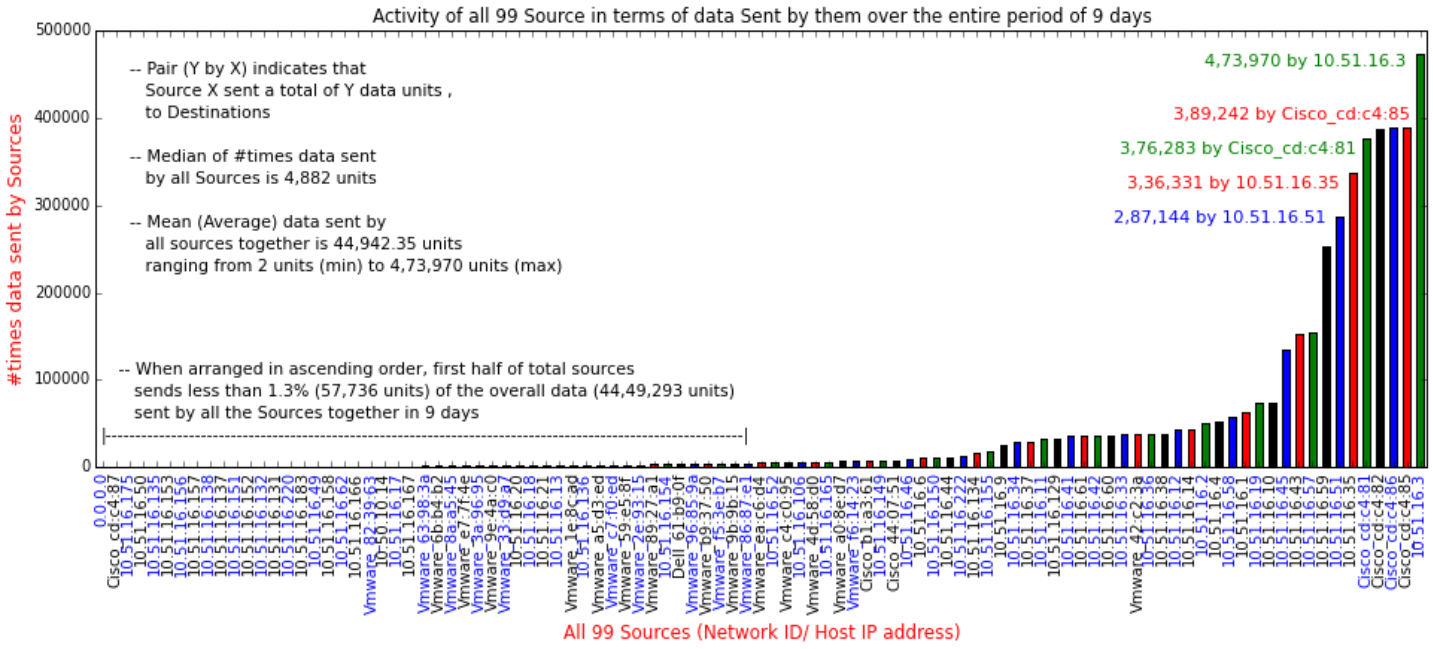
Removing those rows solved the problem of missing data as well. Since there were only 3 row, removing them from data set would not harm much on the overall analysis of data.

2. Finding proper tools to analyze the data.

A. iPython Notebook implementing numpy, matplotlib, pyplot, csv,numpy, networkX and Pandas Package of python.
B. **Delimit** - Since MS-Excel was not able to load all the data due to limitation on number of rows, I had to use another software delimit. It helped me with loading the big csv file at once.
C. **Google Refine** – This helped me in clustering the nodes and protocols in groups.

