GENI Opt-In Working Group: Issues Relating to Data Acquisition, Retention, Use, and Disclosure

Aaron J. Burstein
DRAFT – DO NOT DISTRIBUTE

May 5, 2009

Network experiments that involve "real" data—that is, data obtained from stored data sets, participants in an experiment, or other network users—raise legal and ethical issues that scientists should confront as early as possible in the design of their experiments. Four types of operations on data raise legal and ethical issues that scientists should consider separately: (1) acquisition; (2) retention; (3) use; and (4) disclosure. This document does not provide an exhaustive discussion of any of these topics. Instead, it aims to provide a general overview of the actors who have an interest in real network data, what those interests are, and when a scientist should seek more detailed guidance for a particular experiment.

Interested Parties A fairly diverse range of individuals and organizations can be affected in one way or another by a network experiment. Whether a given party has an interest in the data used in a particular experiment depends, of course, on the details of the experiment.

Participants in Experiments One of the distinguishing features of GENI's architecture is that it allows scientists to enroll users in experiments to test the properties of new applications, services, and protocols. Data that might be useful for carrying out these experiments include packet traces, application use histories or logs, and direct observation of user behavior. This means that information privacy is the primary individual interest to protect where participants in experiments are concerned.

If users are enrolled in a study that involves obtaining "identifiable private information" or obtaining data through intervention or interaction with an individual, then the participants may be "human subjects" who are entitled to protections specified by federal regulations widely known as the "Common Rule".[1] These protections include giving informed consent that clearly states what data will be collected and how it will be used. The general purpose behind these rules is that individuals are and should remain autonomous. Fully informing subjects of how data will be collected, used, retained, and

[1] See 45 C.F.R. § 46.

disclosured, and how the actions might harm their information privacy interests, provides one mechanism for preserving this autonomy.

Even if an experiment is not subject to the Common Rule, researchers should nevertheless analyze how the experiment's data needs implicate informational privacy. More specifically:

- Acquiring data about individuals can reduce or eliminate the buffer that usually separates them from society, thereby causing them to limit or modify their activities.[2]
- Analyzing identifiable data in ways that are not disclosed to data subjects can harm their informational privacy interests.
- Losing control of data through accident or a malicious attack provides another way that individuals can lose informational privacy.

All of these possibilities counsel careful consideration of what data is necessary to conduct an experiment, how long it will be kept, and how it will be used. Giving primacy to information privacy would suggest starting with restrictive practices and graudally expanding them as needed. Typically, however, researchers do not know have precise answers to any of these questions; they may not know what data will be useful without first examining the data. To add to the complexity, there is no bright line that separates "identifiable" from "non-identifiable" information. Given enough data (and metadata), an adversary might be able to identify information (to an individual, an organization, or some other entity that the research would like to dissociate from the data) using data fields that were though to be non-identifying.

Thus, analyzing data with the goal of understanding how all stages of a dataset's life cycle—acquisition, retention, use, and disclosure—could harm any of the broad array of privacy interests

Individual Users The preceding discussion assumed that researchers have obtained informed consent directly from the individuals who are represented in a dataset. Obtaining this level of consent might not always be possible or desirable, and under some circumstances GENI researchers might gain access to data about individual users without informed consent. The suggestions given above for analyzing effects on privacy remain largely the same. The *legal* considerations, however, are far more difficult. This section provides a brief overview of these considerations.[3]

Federal criminal law generally prohibits the real-time interception of the contents of electronic communications.[4] It is also illegal under federal law to collect in real-time

[2] Daniel J. Solove, *A Taxonomy of Privacy*, 154 University of Pennsylvania Law Review 477, 493 (2005).

[3] For a more thorough analysis, see Aaron J. Burstein *Conducting Cybersecurity Research Legally and Ethically*, USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET '08) (Apr. 2008).

[4] The statute that makes this prohibition is commonly referred to as the Wiretap Act, 18 U.S.C. § 2510-22. An electronic communication is "any transfer of signs, signals, writing, images, sounds, data, or intelligence of any nature transmitted in whole or in part by a wire, radio, electromagnetic, photoelectronic or photooptical system that affects interstate or foreign commerce," with a few exceptions that are not relevant here. 18 U.S.C. § 2510(12).

the non-content portions of an electronic communication.[5] Both of these prohibtions are subject to a consent exception as well as a "provider" exception, which permits interceptions that are "a necessary incident to the rendition of his service or to the protection of the rights or property of the provider of that service."[6] What this means in practice is that a research typically needs to coordinate real-time data interceptions of electronic communications with the network operator's operational staff, to ensure that there is a link between the interception and the network service or its protection.

Stored communications data—whether contents or non-content information—is subject to far fewer restrictions. Federal law imposes essentially no restrictions on the *use* of stored communications data within an organization, nor does it regulate how much or how long data may be stored, provided that the acquisition was legal in the first place. Instead, the law generally prohibits the *disclosure* of stored communications: contents generally cannot be disclosed without at least a court order, while non-content information may not be voluntarily disclosed to a governmental entity. This statutory scheme assumes a sharp distinction between "content" and "non-content," which is often missing in practice. Researchers should obtain individualized advice to figure out where a proposed data collection falls within this scheme.

Researchers' Organizations The third category of entities that a researcher should consider in connection with data acquisition, use, retention, and disclosure is the organization with which he or she is affiliated. Organizations do not have privacy interests, but actions taken by researchers can nevertheless harm its interests. This sections examines a few such interests.

- **Security.** Data that reveals details about an organization's security posture could facilitate attacks against the network. Disclosing data without allowing network operations staff to review what is being disclosed and to whom could provide a nasty surprise to staff and make the organization reluctant to approve any future data releases.

- **Reputation; user backlash.** Whether or not an organization has a clear and public policy about what data it collects as part of operating its network, it is likely that most of the network's users are unaware of the policy. Disclosing data in a way that allows it to be identified to the originating institution might therefore be a sensitive issue if it might prompt user backlash or paint the organization as unfriendly to privacy. These considerations might make officials unwilling to share data with researchers at other organizations or to authorize public releases of data.

  A separate element of reputation is an organization's interest in employing, and being perceived as employing, researchers who act ethically. Acquiring or using data in ways that disregard others' interests (e.g., privacy) can tarnish the reputation of the researcher and the organization.

[5] This prohibition is found in the Pen Register/Trap and Trace statute, 18 U.S.C. §§ 3121-27.
[6] 18 U.S.C. § 2511(2)(a)(i).

- **Competitive Interests.** Data releases can harm a number of an organization's competitive interests. For instance, if the data contains the destination of traffic, other might be able to infer the existence of collaborations between organizations that would otherwise have remained confidential. Data might also reveal details about a network's configuration that allow it to provide better service than its competitors. Finally, network data might reveal that an organization is not handling traffic in a manner consistent with a peering or transit agreement. These are all considerations that researchers should be prepared to discuss when requesting to acquire or disclose data.