

Student Name : Teo Chen NingGroup : TS3Date : 12/10/2020**LAB 3: 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 information.

EXERCISE 3A: 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	Organization
1	103.26.47.233	9646	Multimedia Development Corp
2	13.107.4.50	4950	Microsoft Corp
3	155.69.160.78	4563	Nanyang Technological University
4	130.14.250.7	3914	National Library of Medicine
5	173.194.22.215	2896	Google Inc

TOP 5 LISTENERS

Rank	IP address	# of packets	Organization
1	103.22.221.73	9646	National Information Society Agency
2	137.132.228.33	7835	National University of Singapore
3	137.132.228.29	5964	National University of Singapore
4	137.132.228.42	4987	National University of Singapore
5	103.37.198.100	3915	A*STAR

EXERCISE 3B: TRANSPORT PROTOCOL

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

Rank	Header value	Transport layer protocol	# of packets
1	6	TCP	155799 (76.37%)
2	17	UDP	45377 (22.24%)
3	0	HOOPT	1218 (0.60%)
4	47	GRE	891 (0.44%)
5	50	ESP	643 (0.32%)

EXERCISE 3C: APPLICATIONS PROTOCOL

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

<https://www.adminsub.net/tcp-udp-port-finder/>

Rank	Destination IP port number	# of packets	Service
1	443	42975	HTTPS
2	80	11960	HTTP
3	56800	3918	Dynamic and/or Private Ports OR Xsan Filesystem Access (Apple)

4	15000	2697	Dynamic and/or Private Ports OR Hypack Data Acquisition
5	44678	1158	Dynamic and/or Private Ports

EXERCISE 3D: 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 1000)

Total Traffic (Based on ip_size)	
Bytes	199163627000
Megabytes (Binary)	189937.24
Megabytes (Decimal)	199163.63

EXERCISE 3E: ADDITIONAL ANALYSIS (BONUS MARKS)

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.

Analysis below

EXERCISE 3F: SOFTWARE CODE

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

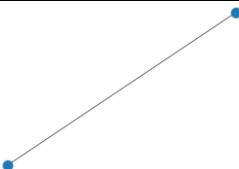
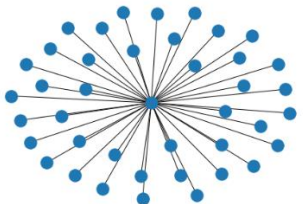
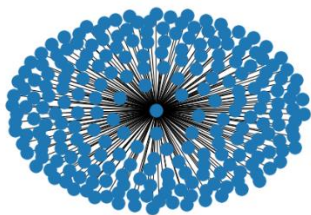
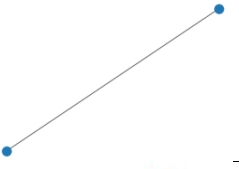
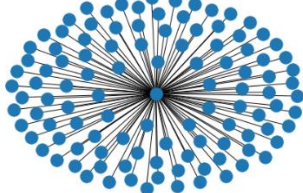
EXERCISE 3E: ADDITIONAL ANALYSIS (BONUS MARKS)
TOP 5 COMMUNICATION PAIRS

Rank	IP Address 1	Organization	IP Address 2	Organization	Count
1	103.22.221.73	National Information Society Agency	103.26.47.233	SDN Network	11092
2	104.44.201.147	Microsoft Corporation	202.21.159.244	Asia Pacific Network Information Centre (APNIC)	4608
3	103.37.198.100	Asia Pacific Network Information Centre (APNIC)	130.14.250.7	National Library of Medicine	4358
4	129.99.230.54	National Aeronautics and Space Administration (NASA)	137.132.22.74	Asia Pacific Network Information Centre (APNIC)	3203
5	128.117.28.212	National Center for Atmospheric Research (NCAR)	155.69.52.27	Asia Pacific Network Information Centre (APNIC)	1572

Judging from the top talkers/listeners as well as the top 5 communication pairs, it can be deduced that this is a network meant for education/research purposes, possibly centered in the Asia Pacific region.

