Final Assignment — From measurements to conclusions

Markus Peuhkuri Tran Thien Thi Yu Fu Weixuan Jiang

2023-10-09

General guidelines for this final assignment

Please note this assignment might require quite an extended 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!

Some tasks are repetitive, i.e. same analysis is 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. Also note that you are able to take advance of scripts and tools used in earlier assignments.

Assessment

Please note that in the review session, the assignment **must include** at least a draft state of most sections even if final graphs, tables and conclusions need not to be available.

Final assignment has a total weight of 60% in the final grade and will be graded on a continuous scale ranging from 0 to 100 points, where points less than 50 are considered as rejected. Both the final assignment and the weekly assignments must be completed successfully (you should get at least a grade 1 for all) in order to pass the course.

The assignment is 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 detailed scoring rules can be found in **Grading standard** section.

Support

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

- Interactive exercise classes
- Review sessions about a week before deadline (schedule will be published a week in advance)
- Course Zulip *finalassignment* stream for questions to course staff and also peer support.

Remember the correct discussion principles: write a descriptive subject in forums and clearly describe your question or problem. Also, describe what you have already tried to do but had problems with. Course staff monitors Zulip channels mainly during office hours but may not be able to give timely responses all the time because of other tasks. For code debug and quick questions, Zulip works nicely. If you have more very long text (more than few tens of lines of output / commands) use some Pastebin services like dpaste.org, fpaste.org, pastebin.ca, paste.ee, gist.github.com or an attachment.

Introduction

This final exercise will cover almost all the concepts taught in this course, ranging from data measurements to deriving results and conclusions from datasets. Upon completing this exercise, students will have a solid understanding of how to obtain the desired data and final results from measured data related to network traffic.

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

- Task 1: Capturing data
- Task 2: Flow data
- [Task 3: Analyzing active measurements]
- Final conclusions

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

Prerequisites

This exercise requires students to have a good understanding and hands-on experience with 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 a *ELEC-E7130 Network capture tutorial* at supporting material section.

Task 1: Capturing data

Data set I is obtained by packet capture, so first you will capture packets on your own. Then this captured data set will be pre-processed in 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 analyzed separately in the data analysis phase.

Acquiring packet capture data

The recommended way to 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 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 the material section of the 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 data there is, 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 on how you can loan a computer that can be used to perform the packet capture. As a last resort, you can use some publicly available traces

For this part, 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 some public 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 analyzed both at packet level and at flow level.

In 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 analyses to avoid confusion. Note that anonymisation will render geo-locating IP addresses impossible (can be problematic in 1.6).

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

For this part, 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 will vary based on the specific requirements of the data analysis tasks. To determine the necessary information on individual packets for different sections, refer to the required tasks in the data analysis section. Clean the collected data to retain only the relevant columns accordingly. In other words, pre-processing PS1 depends on the specific tasks. Remember to thoroughly document the selection process made during the pre-processing.

