Transparency establishes trust

By: Latanya Sweeney, Chief Technologist | Apr 3, 2014 9:53AM

TAGS: Accountability | Data sharing risks | Personal harms

We are amidst an era of *open data* –a period in which we share details of our personal lives widely in exchange for all kinds of services, often trusting companies with our most intimate facts. Sharing information about our personal lives has fostered technological innovations and influenced more transparency in government (e.g., [1,2]) and in science (e.g., [3,4]). However, once personal data are acquired, it may be shared with others without consumer awareness. So how might we add transparency to data sharing? The goal of this blog is to spark discussion and debate.

Before I go any further, let me advise you that I am solely responsible for this blog's content, characterizations, ideas and choice of topic. This blog may not reflect the views of the FTC or any of its Commissioners.

In 2012, the Federal Trade Commission (FTC) issued its Privacy Framework report that urges companies to adopt practices that make information collection and use transparent [5]. The report describes a particular lack of transparency about the practices of companies that often buy, compile, or sell a wealth of highly personal information about consumers who never interact directly with the company. Consumers are often unaware of the existence of these entities, as well as the purposes for which they collect and use personal information.

A lot of my academic work concerning privacy and technology has been in the healthcare arena [6], so let me use healthcare as an example of what we can learn and achieve when flows of personal information become transparent. A patient expects her doctor and those involved in her care to have access to her medical information. What is not transparent are all the other places where a patient's data may go beyond treatment, care or payment. In the paragraphs that follow, I will describe how we learned about some flows of patient information that otherwise would have been

hidden, and using that knowledge, how we assessed risks that inspired solutions.

Mapping Health Data Flows

Under my lead last year, the Data Privacy Lab at Harvard University started theDataMap project (thedatamap.org), which set out to document all the places personal health data goes beyond the doctor-patient relationship [7]. My team mined publicly available sources of information (e.g., breach notices) to document flows of personal health information [8]. We also used record request letters sent to state agencies to inquire about recipients of publicly available personal health information the state agencies held [8]. Our results appear as Figure 1 below. Each circle represents a category of organizations (e.g., companies and agencies) and the lines between them represent documented flows of personal health information. If the line is dashed, the shared information has no explicit personal identity (e.g., has no name or Social Security number). If the line is solid, the data includes the explicit name of the person or has other directly identifying information. At the website, you may click on a circle to see the names of actual organizations involved in the sharing.

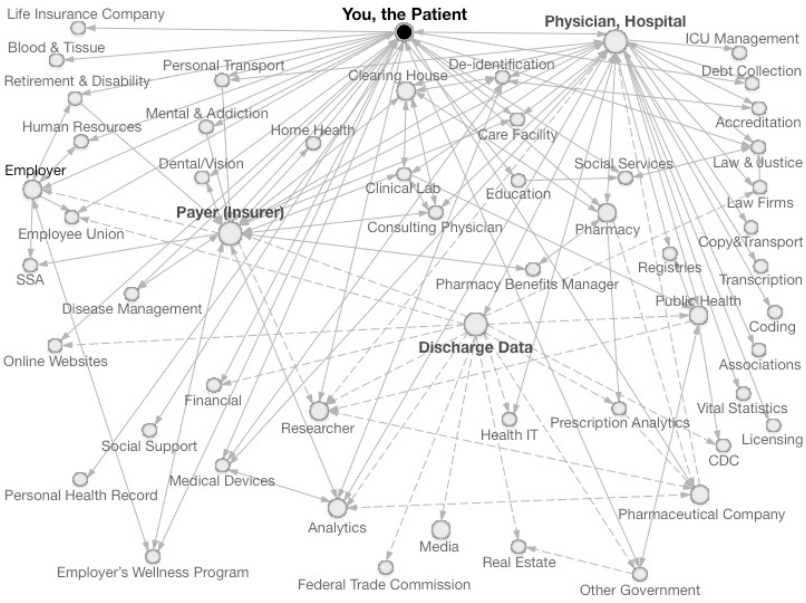

Figure 1. Flows of personal health data documented at theDataMap.org as of March 2014.

The biggest problem is not the extent of sharing, but individuals and authorities having insufficient knowledge of the sharing to be able to assess potential harms. TheDataMap can help. In a regime that uses privacy notices, theDataMap can confirm whether a privacy policy includes entities known to receive the data. In a regime that requires consent for data sharing, theDataMap can identify consented flows. In a regime of breach notices reporting lost or stolen personal data, theDataMap allows a

person to reason whether a breach might contain his data.

