Control Use of Data to Protect Privacy

Susan Landau

We live in an era of an explosion of data. For a variety of reasons, including the massive collection by both the private sector and governments as well as the ease of computing correlations — thus deriving information even about people whose data is not in the set — the old methods for protecting privacy, no longer work. An old protection made new, managing use, now seems the most appropriate way to secure privacy. Controlling use is complex, but combining technology, policy, and law is the best way to control incursions from businesses and governments.

The principles governing data protection are forty years old. The Fair Information Practices (FIPs) were developed in response to the rise in the 1960s of computerized data systems. Coming originally from a report from the U.S. Department of Health, Education, and Welfare [HEW], the FIPs were revised by the Organization of Economic Cooperation and Development [OECD]. The more expansive OECD privacy principles have been the basis for many national and international privacy regulations.

User control sits at the heart of the FIPs. *Transparency/Notice* says there should be no data collection system whose existence is secret, *Access*, that there should be a way for the data subject to find out what information is in her record and how it is used, *Consent* — sometimes called *Choice* — that data collected for one purpose not be used for another without user permission, *Redress*, requires ability for the data subject to correct inaccuracies, and *Integrity* and *Security*, that the data collector keeps reliable records and protects them. In 1998, the US Federal Trade Commission identified these as the "five core principles of privacy protection" [FTC1998, p. 7], and noted that notice was fundamental, calling choice/consent the "second widely accepted core principle" [FTC1998, p. 8].

While the US and Europe have taken different routes to protecting privacy — the US using sector-specific protections (financial data, banking information, health records), Europe pursuing broader data-protection schemes — both emphasized *notice* and *consent.* But while the FIPs made sense when an individual could discern and react to a data-collection event, this is no longer true.

Consider data collection from a smart phone. This may include the speed of the phone; aggregating such information from multiple users can provide information on alternate routes when there is a traffic jam. Tracking user response to apps — e.g., where installs are from or the order in which apps are used — can facilitate improvements in an app's design [Google].

Users directly benefit by providing such data. Their previous accesses on the site, their interests, the specifics of their device's software and hardware makes more personalized search, faster access to information of interest, and simplified access possible. The combination of information from the user and from others (e.g., how her cohort is doing on their marathon training) improves her experience. For companies, such data promotes faster, more targeted services, and ties the consumer more strongly to the business. Data about the user also enables more targeted advertising or services — and higher profits. For researchers, massive data illuminates connections that might not have been apparent (e.g., tracking 130 thousand hospital admissions at three institutions increased understanding of the spread of hospital-acquired infections [Wiens]). Data may uncover correlations that are actually causations.

Because data collection involves compilation of massive amounts of small bits of data, notice and consent are difficult for users to manage. Should collection of location data increase when a traffic accident blocks a popular route? How about when inclement weather threatens? What if the user is on a private assignation that day? That a service that provides up-to-date route information also collects up-to-date location data is not something all users realize (though they should). Frequent queries about permission for collection creates a situation in which the user inattentively clicks "Yes," — not exactly a win for privacy.

Notice simply doesn't make much sense in a situation where collection consists of lots and lots of small amounts of information. Written to cover all contingencies, privacy notices are not designed for human use. A 2008 study showed that the average reader would need 244 hours simply to read the privacy policies for all websites she accessed in a year [McDonald].

Consent is often not an option. Almost a decade ago Fred Cate noted, "If consent is required as a condition for opening an account or obtaining a service, a high response rate can always be obtained," [Cate, p. 366], while a 2014 President's Advisory Committee on Science and Technology (PCAST) report on big data and privacy observed, "Only in some fantasy world do users actually read these notices and understand their implications before clicking their consent." [PCAST, p. xi].

Sometimes the user is not even given a choice about consent. Because of overwhelming complexity, Google, whose Android platform dominates the consumer smartphone market [IDC], decided to put permissions for information access into groups. Thus a user lacks ability to conduct fine-grained decisions on which information to permits apps to access [Google]. The user moves on, rarely examining, — or withdrawing — consent afterwards.

