Student Name : _____Yong Wen Shiuan_____

Group : _____TS7_____

Date : _____16/10/2021_____

LAB 4: ANALYZING 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	NUS Gigapop
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	78.24
2	17	UDP	36852	20.94
3	50 / 47 / 1 / 58 / 41 / 2 / 0	Others	1458	0.83

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 Port
4	15000	2103	Hypack Data Aquisition
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): 169.93475 (assuming 10^6 bytes = 1 MB)

Estimated Total Traffic taking into account the sampling rate (MB)	Assuming “Sflow” data was collected over the monitored period of 15 seconds and that packets not sampled are of similar size to those sampled, 169.93475 * 2048 = 348026.368

EXERCISE 4E: ADDITIONAL ANALYSIS

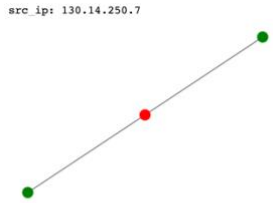
Top 5 communication pairs (bidirectional):

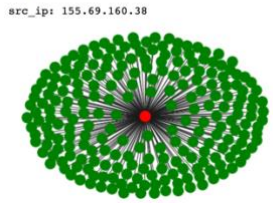
Rank	IP Address 1	Organization	IP Address 2	Organization	Count
1	130.14.250.7	National Library of Medicine	103.37.198.100	A*STAR	4201
2	171.67.77.19	Stanford University	192.122.131.36	A*STAR	3628
3	129.99.230.54	National Aeronautics and Space Administration (NASA)	137.132.22.74	National University of Singapore	2417
4	137.132.228.42	National University of Singapore	137.131.17.212	The Scripps Research Institute	2370
5	104.146.199.27	Microsoft Corporation	202.21.159.246	Republic Polytechnic	1794

Judging from the top talkers/listeners as well as the top 5 communication pairs, it is likely this sFlow data was obtained from a router in Singapore, possibly one involved in a network meant for education/research purposes

VISUALIZING COMMUNICATION BETWEEN IP HOSTS

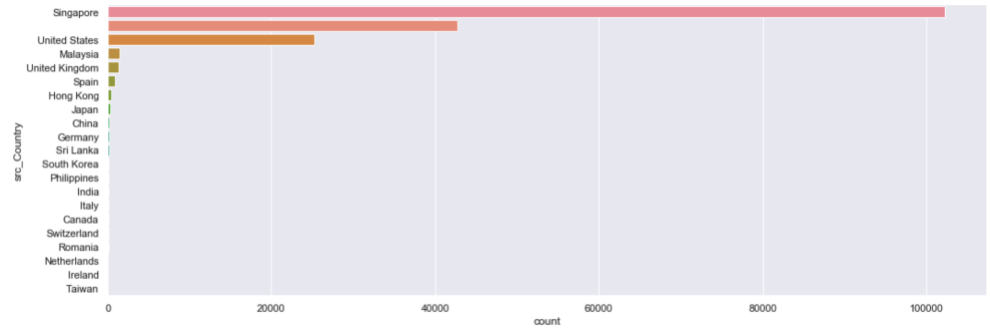
Let us see if we can find out more information about the various IP hosts involved in the network. I visualised the graphs for the top 5 talkers & listeners of the network (for the full graph, can refer to the .ipynb file). With the talker/listener node in red and other nodes it sends/receives packets directly from in green.