All organizations, including commercial entities, can also reap benefits from theDataMap. It identifies sources of data; enables systematic review of inappropriate data access; makes it easier to anonymize data because the data provider can better know what other data the data recipient may hold; improves data quality by identifying paths along which data corrections should propagate; and, supports data use by identifying the originating environment in which the data capture occurred.

TheDataMap can also help us better understand the scope legal frameworks may have on data flows. For example, the Health Insurance Portability and Accountability Act (HIPAA) is the federal regulation in the United States that dictates sharing of medical information beyond the immediate care of the patient, prescribing to whom and how physicians, hospitals, and insurers may share a patient's medical information broadly. Many Americans may falsely believe that HIPAA governs the sharing of *all* personal health data in the United States. As shown in Figure 2, HIPAA covers less than half of the flows documented on theDataMap and few of the non-HIPAA covered data flows come directly from the patient.

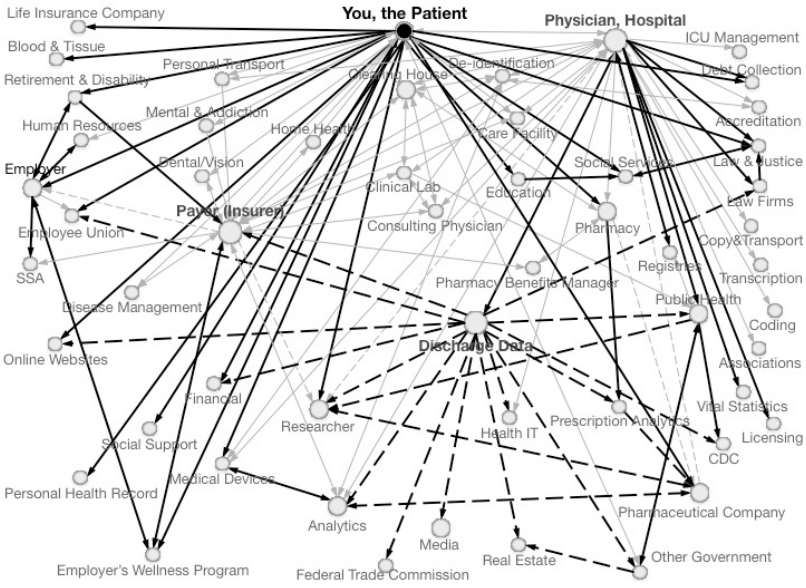

Figure 2. Flows of health data documented at theDataMap.org, not covered by HIPAA.

Risks and Benefits

While the potential for consumer benefits may be great, there may also be important privacy risks to consider. Statewide health data poses an example. A major source of health data sharing on theDataMap is "discharge data"; copies of it go many places. Most Americans have never heard of it even though many Americans are in it. These are statewide

collections of patient health information collected in almost every state, usually under state mandates. Hospitals must forward information about diagnoses, treatments and payments for each hospital visit, and in some states, physicians report this information for each office visit. The state, in turn, may share or sell versions of the data [9].

Sharing data beyond the patient encounter offers many worthy benefits to society. Statewide health data may be particularly useful because they contain a complete set of hospital discharges within the state, thereby allowing comparisons across regions and states such as rating hospital and physician performances and assessing variations and trends in care, access, charges and outcomes (e.g., [10, 11, 12, 13]). Research studies that have used these datasets include: examinations of utilization differences based on proximity [14], patient safety [15, 16], and procedures [17]; and, a comparison of motorcycle accident results in states with and without helmet laws [18]. The very completeness that helps these studies makes it impossible to rely on patients to consent to sharing because the resulting data may not be as complete. Last year a Businessweek article reported that the top acquirers of statewide health data are not researchers, but private companies [19].

While those who want detailed versions of the data have to pass through rigorous application and review procedures and are subject to strong restrictions on how they can use the data, many states also make available to the general public (including commercial enterprises) what are considered less sensitive versions of the data called “public use” data. Obtaining public use data often requires little or no review by the state, and subject to little or no restrictions on use. In this writing, all references to statewide health data (or discharge data) is to the “public use” version unless explicitly stated otherwise.

While almost all states collect discharge data, 33 states sell or share de-identified versions [9]. HIPAA does not cover these data, and only 3 of the states provide the public use data in a way that is as protective as HIPAA warrants [9]. The other 30 states use protections less strict than HIPAA when selling or giving away personal health data to the public. Is the federal standard, HIPAA, too strict? On the other hand, are these states making data more vulnerable to re-identification?

