INFO8011: Network Infrastructures
Network Performance & Measurements

B. Donnet, S. Ben Mariem

1 Overview

Many networks collect NetFlow measurements directly from the routers. For more information, read the Wikipedia and Cisco pages on NetFlow (see eCampus).

In this project, you will analyze six hours of NetFlow records captured from a router in the Université de Liège directly connected on the Belnet network. Note that the NetFlow data has been anonymized (IP source and destination), to protect user privacy. The records have been parsed into CSV (comma-separated variable) format, with the names of the fields listed in the first row of the file.

The flow records to analyze is available here: `http://gofile.me/2PfYg/ck36XkKi0` (1.63GB, compressed).

2 Netflow

The important fields in the NetFlow data are[1] provided by Table 1.

For example, looking at the first two lines of the file, we have[2]

```
#:unix_secs,unix_nsecs,sysuptime,exaddr,dpkts,doctets,first,last,engine_type,
engine_id,srcaddr,dstaddr, nexthop,input,output,srcport,dstport,prot,
tos,tcp_flags,src_mask,dst_mask,src_a s,dst_as
1285804501,0,2442636503,127.0.0.1,1,40,2442590868,2442590868,0,
0,128.103.176.0,24.8.80.0,64.57.28.75,213,225,80,51979,6,
0,17,16,0,1742,0
```

In this example, we have a flow with one 40-byte packet that arrived at time 2442590868. The packet was sent by source 128.103.176.0 to destination 24.8.80.0.[3] The source port is 80 (i.e., HTTP) and the destination port is 51979 (i.e., an ephemeral port), suggesting this is traffic from a Web server to a Web client. The protocol is 6 (i.e., TCP). The `tcp_flags` of 17 suggests that the `ACK` and `FIN` bits were set to 1, suggesting this is a `FIN-ACK` packet; the other packets of the Web transfer were presumably not included in the flow record due to packet sampling. The source and destination masks were 16 and 0, respectively, meaning that the source prefix 128.103.0.0/16 and the destination prefix was either unknown or 0.0.0.0/0. The source AS was 1742 (Harvard, according to `whois -h whois.arin.net 1742`), and for whatever reason the destination AS was not known.

[1] Note that fields names might slighty differ in the file provided but they are sufficiently explicit so that you can link them with this document.

[2] each line has been cut to fit the page size.

[3] Remind that the last 11 bits are set to 0 due to the data anonymization.

Field	Description
`unix_secs`	Current count of seconds since 0000 UTC 1970
`unix_nsecs`	Residual nanoseconds since 0000 UTC 1970
`sysuptime`	Current time in milliseconds since the export device booted
`exaddr`	Export device IP address
`dpkts`	Number of packets in the flow
`doctets`	Total number of Layer 3 bytes in the packets of the flow
`first`	SysUptime at start of flow
`last`	SysUptime at the time the last packet of the flow was received
`engine_type`	Type of flow-switching engine
`engine_id`	Slot number of the flow-switching engine
`srcaddr`	Source IP address
`dstaddr`	Destination IP address
`nexthop`	IP address of next hop router
`input`	SNMP index of input interface
`output`	SNMP index of output interface
`srcport`	TCP/UDP source port number or equivalent
`dstport`	TCP/UDP destination port number or equivalent
`prot`	IP protocol type (for example, TCP = 6; UDP = 17)
`tos`	IP type of service (ToS)
`tcp_flags`	Cumulative OR of TCP flags
`src_mask`	Source address prefix mask bits
`dst_mask`	Destination address prefix mask bits
`src_as`	Autonomous system number of the source, either origin or peer
`dst_as`	Autonomous system number of the destination, either origin or peer

Table 1 – Fields in the NetFlow CVS file.

3 Questions

In this project, you will have to provide an answer to the following questions.

Question 1

Plot, as a Cumulative Distribution Function (CDF) the packet size distribution, across all traffic in the trace. Describe the plot and explain what you can learn from the plot.

In addition, what is the average packet size, across all traffic in the trace?

Question 2

Plot the Complementary Cumulative Distribution Function (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). 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? What artifacts of Netflow and of network protocols could be responsible for these features? Why is it useful to plot on a logarithmic scale?

Question 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 numbers reference for details.) Explain any significant differences between the results for sender vs. receiver port numbers.

Question 4

Aggregate and plot (appropriately) the traffic volumes based on the source IP prefix (it is up to you to find the most suitable way to agregate anonimized IP addresses into meaningful prefixes). 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? Some flows will have a source mask length of 0. Report the fraction of traffic (by bytes) you cannot aggregate, and then exclude this traffic from the rest of the analysis. That is, report the top 0.1%, 1%, and 10% of source prefixes that have positive mask lengths.

Question 5

Université e Liège owns the 139.165.0.0/16 address block. Within this block, the Montefiore Institute has the 139.165.223.0/24 address block, while the RUN (Research Unit in Networking) team has the 139.165.222.0/24 address block.

What fraction of the traffic (by bytes and by packets) in the trace is sent by (*i*) Montefiore and (*ii*) RUN? To (*iii*) Montefiore and (*iv*) RUN? What do you observe?

4 Achievements

For this assignment, it is expected to write a report using the LaTeXtemplate (both LaTeXsource and BibTeXsource file) provided on eCampus. A few macros are provided in the LaTeXheader. Modify them so that the report is carefully generated (consider using `pdflatex`). Note that your report must be **6 pages** long maximum.

For processing the dataset and plotting results, we want you to use Python (in particular, the Pandas and Matplotlib libraries, as we did during Lab 1). Python files you have written must be included in the archive to submit (see Sec. 6 for details).

5 Gradings

This assignment accounts for 15% of the final grade. The mark will be distributed as follows:

- **Coding style**. This assignment is not given in a programming course context. However, at this step of your studies, we assume you are able to provide elegant solutions to complex problems. The coding style (and, consequently, your solution elegance) will be part of the grading.
- **Documentation**. In the fashion of other popular programming languages, Python comes with a powerful tool for documenting code: `Docstring`.[4] We ask you to fully document your code with respect to Docstring standards.
- **Report**. You must write a report[5] that provides answers to questions raised in Sec. 3. The grade will be based on the report quality, i.e., the capacity to provide a scientific document (abstract, introduction, content, conclusion[6]), the plots quality, and your ability to discuss and put into perspective your results. If, during data processing, you have implemented efficient and unusual algorithm(s), do not hesitate to describe it/them in the report.

6 Submission

The submission of your lab is subject to the following rules:

1. an archive named **Team_xx.zip** (where **xx** must be replaced by your TeamID), which contains all required files,must be uploaded on the submission platform (see `https://submit.montefiore.ulg.ac.be`)
2. the deadline is **October, 25th, 08h00** (hard deadline).

[4] See, for instance, `https://www.datacamp.com/community/tutorials/docstrings-python`

[5] See LaTeXtemplate on eCampus.

[6] Summarize here what you have learned with this assignment.