Getting to “Join” in Privacy Aware Data Sharing

John Heidemann (USC/ISI and CS Dept.)

joint work with Wes Hardaker (USC), Michalis Kallitsis (Merit Network), and Jelena Mirkovic (USC)

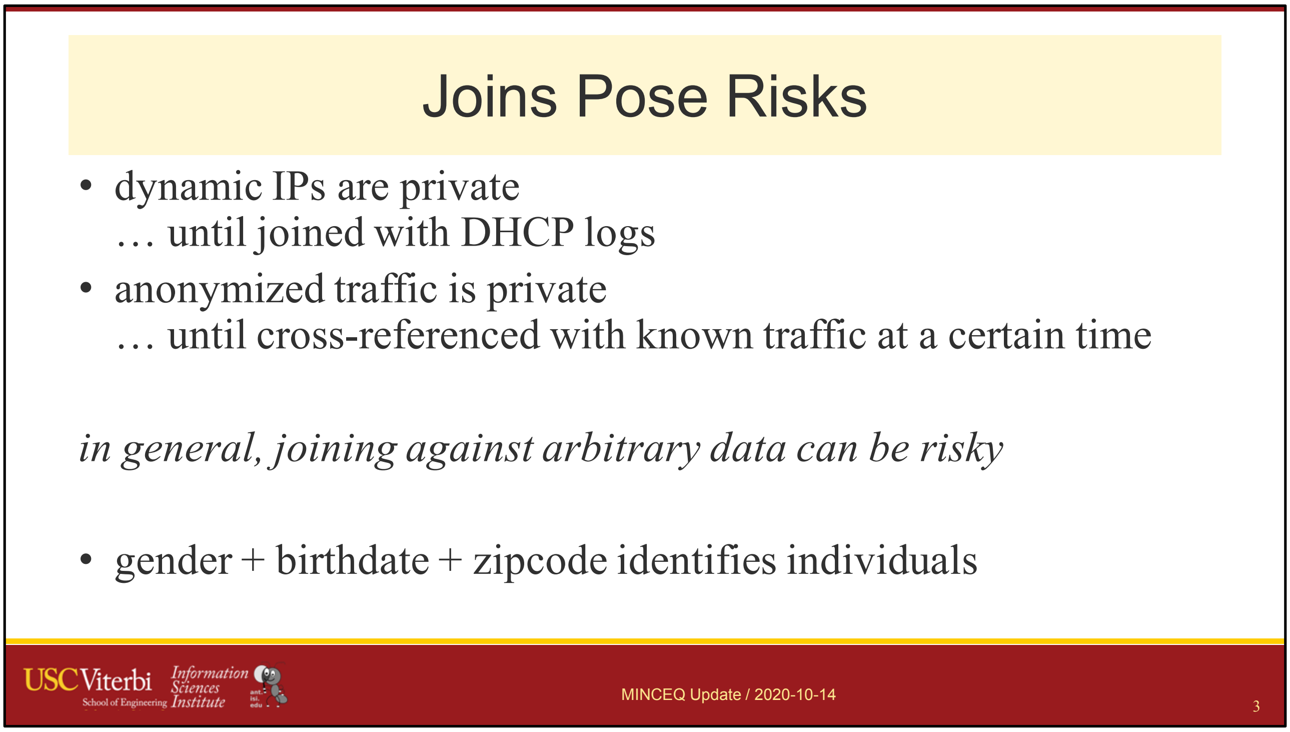
NSF Workshop on Overcoming Measurement Barriers to Internet Research

2021-01-11

Research Challenges Must “Join”

many interesting projects require joining two datasets:

- risks of climate change [Durairajan et al, IMC 2018]
  - requires sea-level rise + network’s physical geography
- risks of route hijacking and eavesdropping [Ballani et al, SIGCOMM 2007]
  - requires countries + routing paths
- detecting Covid work-from-home [Song et al, 2021]
  - requires network changes + IP physical location
- applying ML to network data [several other talks in this session]
  - requires labels from multiple systems: n-way join

