Student Name : Wilson Thurman Teng

Group : TS6

Date : 1st April 2020

LAB 4: ANALYZING NETWORK DATA LOG

You are provided with the data file, in .csv format, in the working directory. Write the program to extract the following information.

EXERCISE 4A: TOP TALKERS AND LISTENERS

One of the most commonly used functions in analyzing data log is finding out the IP address of the hosts that send out or receive large amounts of packets, usually known as TOP TALKERS and LISTENERS respectively. Based on the IP address, we can obtain the organization who owns the IP address.

List the TOP 5 TALKERS

Rank	IP address	# of packets	Organisation
1	193.62.192.8	3041	JANET
2	155.69.160.32	2975	Nanyang Technological University
3	130.14.250.11	2604	National Library of Medicine
4	14.139.196.58	2452	National Knowledge Network
5	140.112.8.139	2056	National Taiwan University

TOP 5 LISTENERS

Rank	IP address	# of packets	Organisation
1	103.37.198.100	3841	A*STAR
2	137.132.228.15	3715	National University of Singapore
3	202.21.159.244	2446	Republic Polytechnic, Singapore
4	192.101.107.153	2368	ESnet
5	103.21.126.2	2056	IITB

EXERCISE 4B: TRANSPORT PROTOCOL

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

Total Number of Packets = 69370

	Header value	Transport layer protocol	# of packets	Proportion of protocol (in %)
1	6	TCP (6)	56064	80.82%
2	17	UDP (17)	9462	13.64%
3	50	ESP (50)	1698	2.45%
4	0	HOPOPT (0)	1261	1.82%
5	47	GREs (47)	657	0.95%

EXERCISE 4C: APPLICATIONS PROTOCOL

Using the Destination IP port number to determine the most frequently used application protocol.
(For finding the service given the port number https://www.adminsub.net/tcp-udp-port-finder/)

Rank	Destination IP port number	# of packets	Service
1	443	13423	https
2	80	2647	http
3	52866	2068	others
4	45512	1356	others
5	56152	1341	others

EXERCISE 4D: TRAFFIC

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 (in MB)	7722.12 MB (***Assuming IP packet size given in the csv file is in bits***)

EXERCISE 4E: ADDITIONAL ANALYSIS

Please append ONE page to provide additional analysis of the data and the insight it provides.
Examples include:
Top 5 communication pairs;
Visualization of communications between different IP hosts;
etc.

Please limit your results within one page (and any additional results that fall beyond one page limit will not be assessed).

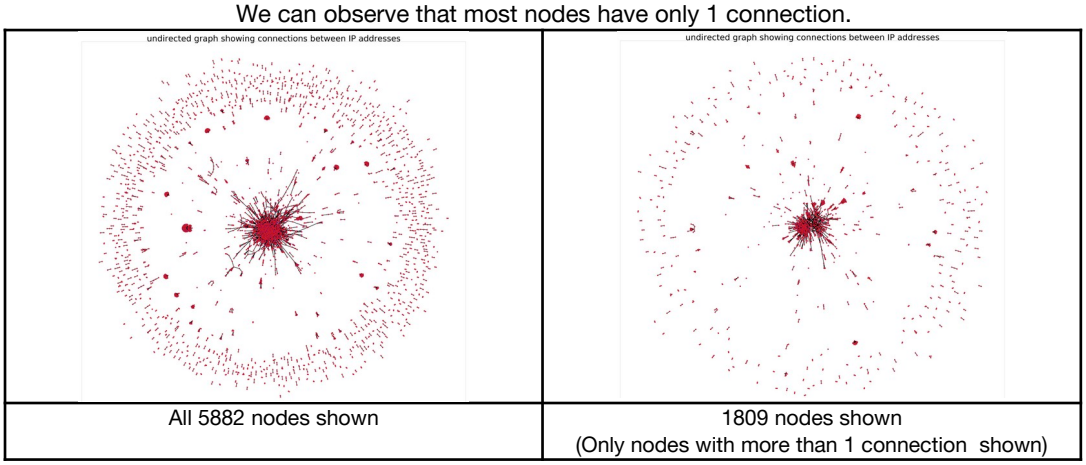
EXERCISE 4F: SOFTWARE CODE

Please attach a copy of your code in an appendix.

4E: ADDITIONAL ANALYSIS

1) Visualization of Data

We can observe that most nodes have only 1 connection.