3. Analyzing the network activity at Source and Destination nodes.

A. At Source →



Observations –

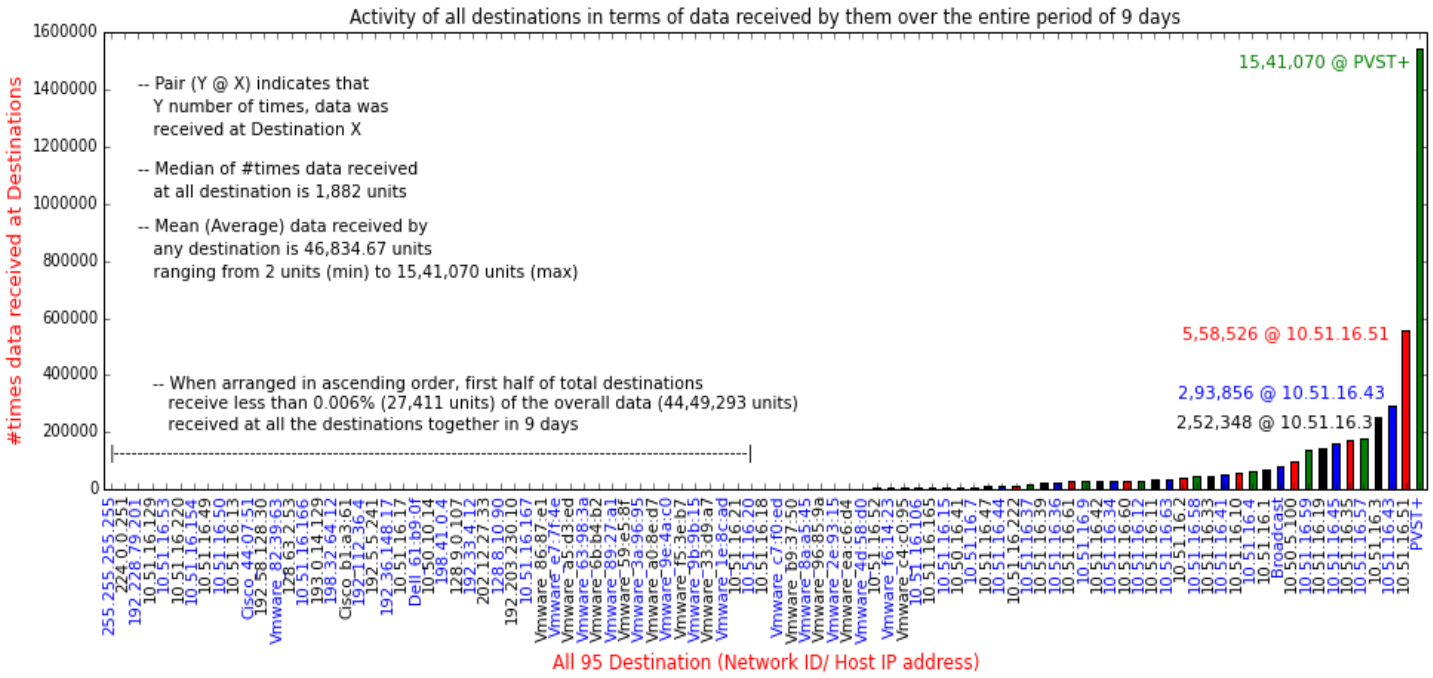
Most of the observations are mentioned in the graph itself. Some are mentioned below.

- The highest number of data units were sent from 10.51.16.3.

Conclusion –

- There are many sources taking part in the network almost equally and the percentage of contribution drops gradually and not suddenly from the highest to the lowest contributing source node.

B. At Destination →



Observations –

Most of the observations are mentioned above in the graph itself, rest are mentioned below -

- The sending pattern and receiving pattern are different in terms of data sent and received by highest and second highest contributing nodes.