A fundamental problem is that seemingly innocuous data may trigger a privacy incident. Discovering that a woman has bought "a purse large enough to double as a diaper bag, zinc and magnesium supplements and a bright blue rug" is enough for Target to conclude that she has a baby due five months later [Duhigg]. Using the history of buying patterns of other customers, Target predicted a teenager's pregnancy from her vitamin purchases [Duhigg], while the ride-share firm Uber claimed to be able to discern one-night stands from the usage patterns of rider pick-up and drop-off data [Voytek]. Solon Barocas and Helen Nissenbaum noted, "The willingness of a few individuals to disclose information about themselves may implicate others who happen to share the more easily observable traits that correlate with the traits disclosed." [Barocas, p. 32]

Context matters in privacy. That idea was first espoused by Nissenbaum a decade ago [Nissenbaum] and is now gaining support in policy circles, including in the White House Consumer Bill of Rights [WH] and a recent Federal Trade Commission report [FTC]. Massive amounts of data create such personal and societal benefits that collection is unlikely to stop. The FIPs protected privacy through notice and consent, but for reasons of complexity (too many tiny collections, too many repurposings), those are no longer effective. Nonethless notice and consent provide benefits: notice for transparency, consent, for certain types of data or use, and for controlling context [Cate2013]. But the value of big data means we must directly control use rather than using notice and consent as proxies [PCAST]. That is true no matter who the collector is.

This is easier said than done. Big data provides the patterns that allow us to use resources efficiently. Determining how to continue to collect and use big data but control its use is complex. The tools are technology, policy, and law, and there are some examples that can illuminate the way.

Once the most solitary of activities, reading is losing the privacy between the reader and the page. Amazon and other purveyors of ebooks have discovered multiple ways of tracking activity: where readers start, what they reread, whether they mark a passage, if they finish the text [Alter].

There are other approaches that make tracking user reading more difficult. 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 email address, but by her right to the resource. This could be because she is a member of the University of Michigan community (student, staff), a participant in a particular course, 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 ID. The Family Educational Rights and Privacy Act, which protects 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 [Seneviratne]. The system consists of a user who wishes to access data that has 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 US 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, to conduct business at a bank or online broker. Though the need for secure, interoperable, and easy-to-use credentials for on-line identities was clear, development and adoption of such tools was slow.

The US 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 [NSTIC].

Tracking when a user goes on a .gov website can reveal their private information: 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 [GTRI]. 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. Here is a case where policy controls data usage and is then manifested in technical design.

Laws can provide shields against inappropriate data usage. The 1970 US Fair Credit Reporting Act (FCRA), which predates the FIPs, 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) [FCRA].

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 55-65% and 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 health information of individuals; it also governs researchers who interact with such data [HHS]. There are a number of ways this is done: through the law itself, its implementation in regulations [HHS], Institutional Review Boards which examine researchers access and use of patient data, as well as social controls. A researcher who is careless about the privacy of health records will find future access to such records very difficult.

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, though 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 [PPD28]. Bulk collection, specifically of telephony metadata — NSA receives daily downloads of telephony metadata (to, from, time, data, length of call data) from major service providers — has generated much consternation. Metadata is data about the call, not its content, but mobile phones and the fact that cellphones are usually associated with a single individual means metadata itself is remarkably revelatory [Landau, pp. 99-101], [Mayer]. 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 [NSAReview], [PCLOB].

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 [PPD28]. 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 [PPD28]. 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. 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 was being collected. Nonetheless we found 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 [PPD28]. 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 [PPD28]). Automated controls on data usage will also provide transparency. In the PPD28 report, we also proposed research into privacy-protective methods of auditing by outsiders [PPD28].

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 68% of searches in the US [ComScore] and over 90% in Europe [Kanter], Internet-connected sensors in automobiles, bridges, cargo trucks, etc. — are novel, but the fact that technologies and changing social morés 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 US 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." [Warren] Today is such a time. The nature and extent will be on 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.

[HEW] US Department of Health, Education, and Welfare, Secretary's Advisory Committee on Automated Personal Data Systems, *Records, Computers and the Rights of Citizens*, 1973.

[OECD] Organization for Economic Cooperation and Development, "OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data," 1980.

[FTC1998] Federal Trade Commission, *Privacy Online: A Report to Congress*, June 1998.

[Google] Google, *Optimize Your App*, https://developer.android.com/distribute/essentials/optimizing-your-app.html [last viewed December 4, 2014].

[Wiens] Jenna Wiens, John Guttag, Eric Horvitz, "A Study in Transfer Learning: Leveraging Data from Multiple Hospitals to Enhance Hospital Specific Predictions," *Journal of American Medical Informatics Association*, January 2014.

