Open data for networking education

Date: March 7, 2022
Author: Phillipa Gill, Google

Measurement Lab (M-Lab) provides a range of open data sets that are ideal for students to learn networking concepts by analyzing real networking data. At the heart of our data is results of Internet speed tests run from around the world (~3M tests per day). Data is available for various aspects of these speed tests at varying levels of aggregation to make assignments that can adapt depending on the level of the students. From most to least aggregated these are some datasets that might be of interest for computer networking educators:

- NDT unified views[1]: This data is prefiltered to remove tests that ended erroneously and provides data about throughput, latency and loss for speed tests from clients around the globe.
- TCP_INFO[2]: This dataset includes data from the tcp_info (and bbr_info) kernel structure for the duration of a speed test.
- Packet headers[3]: This dataset includes packet captures (of headers) for the speed tests.
- Traceroute[4]: For each speed test, a traceroute (using MDA traceroute which can reveal multiple load balanced network paths) is logged and stored in a database.

Potential assignments/activities:
One benefit of our data mostly being available in bigquery is that it allows students to access the data using database queries from standard python notebook-style tools (eg., Colab[5]). Within this framework one could imagine assignments looking at the following:

- Characterize network performance for different types of networks; this could include looking at different countries, identifying different ISPs, or even looking at data for a single ISP over time to observe performance anomalies.
- Understanding the evolution of TCP state over the course of a connection. One could use the tcp_info data to observe how the TCP state evolves throughout the connection. Another variant of this idea would be to look at the packet headers and plot data transmitted over time to illustrate the concepts of congestion control.
- Understanding network paths. With traceroute data the students can see how many different networks the traffic flows through during the speed test.
- One could also imagine more advanced/open ended assignments correlating speed tests results with AS-level topology for certain countries (eg., those with low Internet freedom), or relating the M-Lab data with other sources of open data (eg., RIPE, CAIDA)

Next steps. We are interested in working with faculty members who would like to use M-Lab data in their courses. Ideally this would create a repository of activities and examples of analysis that could be shared across multiple institutions.

---

[1] https://www.measurementlab.net/tests/ndt/#ndt-data-in-bigquery
[2] https://www.measurementlab.net/tests/tcp-info/
[3] https://www.measurementlab.net/tests/pcap/
[4] https://www.measurementlab.net/tests/traceroute/
[5] https://research.google.com/colaboratory/