CS5313: Computer Networks **Instructor: Dr. Deepak Tosh**
Fall 2020 **Due Date:** 11/29/2020

Assignment # 5: Netflow Packet Analysis

Aim:

This assignment aims to provide the students an experience on analyzing the publicly available network measurement data to understand various properties of Internet. You are free to use any plotting tool like matplotlib, Gnuplot, or Excel to generate graphs/plots to the dynamics. In addition, you may be required to write small python scripts to gather information relevant to various questions and calculate the network statistics.

Dataset Details:

Many networks collect Netflow measurements directly from the routers. There are many tools and protocols that network administrators use and Netflow is one of them. NetFlow is a network protocol system created by Cisco that collects active IP network traffic as it flows in or out of an interface and this data can be analyzed to create a picture of network traffic flow and volume. In this assignment, you are given with a 5-minute packet capture data (in a CSV format) from Netflow that was extracted from a router deployed in core network. You will notice that the last 11 bits of the source and destination IP in the dataset is anonymized in order to protect user privacy.

Note: The Netflow measurement dataset (*Netflow_dataset.csv*) is provided in BB.

The fields in the above dataset are:

1. dpkts and doctets:- # of packets and bytes in the flow, respectively
2. first and last:- Timestamps of the first and last packets in the flow, respectively
3. srcaddr and dstaddr:- Source and destination IP addresses, respectively
4. srcport and dstport:- Source and destination transport port numbers, respectively
5. prot:- Transport protocol, e.g., TCP, UDP
6. src_mask and dst_mask:- the length of the longest matching IP prefix for the source and destination IP addresses, respectively
7. src_as and dst_as:- Autonomous system (AS) that originated the IP prefixes matching the source and destination IP addresses, respectively.

Task Details:

Your tasks are to answer the following questions by analyzing the above Netflow dataset. Please note that you might have to write small (python) scripts and plotting tools to provide response to the questions. You may use libraries like numpy, scipy, and matplotlib.

a) What is the average size of the packets across all the traffic captured in the dataset? Provide the description on how you calculated this number.

b) Plot the Complementary Cumulative Probability Distribution (CCDF) of flow durations (i.e., the finish time minus the start time) and of flow sizes (i.e., number of bytes, and number of packets).
   i. First plot each graph with a linear scale on each axis, and then a second time with a logarithmic scale on each axis.
   ii. What are the main features of the graphs?
   iii. Why is it useful to plot on a logarithmic scale?

c) Summarize what kind of traffic is going through this router the most.
   i. Create two tables, listing the top-ten port numbers by sender traffic volume and by receiver traffic volume including the percentage of traffic (by bytes) they contribute.
   ii. Explain what applications are likely be responsible for this traffic. (See the IANA port numbers reference for details.) Explain any significant differences between the results for sender vs. receiver port numbers.

d) Summarize the traffic volumes based on the source IP prefix.
   i. What fraction of the total traffic comes from the most popular 0.1% of source IP prefixes? (count by number of bytes)
   ii. What fraction of the total traffic comes from the most popular 1% of source IP prefixes?
   iii. What fraction of the total traffic comes from the most popular 10% of source IP prefixes?
   iv. Some flows will have a source mask length of 0. What fraction of traffic (by bytes) that has a source mask of 0?
   v. Now, exclude this traffic (mask=0) from the rest of the analysis and answer d(i), d(ii), and d(iii).

e) Assume an Institute-A has the 128.112.0.0/16 address block.?
   i. What fraction of the traffic in the dataset is sent by A? Measure both in terms of bytes and packets.
   ii. What fraction of the traffic in the dataset is sent to A? Measure both in terms of bytes and packets.

What to be submitted:

1. A **PDF formatted report** with all required answers, and necessary evidences as asked above including scripts, execution samples, and references used. The filename should be formatted as “*Lastname_firstname_Assign5.pdf*”.

Submission:

Submit your **assignment** on **Blackboard only.**