Title: Internet Quality Assessment and Measurement Technique Challenges

Authors: Udit Paul, Elizabeth Belding, Arpit Gupta (UC Santa Barbara) and Matt Calder (Microsoft, Columbia)

The availability and quality of Internet access vary across demographic attributes and locations (e.g., income levels, urban vs. rural, etc.). Using public and proprietary active-measurement data available from services, such as Measurement Lab and Ookla, one can characterize the relationship between the Internet's quality with the location- and demography-specific features. Poor internet access affects several facets of human lives including economic, education, health, and even the ability to self-isolate to prevent the spread of COVID-19. The active measurement data, particularly those collected from GPS-enabled mobile devices, are extremely valuable in studying Internet availability and quality. However, the user-initiated-tests' very nature introduces bias associated with non-random sampling in these active measurement tests.

In contrast, passive network measurements are unaffected by such bias as they do not require voluntary participation from the users. Network metrics (e.g., RTT, TCP throughput, etc.) passively collected by private organizations with extensive and diverse user bases such as Microsoft and Facebook can shed light on the Internet's quality. Such passive measurement data can help characterize the Internet's quality for areas under-represented in data collected by active measurement services, such as Ookla's Speedtest, MLab, etc. For areas where both forms of measurements have adequate representations, passive measurement data can help characterize the bias in data collected by active measurement services.

In an ideal world, performing a 1-1 comparison between network measurement data collected using different techniques (active vs. passive) and goals should be straightforward. However, in practice, comparing these two diverse forms of measurement techniques is difficult. More specifically, cloud services or content providers such as Microsoft and Facebook collect passive network measurements to aid traffic engineering and fault detection/localization. Additionally, to mitigate user privacy concerns and avoid the risk of disclosing personal information, user location is often inferred from passive measurements using IP geolocation techniques, which are known to be extremely unreliable for small geographic units such as census block or census block group. As the underlying demographic can change significantly with coarser levels of geographic granularity, establishing relationships between coarsely located passive network measurements and demographic attributes poses significant challenges.

In addition to their respective shortcomings, both active and passive measurement entities are further bounded by data-protection laws that prevent them from storing user information such as these measurements over an extended period. These limitations also affect the ability to study network behavior over long timeframes. To leverage the advantages offered by these diverse techniques of Internet measurement to discover potential relationships between attributes such as location and sociodemographics on Internet quality, we need to answer the following questions:

1. How can the inherent bias present in active measurement tests across different areas and demographics be accurately quantified?
2. How can the difference between active and passive measurement techniques be reconciled to make fair comparisons?
3. How can spatially fine-grained measurements be collected to understand more detailed Internet quality patterns without violating individual user privacy and data protection laws?