CSDS 325/425: Computer Networks
Project #5
Due: December 5, 11:59 PM

The last project of the semester involves conducting active measurements of some network facet / behavior and reporting the results. The goal of this project is to move beyond learning about networks from the text book and towards learning about networks via observations of how they work in practice. This is a student-directed project. The broad contours of the project are outlined below, but part of the project is arriving at the theme and the details yourself.

The active measurement tools you may choose from are: (*i*) *ping* to measure latency, (*ii*) *curl/wget* to measure various facets about web transactions, (*iii*) *dig* to measure various facets about DNS lookups and (*iv*) *traceroute* to measure Internet routes and associated information. You may use alternate / additional tools, but you *MUST* discuss this with Mark prior to doing so.

Project #5 submissions will take the form of a tarball that includes a (*i*) project report in PDF (called "report.pdf"), (*ii*) any tools you wrote for the project to collect and/or analyze data (*iii*) a pointer to the data you collected and/or used in the project (in a text file called "data.txt"). More details below.

Report

Your report will consist of three sections, as follows:

1. **Introduction / Motivation:** The first section of your report should sketch what you intend to measure in high-level terms. Your project must be centered around some theme. E.g., understanding web servers. You may ultimately take different sorts of measurements, but they should all be in service of your theme. In other words, there must be some connective tissue between your measurements. The discussion in this section must be in broad terms (e.g., "delay" not "*ping* measurements"). This section should argue and ultimately motivate why this is an important aspect of the network to understand.

2. **Procedure:** This section should discuss how you go about conducting your study. This will include two sub-sections:

   2.1 **Gathering Data:** This sub-section of the report will discuss your data collection. In particular, while you will use standard tools to collect the data, you will need to write additional software to run these tools at scale (see below). These collection tools you develop must be described in this section. Further, the tools you create must be included in the project submission.

   2.2 **Analyzing Data:** This sub-section of the report will discuss *how* you analyzed your data and what tools you created to do so. These tools must also be included in the project submission.

3. **Results:** This section of your report will discuss the results of your measurements. You must include at least N distinctive results about your chosen theme.[1] Use tables and figures as appropriate—but these must be discussed and referenced in the text (e.g., to describe what the table/plot shows and highlight particularly important aspects). Your results should not be speculation, but instead be rooted in the data you have collected.[2]

Further report guidelines:

- Writing clarity is of the utmost importance. Grades will be reduced for unclear or sloppy reports. Grades will be reduced if crucial aspects of the measurement or analysis are not included. *Leave plenty of time to write a quality report.*
- Per syllabus, do not use any artificial intelligence tools to write your report.

[1] For 325 students $N = 5$. For 425 students $N = 7$.

[2] It is fine to speculate about why something may have happened, or when discussing possible measurements that could shed more light on the situation. But, you must have at least N results that are based on observations from the data you have collected.

- There is no page requirement or limit for the report. Use as much space as is appropriate.
- You must use at least a 10pt font.

Report Outline

Per the above the report will have a number of sections. The report outline looks like this:

Report Title
Name (Case Network ID)

1 Introduction
2 Procedure
2.1 Data Collection
2.2 Data Analysis Methods
3 Results
3.1 Delay Characterization (this is an example; the title of each section will depend on your project)
...
3.N Description of Result N

Reports that do not follow this specific outline will lose points.

Additional Considerations

Amount of Data

Conclusions based on only a single or a few data points are inherently weak. Such anecdotes are sometimes interesting and can be illustrative and as such, you can include them in your report to provide color. However, your N observations should be based on a decent amount of data. As a touchtone, if you can analyze the data by hand, you no doubt have too little. The amount of data we are interested in seeing will require a tool to analyze. At a minimum you'll want to invoke data collection tools hundreds of times and make use of thousands of data points for each of your N results.

Developed Tools

While you will use standard tools to do the network measurements, you will need to develop programs to run these measurement tools at scale. Further, you will need to develop programs to analyze the data. You may write your data collection and analysis in any programming/script language.