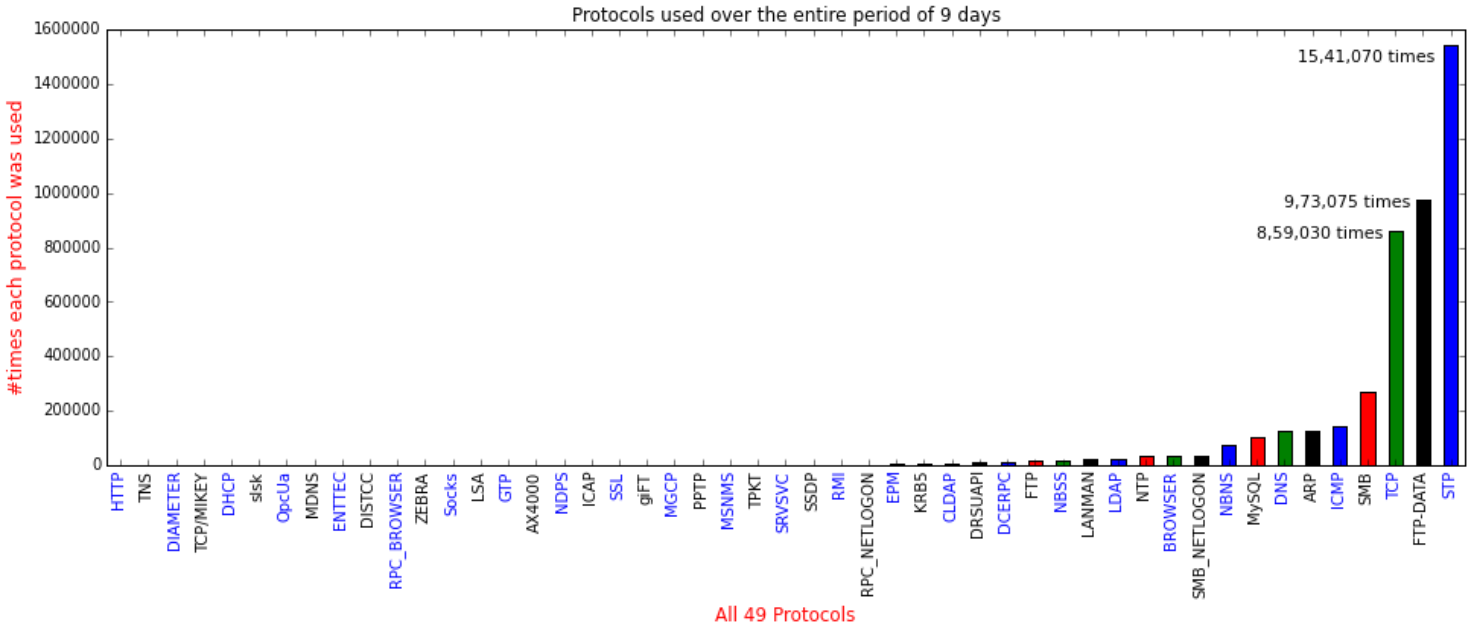
	% Contribution of largest contributing node	% Contribution of second largest contributing node	% Difference in data being sent/received by the largest and second largest contributing nodes
Source	10.6	8.74	1.86
Destination	34	12.5	21.5 (Too much in comparison to source)

- After comparing the sending pattern of sources and receiving pattern of destinations, it can be concluded that there are many sources taking part in the network almost equally and the percentage of contribution drops gradually and not suddenly from the highest to the lowest contributing source node. However, in case of Destination nodes, the percentage of contribution drops suddenly (and not gradually) from highest to the second highest contributing destination node and then the drop becomes gradual.

Conclusion –

Taking down the highest data receiving node (PVST+) with any sort of attack may impact a lot in terms of data packet lost throughout the network in comparison to taking down the highest data sending node (10.51.16.3).