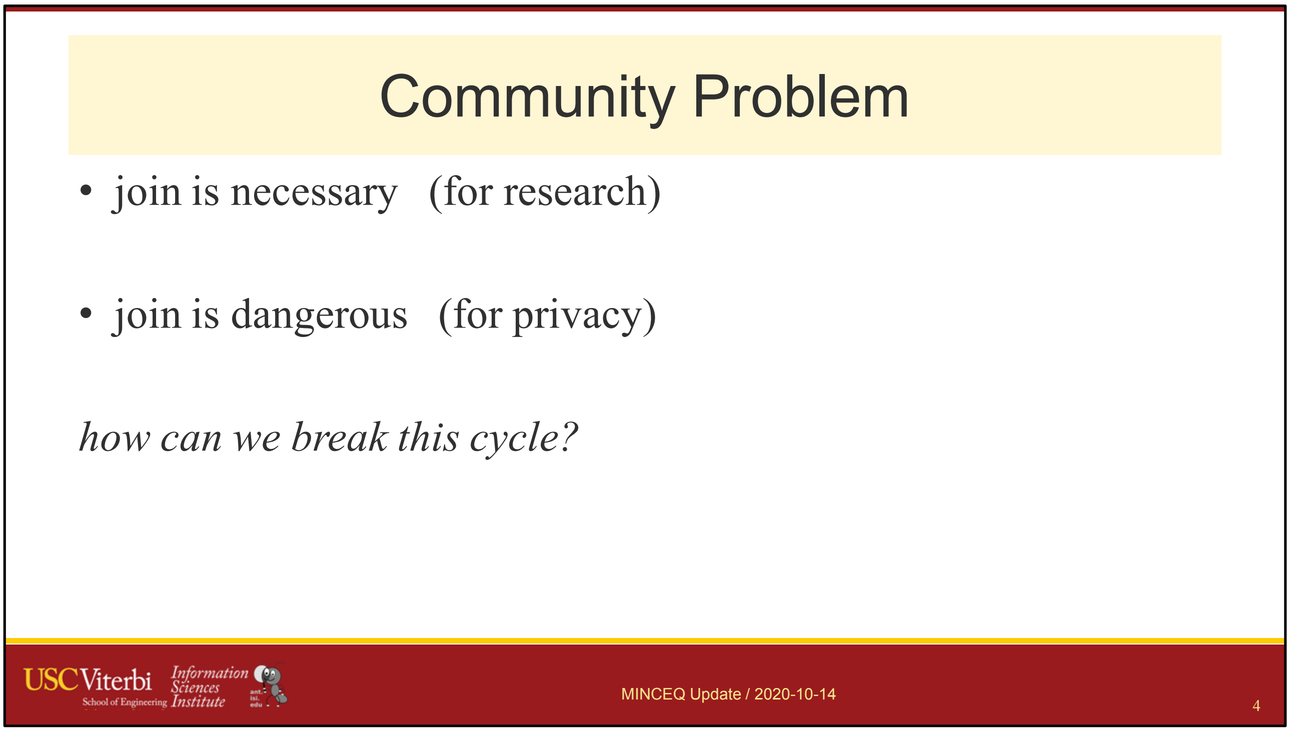


Joins Pose Risks

- dynamic IPs are private
  … until joined with DHCP logs
- anonymized traffic is private
  … until cross-referenced with known traffic at a certain time

in general, joining against arbitrary data can be risky

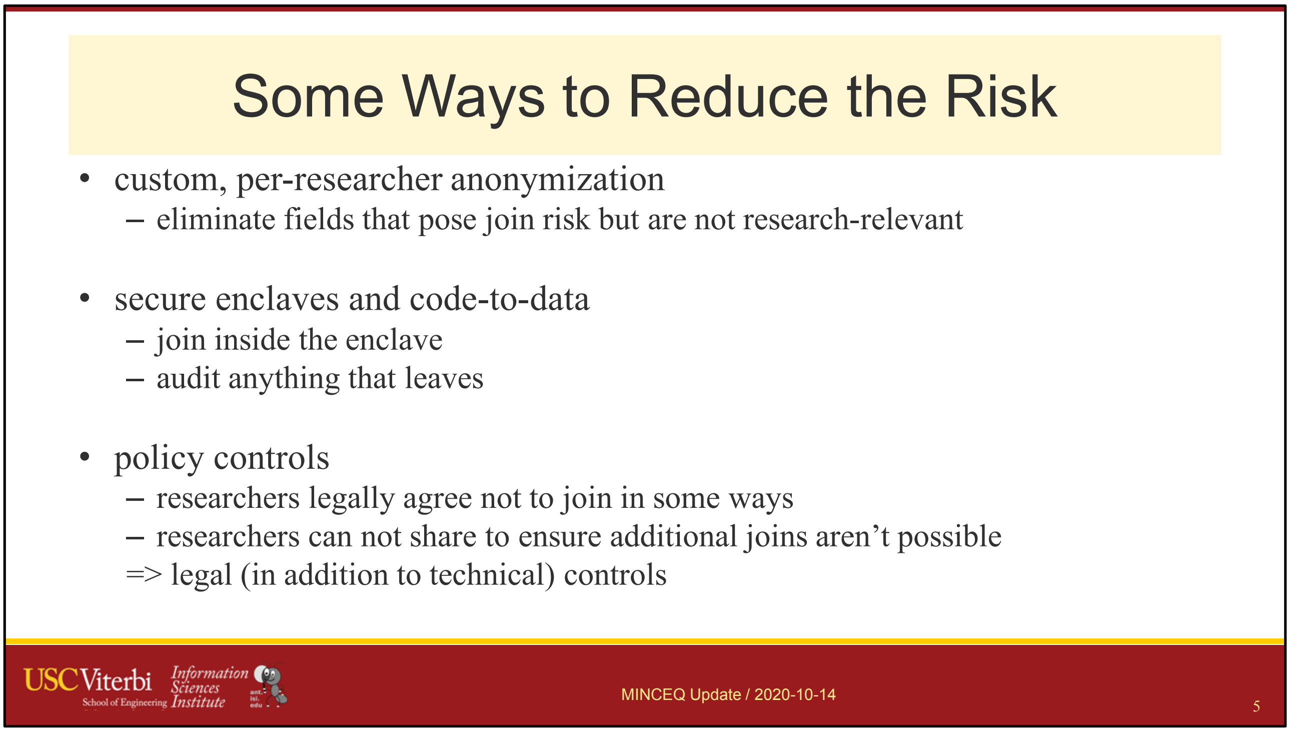
- gender + birthdate + zipcode identifies individuals



