Final Assignment — From measurements to conclusions

Markus Peuhkuri Tran Thien Thi

2020-10-26

General guidelines for this final assignment

Please note this assignment might requires quite long amount of time and work especially if you are not familiar with software oriented analysis methods and tools so please take this into your considerations when you are planning for your schedule and deadlines!

Report

You should prepare a report based on your analysis by including all the details of the results in a written report. Submission of the report consists of two phases:

- Mandatory participation on review with assistants. You must enrol to one of sessions at MyCourses. By that time you should have **at least an initial draft** and some of the analysis done. The sessions will follow format of weekly assignments i.e. discussion in groups and joint review and discussion about matter.
- The report will be returned via MyCourses before the end of November. A late submissions will only get grade 1 maximum.

The report should have two parts:

1. Main document explaining results and findings without technical details. This is like information that would be given to the customer hired you to make analysis.
2. Appendix contains detailed explanations on what have been done supplemented by commands used to get a result or draw a figure, if appropriate. Plain commands, scripts, or codes without comments are not sufficient. This is like information you would hand out to your colleague who needs to do similar analysis for an another customer.

   Also include samples of data sources, like 5-10 first relevant lines when appropriate. Do not include full data.

When you are asked to plot or visualise a certain parameter, make sure that your figures are as informative as possible and are really visualising a parameter(s) in question by selection of appropriate plot, units and scales (linear vs. logarithmic, ranges) and not just plotting some numbers and figures with default setting.

It is recommended to go through following processes for each dataset:

- Initial observations
- Pre-processing
- Analysis
- Conclusions

Address all the sections carefully and in the order where they come. Organise your report clearly, using sections for data sets, subsections for pre-processing, analysis, and conclusions for each data set. It is recommended that each plot contains a short description and also descriptive labels for the axis. Pay enough attention to the conclusions as they are considered to be one of the most important parts in evaluations.

Some tasks are repetitive, i.e. same analysis are done for multiple distinct data sets. It is much easier to do, if you create small functions or scripts that will just take different data. Or even run all analysis one go for all the data.

Assessment

Please note that in the review session the assignment must include at least draft state of most sections if final graphs, tables and conclusions are not available.

Final assignment has a total weight of 70 % in the final grade and will be graded with a continuous scale ranging from 0 to 5 where grades less than 1 is considered as the rejected. Both the assignment and the weekly assignments must be completed successfully (you should get at least a grade 1 from all) in order to pass the course.

The assignment is an individual work. You may cooperate with others by discussing the tasks - this is in fact *encouraged*, but **all output should be produced by yourself**. The assignment grade is composed of:

- Correct and insightful answers (weight 80%)
- Readability, clarity and style of the report (weight 20%)

Support

The assignment is meant to be individual work, but there are two kinds of support available for the students:

- Interactive exercise classes will be arranged on in October as needed.
- Review sessions later in October

- Assistant support sessions on exam week (week starting 22nd October)
- Slack for questions to course staff and also peer support.
- Discussion forum at course pages at http://mycourses.aalto.fi is also useful for two-way discussion and guidance.

Remember the correct discussion principals: in forums write a descriptive subject and describe your question or problem clearly. Also describe that you have already tried to do but had problems with. Course staff monitors forums and slack channels mainly on office hours but may not be able to give timely responses all the time because of other tasks. Course forum questions should address general issues such as clarifications on tasks, hints on tools. For code debug and quick questions slack may work better! If you have more than few lines of output / commands use some pastebin services like [dpaste.org](https://dpaste.org, fpaste.org, pastebin.ca, paste.ee, gist.github.com or an attachment.

Introduction

This final exercise will cover almost all the concepts covered so far in this course ranging from data measurements to driving the results and conclusion from the datasets. After the completion of this exercise the students shall have a solid understanding of how to get the desired data and final results out of the measured data from the network traffic.

This final assignment contains three main tasks both with several sub-tasks and final conclusions:

- Task 1: Flow data
- Task 2: Capturing packets
- Task 3: Analysing active measurements
- Final conclusions

In Task 1, you will be provided "data set FS" which you will need use to solve the required tasks. In Task 2, you will capture own "data set PS" which you will utilise to solve the required tasks. In Task 3 you will analyse active measurement data, "data set AS".

