Student Name : Ng Chi Hui

Group : FDDP1

Date : 21/10/2021

LAB 4: ANALZING NETWORK DATA LOG

You will be provided with the data file, in .csv format, in the working directory. Write the program to extract the following informations.

EXERCISE 4A: TOP TALKERS AND LISTENERS

One of the most commonly used function in analyzing data log is finding out the IP address of the hosts that send out large amount of packet and hosts that receive large number of packets, usually know as TOP TALKERS and LISTENERS. Based on the IP address we can obtained the organization who owns the IP address.

List the TOP 5 TALKERS

Rank	IP address	# of packets	Organisation
1	13.107.4.50	5960	Microsoft Corporation
2	130.14.250.7	4034	National Library of Medicine
3	155.69.160.38	3866	Nanyang Technological University
4	171.67.77.19	2656	Stanford University
5	155.69.199.255	2587	Nanyang Technological University

TOP 5 LISTENERS

Rank	IP address	# of packets	Organisation
1	137.132.228.33	5908	National University of Singapore
2	192.122.131.36	4662	A*STAR
3	202.51.247.133	4288	Nusgp
4	137.132.228.29	4022	National University of Singapore
5	103.37.198.100	3741	A*STAR

EXERCISE 4B: TRANSPORT PROTOCOL

Using the IP protocol type attribute, determine the percentage of TCP and UDP protocol

	Header value	Transport layer protocol	# of packets	%
1	6	TCP	137707	77.698723
2	7	UDP	36852	20.793085
3	Others	Others	2673	1.508193

EXERCISE 4C: APPLICATIONS PROTOCOL

Using the Destination IP port number determine the TOP 5 most frequently used application protocol.

Rank	Destination IP port number	# of packets	Service
1	443	43208	https
2	80	11018	http
3	50930	2450	Dynamic and/or Private Ports
4	15000	2103	hydap
5	8160	1354	patrol

EXERCISE 4D: TRAFFIC INTENSITY

The traffic intensity is an important parameter that a network engineer needs to monitor closely to determine if there is congestion. You would use the IP packet size to calculate the estimated total traffic over the monitored period of 15 seconds. (Assume the sampling rate is 1 in 2048)

Total calculated sampled traffic (MB):

Estimated Total Traffic taking into account the sampling rate (MB)	169.93475 MB

EXERCISE 4E: ADDITIONAL ANALYSIS

Please described additional analysis of the data and how it is useful. Please use a separate sheet to submit your new graphs and observations. Your report for this exercise is limited to 2 pages. The answer template and the two page additional analysis are to be submitted to your e-learning drive.

Examples

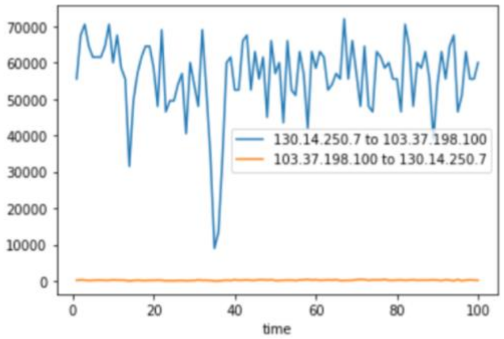
- Visulisation using scatter graph of port and IP address to determine if a specific node been port scanned by another node.
- Visualisation using network graph
- Other methods

You must analyse and explain the graphs. Please do not be limited by the above examples.

Top 5 Communication Pairs

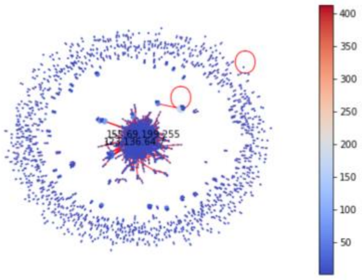
Rank	Source Organization	Source IP	Destination Organization	Destination IP	# of Packets
1	National Library of Medicine	130.14.250.7	A*STAR	103.37.198.100	3739
2	Stanford University	171.67.77.19	A*STAR	192.122.131.36	2656
3	National Aeronautics and Space Administration	129.99.230.54	National University of Singapore	137.132.22.74	2097
4	National University of Singapore	137.132.228.42	The Scripps Research Institute	137.131.17.212	1553
5	Nanyang Technological University	155.69.252.133	M1 LIMITED	138.75.242.36	1475

The table above shows that top 5 communication pairs, it can be observed that communication is done mainly between government and education organisations.



The graph above shows the traffic between the top 1 communication pair. It can be observed that traffic flow from 130.14.250.7 to 103.37.198.100 is higher that the opposite way.

Visualisation of Communication between Hosts

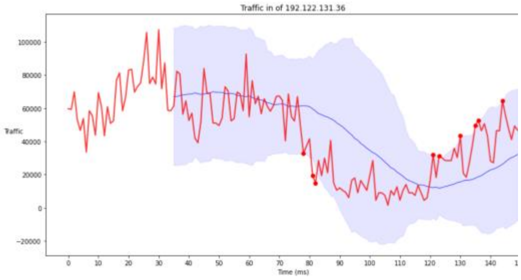


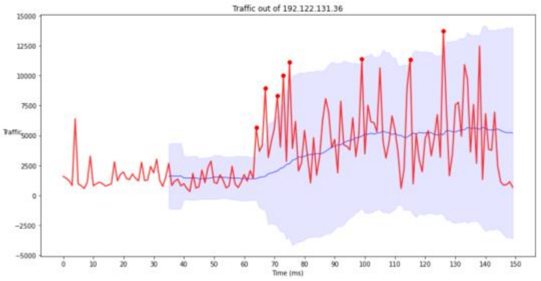
The network graph above visualises the communication between the connection between the source IP and destination IP. We are able to observe that the traffic on the network are centred around a few locations.

The node with the most unique connections and highest degree of link has an IP address of “155.69.199.255” which belongs to NTU. Other highly communicated nodes include “123.136.64.7” and “155.69.160.78” which belongs to AStar and NTU respectively. All the nodes shows a common trend whereby they are educational and research institutions. Upon further analysis, it has been shown that the most common application that is used for communication is port 443 (https) which allows for secure data transfer. Port 443 was used in 56% of the connections and 62% of communication for NTU and Astar respectively.

Analysis of Traffic In and Out

The Ip address chosen for this analysis is ‘192.122.131.36’ which belongs to Astar. This IP address was picked as it is the node with the most number of data packets send and received. Analysis was performed by splitting 150 chunks into 15 second time frame and plotting the sum of the traffic at each chunk. A rolling mean window was used as a means of anomaly detection. In this example, I took the previous 35 values, calculated the mean, and added an upper/lower bound of +/- 2.5 standard deviations. Values that fall outside this range will be considered anomalous. It's identified that during this anomalous period for this IP address, most of traffic arrived from 171.67.77.19 which belongs Sandford University thus Astar may want to consider prioritising this network





EXERCISE 4F: SOFTWARE CODE

Please attach a softcopy of your code to the e-learning drive.

Please refer to Ng Chi Hui.ipynb for the code