All 5882 nodes shown	1809 nodes shown (Only nodes with more than 1 connection shown)

2) Top 5 communication pairs

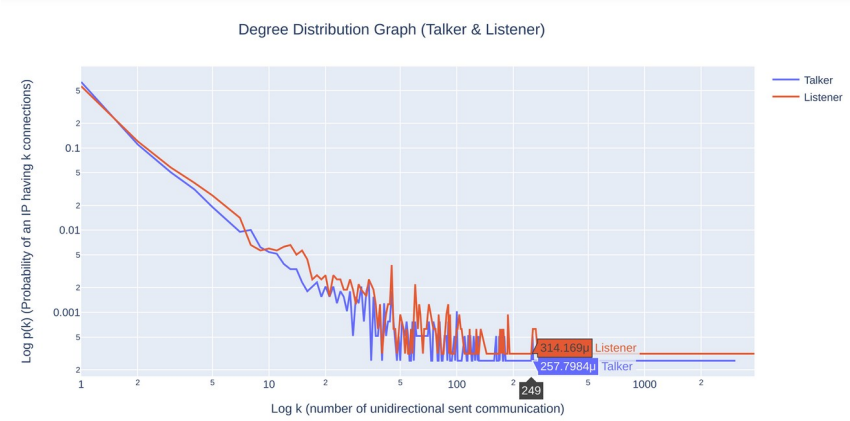
Top 5 One-Way Communication Pair

	src_IP	dst_IP	number_of_times_communicated
0	193.62.192.8	137.132.228.15	3041
1	130.14.250.11	103.37.198.100	2599
2	14.139.196.58	192.101.107.153	2368
3	140.112.8.139	103.21.126.2	2056
4	137.132.228.15	193.62.192.8	1910

Top 5 Two-Way Communication Pair

	IP_addr_1	IP_addr_2	correct_communication_count
0	137.132.228.15	193.62.192.8	4951
1	103.37.198.100	130.14.250.11	2842
2	14.139.196.58	192.101.107.153	2368
3	140.112.8.139	103.21.126.2	2056
4	140.90.101.61	167.205.52.8	1752

3) Which organisations to protect? (In the event of a cyberattack)



To determine which organizations are the most important, the principles of network science are used and a Log-Log plot of the “number of communications made, K” against the “probability of a node having k connections, p(k)”.

We can observe from the graph that the communications distribution follows a power-law distribution and hence, we can deduce which are the most important communication hubs from where there is a cutoff of the gradient.

In this dataset, the cutoff for both the talker & listener is 249 as shown in the graph above. With this cutoff, we can narrow down and determine which organizations to protect in the event of a cyber-attack. In this dataset, using the cutoff method explained, I have managed to narrow down from 848 organizations to 28 and 21 organizations to protect for talkers & listeners respectively.

[Graphs available in source code | Organisations to protect available in Appendix in the following page]

Appendix

Talker organisations to protect:	Listener organisations to protect/be wary of:
1. 'Microsoft Corporation' 2. 'A*STAR' 3. 'A-STAR' 4. 'National Library of Medicine' 5. 'Rutherford Appleton Laboratory' 6. 'Kyushu University' 7. 'National University of Singapore' 8. 'Asia Pacific Network Information Centre' 9. 'National Knowledge Network' 10. 'National Taiwan University' 11. 'Taiwan Academic Network (TANet) Information Center' 12. 'NOAA' 13. 'Japan Advanced Institute of Science and Technology' 14. 'Duke University' 15. 'Nanyang Technological University' 16. 'University of Chicago' 17. 'JANET' 18. 'Internet2' 19. 'FDU' 20. 'Singapore Telecommunications Ltd, Magix Services' 21. 'Republic Polytechnic, Singapore' 22. 'THAINET' ' 23. 'JPNIC' 24. 'Multimedia Development Corporation' 25. 'National Climatic Data Center' 26. 'Internet Archive' 27. 'Facebook' 28. 'National Information Society Agency'	1. 'Republic Polytechnic, Singapore' 2. 'Asia Pacific Network Information Centre' 3. 'Singapore Telecommunications Ltd, Magix Services' 4. 'A*STAR' 5. 'TANET' 6. 'National Knowledge Network' 7. 'JANET' 8. 'Multimedia Development Corporation' 9. 'ESnet' 10. 'IITB' 11. 'Institut Teknologi Bandung' 12. 'Nanyang Technological University' 13. 'Internet2' 14. 'Google LLC' 15. 'National University of Singapore' 16. 'A-STAR' 17. 'Duke University' 18. 'Microsoft Corporation' 19. 'SingAREN' 20. 'National Information Society Agency' 21. 'International Islamic University Of Malaysia'