Control use of data to protect privacy

Susan Landau*

Massive data collection by businesses and governments calls into question traditional methods for protecting privacy, underpinned by two core principles: (i) notice, that there should be no data collection system whose existence is secret, and (ii) consent, that data collected for one purpose not be used for another without user permission. But notice, designated as a fundamental privacy principle in a different era, makes little sense in situations where collection consists of lots and lots of small amounts of information, whereas consent is no longer realistic, given the complexity and number of decisions that must be made. Thus, efforts to protect privacy by controlling use of data are gaining more attention. I discuss relevant technology, policy, and law, as well as some examples that can illuminate the way.

User control sits at the heart of the FIPs. Trans-

control incursions from businesses and governments.
There are other approaches that make tracking user reading more difficult, rather than less so. One such is Shibboleth, software that enables a user at one participating institution, say, the University of Michigan, to access electronic resources at another, say, the University of Illinois. A user authenticates on the University of Michigan. The user, however, is identified to the University of Illinois not by personal identifier such as name or e-mail address but by her right to the resource. This could be because she is a member of the University of Michigan community (student or staff), a participant in a particular course, or one of a set of users authorized to access particular resources. Unless the information is specifically needed, the University of Illinois does not learn the user’s actual online identity. The Family Educational Rights and Privacy Act, which protects the privacy of student educational records, and the fact that librarians view reader privacy as fundamental motivated this privacy-protective architecture.

A potentially powerful approach to controlling data usage is “accountable http,” a variant of the http protocol. Proposed by MIT researchers Oshani Seneviratne and Lalana Kagal, httpa creates a system to track information usage (17). The system consists of a user who wishes to access data that have usage restrictions (e.g., no sharing, no sharing without informing the data owner, etc.); a data provider using an httpa server; and a Provenance Tracking Network (PTN). The PTN is a network of servers that log each data access and usage, either from the original data provider or any user downstream.

The magic behind the system is httpa, a protocol that conveys usage restrictions between the data providers and data users, creating a log in the PTN for each time a protected resource is accessed. These logs do not enforce compliance but can be used to determine it. This general approach to controlling data usage has only been tested in a small-scale effort; whether it can scale to the Internet is unclear. But it might be valuable in limited settings, such as patient health data, where a motivator might be the Health Insurance Affordability and Accountability Act (HIPAA), the U.S. law that restricts the sharing of patient medical data.

Online identities are used ubiquitously across the Internet to access restricted resources (e.g., pay-for-use subscriptions or library memberships confined to a university community), to post comments in restricted settings such as YouTube, and to conduct business at a bank or online broker. Although the need for secure, interoperable, and easy-to-use credentials for online identities was clear, development and adoption of such tools was slow.

The U.S. federal government stepped in, creating the National Strategy for Trusted Identities in Cyberspace (NSTIC) to provide funding for pilot programs and standards efforts that would provide both privacy and security. Using access to federal government sites as a lever, NSTIC requires that private-sector identity providers protect the privacy of information regarding user activities on federal sites (18).

Tracking when a user goes on a .gov website can reveal their private information, e.g., interest in HIV/AIDS or in penalties for unpaid taxes. Federal rules prevent identity providers from using tracking information from federal sites.
for anything but authentication, audit, or complying with the law (29). In other words, no ads, no sharing the information with a third party, and no using the information to promote the identity provider’s products. A signed-on user has greater privacy protections when visiting the National Cancer Institute website than when visiting the American Cancer Society site. Policy controls data usage and is then manifested in technical design. Laws can provide shields against inappropriate data usage. The 1970 U.S. Fair Credit Reporting Act (FCRA), which precludes the FIP’s, does not control collection. Instead the FCRA strictly limits the circumstances under which a person’s credit information can be accessed (essentially for credit, employment, and in response to court orders) (20).