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 generate flow data, you could use the `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. These choices offer effective ways to preprocess the PS2 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
```

The provided command will generate statistics for each TCP connection observed in the captured file. If you omit the `--csv` option, you will receive more detailed output (feel free to try it to get an overview of the data items, but keep in mind that the CSV format is easier to parse by programs). For additional information, you can refer to the manual page of the tcptrace command by using `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 contain a short description and also descriptive labels for the axis.

Packet data PS1

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

Flow data PS2

- 1.4: Visualise flow distribution by port numbers.
- 1.5: Plot traffic volume as a function of time with at least two sufficiently different time scales.
- 1.6: 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).

- 1.7: Plot origin-destination pairs both by data volume and by flows (Zipf type plot).
- 1.8: Plot flow length distribution, its empirical cumulative distribution function and key summary statistics.
- 1.9: Fit a distribution for the flow lengths and validate the model.
- 1.10: Compare the number of flows with 1, 10, 60, 120 and 1800 second timeouts. In this, you need to generate flow data multiple times.

TCP connection data PS3

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

- 1.11: Round-trip times and their variance.

  > **NOTE:** Among the various columns, you might find **RTT_{avg, min, max}_{a2b, b2a}** particularly relevant for your analysis. These columns provide information about the average, minimum, and maximum round-trip times for the respective directions of communication. Consider focusing on these columns to gain insights from the data.

- 1.12: 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 identifiable patterns or trends that you observed?
- The top 5 most common applications based on their port numbers. Identify the corresponding applications (e.g., HTTPS application) and analyze their characteristics.
- Differences of flow and packet measurements in the example case.
- Your findings on retransmissions.

Task 2: Flow data

In task 2, we will use data set II, which will be provided to you. First, you need to obtain access to the dataset. Once you have access, you will pre-process the data set to extract only the relevant subnetwork data. After completing the pre-processing, you will proceed with the data analysis and work on solving the required tasks.

Acquiring flow data

Data set II 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 has 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 the 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-log`: 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 or `/tmp` directory if your home folder does not

have enough space. Note that files in the `/tmp` folder can be deleted at any time, so use it only for intermediate files, not your code files.

Data pre-processing

The given data set **FS1** contains flow data from an entire day, which can be quite large. For your analysis, you do not need to examine the entire data set (except for task 2.3). Instead, you can select one of the three directories that best suits your analysis type. Please focus on a single /24 network from the list below, based on the last digit of your student number. This selected data set will be referred to as **FS2**.

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

digit	subnetwork
0	163.35.10.0/24
1	163.35.158.0/24
2	163.35.94.0/24
3	163.35.139.0/24
4	163.35.138.0/24
5	163.35.93.0/24
6	163.35.92.0/24
7	163.35.250.0/24
8	163.35.235.0/24
9	163.35.116.0/24

As an example, let's assume you have selected the `1200.t2` file from the `tstat-log` directory. If you want to extract relevant data for your own network with the IP address range of 192.0.2.0/24, you can use the `gawk` command as follows:

