Getting to “Join” in Privacy-Aware Data Sharing

John Heidemann Wes Hardaker Michalis Kallitsis Jelena Mirkovic

1 The Challenge: Making Measurement Data Relevant

Network data is essential for broad classes of research on Internet performance and security. As the Internet continues to evolve, we must have data to understand the trends that face us today and tomorrow.

But increasingly, no *single* dataset is enough to answer important questions. Researchers nearly always must “join” (in the sense of relational databases) data from multiple sources to answer important questions:

- Joining data about flooding and sea-level rise with the location of network infrastructure helps quantify how climate change threatens the Internet [4].
- Joining political geography data with Internet routing data helps identify challenges in handling route hijacking, traffic eavesdropping, and national data sovereignty [1].
- Knowledge about user populations can make interpreting latency of services more meaningful [5, 2].
- Geography is important to interpret network changes as effects of Covid-19 [6].

Privacy concerns increasingly prevent joining data and therefore inhibit research. Datasets cannot be joined when key parameters such as IP addresses are anonymized differently or heavily truncated. Datasets cannot be joined when data is siloed by isolation on a different systems, or locked within a closed web portal.

Yet respecting individual privacy is important—we must find ways to perform research ethically, and must balance risks of research with the benefits [3].

2 Potential Approaches to Balance Join and Privacy

Addressing the need to “join” but manage privacy will require new technical approaches and also changes to how we do research. Here we consider some components that begin to address the problem.

Flexible, per-research anonymization is important. What is important to one researcher can be irrelevant to another. The principle of least privilege suggests that we anonymize everything we can, but not bits that are required to answer the research question. As an example, research projects may have to choose between selecting data with anonymized IP addresses or other anonymized information (perhaps payloads and ports) to reduce risks.

Policy controls around data will be needed to complement technical controls. While technical controls (anonymization or approaches such as differential privacy) are ideal when they can be applied, they are often too restrictive for meaningful research. Access to data under a legal agreement on how it is used or what information can be extracted can unblock research when technical solutions constrain it too much to be useful.

Analysis in secure enclaves and code-to-data are technical approaches that formalize a more limited access to data. A secure enclave (perhaps a virtual machine with the data for analysis) can limit what other data is brought in, and allows auditing of what information flows out. Code-to-data allows a researcher to apply a function (code) to sensitive data, receiving only the results, which can be vetted to evaluate what leaks, and rate-limited to avoid repeated query attacks.

3 Implications of “Join”

These ideas have implications for the research community, research sponsors, and data providers.

The research community should expect that some future research may be more encumbered than it has been in past (when one could get data based on private agreements). *Implications:* Researchers should expect to deal with legal agreements. We can minimize this cost by standardizing those agreements. (We *must* get away from bespoke legal contracts and the need to have every organization’s lawyers tweak the agreement.)

Researchers should also expect to see *greater use of secure enclaves and code-to-data*, and data providers should understand what these tools offer. Coordination between researchers and data providers is key in helping both parties comprehend what data processing tools are required, what external data sources can (or cannot) be brought in and what compute resources are necessary for the task at hand. Working “at-arms-length” with sensitive data can help manage privacy concerns, even if it can be more difficult.

Finally, we need *research on data sharing methods and best practices.* While specific technical solutions are helpful, there is no one silver bullet—be it new anonymization, a general technique like differential privacy, a single web-portal or enclave—that will address the range of challenges.

References

[1] Hitesh Ballani, Paul Francis, and Xinyang Zhang. A study of prefix hijacking and interception in the Internet. In *Proceedings of the ACM SIGCOMM Conference*, pages 265–276, Kyoto, Japan, August 2007. ACM.

[2] Matt Calder, Manuel Schröder, Ryan Gao, Ryan Stewart, Jitendra Padhye, Ratul Mahajan, Ganesh Ananthanarayanan, and Ethan Katz-Bassett. Odin: Microsoft's scalable fault-tolerant CDN measurement system. In *Proceedings of the USENIX Symposium on Network Systems Design and Implementation*, pages 501–517, Renton, WA, USA, April 2018. USENIX.

[3] David Dittrich and Erin Kenneally (editors). The Menlo report: Ethical principles guiding information and communication technology research. Technical report, United States Department of Homeland Security, September 2011.

[4] Ramakrishnan Durairajan, Carol Barford, and Paul Barford. Lights out: Climate change risk to internet infrastructure. In *Proceedings of the Applied Networking Research Workshop*, Montreal, QC, Canada, July 2018. ACM.

[5] Ricardo de O. Schmidt, John Heidemann, and Jan Harm Kuipers. Anycast latency: How many sites are enough? In *Proceedings of the Passive and Active Measurement Conference*, pages 188–200, Sydney, Australia, March 2017. Springer.

[6] Xiao Song and John Heidemann. Measuring the Internet during Covid-19 to evaluate work-from-home (poster). Poster at the NSF PREPARE-VO Workshop, December 2020.