Prerequisites

This exercise requires students have a good understanding and hands on experience on all concepts and techniques mentioned so far in this course to properly answer the questions.

More information about available tutorials be found from material section of course web page on MyCourses.

There is an updated *ELEC-E7130 Network capture tutorial* at Supporting material section.

Task 1: Flow data

In task 1, we will use data set I that will be provided. First, you need to get access to it. Then, you will pre-process the data set so that you only have one subnetwork data. After that, the actual data analysis will happen and you will solve the required tasks.

Acquiring flow data

Data set I consists of anonymised flow measurements from an access network (if interested, see how they were created in the *Network capture tutorial*). A sample of users have been selected for the data collection. The time stamps on the flows are given in terms of UNIX epoch time.

This flow data is available at `/work/courses/unix/T/ELEC/E7130/general/trace` under three directories (please note the file sizes!). After sourcing use script, directory is in environment variable `$TRACE`.

Directories contain following data:

- `flow-continue`: output generated with `crl_flow` tool using 60 second timeout to expire flow. Time intervals are aligned as one hour.
- `flow-expire`: same as above, but all flows are expired when reporting period (one hour) ends.
- `tstat-stat`: output generated with tstat tool.

Note: Performing any file-handling operations in these directories is not possible with normal user privileges. You will need to redirect all operations to, for example, your home directory.

Data pre-processing

The given data set **FS1** contains all flow data from one whole day, which can be too massive. You do not need to analyse whole data set (except in task 1.6) but you should focus your analysis on single /24 network is based on following list. Select item based on the last digit on your student number. This data set is **FS2**.

Table 1: Subnetwork based the last digit of student number.

digit	subnetwork
0	133.60.164.0/24
1	133.60.165.0/24
2	133.60.168.0/24
3	133.60.169.0/24
4	133.60.170.0/24
5	133.60.171.0/24

digit	subnetwork
6	133.60.172.0/24
7	133.60.173.0/24
8	133.60.174.0/24
9	133.60.175.0/24

You can extract relevant data using e.g. `gawk` command. Let's assume you network is `192.0.2.0/24` and the tstat `log_tcp_complete` contains IP addresses in fields 1 and 15.

```
gawk '$1~/^192\.0\.2\./||$2~/^192\.0\.2\./' 1200.t2 > ~/my_1200.t2
```

The `gawk` command above seeks all the rows that have IP-address pattern of "192.0.2." in 1st column or 2nd column from the file `1200.t2`. Such rows that match this IP pattern will be outputted to new file `my_1200.t2` in your home directory.

In tstat log files files, IP addresses are in 1st and 15th fields.

In addition to this, other pre-processing may be needed. Document for your notes

- Commands or code that is used in pre-processing.
- Short samples (10 lines or so) taken from the distilled data.

Data analysis

After pre-processing, analyse the data set **FS2** carefully. The minimum requirements are detailed below, but additional insight and plots supporting those are welcomed. Each plot should contains a short description and also descriptive labels for the axis.

1.1: Plot traffic volume

Plot traffic volume as a function of time with at least two sufficiently different time scales (for example, bits per second in function of seconds and bits per second in function of minutes).

1.2: Flows by port numbers

Visualise flow distribution by port numbers. If you use histograms, consider if port numbers are continuous or discrete values.

1.3: OD-pairs

Plot origin-destination pairs by both by data volume (=bytes) and by number of flows (Zipf type plot).

1.4: Per user data volume

Compute the aggregate data volume for each user and draw a histogram to visualise distribution of user aggregated data. In other words, make one histogram that contains all users, no need to identify users from each other. (*user* would be one IP address within your assigned subnetwork)

1.5: Flow length distribution

Plot flow length distribution, its empirical cumulative distribution function, and key summary statistics. Fit a suitable distribution for the flow lengths and validate it.

1.6: Flow sampling

For this task, use **FS1** and take ALL flow data into account (i.e., not limiting the scope solely on your subnetwork).

Make two random selections from all flows by sampling flows from the 24h flow data: first selection to only include IPv4 traffic and the other only IPv6. Define your sampling process such that you will get about the same number of flows for this all flow data as in your assigned subnetwork. Document your selection process.

