Before the
Department of Commerce
National Telecommunications and Information Administration

Docket No. 120214135-2135-01

Multistakeholder Process to Develop Consumer Data Privacy Codes of Conduct

Request for Comments

Comments of Peter Swire
C. William O'Neill Professor of Law
Moritz College of Law
The Ohio State University
and
Senior Fellow
Center for American Progress Action Fund

The National Telecommunications and Information Administration, or NTIA, has asked for comments on what issues should be addressed through a privacy multistakeholder process. Based on my experience in privacy law and policy, I believe an early and prominent candidate should be the definition of what counts as "de-identified" information. As discussed below this topic has multiple advantages, including heightened protection for consumers, positive effects on innovation and the broader economy, and likelihood of concrete, enforceable success for the process itself.

These comments provide background for the discussion and then explain the importance of the topic of de-identified data. The comments explain how the recent Federal Trade Commission privacy report provides a new and useful set of proposals for how to handle de-identified data, and concludes with an analysis of why the topic of de-identified data is a good candidate for early consideration in a multistakeholder process.

Background

As background for these comments, I am the C. William O'Neill Professor of Law at the

Moritz College of Law of the Ohio State University, and Senior Fellow at the Center for American Progress Action Fund and the Future of Privacy Forum. Under President Bill Clinton I served as chief counselor for privacy in the U.S. Office of Management and Budget. Under President Barack Obama I was special assistant to the president for economic policy in 2009 and 2010. Further information is available at www.peterswire.net.

This February the administration issued its white paper, "Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy."[i] This privacy framework defined a Consumer Privacy Bill of Rights. To implement this bill of rights, the framework called on the Department of Commerce to foster the development of enforceable codes of conduct for consumer privacy. These codes of conduct will be developed through multistakeholder processes, so that the range of relevant stakeholders can convene and develop codes of conduct even in the absence of binding legislation or regulation. Consumer privacy legislation has been difficult to enact in the United States, so consumer protection will advance more quickly through initiatives, such as the multistakeholder process, that do not depend on passage of such legislation.

Along with the administration's framework, the Federal Trade Commission, or FTC, has continued its vital role in U.S. privacy policy and enforcement. On March 26, 2012, the FTC issued "Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers."[ii] This report reflected intensive FTC efforts on a wide range of privacy topics. The comments here, building on a short previous statement,[iii] focus on the FTC's recommendations about how to approach the important issue of de-identified data.

The importance of de-identified data

The title of the administration's white paper reflects two principal goals for policy concerning the data of individual consumers: "A Framework for Protecting Privacy and Promoting Innovation." This title reflects the risks to individuals if privacy is not protected effectively. It also reflects the importance of creating good information rules in order to foster innovation and growth in our information economy.

The issue of de-identified data creates a vital opportunity to meet both goals—use data for innovation and growth while also protecting privacy. At least in theory, de-identified data allows us to have our cake and eat it, too. With de-identified data, we strip out the name and other information that reveals identity, but we nonetheless can process the data, do research, discover patterns, and innovate in how we respond to the information.

In any statute or other legal obligation, such as a company's enforceable promise to protect privacy, the most important definition is what counts as covered by the law or obligation. Defining what counts as "de-identified" is crucial because it draws the line between what data is covered by privacy protections (still "identified") and what data is not ("de-identified").

In U.S. law de-identified data was first defined as part of the Health Insurance Portability and Accountability Act, or HIPAA, medical privacy rule drafted in the late 1990s. I was very involved in drafting the proposed and final HIPAA rule and paid particular attention to defining what counted as "de-identified." In HIPAA "identified" data is considered personal health information, subject to the full range of privacy protections. If the data is scrubbed hard enough, however, then it becomes de-identified data and no longer subject to the regulatory requirements.

The final HIPAA medical privacy rule provided two ways to show that data was de-identified. First, the holder of the data could remove a list of at least 17 data fields that could identify a person, such as name, address, or Social Security number. Second, a statistical expert could certify that the risk is very small that the information could be used, alone or in combination with other reasonable available information, to re-identify the individual. Since HIPAA went into effect nearly a decade ago, health care entities have been able to publicly release health data if it has been scrubbed well enough to meet the regulatory requirements for de-identification.

Finding a Goldilocks solution for de-identified data

Since the HIPAA de-identification provisions were proposed in 1999, we have learned a lot about when and how it is possible to "re-identify" data—to link a person's name with the supposedly de-identified data. Two big trends have made it harder to keep information de-identified. First, search on the Web has gotten much better. Google was not incorporated until 1998, and today's search engines let anyone link together tidbits from previously hard-to-link data sources. Second, the amount of information on the Web about a typical person has grown astronomically, including all of the personal details on a person's blog or Facebook page.