4. **Analyzing the protocols which were being used to send/receive the data units.**



Observations –

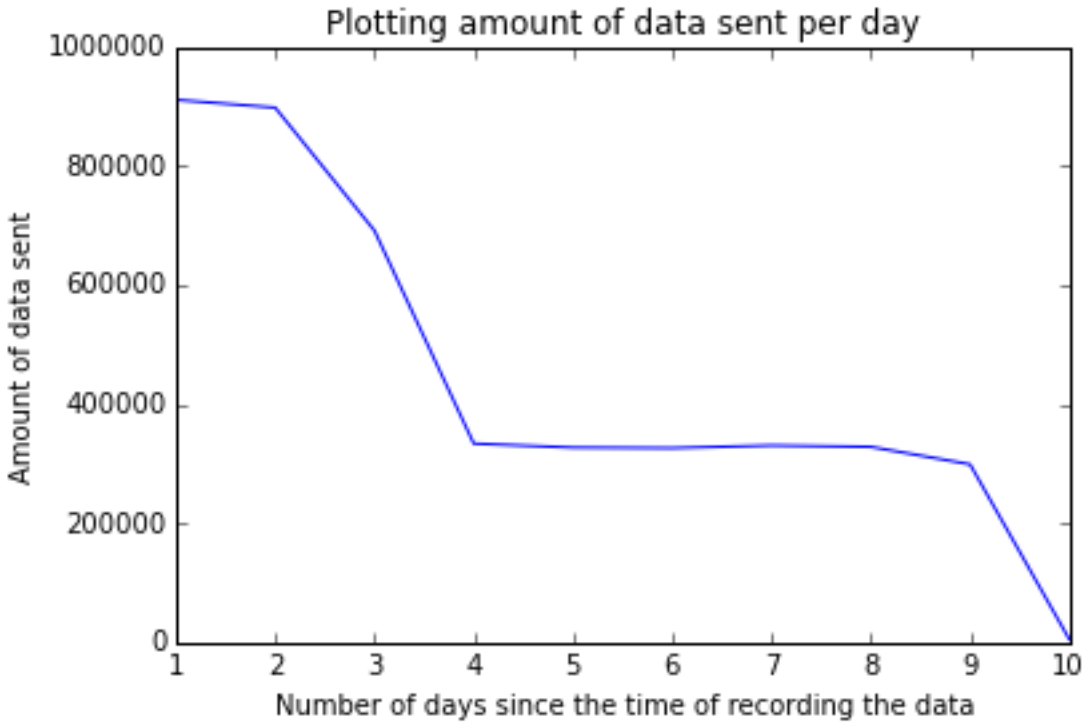
STP (Spanning Tree Protocol) is the protocol that has been most commonly used in this network for the exchange of data units followed by FTP (File Transfer Protocol) and TCP (Transmission control Protocol).

Conclusion –

1) Since STP is being used most commonly, it can be inferred that the benefit of network segmentation (one of the primary functions of STP) will result into amount of competition for use of the network path being reduced by half and the possibility of the network coming to a halt is significantly reduced.
Reference - http://searchnetworking.techtarget.com/definition/spanning-tree-protocol

2) FTP provides encryption of content being shared and also enables secure transmission over the network yet has security weaknesses. It is vulnerable to attacks like Brute Force Attack, Spoofing attack. FTP being used second highest number of times in this networks indicates that the network is prone to attacks and to prevent the attacks it could be replaced with more secure protocols like SFTP(Secure FTP) or SSH (Secure Shell).
Reference –
1) https://www.eldos.com/security/articles/4672.php?page=all
2) https://en.wikipedia.org/wiki/Secure_Shell

5. **Analyzing data transmission with respect to time.**
    - Since the amount of data is very huge and time is incremented by real-valued incremental seconds to 6 decimal places, I have broken down the time to days.
    - The data has been recorded for a bit more than 9 number of days (217 hours).