As I reported last year, I purchased a public version of patient-level hospital discharge data from a state for $50 and conducted an experiment to determine the strength of the de-identification of public use data [20]. This publicly available dataset had virtually all hospitalizations occurring in the state in the year, and included patient demographics, diagnoses, procedures, attending physician, hospital, a summary of charges, and how the bill was paid. It did not contain patient names or addresses (only residential postal codes known as ZIPs). Newspaper stories printed in the

state for the same year that contained the word “hospitalized” often included a patient’s name and residential information and explained the reason for the hospitalization, such as vehicle accident or assault. A sample of news information uniquely and exactly matched medical records in the state database for 35 of the 81 sample cases (or 43 percent) found in 2011, thereby putting names to patient records. An independent news reporter verified matches by contacting patients and found them all correct (editors agreed not to publish any names without the explicit consent of the patient) [21]. Matches included high profile cases, such as politicians, professional athletes, and successful businesspeople. Some of the codes included sensitive information beyond the purpose of the visit, such as drug and alcohol use and sexually transmitted diseases.

This experiment generalizes beyond news stories. The kind of information appearing in the newspaper articles is the same kind of information an employer may know about employees who are absent from work for medical reasons and a banker may know about debtors who give medical reasons as a basis for late payments. I am not saying any of the organizations listed on theDataMap are engaged in this practice, but merely noting that employers, financial organizations, and even friends and family members know the same kind of information as reported in news stories making it just as easy for them to identify the medical records of employees, debtors, and others.

After becoming aware of the experimental results, states immediately began solving the problem by improving the protections of publicly available statewide databases [22, 23]. States continue to impose stringent requirements on data requesters who need more identifiable data than these public versions.

In summary, knowing about flows of personal health information allowed us to spot a concern, assess its risk, and help reduce potential harms. Data sharing from the states is transparent because states have requirements to report with whom they share or sell data. Putting the pieces together using theDataMap helped identify a risk, and the experiment quantified that risk and led to improvements. How do we accomplish this kind of transparency generally, beyond health data and with data sharing other than government agencies? There are few, if any, requirements for commercial enterprises, for example, to report data sharing arrangements. How can we know what is going on and where the risks and remedies may be when sharing personal information beyond the person's knowledge?

Audit Logs

Audit logs can help but alone are not sufficient. HIPAA requires health plans, health care providers, and health care clearinghouses to maintain audit logs on electronic access to patient information. Audit logs record

who accessed which patient's data and when the access occurred. Hospitals have rotating staffs with dynamic role assignments, making it difficult to identify inappropriate access at the time of occurrence but in hindsight, audit logs have been helpful. Audit logs documented hospital workers snooping at former President Clinton's record when he was undergoing heart surgery [24] and allegedly providing sensitive medical information about basketball player Kobe Bryant to a newspaper [25]. Of course, audit logs do not necessarily document flows outside the organization.

Approaches

Should companies that hold, buy, or sell personal information about consumers publicly describe the information they hold? Should consumers have reasonable access to the data that companies maintain about them? What should the best practices be? Below are four approaches to ignite brainstorming on possible ways to add transparency to data sharing in the commercial sector.

Approach #1 Public Registry

Each time a company sells or shares a substantial amount of sensitive personal data, information about the data sharing arrangement appears in a publicly available log maintained at the company's website. The registry would include the name of the party receiving the data and aggregate information about the date, number of records and kinds of data fields shared. The registry would not contain any actual personal data. Examples of public data sharing registries exist for statewide health discharge data in Maine [26] and Texas [27].

Approach #2 DataMaps in Privacy Policies

Companies augment privacy policies to include a datamap that shows flows of personal information to and from the company. The flows would use dashed or solid lines to indicate whether the information explicitly identifies the person. Rather than listing the specific names of the parties, the datamap would show the kinds of entities involved.

Approach #3 Personal Copy

A person can acquire a copy of her own data from any company holding a copy of personal information about her. The FTC's's Privacy Framework report considered consumer access to data as a means to promote transparency [5]. It recommends that companies provide consumers with reasonable access to the personal data companies maintain about them, proportionate to the sensitivity of the data and the nature of its use. In addition, the U.S. Department of Health and Human Services requires qualifying healthcare organizations to provide patients with a way to view, download and transmit their own personal health data in a format that is

both machine and human readable [28].

Approach #4 Personal DataMap

Imagine having technology that would produce a personal datamap for a person, showing the actual flows of the person's information across organizations over time. One way to accomplish this goal is to have companies that hold personal information maintain a log of all copies of personal information shared or sold, as described in the first approach, but more detailed to include which fields of which people were shared or sold. The result is like an audit log, but rather than recording access to a person's information within an organization, it would log flows of personal data outside the organization. Making these logs electronically available to the person who is the subject of the data (similar to the Blue Button campaign with health data [28]), provides an opportunity for technology to walk through transactions on behalf of a person, iteratively asking each company in turn for a list of data sharing transactions to construct a person's datamap automatically.