The combination of efficient search tools and lots of data means that there is a higher likelihood today that a person's medical or other records can be re-identified even if the name and other traditional identifiers are deleted. For instance, the de-identified medical record might state that a person in Ohio had minor hand surgery on April 3. In the past, it would have been difficult or impossible for an outsider to figure out the name. Today, online search might turn up a social network thread about the hand surgery—there are multiple such surgeries in Ohio each day, but not that many. A bit of follow-up research, using the rest of the supposedly de-identified information, might easily pinpoint the person who had the surgery.

As academics have analyzed these facts about re-identification, some have concluded that the entire effort to de-identify data has failed, because of the risk of linking information back to the individual. [iv] Others have emphasized the limited actual success of re-identification efforts in practice, and found that the benefits to research and innovation are so great that they outweigh the privacy risks.[v]

The preliminary FTC report, issued in 2010, received strong criticisms from both of these

perspectives. The earlier report would have applied privacy protections to "consumer data that can be reasonably linked to a specific consumer, computer, or other device." The debate centered on what the FTC meant by "reasonably linked." Consumer groups correctly emphasized that it is easier now to search on the Web and re-identify data, at risk to privacy. Researchers and other users of data focused on the problems that come with an over-broad definition of "reasonably linked," which could extend privacy rules to an almost unlimited range of data processing, if enough effort is put into tracking down and re-identifying data.

Responding to these critiques, the FTC looked at the technical de-identification issues,[vi] and found what I believe is a Goldilocks solution for the problem of de-identified data. The FTC provides what amounts to a safe harbor where: "(1) a given data set is not reasonably identifiable; (2) the company publicly commits not to re-identify it, and (3) the company requires any downstream users of the data to keep it in de-identified form."

The FTC approach responds to the technical experts who correctly say that it is easier today to find data on the Web that helps us re-identify data. To address the privacy concerns the FTC approach first requires a company to make a data set reasonably de-identified. We can think of this as "good but not foolproof de-identification." Then, in addition, the FTC requires administrative protections. The company has to commit publicly that it won't re-identify the data. The company also has to get similar promises from anybody downstream who receives the data. These promises are enforceable because Section 5 of the FTC Act prohibits deceptive practices, such as broken privacy promises. Privacy is protected through the combination of technical measures, having reasonably de-identified data, and backup administrative measures, so that the only people who receive the data have made binding promises not to re-identify.

The FTC approach also responds to those who want to study data for research, innovation, and related purposes. Data must be scrubbed pretty hard but not incredibly hard—the dataset need merely not be "reasonably identifiable." That data should still often be detailed enough to be useful for a variety of purposes, protected by the enforceable promises not to re-identify.

I have long believed that technical controls alone are not enough to protect consumers against possible re-identification, as shown in a 2009 report by the Center for Democracy and Technology[vii] and my December talk on de-identified data.[viii] The best path is to have reasonably strong technical protections, supplemented by the sorts of enforceable promises that the FTC report supports.

Why defining de-identified data is a good fit for the multistakeholder process

The combination of the importance of de-identified data and the FTC's support for the mix of technical and administrative protections makes the de-identification issue a top candidate for early use of the multistakeholder process. This issue has multiple advantages, including heightened protection for consumers, positive effects on innovation and the broader economy, and likelihood of concrete, enforceable success for the process

itself.

Consumers benefit if and when companies implement the FTC de-identification safe harbor. Privacy risks are lower if companies hold data in reasonably de-identified form, compared to holding that data in fully identified form. Within the company, reasonably de-identified data is less likely to be subject to peeping by employees who are not authorized to see the data, as has happened for instance to the passport records of presidential candidates and the medical records of numerous celebrities.[ix] Reasonably de-identified data also reduces the risk from a data breach, because the chances of identity theft and other harms to consumers will be lower if their names and other identifying information have been masked. In addition, the enforceable privacy promises by the companies mean that a new layer of administrative protections will exist on top of current technical de-identification protections.

Researchers and others who use data will benefit from the de-identification safe harbor. Good-faith researchers already implement privacy and security measures to protect the confidentiality of the data about individuals. For instance, medical researchers who agree to "data use agreements" under HIPAA get enhanced access to personal health information while promising to implement good confidentiality protections. By creating a clear legal mechanism to enable research and similar uses, the FTC de-identification safe harbor encourages responsible and innovative use of information.

With the safe harbor, companies also gain an important new incentive to implement reasonable de-identification procedures. The safe harbor makes it worth the while of companies to implement reasonable de-identification procedures—the company faces lower data breach and other risks from disclosure or use of the data. Under the Consumer Privacy Bill of Rights, responsible companies would have specific compliance responsibilities, such as to provide access to certain personal data and to use personal data consistent with the context in which it was collected. De-identified data, however, is outside the scope of these compliance responsibilities. Companies thus can make compliance easier by using the de-identification safe harbor.