Measurement Vantage Point

You may take and analyze measurements on the class servers. However, you also may also use other machines you have access to, if that is more convenient or is helpful in the measurement process. The class servers have all measurement tools listed above installed. If you need a tool or package that would somehow benefit from being properly installed on the machine, let me know. Part of the data collection section should be a discussion of where the measurements were taken.

Limitations

Network measurements can sometimes cross into a grey area where ethical analysis is necessary to determine the appropriateness of the measurement. Further, your goal should be to be polite to the network and the operators. Full fledged measurement studies are often conducted with the blessing of network operators to ensure the measurements do not unduly impact the network and the operators have a full understanding of what is going on. In lieu of cooperating with the Case network operators we will use these conservative guidelines:

- Your project will not involve passive capture of network traffic.

- We will follow the "one at a time" rule. That is, you may only run one measurement at a time. Once that measurement is finished, you may conduct another measurement. E.g., if you are using *dig* to conduct 100 DNS lookups, only one invocation of *dig* may be running at a given time. You may not run 100 lookups in parallel.
- If you are transmitting packets at some rate into the network—e.g., using *ping*—the average rate can be no more than five packets per second. If your measurement requires a quicker "burst" or "flood" of traffic, you must follow that with a silent period. E.g., you may send 100 packets into the network within one second, but then must pause for 19 seconds before your next burst, as to ensure the average sending rate is at most five packets per second.

These limitations are conservative. Please be polite and follow the spirit of these guidelines. Do not try to find "loopholes". Measurements that end up being intrusive will result in a grade reduction.

Note: Even if you're using a non-Case networks for your measurements, please adhere to these limitations.

Measurement Targets

There will be a list of popular hostnames on the project 5 web page. You may use this—or likely a subset—to find a list of targets for your measurements. This is a resource not a requirement.

Database Lookups

Caution: There are databases on the Internet to tell you about hosts. E.g., you can provide an IP address and obtain its geographic location.[3] These database lookups can provide insights about why network measurements show some phenomenon. E.g., longer delays may correlate with traveling to further geographic locations. You may use these databases to help explain your results. However, lookups in these databases are *not* network measurements and cannot be used as one of your N results.

Sample Questions

The following are a set of notions to help guide your thinking when formulating your project. These are samples and you are encouraged to branch out.

- Think about probing multiple targets and comparing and contrasting the results.
- Think about probing multiple times to the same destination and then comparing and contrasting across these different measurements.
- Think about probing at different times of the day or different days of the week to determine if there are differences.
- Think about probing from different locations in the network to try to determine what is different.
- Think about probing for network characteristics (routes, delays, packet loss, etc.).
- Think about probing for characteristics of servers (e.g., what can we learn from HTTP headers?).

[3] Or, what some scheme has determined to be the geographic location—which may or may not be correct.

Final Bits

1. Submission specifications:
   (a) All project files must be submitted to Canvas in a gzip-ed tar file called "[CaseID]-proj5.tar.gz".
   (b) Your submission must contain a report in PDF format called "report.pdf".
   (c) Your report must include a title, your name and Case network ID.
   (d) Do not include multiple versions of your report in your submission.
   (e) Your submission must contain all code you wrote for this project.
   (f) Do not include raw data in your submission. However, you must make your raw data available when you submit your project. Place your raw data somewhere (Google Drive, Dropbox, your web server, etc.) and put a link in a "data.txt" file you include in your tarball. If this is problematic, please contact me and we will figure out a place to put your data.
   (g) Do not include directories in your tarball.
   (h) Do not use spaces in file names.
2. If you are in doubt about your project idea, please feel free to describe it in email or come chat about it during office hours.
3. Your submission may include a "notes.txt" file for any information you wish to convey during the grading process. We will review the contents of this file, but not of arbitrary files in your tarball (e.g., "readme.txt").
4. *WHEN IN DOUBT, ASK QUESTIONS!*