In comparison, the first approach provides information that immediately helps authorities, policy makers and researchers identify possible risks of harms. It also helps with breaches if a person knows a company that had a breach was holding her information, she could use the registry to reason whether her information was in the breach and to ascertain possible consequences or personal harms. The second approach is an incremental improvement to privacy policies but the information is not as detailed as in the first approach. The third approach helps a person learn more about the information others hold about him, but is only useful in cases where he knows the companies holding his personal information. The fourth approach combines the first and third approaches with technical innovations to provide a personal datamap that may be useful for assessing harm.

What Do You Think

This inquiring mind wants to know what you think. Perhaps you have your own approach to describe, an experiment to report, or a comment to make.

Acknowledgements

I presented parts of this post at the first two White House Workshops on Big Data hosted by the White House Office of Science and Technology Policy. Thanks to Daniel Weitzner and Danah Boyd for organizing the workshops and to the participants for the rousing discussion.

References

1. Open Government Initiative. United States. The White House. http://www.whitehouse.gov/open

2. Opening up Government. http://data.gov.uk/

3. Holdren J. Memorandum for the Heads of Executive Departments and Agencies. Office of Science and Technology Policy. Executive Office of the President of the United States. February 22. 2013. www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

4. Open Science. Champions of Change. http://www.whitehouse.gov/champions/open-science

5. Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers. Federal Trade Commission. March 2012. http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf

6. Sweeney L. Research Accomplishments of Latanya Sweeney, 1996-2009. http://dataprivacylab.org/people/sweeney/work/index.html See also the preamble of the Health Insurance Portability and Accountability Act (HIPAA). *Federal Register* v65(250) December 28, 2000. http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/prdecember2000all8parts.pdf. See also the preamble of the Health Breach Notification Rule, *Federal Register* v74(163) August 25, 2009. http://www.gpo.gov/fdsys/pkg/FR-2009-08-25/pdf/E9-20142.pdf

7. TheDataMap Project. http://thedatamap.org

8. TheDataMap Project History. http://thedatamap.org/history.html

9. Hooley S and Sweeney L. Survey of Publicly Available State Health Databases. Harvard University. Data Privacy Lab. 1064-1. June 2013. http://thedatamap.org/1075-1.pdf

10. Balan D and Romano P. A Retrospective Analysis of the Clinical Quality Effects of the Acquisition of Highland Park Hospital by Evanston Northwestern Healthcare. Federal Trade Commission. Working Paper 307. November 2010. http://www.ftc.gov/reports/retrospective-analysis-clinical-quality-effects-acquisition-highland-park-hospital-evanston

11. Garmon C. Hospital Competition and Charity Care. Federal Trade Commission. Working Paper 285. October 2006. http://www.ftc.gov/reports/hospital-competition-charity-care

12. Simpson J and Shin R. Do Nonprofit Hospitals Exercise Market Power? International Journal of the Economics of Business. 5(2) 1998. http://www.tandfonline.com/doi/abs/10.1080/13571519884486#.Uznj6V7TZQg

13. The Federal Trade Commission (FTC) is on theDataMap because it acquires discharge data to advance its law enforcement and policy activities, which include promoting competition in healthcare markets (e.g. [10, 11, 12]). The FTC has specific safeguards and policies in place for handling, storing, and working with discharge data, including public use versions. The FTC does not use discharge data to identify individual patients, and none of the discharge data acquired by the FTC was used in any of the re-identification experiments reported in this blog post.

14. Basu J, Friedman B. A Re-examination of Distance as a Proxy for Severity of Illness and the Implications for Differences in Utilization by Race/Ethnicity. Health Economics 2007;16(7):687-701.

15. Li P, Schneider J, Ward M. Effect of Critical Access Hospital Conversion on Patient Safety. Health Services Research 2007;42(6 Pt 1):2089-2108.

16. Smith R, Cheung R, Owens P, Wilson R, Simpson L. Medicaid Markets and Pediatric Patient Safety in Hospitals. Health Services Research 2007;42(5):1981-1998.

17. Misra A. Impact of the HealthChoice Program on Cesarean Section and Vaginal Birth after C-Section Deliveries: A Retrospective Analysis. Maternal and Child Health Journal 2007;12(2):266-74.

18. Coben J, Steiner C, Miller T. Characteristics of Motorcycle-Related Hospitalizations: Comparing States with Different Helmet Laws. Accident Analysis and Prevention 2007;39(1):190-196.

