Collecting and Sharing Better Internet Data

Anita Nikolich (UIUC)

Co-PI FABRIC, PI FAB
Co-Organizer DEFCON AI Village

Goal

Share data that currently exists in separate verticals (security/networking/infrastructure operators/social science/econ) with varying security and privacy concerns

Problems

1. Limited or No Access to 'Ground Truth' data
   - NDAs 1:1 with researchers (even in R&E)
   - Proprietary Data sets
   - Social media data; censorship data
2. Incorrect or Non-Existent 'Ground Truth'
   - Form 477 – no incentive to report accurate data; physical
   - ISPs may not have accurate data
   - Unknown infrastructure interconnectivity – especially in urban areas
3. Limited Ability to Include 'Grey Data'
   - **Non-traditional researchers**
   - **Citizen Science**
4. Artificial social barriers
   - Network vs Security vs Other see themselves as distinct communities

Opportunities for Insights

- Enable Privacy-Preserving Data Sharing
  - Don't need to centralize data into a repository – let people curate
  - Less aggregating and making less meaningful
  - Democratization of Data and Analysis
- Better anomaly detection for operators due to access to more data
- Increased infrastructure resilience
- Combine Data Sets
  - Richly layered maps from physical up to user
- Eliminate time and money spent on reverse engineering ground truth to get maps
- Updated Menlo Report and/or more prescriptive guidance on data that can be shared

Privacy-Preserving Data Sharing & Computation

- Multi-Party Computation (MPC) & it’s derivatives
- Federated Learning
- NSF/R&E Community should lead*. Already Cloud provider offerings (ie Google Private Join and Compute; Azure confidential computing; IBM hosted HSMs)
- NSF-funded FABRIC project** can serve as a testbed for such collection and sharing techniques

*See problems with Google + hospital sharing

** **https://whatisfabric.net/**