[McDonald] Aleecia McDonald and Lorrie Faith Cranor, "The Cost of Reading Privacy Policies," *I/S: A Journal of Law and Policy for the Information Society*, 2008 Privacy Year in Review issue.

[Cate] Fred Cate, "The Failure of Fair Information Practices Principles" in "Consumer Protection in the Age of the 'Information Economy'," Jane Winn, ed. 2006.

[PCAST] President's Council of Advisors on Science and Technology, *Big Data and Privacy: A Technological Perspective*, May 2014.

[IDC] IDC, "Smartphone OS Market Share, Q2, 2014," http://www.idc.com/prodserv/smartphone-os-market-share.jsp [last viewed September 4, 2014]

[Google] Google, Review app permissions, https://support.google.com/googleplay/answer/6014972?hl=en [last viewed on September 3, 2014].

[Duhigg] Charles Duhigg, "How Companies Learn Your Secrets," *New York Times Sunday Magazine,* February 19, 2012.

[Voytek] Bradley Voytek, "Rides of Glory," March 28, 2012, now posted at: https://web.archive.org/web/20140827195715/http://blog.uber.com/ridesofglory

[Barocas] Solon Barocas and Helen Nissenbaum, "Computing Ethics: Big Data's End Run Around Procedural Privacy Protections," Communications of the ACM, Vol. 57, No. 11 (November 2014), pp. 31-33.

[Nissenbaum] Helen Nissenbaum, "Privacy as Contextual Integrity," Washington Law Review, Vol. 79 (2004), pp. 119-157.

[FTC] Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers*, March 2012.

[WH] White House, *Consumer Data Privacy in a Networked World: a Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy,* February 2012.

[Cate2013] Fred Cate and Viktor Mayer-Schonberger, "Notice and Consent in a World of Big Data," International Data Privacy Law, Vol. 3, No. 2 (2013), pp. 67-73.

[Alter] Alexandra Alter, "Your E-Book is Reading You," *Wall Street Journal,* July 19, 2012.

[Seneviratne] Oshani Seneviratne and Lalana Kagal. "Addressing Data Reuse Issues at the Protocol Level," *IEEE International Symposium on Policies for Distributed Systems and Networks,* pp. 141-144, 2011.

[NSTIC] National Institute for Standards and Technology, *National Strategy for Trusted Identities in Cyberspace*, http://www.nist.gov/nstic/

[GTRI] Georgia Tech Research Institute, "GTRI NSTIC Trustmark Pilot," October 7, 2014, https://trustmark.gtri.gatech.edu/operational-pilot/trustmark-definitions/ficam-privacy-activity-tracking-requirements-for-csps-and-bae-responders/1.0/

[FCRA] Fair Credit Reporting Act, 15 USC § 1681.

[HHS] Center for Disease Control, HIPPA Privacy Rule and Public Health, http://www.cdc.gov/mmwr/preview/mmwrhtml/m2e411a1.htm [last viewed November 29, 2014].

[PPD28] Committee on Responding to Section 5(d) of Presidential Policy Directive 28: The Feasability of Software to Provide Alternatives to Bulk Signals Intelligence Collection, *Bulk Collection of Signals Intelligence: Technical Options*, National Academies Press, 2015.

[Landau] Susan Landau*, Surveillance or Security? The Risks Posed by New Wiretapping Technologies*, MIT Press, 2011.

[Mayer] Jonathan Mayer, "Metaphone: The Sensitivity of Telephone Metadata," March 12, 2014, http://webpolicy.org/2014/03/12/metaphone-the-sensitivity-of-telephone-metadata/

[NSAReview] President's Review Committee on Intelligence and Communications Technologies, *Liberty and Security in a Changing World*, December 13, 2013.

[PCLOB] Privacy and Civil Liberties Oversight Board, *Report on the Telephone Records Program Conducted Under Section 215 of the USA PATRIOT Act and on the Operations of the Foreign Intelligence Surveillance Court*, January 23, 2014.

[ComScore] comScore Releases March 2014 U.S. Search Engine Rankings, https://www.comscore.com/Insights/Press-Releases/2014/4/comScore-Releases-March-2014-U.S.-Search-Engine-Rankings [last viewed November 23, 2014].

[Kanter] James Kanter, "Opposition Grows in Europe to Google Antitrust Proposal," New York Times, September 4, 2014.

[Warren] Samuel Warren and Louis Brandeis, "The Right to Privacy," *Harvard Law Review*, Vol. 4 (1890), p. 193.