Collecting, Aggregating and Sharing Better Internet Maps Using Multiparty Computation

Multi-disciplinary measurement research encompasses three major communities - network, security and infrastructure operators (I'll include content providers as a subset here) - which don't (often) willingly collaborate by sharing raw research data. Additionally, crucial *existing* pieces of data that would allow more comprehensive analyses are unavailable to researchers either because the data is proprietary, the source is unwilling to share it, or because there are (often confusing) varying notions of user and data privacy preventing sharing. **Network** measurement research questions are often revisited – how broken is BGP? where are broadband speeds and latencies bad? how prevalent is IPv6? – but the answers (while interesting) lead to incremental discoveries and rarely improve the end user experience or the resilience of the Internet infrastructure. **Security** (and privacy) researchers tackle related, overlapping questions - does DoH prevent effective security operations? is malware affecting IoT devices and traversing the Internet? did the Solar Winds breach have a significant impact? - but collect slightly different data using different methods and tools. (Shodan maps are an amazing complement to infrastructure maps). **Infrastructure operators** and content providers are a critical piece of the puzzle. They collect (unvetted) data on the physical Infrastructure (FCC Form 477) and end user access patterns as well as actual data like cell tower data dumps - data they have but are restricted from sharing with researchers (though unironically will sell it to third parties) except via one off agreements combined with university-approved IRBs. One unspoken point is that these communities view themselves as distinct. And as a result, data that can be used across sectors just isn't.

How can data sharing be accelerated, and for what real world problems? 2020 has proven (what most of us know) that the US Internet infrastructure is inadequate. ISPs don't willingly share (real) information on their own actual capacity, facilities and user patterns. How populations use the Internet (traffic and access patterns of students vs remote workers) and granular data about how they access Internet infrastructure (DSL? Starbucks WiFi?) are critical questions for which researchers still don't have accurate answers but spend a lot of research funding reverse engineering. It can't be emphasized enough that this data exists, but is unavailable to academic researchers or competitors.

It's challenging to provide Telecom, AI and Cybersecurity policy recommendations for a new administration with the same open questions as we've asked for 30 years. Three challenges can be addressed in the near term: 1. ***Construction of a non-partisan, non-corporate, authoritative, accurate map of the physical Internet infrastructure, speeds, coverage and services. This only exists in pieces and not in useful formats for every community. 2. Applications of privacy preserving techniques such as Multi Party Computation (MPC) and Federated Learning to perform research on private data sets or across edge devices with sensitive information. 3. Data curation and storage in a trusted environment.*** Not only could more robust data be collected, but multiple research and operator communities could benefit.

The data and associated challenges:
1. Collecting accurate end point data - ideally, what devices (IoT, phone, PC, critical infrastructure) exist and what are the general user patterns. (This isn't to say the user data is collected.) Cell Carriers and Phone manufacturers have this already.
2. Collecting accurate broadband maps - self reporting by ISPs clearly is inaccurate, especially given that Microsoft was able to draw a more accurate map of speeds.

3. Sharing the data. MPC has been done in practice for a decade for sensitive data sets and use cases. Privacy Preserving Machine Learning (PPML) and Federated learning are newer techniques but are being successfully used by the medical community. MPC techniques are a decade old and are best used for joint computations vs data sharing but can be ideal for either. A successful real world example is the 2017 Boston Women's Workforce Pay Study in which companies didn't have to reveal salaries to each other, yet MPC computed the varying wages by gender. MPC solutions now exist for cryptocurrency key signatures such that one doesn't have to know the identity of other signatories of the single (split) key. The related area of PPML is being used for AI inference at the edge and for medical data sharing. MPC and PPLML in many cases of network data sharing may even be overkill, but given sensitivities around the data, it seems like a reasonable way to share to draw larger research conclusions that are prevented by the inability to share data.

The barriers are legal, ethical, policy and Institutional. IRBs often are strict, FCC data is not vetted, and data collected has to be strictly defined.

The benefits of this approach are:

-Cross-disciplinary work with the security/crypto community on MPC methods of storing, computing and accessing private data.

-Multi-disciplinary work with the AI/ML community (especially those that work at the literal edge with sensitive data) including commercial clouds, on privacy preserving machine learning techniques for data sharing.

-More complementary, comprehensive mapping that combines security, network infrastructure and data flows. This enables researchers to focus on designing a future, more user friendly and efficient Internet. This would be a great project for testing out on the new NSF-funded FABRIC test bed!