A different example of control is the Genome Information Nondiscrimination Act (GINA) of 2008, which protects against discrimination in health insurance and employment based on genetic data. But GINA, too, has its limits. If a woman discovers through genetic testing that she is BRAC1– or BRAC2-positive, with a 55 to 65% or 45% chance, respectively, of developing breast cancer by age 70, GINA protects her ability to obtain health insurance and employability but says nothing about her ability to obtain disability, long-term care, or life insurance in the face of the BRAC1 and BRAC2 data.

There are other examples of how technology, policy, and law combine to control use. A well-known one is in medical research. The HIPAA privacy rule governs how researchers within health care organizations handle the health information of individuals; it also governs researchers who interact with such data (27). There are a number of ways this is done: through the law itself; its implementation in regulations (27); and Institutional Review Boards, which examine researchers’ access and use of patient data, as well as by social controls. A researcher who is careless about the privacy of health records will find future access to such records very difficult.

**Privacy and national security**

An example that doesn’t tend to appear when discussing privacy and big data is national-security collection. Yet the Snowden leaks disclosed massive collection both domestically and abroad. These disclosures started a national discussion on collection and use, although not, for obvious reasons, on notice and consent, which have little role in national-security collection. I recently participated in a National Academies study on software alternatives to bulk signals intelligence collection (22). Bulk collection, specifically of telephony metadata—NSA receives daily downloads of telephony metadata (to, from, time, data, and length of call data) from major service providers—has generated much consternation. Metadata are data about the call, not its content, but mobile phones and the fact that cell phones are usually associated with a single individual mean that metadata themselves are remarkably revelatory (23, 24). Both a presidentially appointed review group on intelligence and communications technologies and the Privacy and Civil Liberties Oversight Board, an executive-branch oversight board, recommended ending the government telephony metadata program (25, 26).

Our charge was somewhat different—technical alternatives to the collection—and our conclusion was also somewhat different. Because the program provides information that cannot be found in other ways, we believe there are no alternatives providing the same information (22). In particular, if past events become interesting in the present—a non-nuclear nation is discovered to be pursuing nuclear weapons or a new target is identified as a terrorist—past history may present new leads (22). Such past history will be available, in general, only if there were bulk collection in the past.

We made no judgment on whether the bulk collection program should continue; that is a policy decision, not a technical analysis, and out of scope for the study. We observed that the only way to protect privacy in the face of bulk collection is to control use—the same solution as the one to the private-sector big data collection issue.

We had no evidence that NSA was inappropriately using bulk data that were being collected. Nonetheless, we found that there were possible improvements for controlling use. We recommended increased utilization of automated controls and auditing, noting that manual controls and auditing will also always be necessary (22). Automating controls on data usage means data usage rules must be stated with great precision. That has its own advantages, including preventing inconsistencies (one such, on the meaning of archive, resulted in inappropriate access to the database) (22). Automated controls on data usage will also provide transparency. We also proposed research into privacy-protective methods of auditing by outsiders (22).

**Controls on use**

Our point was that “Controls on use ... offer an alternative to controls on collection as a way of protecting privacy.” The same is true outside the national-security domain. Privacy intrusions are everywhere. The technologies—smart phones and their apps; the ubiquity of Google, which performs 88% of searches in the United States (27) and over 90% in Europe (28); and Internet-connected sensors in automobiles, bridges, cargo trucks, and so forth—are novel, but the fact that technologies and changing social mores create privacy intrusions is not. In 1890, a similar situation—yellow journalism and hand-held cameras—induced Samuel Warren and Louis Brandeis to write “The right to privacy,” which laid a foundation for U.S. privacy protections. Warren and Brandeis observed that, “it has been found necessary from time to time to define anew the exact nature and extent of such protection” (29). Today is such a time. The nature and extent of redefinition will be of control on use, and determining the right controls, and the right ways to exercise them, will be challenging—but that is what we must do.

**REFERENCES**

30. 2012/aa94961.