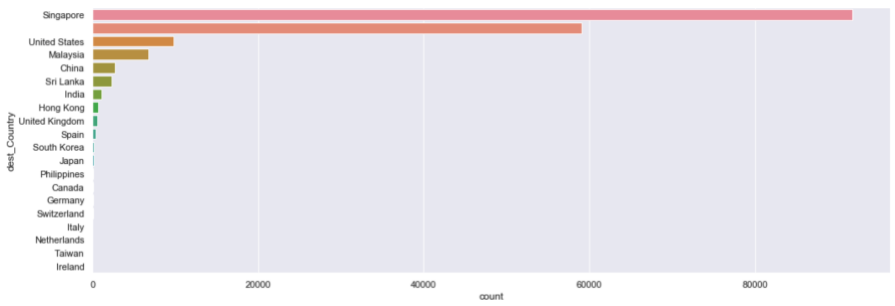


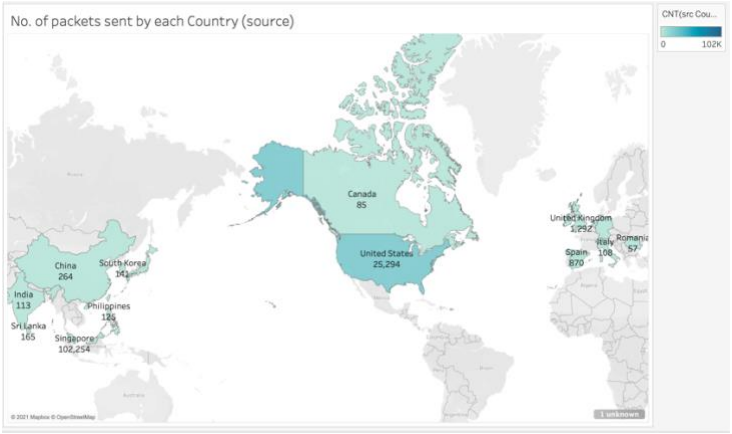


For the top talkers, there may interact with many or only a few. For example, even though 130.14.250.7 sends more packets than 155.69.160.38, 155.69.160.38 sends to many different addresses, whereas 130.14.250.7 sends to only 2 different addresses.

Next, we can try to map each IP address to a country to observe geographically where the packets are sent to and where packets are received.
I used the ipinfo module, however there was a request limit so was unable to map all IP addresses to their respective countries. Given the limited data, I was still able to come up with some maps to visualise the geographical distribution, they can be seen in the src_map.png and dest_map.png. Otherwise, analysis can be performed on the bar charts shown below:

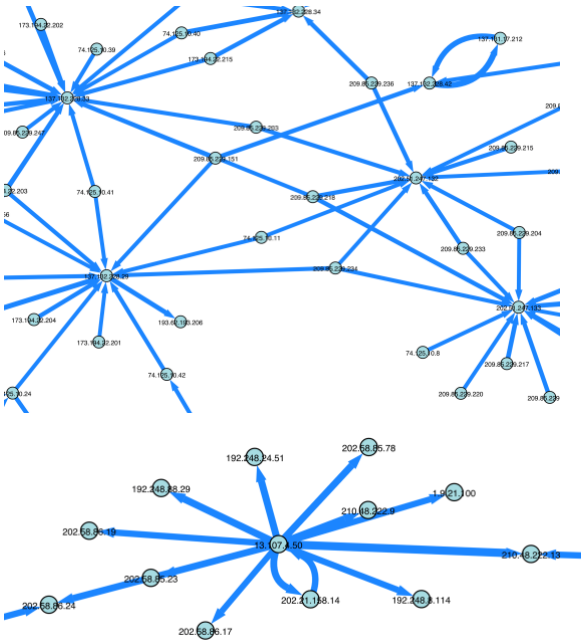






-For source country, most of it is from Singapore, followed by the United States. Likewise for destination country, most of it is also sent to Singapore, followed by United States. After that, a relatively sizeable number of the packets were sent to Malaysia and China.

Moving on to graph visualisation, I tried to filter out those communications with fewer packets sent, to display the ones with many packets sent from a certain source to a particular destination. So that the graph would not be too cluttered and the labels may be more visible.



Zooming in, one may observe that certain nodes have many different addresses that sent packets to it, namely addresses 137.132.228.33, 137.132.228.29, 202.51.247.133, 202.51.247.132. We can also see that 13.107.4.50 sends quite a few packets to different addresses. It turns out that the above 4 IP addresses are under NUS, whilst 13.107.4.50 is registered under Microsoft. Further analysis for port scanning reveals that actually, 13.107.4.50 (Microsoft) sends packets to several different ports, typically ephemeral ports, to different IP addresses. It could indicate that Microsoft is routing data through from various sources, serving as an intermediary node. Alternatively, it could be that they are using Microsoft software to perform port scanning, possibly to detect for any network vulnerabilities. Google is another organisation whose IP address was used to perform port scanning. In total, a whopping 552 different ports from 210.48.222.13 (International Islamic University Of Malaysia), most of which I are ephemeral/dynamic ports, were scanned by 13.107.4.50 (Microsoft).

EXERCISE 4F: SOFTWARE CODE

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

Accessible on Google Drive:
https://drive.google.com/drive/folders/1EeflhTsj4SNiDudfo0r3TPGBW9xIT66m