CS 455 Introduction to Computer Networks
Homework 3
Due midnight 12/3/2014 on Angel

In this homework, you will make use of several tools including *cut, sort, grep, uniq,* and *awk* to analyze an Internet trace. The goals of this homework is to get familiar with some basic (network) analysis tools and to understand some important properties of the Internet.

Before going into the details, you should review/learn some important concepts and tools.

- Understand how to use cut, sort, grep, uniq, and awk tools. In linux, use *man [cut/sort/grep/uniq/awk]* for their description.
- Understand the cumulative distribution function (CDF - http://en.wikipedia.org/wiki/Cumulative_distribution_function) and how to plot it. For plotting, you can use gnuplot, matlab, excel, or whatever tool you know.

1. Internet Traffic Analysis

You are going to analyze a 5-minute trace of Netflow records captured from a router in the Internet2 (internet2.edu) backbone network that connects major universities, including Washington State University, in United States. The trace is provided on the class website. Measurements were taken with 1/100 sampling, so the data reflect 1% of the traffic at the router. The trace is in the *csv* format; fields are separated by commas. Loosely defined, a *flow* summarizes traffic from a source to a destination. Important fields are described below:

- *unix_secs*: unix seconds
- *unix_nsecs*: unix nano seconds
- *sysuptime*: system uptime
- *dpkts*: number of packets
- *doctets*: number of bytes
- *first*: timestamp of the first packet in the flow
- *last*: timestamp of the last packet in the flow
- *srcaddr*: source address
- *dstaddr*: destination address. The last 11 bits of the source and destination IP addresses have been anonymized to protect user privacy
- *srcport*: source port number
- *dstport*: destination port number
- *prot*: transport layer protocol (e.g, 6 for TCP, 17 for UDP). Check http://www.iana.org/assignments/protocol-numbers/protocol-numbers.xhtml for the full list
- *src_mask*: length of the longest matching IP prefix for source IP
- *dst_mask*: length of the longest matching IP prefix for destination IP

- *src_as*: The Autonomous System (AS) that originate the IP prefixes matching the source IP
- *dst_as*: The AS that originates the IP prefixes matching the destination IP

For example, the first two lines of the trace look like following

```
#:unix_secs,unix_nsecs,sysuptime,exaddr,dpkts,doctets,first,last,engine
_type,engine_id,srcaddr,dstaddr,nexthop,input,output,srcport,dstport,pr
ot,tos,tcp_flags,src_mask,dst_mask,src_as,dst_as
1285804501,0,2442636503,127.0.0.1,1,40,2442590868,2442590868,0,0,128.10
3.176.0,24.8.80.0,64.57.28.75,213,225,80,51979,6,0,17,16,0,1742,0
```

In this case, there is a flow with 1 (dpkts =1) 40-byte (doctets=40) packet. The packet arrives at time `2442590868.` The packet was sent from source `128.103.176.0` to destination `24.8.80.0`. The source port number is 80 (e.g., http) to the destination port number 51979. The protocol number is 6 (TCP). The tcp_flags is 17, FIN and ACK bits are set (e.g, a FIN-ACK packet). The source AS was 1742 while the destination AS was not known.

2. Homework 3

Please write scripts (e.g., bash) to produce answer the following questions

2.1. What is the *average packet size*, across all traffic in the trace? Describe how you computed this number.

2.2. Plot the Cumulative Probability Distribution (CDF) 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). First plot each graph with a *linear* scale on each axis, and then a second time with a *logarithmic* scale on each axis. What are the main features of the graphs? Why is it useful to plot on a logarithmic scale?

2.3. Summarize the traffic by which TCP/UDP *port numbers* are used. Create two tables, listing the top-ten port numbers by *sender* traffic volume (i.e., by source port number) and by *receiver* traffic volume (i.e., by destination port number), including the percentage of traffic (by bytes) they contribute. Where possible, explain what applications are likely responsible for this traffic. (See the IANA port number reference for details http://www.iana.org/assignments/service-names-port-numbers/service-names-port-numbers.xhtml) Explain any significant differences between the results for sender versus receiver port numbers.

2.4. Aggregate the traffic volumes based on the source IP prefix. What fraction of the total traffic comes from the most popular 0.1% of source IP prefixes? The most popular 1% of source IP prefixes? The most popular 10% of source IP prefixes?

2.5. WSU has the *134.121.0.0/16* address block. What fraction of the traffic (by bytes and by packets) in the trace is sent by WSU? To WSU? (Note: should use only 16 prefix digits for searching)

3. Tips and Submission

- Spend some time to get familiar with the tools first. Note that you are asked to write scripts for analyzing the trace. You cannot use excel for analyzing except plotting. Try with few lines first (e.g., using head –n 40) before trying the whole trace.

For example:

```
head –n 40 flow.csv | cut –d "," –f6 | sort | uniq –c | sort –nr | awk
'{ bytes +=$1*$2; packets+=$1}; END {print bytes, packets}'
```

extracts the sixth comma-separated field (i.e., number of bytes in the flow), counts the number of occurrences of each value, lists the frequency counts from most-popular value to least-popular, calculates to total number of bytes and total number of packets, and finally prints both total numbers out.

- Please submit both the scripts and the written up answers (with graphs of course) for the five questions above on Angle by midnight of the due date.