Not all paths are created equal: The case for datasets weighted by traffic volume

Matt Calder (Columbia) and Ethan Katz-Bassett (Columbia)

Internet researchers have produced exciting new research due to several improvements in measurement coverage. RIPE Atlas offers widely deployed vantage points. ZMap has commoditized Internet-wide scans. The cloud provider ecosystem allows easy world-wide deployment of vantage points.

While these improvements have increased our visibility, we may fail to correctly interpret measurements because, for a given scenario, not all paths are relevant or equivalent, and we lack the data to weight results properly. As a simple example of the impact of weighting, an IMC 2015 paper examined AS path lengths on the Internet [5]. When considering paths from PlanetLab (at the time, a popular testbed) to all prefixes with responsive destinations, only 2% are two ASes long. When instead issuing paths from Google cloud (responsible for more traffic than PlanetLab sites), 41% are two ASes long. When considering paths from Google to end-user destinations (the source of most connections for Google cloud), 61% are two ASes long. When weighting by query volume, 72% are two ASes long. The answer to a simple question like "how long are Internet paths" can vary dramatically depending on how exactly paths are counted. Research focused on user experience might care more about the paths weighted by query volume rather than the paths from PlanetLab to the Internet, many of which might carry very little traffic in practice. Similarly, the impact of a circuitous or unreliable path depends on how much that path is used. Information on which networks host how many users and source how many queries can also help calibrate and unbias measurements from platforms such as RIPE Atlas, which are skewed towards certain types of networks and regions. The IMC 2015 paper pointed out the issues using a private, sensitive dataset of query volumes, so did not provide a solution.

Motivated by these issues, we ask: **How can we weight relative network activity on a global scale to reflect the differences of user activity**? Depending on the research topic, it may be useful to weight networks/paths by bytes, users, or number of requests/connections. In others, it makes sense to use cost, utilization, prevalence, or service. The APNIC ISP Customer Estimate dataset is an example weighting that estimates the user population of every ISP using Google Ads [1], but has never been publicly validated. Ono and follow-up work [2,3,4] collected IP addresses of P2P clients, one measure of user activity. Bittorrent usage has dropped dramatically from ten years ago, and a rise in the popularity of VPN services is believed to obfuscate the true numbers by hiding the user's real network information. We require new techniques to overcome these limitations. We propose the following areas of investigation:

- New data collection techniques. Three directions can be pursued: privacy-aware sharing of private datasets (e.g., from large cloud providers); generation of new public data (e.g., a measurement plugin hosted on a wide range of sites by the community); or development of techniques that combine public datasets to approximate private ones. Funding for investigation into new techniques should focus on cost-effectiveness, and service and application diversity.

- Validation of new and existing approaches. As more research uses the APNIC data, it is important that the community understand its strengths and limitations, as well as those of new techniques.

- Guidelines for usage. Digital equity and the need to encourage innovation require that paths with little traffic should not be ignored. It will require careful consideration to balance a focus on popular paths with research that considers other paths.

The NSF could support this research by (a) funding long term infrastructure for collecting measurements; (b) facilitating data sharing with industry and (c) within academia; and (d) encouraging adoption of measurement plugins or other crowd-sourced approaches.

[1] APNIC Labs. Customers Per AS Measurements. Description: https://labs.apnic.net/?p=526 Data: https://stats.labs.apnic.net/aspop

[2] Chen, Kai, et al. "Where the sidewalk ends: Extending the Internet AS graph using traceroutes from P2P users." CoNEXT. 2009.

[3] Otto, John S., et al. "On blind mice and the elephant: understanding the network impact of a large distributed system." *Proceedings of the ACM SIGCOMM 2011 conference*. 2011.

[4] Choffnes, David R., and Fabian E. Bustamante. "Pitfalls for testbed evaluations of Internet systems." *ACM SIGCOMM Computer Communication Review* 40.2 (2010): 43-50.

[5] Chiu, Yi-Ching, et al. "Are we one hop away from a better internet?." *Proceedings of the 2015 Internet Measurement Conference*. 2015.

[6] Arnold, Todd, et al. "Cloud provider connectivity in the flat internet." *Proceedings of the ACM Internet Measurement Conference*. 2020.