VISUALIZING COMMUNICATION BETWEEN IP HOSTS

Let us see if we can find out more information about the various IP hosts involved in the network. For now, we will focus on the top 5 talkers of the network. We graph their connections to their corresponding destination IPs:

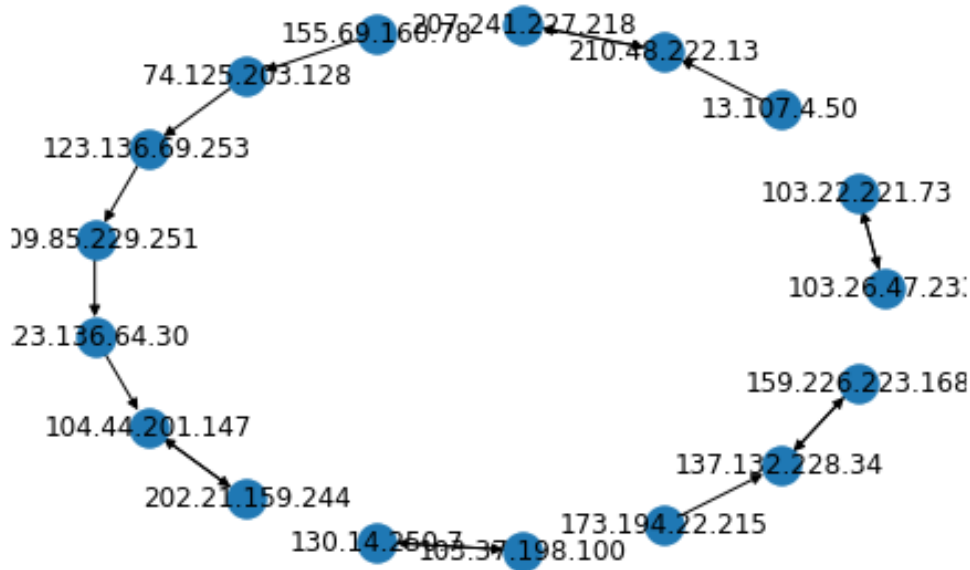
Source IP	Organization	Connection Graph
103.26.47.233	Multimedia Development Corp	
13.107.4.50	Microsoft Corp	
155.69.160.78	Nanyang Technological University	
130.14.250.7	National Library of Medicine	
173.194.22.215	Google Inc	

As we can see, some high-volume senders, such as Malaysia's Multimedia Development

Corporation (103.26.47.233) and the National Library of Medicine (130.14.250.7) only send to one destination. Their corresponding ports that they communicate on is 36296 and 56800 respectively, both of which are private ports. This suggests that they may be sending packets for a private API or within an internal network (routed through an SFlow agent).

On the other hand, sparsely connected sender nodes like Microsoft and (to a lesser extent) Google, may signal that they are routing data through from/to various sources, serving as an intermediary node.

Finally, densely connected sender nodes like Nanyang Technological University may represent an actual sender node that is sending data over the internet to many addresses.



By tracing the route each IP address sends its packets to, we can see how the top 5 talkers are sending their packets. Here, each directed arrow represents a sending from src_ip to dest_ip.

For Multimedia Development Corp (103.26.47.233) and National Library of Medicine (130.14.250.7), we see that a route trace only bounces between two nodes in the network, perhaps showing how their data remains within an internal network.

For Google (173.194.22.215) and Microsoft (13.107.4.50), we see that they forward the packet data between up to 3 nodes. A simple WHOIS search along the IP address of the route shows that they both forward packets along the Asia Pacific Network Information Centre. We can thus presume that Google and Microsoft serve as intermediaries in the network for geo-replication and/or transferring data.

Finally, for NTU (155.69.160.78), we have a long route. Following it, we see that packet data gets bounced along various Google, Microsoft and APAC nodes.

For all the routes, we see a two way directional data transfer, perhaps hinting at how the data is routed to the destination node (whether it is a private node or APAC), and then the corresponding node retrieves data from some other router that is not logged in SFlow, and sends the data back.

The same tracing can be done for the top 5 listeners to find out what they are listening for (not shown here due to lack of space). For the top 5 listeners, we can see that the National Information Society Agency and A*Star only listens from 2-3 nodes, presumably over some internal network or the like.

Meanwhile, the National University of Singapore listens from a large number of nodes, signalling open traffic from various sources.