Select one of previous tasks (1.1-1.5) and perform same analysis for the both sampled data sets you just collected. Compare the results to the original task where you used your subnetwork (**FS2**) only. Can you say characteristics of your subnetwork is representative? Is there difference between IPv4 and IPv6?

1.7: Conclusions

Based on results above, explain your conclusions on data for:

1. Traffic volume at different time scales. Are there any recognisable patterns?
2. What are the 5 most common applications (study the port numbers)?
3. What kind of users there are in the network? Speculate on what kind of network this network could be based on traffic volumes and user profiles. Is your subnetwork different from larger population?

Please feel free to use additional visualisations to support your claims and conclusions if necessary.

Task 2: Capturing packets

Data set II is obtained by packet capture, so first you will capture packets on your own. Then this captured data set will be pre-processed three different ways so that at the end of the pre-processing, you will have three data sets: **PS1**,

PS2, and **PS3**. PS1 will contain packets, PS2 will contain flows, and PS3 will only contain TCP connections. All these data sets will be analysed separately in data analysis phase.

Acquiring packet capture data

The recommended way get the packet trace is to carry out your own measurements. You will need to use your own computer or a network where you have an access and the right permission to perform packet capture to get the data.

You can use **dumpcap** (Wireshark) or **tcpdump** for getting those data. More information about the Wireshark and TCPdump can be found from material section of course web page on MyCourses.

The measurement period should be at least two hours long, while a day-long trace is much better as the more there is data the more interesting it is. You can use your own computer to perform the packet capture. In a case where you do not have a personal computer to do so, you can ask course staff for instructions how you can loan for a computer which can be used to perform the packet capture. As the last resort, you can use DEC traces from Internet Traffic Archive

Please note your report must clearly include packet capture metadata:

- What kind of trace file and tool/s you are using to perform the packet capture.
- Date, time, duration, measurement setting (in terms of profile if you are using the Wireshark) or file name if you are using the DEC-traces.
- Provide a short sample (10 lines or so) of the data taken from your capture file.

Data pre-processing

After you have the raw packet data, you need to convert it to a suitable format. The data will be analysed both at packet level and at flow level.

At the first phase, you *can* anonymise your traces using `crl_to_pcap` utility. This is not mandatory but if you choose to anonymise the trace, use the anonymised trace consistently in all your analysis to avoid confusions. Note that anonymisation will render geo-locating IP addresses impossible (can be problematic in 2.5).

Three (3) data sets will be distilled from the raw data. We refer to these as **PS1**, **PS2**, and **PS3**, respectively.

Your report must include:

- Commands or code that is used in pre-processing for each case.

- Short samples (10 lines or so) of the distilled data in each case (for PS3, one connection summary is enough).

Following is the precise structure we need for each dataset:

Cleaning the data packets (PS1)

Regarding pre-processing of PS1, it depends. Have a look at the data analysis section of required tasks in order to get an idea about which information on individual packets are needed in the different sections. Then clean the collected data to contain only the relevant columns. In other words, pre-processing PS1 depends on the required tasks. Document what you have selected.

Converting packet trace to flow data (PS2)

Regarding pre-processing of PS2, you have multiple options to convert the captured packets into flow data. To produce flow data, you could use use `crl_flow` utility from CoralReef package with time-out of 60 seconds, you could use tstat, or you could use your own script to extract the flow data.

TCP connection statistics (PS3)

Regarding pre-processing of PS3, you can use `tcptrace` command on your captured file to produce statistics from TCP connections as follows:

```
tcptrace -l -r -n --csv myown.pcap > myown-tcp.csv
```

Above command will produces statistics about every TCP connection seen. You get more verbose output if you omit `--csv` option (try it to get idea of data items, csv is easier to parse). You can find more details from manual page of the command `man tcptrace`.

Data analysis

Analyse the data set carefully. The minimum requirements are detailed below, but additional plots and insights are welcomed. Each plot should contains a short description and also descriptive labels for the axis.

Packet data PS1

- 2.1: Visualize packet distribution by port numbers.
- 2.2: Plot traffic volume as a function of time with at least two sufficiently different time scales.
- 2.3: Plot packet length distribution (use bins of width 1 byte), its empirical cumulative distribution function and key summary statistics.