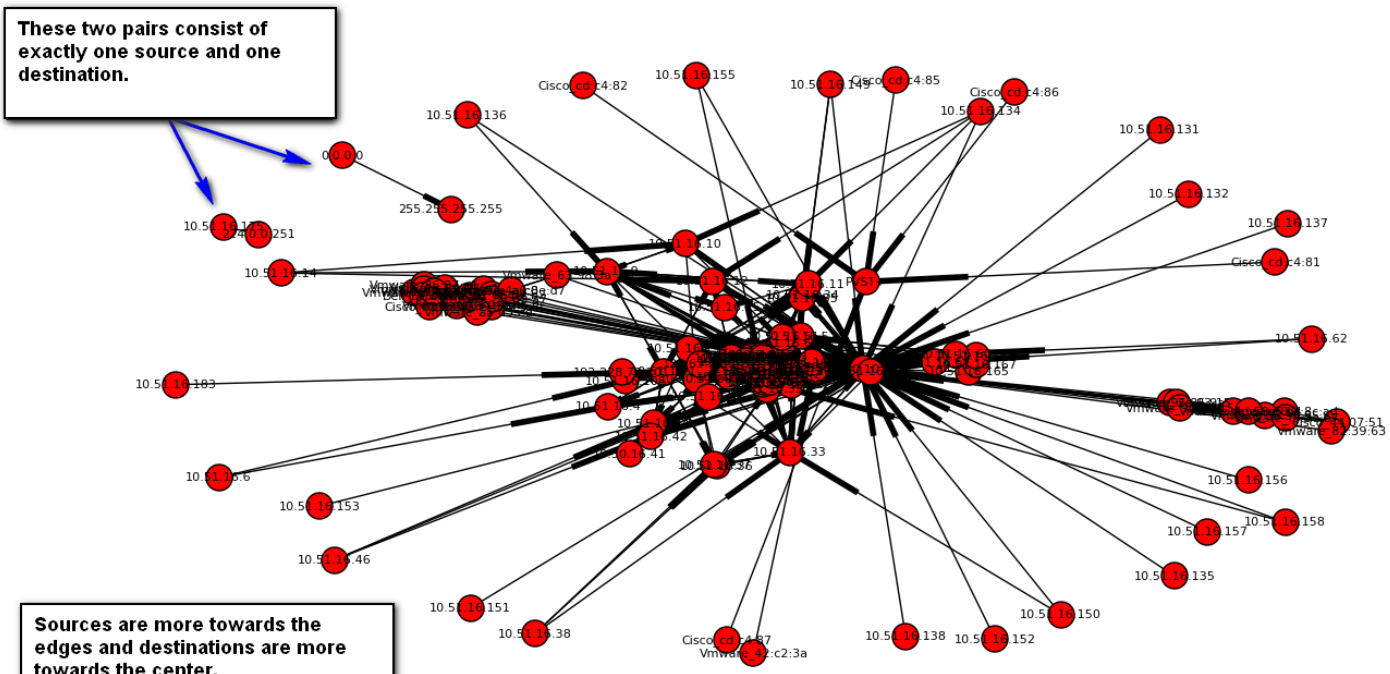


Conclusion –

- Looking at above figure, it can be concluded that data being transmitted in the first 24 hours was the maximum among all the days.
- The amount of data have decreased as the time progresses and it has never surged from low to high.

6. Analyzing the Network in terms of connectivity among the nodes.

I plotted a directed graph connecting the nodes of the network. The direction from source to destination is visible by the edge getting thickened near the destination.



A Directed graph of the network

Observations –

- As shown in the graph, there are exactly two pairs of source (X) and destination(Y) nodes such that-
  a) Source X sends data only and only to destination Y and not to any other destination in the network in the entire duration of 9 days.
  b) Destination Y receives data only and only from Source X and not from any other Source in the network in the entire duration of 9 days.

  The two pairs are –

	Source	Destination
Pair 1(X→Y)	0.0.0.0	255.255.255.255
Pair 2(X→Y)	10.51.16.175	224.0.0.251

Conclusion –

- The nodes in the graph except the two pair mentioned above violates at least one of the condition mentioned in a) or b)
- Source nodes are sparsely distributed in comparison to destination nodes. Other way of saying this is that in contrast with 95 destination nodes, more and more source nodes out of 99 sources are contributing significant amount of data towards the overall exchange of data.