```
gawk '$1~/^192\.0\.2\./||$15~/^192\.0\.2\./' 1200.t2 > ~/my_1200.t2
```

In this command, the `gawk` program searches for rows in the `1200.t2` file where the IP address in either the 1st column or the 15th column matches the pattern "192.0.2.". The matched rows are then saved into a new file named `my_1200.t2` in your home directory.

Please note that in tstat-log files, IP addresses can be found in the 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 contain a short description and also descriptive labels for the axis.

2.1: Plot traffic volume

Select one of the previous tasks (1.4-1.5, 1.7-1.9) and perform the same analysis for the **FS2** data set. This means that you should choose either tasks 1.4 and 1.5, or tasks 1.7, 1.8, and 1.9. Once you have chosen the task, apply the analysis steps to the **FS2** data set.

2.2: 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)

2.3: 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 the previous tasks (2.1-2.2) and perform the same analysis for both sampled data sets you just collected. Compare the results to the original task where you used your subnetwork (**FS2**) only. Can you say the characteristics of your subnetwork is representative? Is there a difference between IPv4 and IPv6?

2.4: Conclusions

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

- Traffic volume at different time scales. Are there any identifiable patterns or trends that you observed?
- Identify the top 5 most common applications by studying their port numbers.
- 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?

- Comparison of the above results with the result from data set PS2.

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

Task 3: Analysing active measurements

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

- Latency measurements (data sets **AS1.x**), where **x** includes:
  - AS1.d1: Name server1 with DNS
  - AS1.d2: Name server2 with DNS
  - AS1.d3: Name server3 with DNS
  - AS1.n1: Name server1 with ICMP
  - AS1.n2: Name server2 with ICMP
  - AS1.n3: Name server3 with ICMP
  - AS1.r1: Research server1
  - AS1.r2: Research server2
  - AS1.r3: Research server3
  - AS1.i1: Iperf server1
  - AS1.i2: Iperf server2
- Throughput measurements (data sets **AS2.x**), where **x** includes:
  - AS2.i1: Iperf server1
  - AS2.i2: Iperf server2

3.1 Latency data plots (AS1.x)

- Please create box plots for all successful latency measurements from the **AS1.x** data sets. Each data set should have its own box plot, and any lost packets should be excluded from the analysis. Ensure that the numerical values are clearly visible on the plots. What notable observations can be identified after examining the plots of successful latency measurements? Were there differences in **AS1.d**_N_ and **AS1.n***N?
- Another graph but this time also consider the lost packets. One option is to define all lost packets to have some maximum delay (like 2 seconds, also any packet delayed more than 2 seconds would be shown as 2s) and make a single box plot for each dataset. There can be other options too.
- Provide PDF and CDF plots including all **AS1.x** delay distributions.
- Create a table summarizing delay distributions for all **AS1.x** data sets, following the guidelines of **ITU-T Y.1541**. The table should include the following parameters:
  - First packet delay
  - Mean delay
  - Proportion of packets outside an acceptable delay interval (predefined by you in advance)

  - Distance between two quantiles like 0.95 and 0.5

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?
- Choose **AS1.i2** along with at least two other data sets of interest from **AS1.x**. Create an autocorrelation plot for these data sets. Are there any observations or patterns that stand out from the plot?

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 Throughput 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 the following topics:

- Describe the system from which you obtained measurements and any challenges you encountered during the process.
- Was there any correlation between the path length (number of routers, which can be checked using **traceroute** and/or the TTL value of ICMP Echo Responses) and the stability of measurements? If you also recorded the TTL value, did it change over time?
- Was there any observable correlation between the throughput and latency?

Final conclusions

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

- How was your own traffic (Task 1) different from the data provided (Task 2)? What kind of differences can you identify? What could be a reason for that?

- Comparing RTT latency about TCP connections (3.1), were active latency measurements around the same magnitude or was another much larger than the other?
- Discuss how data protection needs to be taken into account if you as a network provider employee were doing similar measurements as in this assignment in a network provider network (traffic generated by customers that may be private persons or companies).
- Discuss how data protection needs to be taken into account if you as a company ICT support group employee were doing similar measurements as in this assignment in a company network (traffic generated by employees and customers).
- How do you rate the complexity of different tasks? Were some tasks more difficult or laborious than others? Did data volume cause any issues with your analysis?

Grading standard

To pass this course, you need to achieve at least 50 points in this assignment. And if you submit the assignment late, you can get a maximum of 50 points.

You can get up to 100 points for this assignment:

Task 1

- Describe clearly the method used for the measurement. And provide the conclusion of preliminary observation. (2p)
- For data pre-processing, describe the methods you use (2p for each dataset generated). (6p)
- For each analysis sub-task successfully completed, you can get 2 points. (24p)
- Answer the questions raised in the conclusion section and provide your own opinions. (10p)

Task 2

- Describe the pre-processing methods and steps. (2p)
- For each subtask, you can get 2.5 points after completion. (7.5p)
- Answer the questions raised in the conclusion section and provide your own opinions. (10p)

Task 3

- Describe the measurement method and environment. (2p)
- Describe the pre-processing methods and steps. (4p)
- For each subtask, you can get 2.5 points after completion. (10p)
- Answer the questions raised in the conclusion section and provide your own opinions. (7.5p)

Final conclusion

- For each question, you can get 3 points (15p)

The quality of the report (bonus 5p)

- Good explanations
- Interesting findings and conclusions
- Beautiful structure
- etc.

The instruction of assignment

For the assignment, your submission must contain (Please don't contain original data in your submission):

- A zip file that includes your codes and scripts.
- A PDF file as your report.

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 enroll to one of the sessions at MyCourses. By that time, you should have **at least an initial draft** and some of the analysis done. The sessions will follow the 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 deadline. 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 who hired you to make an analysis.
2. Appendix contains detailed explanations on what has 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 a similar analysis for 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 a selection of appropriate plot, units, and scales (linear vs. logarithmic, ranges) and not just plotting some numbers and figures with the default setting.

It is recommended to go through the 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. **Always refer to task number in your report**. Easiest way is to use same numbering scheme in chapters.

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 of evaluations.

Of course, you need a cover page indicating your name, student ID, and e-mail address.