Flow data PS2

- 2.4: Visualise flow distribution by port.
- 2.5: Visualise flow distribution by country. **Hint:** use geoip to transform IP addresses to countries. If you have anonymised IP addresses, the results can be misleading (depending on level of anonymisation).
- 2.6: Plot origin-destination pairs by both by data volume and by flows (Zipf type plot).
- 2.7: Plot flow length distribution, its empirical cumulative distribution function and key summary statistics.
- 2.8: Fit a distribution for the flow lengths and validate the model.
- 2.9: Compare the number of flows with 1, 10, 60, 120 and 1800 second timeouts. In this you need to generate flow data with multiple times.

TCP connection data PS3

For the TCP connection statistics we are interested in retransmissions. Study the association of retransmissions to:

- 2.10: Round-trip times and their variance.
- 2.11: Total traffic volume during the connection (you get the volume from PS2).

Conclusions

Explain your conclusions for:

- Traffic volume at different time scales. Are there any recognizable patterns?
- Characteristics of top 5 most common applications used (studies of the port numbers).
- Comparison of above results with result from data set **FS2**.
- Differences of flow and packet measurements in the example case.
- Your findings on retransmissions.

Task 3: Analysing active measurements

As a result from the *Basic Measurements*, you should have at least two weeks worth of measurement data:

- Latency (data sets **AS1.x**), where **x** includes 3 name servers with DNS (d1, d2, d3) and ICMP (n1, n2, n3), 3 research servers (r1, r2, r3) and 2 iperf servers (i1, i2).

- Throughput measurements (data sets **AS2.x**). where **x** is i1 (`ok1`) and i2 (the other, far away).

Remember to describe where you made the measurements from, i.e. from Aalto servers, your own laptop or from some other environment.

3.1 Latency data plots (AS1.x)

- Provide box plots including all successful latency measurements from **AS1.x** data sets (one box per data set; ignore lost packets). Make sure numerical values could be seen. What observations can be made, for example differences between sites? Were there differences in **AS1.d**_N_ and **AS1.n***N?
- Another graph but this time consider also the lost packets. One option is define all lost packets to have some maximum delay (like 2 seconds, also any packet delayed more than 2 seconds would be shown as 2 s) and make single box plot for each dataset. There can be other options too.
- Provide PDF and CDF plots including all **AS1.x** delay distributions.
- Characterise delay distributions according to ITU-T Y.1541 in a tabular form for all **AS1.x**.

3.2 Latency data time series

- Plot time series of each data set **AS1.x**. Consider appropriate scaling for comparison. Any observations for e.g. diurnal patterns?
- Select **AS1.i2** and minimum two other most interesting data sets from **AS1.x**. Make an autocorrelation plot. Any observations?

3.3 Throughput

- Plot throughput measurements as box plots for both **AS2.x** data sets
- From throughput, compute and tabulate for both data sets representative values using
  - mean
  - harmonic mean
  - geometric mean
  - median

3.4 Thoughput time series

- Plot time series of each data set **AS2.x**. Consider appropriate scaling for comparison. Any observations for e.g. diurnal patterns?
- Make autocorrelation plot on **AS2.x** data sets. Any observations? Compare also to 3.2.

Conclusion

Discuss on conclusion on Task 3 for at least following topics:

- Describe the system your made measurements from measurement. What kind of impact it had for measurements?
- Did there exists some correlation between path length (number of routers, it can be check with **traceroute** and/or with TTL value of ICMP Echo Responses) and measurement stability? If you happened to record also TTL value, did it change over time?
- Did throughput and latency have any correlation?

Final conclusions

After you have completed the Task 1-3, you are now almost done. Based on these tasks, answer the following questions.

- How was your own traffic (Task 2) different from the data provided (Task 1)? What kind of differences you can identify? What could be a reason for that?
- Comparing RTT latency about TCP connections (2.10), were active latency measurements around the same magnitude or was another much larger than the other?
- How do you rate complexity of different tasks? Was some tasks more difficult or laborious than others? Did data volume cause any issues with your analysis?