These benefits to consumers, companies, and the economy create an opportunity for a win/win outcome from the multistakeholder process. In self-regulatory approaches, I have long argued that a credible threat of regulation is often important for convincing participants that it is better to agree to a code of conduct than to leave the status quo in place.[x] Because Congress has long been divided on how to address privacy protection, the likelihood of legislation is not very high in the short term. It may thus be difficult to achieve consensus in the multistakeholder process for issues where stakeholders have sharply differing views.

Instead of facing the threat of legislation on de-identification, companies today face uncertainty in practice about what constitutes "reasonably de-identified" under the FTC safe harbor and what will count as sufficiently strong commitments not to re-identify. Multistakeholder consensus about these issues can provide valuable clarification about what it takes for a company to qualify for the safe harbor, with the accompanying

benefits to companies, consumers, and other users of the data.

A related advantage is that there are well-defined pieces that would benefit from the multistakeholder process. My suggestion is not to seek a global definition of "reasonably de-identifiable" for all types of data. Instead, early efforts can focus on situations that arise often and are near the line between identified and de-identified data. For example, legal regimes have varied about how to treat IP addresses, which are the Internet addresses used by your computer or smartphone when communicating with a Web site. Web sites automatically log these IP addresses when you visit their web page, for reasons including the need to know where to send the pages you select to read. These IP addresses don't list you by name, but with more or less effort a website may be able to link the address to a user's name. I suggest that this topic of IP address could be an early candidate for a multistakeholder process, to define what counts as "reasonably de-identified" for IP addresses, and what sorts of privacy promises effectively reduce the risk of re-identification. A code of conduct here could help everyone who runs a website in order to highlight which activities deserve full protections as personal data and which ones instead qualify for the FTC safe harbor and thus don't trigger the requirements of the Consumer Privacy Bill of Rights.

A second candidate could be how to draw the line between identified and de-identified for other information kept in routine website logs, perhaps including the use of cookies. Like IP addresses, cookies don't list a user's name, but with more or less effort a website can often figure out a way to re-identify the user. In my view, cookie data quite possibly can be scrubbed enough so that at some point it should be considered "reasonably de-identified." When the holder of the cookie information also enforceably promises not to re-identify the user, then the privacy risks from that cookie information become lower. The combination of technical and administrative measures may be an important way to find greater areas of consensus in how cookies are used in connection with targeted online marketing. Even if complete consensus is not reached on an issue as contested as cookies, the process may provide important information about uses that are clearly on one side or the other of the identified/de-identified line.

Conclusion

These comments have explained reasons for the FTC de-identification safe harbor to be the basis for early use of the privacy multistakeholder process. I commend the NTIA and the Department of Commerce for its leadership on privacy issues, and look forward to the continued efforts in this area.

Peter Swire
240.994.4142
peter@peterswire.net
www.peterswire.net

[i] The White House, "Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy" (2012), available at http://www.whitehouse.gov/sites/default/files/privacy- final.pdf.

[ii] Federal Trade Commission, "Protecting Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers" (2012), available at http://www.ftc.gov/opa/2012/03/privacyframework.shtm.

[iii] Peter Swire, "FTC Deserves Praise for Its De-Identification Safe Harbor," Future of Privacy, March 26, 2012, available at http://www.futureofprivacy.org/2012/03/26/fpf-senior-fellow-peter-swire-ftc-deserves-praise-for-its-de-identification-safe-harbor/.

[iv] Paul Ohm, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization" *UCLA Law Review* 57 (1701) (2010), available at http://ssrn.com/abstract=1450006.

[v] Jane Yakowitz, "Tragedy of the Data Commons," *Harvard Journal of Law and Technology* 25 (2011), available at http://ssrn.com/abstract=1789749.

[vi] Ed Felton, chief technology officer of the FTC, listed de-identification as the top issue of "special interest to techies" in the FTC report. Ed Felton, "Tech Highlights of the FTC Privacy Report" (Washington: Federal Trade Commission, 2012), available at http://techatftc.wordpress.com/2012/03/26/tech-highlights-of-the-ftc-privacy-report/.

[vii] Center for Democracy and Technology, "Encouraging the Use of, and Rethinking Protections for De-Identified (and "Anonymized") Health Data" (2009), available at https://www.cdt.org/healthprivacy/20090625_deidentify.pdf.

[viii] Peter Swire, "Keynote – Setting the State: How De-Identification Came into U.S. Law and Why the Debate Matters Today," Future of Privacy Forum, Conference on De-Identification, 2011, available at http://www.peterswire.net/psspeeches2011.htm.

[ix] Peter Swire, "Peeping," *Berkeley Technology Law Journal* (2009), available at http://ssrn.com/abstract=1418091.

[x] Peter Swire, "Markets, Self-Regulation, and Government Enforcement in the Protection of Personal Information," in U.S. Department of Commerce, "Privacy and Self-Regulation in the Information Age" (1997), available at http://ssrn.com/abstract=11472.