Community Problem

- join is necessary (for research)
- join is dangerous (for privacy)

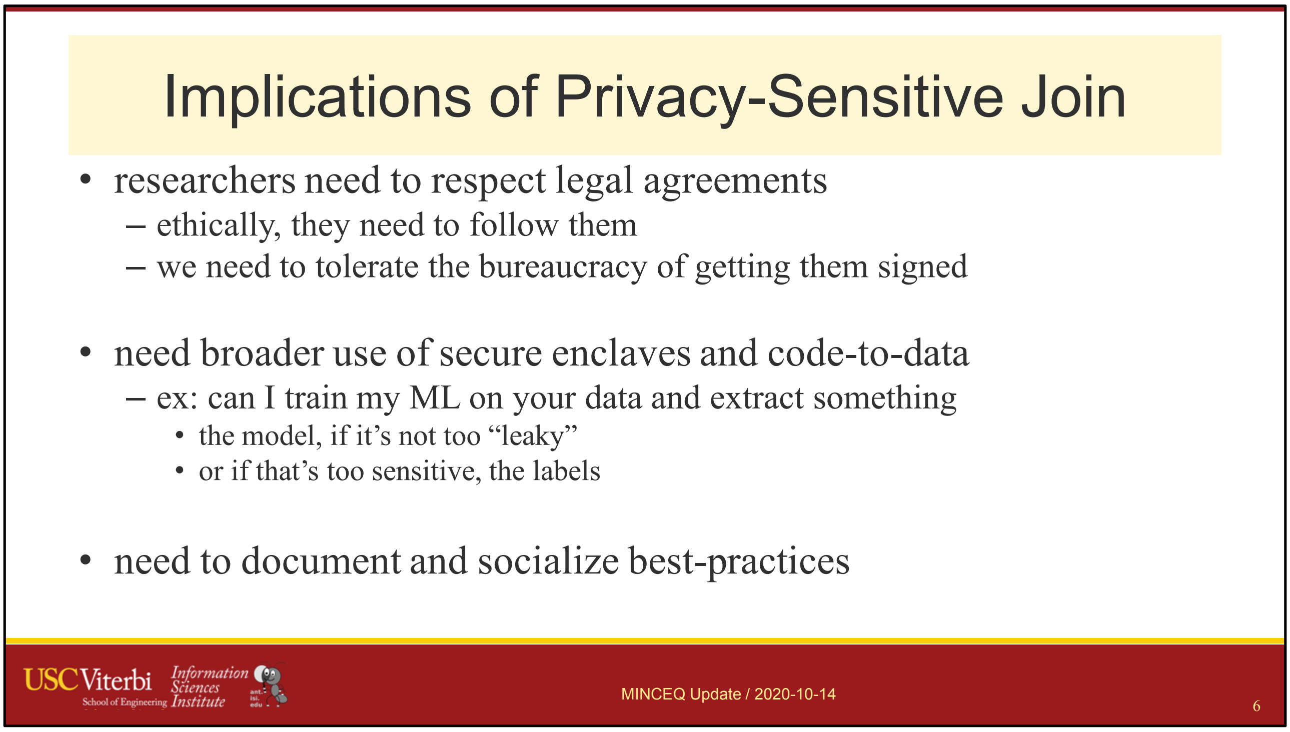
how can we break this cycle?



Some Ways to Reduce the Risk

- custom, per-researcher anonymization
  - eliminate fields that pose join risk but are not research-relevant
- secure enclaves and code-to-data
  - join inside the enclave
  - audit anything that leaves
- policy controls
  - researchers legally agree not to join in some ways
  - researchers can not share to ensure additional joins aren't possible

  => legal (in addition to technical) controls



Implications of Privacy-Sensitive Join

- researchers need to respect legal agreements
  - ethically, they need to follow them
  - we need to tolerate the bureaucracy of getting them signed
- need broader use of secure enclaves and code-to-data
  - ex: can I train my ML on your data and extract something
    - the model, if it's not too "leaky"
    - or if that's too sensitive, the labels
- need to document and socialize best-practices