19. Robertson J. Who's Buying Your Medical Records. Bloomberg News. June 5, 2013. http://www.businessweek.com/news/2013-06-05/states-hospital-data-for-sale-leaves-veteran-s-privacy-at-risk

20. Sweeney L. Matching Known Patients to Health Records in Washington State Data. Harvard University. Data Privacy Lab. 1089-1. June 2013. http://thedatamap.org/1089-1.pdf

21. Robertson J. States' Hospital Data for Sale Puts Privacy in Jeopardy. Bloomberg News. June 05, 2013. http://www.businessweek.com/news/2013-06-05/states-hospital-data-for-sale-leaves-veteran-s-privacy-at-risk

22. Engrossed Substitute Senate Bill 6265. State of Washington. 63rd Legislature. 2014 Regular Session. http://apps.leg.wa.gov/documents/billdocs/2013-14/Pdf/Bills/Senate%20Passed%20Legislature/6265-S.PL.pdf

23. Patient Discharge Data. Healthcare Information Division. Office of Statewide Health Planning and Development. State of California.

http://www.oshpd.ca.gov/HID/Products/PatDischargeData/PublicDataSet/index.html

24. Stein, T. How Safe Are Your Computers. Hack Attack. Physicians Practice. February 1, 2005. http://www.physicianspractice.com/display/article/1462168/1588200

25. Miller, M. Issues of Privacy in the Bryant Case. Los Angeles Times. September 8, 2003. http://articles.latimes.com/2003/sep/08/health/he-court8

26. Current Data Requests. Maine Health Data Organization. https://mhdo.maine.gov/datarequest.aspx

27. Organizations Receiving Chapter 8 Data: Chapter 108 Data Recipients List 2003-November 21, 2011, Center for Health Statistics, State of Texas, http://www.dshs.state.tx.us/thcic/DataPurchasers.pdf

28. About Blue Button. U.S. Department of Health and Human Services. http://www.healthit.gov/patients-families/blue-button/about-blue-button

Add new comment

COMMENTS

Raleigh replied on Apr 3, 2014 3:14PM PERMALINK

I like the data map concept. I also like the public registry idea and the ability to view a personal copy of the specific personal data being shared. I wanted to add another idea to the mix: a "reverse" of the Do Not Call registry. This would be an opt-in registry where people could give their informed consent to their personal data being shared at all. Without such informed consent demonstrated in this registry, then the most severe restrictions on sharing of personal data would be in force. The entities viewing the registry would include the data sharing entities as well as the affected persons who decide whether or not to give informed consent.

reply

FUBAR_53 replied on Apr 3, 2014 4:15PM PERMALINK

My "Major" problem as a consumer is "Undisclosed" third parties that information is disclosed too as routine procedure(s). These include websites that promote "Credit cards for students" after applying for registration at a community college; to, collection agencies applying for AMEX credit cards to get credit reports when I don't use credit accounts! Well, maybe layaway store plans. Privacy in the USA is "An Endangered Species" about to go extinct.

reply

Novelreader replied on Apr 3, 2014 7:32PM PERMALINK

The proposals bring to mind a quote from Alexander Solzhenitsyn's Cancer Ward: "As every man goes through life he fills in a number of forms for the record, each containing a number of questions . .. There are thus hundreds of little threads radiating from every man, millions of threads in all. If these threads were suddenly to become visible, the whole sky would look like a spider's web, and if they materialized as rubber bands, buses; trams and even people would all lose the ability to move, and the wind would be unable to carry torn-up newspapers or autumn leaves along the streets of the city. They are not visible, they are not material, but every man is constantly aware of their existence… Each man, permanently aware of his own invisible threads, naturally develops a respect for the people who manipulate the threads." A personal data map (proposal 4) would make the invisible threads visible, and the resulting map for each individual would be so dense that a page or screen would be unreadable to all but those with the largest monitors made today.

reply

ADD NEW COMMENT

Commenting Policy

Username

Please enter a username. Don't use your email address.

Comment *

What code is in the image? *

Enter the characters shown in the image.

Post

PRIVACY ACT STATEMENT

It is your choice whether to submit a comment. If you do, you must create a user name, or we will not post your comment. The Federal Trade Commission Act authorizes this information collection for purposes of managing online comments. Comments and user names are part of the Federal Trade Commission's (FTC) public records system (PDF), and user names also are part of the FTC's computer user records system (PDF). We may routinely use these records as described in the FTC's Privacy Act system notices. For more information on how the FTC handles information that we collect, please read our privacy policy.

List a Number on the National Do Not Call Registry

Report An Antitrust Violation

Strategy

Open Government