Frameworks for Companies to Share Data With Researchers

By DARAKHSHAN MIR
Department of Computer Science, Wellesley College, Wellesley, MA 02481 USA

I. INTRODUCTION

Our lives in the 21st century are characterized by being data-rich—in the process of constantly interacting with technology we leave a rich trail of our lives in data. These interactions capture varied aspects of our lives, such as our shopping behavior, our mental and physical health, our likes and dislikes, political and religious opinions, products we shop for, how we drive our cars, how much energy we use, and even more nebulous things such as our hopes and aspirations. Credit card companies record every purchase we make, with the time and place of purchase; companies like Amazon and Netflix not only record our purchases but also our interest in things we merely peruse on their websites. Loyalty cards and digital profiles of our purchases even in brick-and-mortar stores such as Target and Walmart and our local supermarkets create a record of all our purchases, easily being able to infer our shopping patterns. In addition to location data captured from using products like Google Maps and FourSquare, call detail records (CDRs), associated with our cell phone usage, generate a detailed account of where we go, making it easier to infer what we do. Personal health devices such as FitBits and wearable personal sensors record information about our daily activities, including information related to our health, sleep, and diet, often with the times and locations of those activities [3] and present new opportunities for healthcare and research [11]. Taken to the extreme, the *Quantified Self* movement aims at incorporating technology in all aspects of an individual's daily life, recording every aspect of one's life in data [22].

Such data, obviously, constitute a rich resource for studying aspects of our lives, habits, and behaviors, both individually and collectively. Furthermore, technological advancements enable indefinite retention of these data. A large amount of these data are captured and retained by corporations, in two broad contexts. On the one hand, companies sell a concrete service or content gathering data as a byproduct—for example, Target, Netflix, Amazon, and mobile phone providers. On the other hand, companies like Google and Facebook gather data in exchange for some of their "free" services we enjoy. Researchers of various hues could find such data crucial for studying different aspects of our lives. Are there frameworks by means of which companies could share such data with researchers? While sharing such data, society at

Digital Object Identifier: 10.1109/JPROC.2015.2461272

large, and companies in particular, have to contend with several ethical and proprietary concerns; are there arguments and procedures that could allay these concerns? This article discusses the ways in which companies use such data, the manner in which they already share data (primarily for profit and not just with researchers), and while recognizing the complex ethical issues raised by the generation, collection, sharing, and questions of ownership of individuals' data, proposes a way forward to share data with researchers. If willing, companies can employ data-sharing strategies that are respectful of norms of information flow people expect in everyday life to share data with researchers. This article argues for a middle ground between the two extremes of keeping data locked away and making it "free for all."

II. INTERNAL AND EXTERNAL USE OF PERSONAL DATA BY COMPANIES

Entities—including corporations—holding people's data use them internally for a variety of analyses. For example, Target assigns each shopper who has shopped, either in person or online, a guest ID, linking their shopping records using a credit card, a home address, a phone number, and/or a loyalty card. In 2012, the *New York Times* [7] published an article on how Target routinely harnesses these data to predict customers' shopping and consumption behavior to serve them customized ads in order to maximize revenue. Often, this is done in combination with traditional modes of advertising. The article illustrates how Target uses these data to predict "life-changing" events in consumers' lives, such as the birth of a child, when they are known to change deeply set-in shopping habits. Using a woman's buying patterns, Target could predict whether she was pregnant and send her ads and coupons that could possibly set in a new shopping pattern at a crucial "life-changing" moment. Target can also buy external demographic information about customers from data brokers and other sources to correlate it with their shopping habits and discover more about its customers.

Data-brokering companies like Axciom collect and combine information about people from a variety of sources and sell it to companies that want to learn more about current or potential customers. Companies like Facebook hold a vast amount of information about individuals, not only via the information, likes, and behavioral patterns that its users share via the social network, but also by partnering with these data brokers. Using this information from brokers like Acxiom and Epsilon [9], an individual's profile on Facebook can be matched with other outside information such as in-store purchases they might have made, to further improve targeted advertising.

Undoubtedly, these data are invaluable to companies because of the extent to which data capture aspects of individuals', and by extension social, lives. However, there is also an inherent investigative potential in these data that holds benefits for society in general. Social scientists, epidemiologists, and other researchers and public policy makers could harness these sources of data to gain novel insights about aspects of our lives. In some cases, such as predicting the spread of an epidemic like Ebola, one could argue that there is a moral imperative to use such data to improve human lives.

III. DATA PRIVACY PIPELINE

The abundance of data and the ease with which varied sources of data can be correlated to reveal a great deal about individuals' lives—the exact reason why it is so valuable to corporations—has caused much consternation among individuals, civil liberties group, policy makers, and legal and governmental organizations. The process of generation, collection, storage, and analyses (including cross-correlational analyses) of individuals' data and their potential sharing constitutes what I call a "data privacy pipeline."

This pipeline begins with any process that generates and/or gathers individuals' data, and spans the stages of: 1) collecting data; 2) storing and using collected data for analyses at the individual and/or population level; and 3) finally, potentially sharing the data or analyses with other external entities whether for research or commercial purposes. Each stage of this pipeline raises several questions about individuals' privacy, such as their attitudes toward such a collection, the control, or lack thereof, they possess over the data they generate, concerns regarding the manner in which such data are used, and which parties have access to this information.

There is a rich collection of literature that investigates each stage of the pipeline from the perspective of various disciplines that intersect with privacy such as law, ethics, philosophy, biology, epidemiology, public health, commerce, and economics. Recent workshops organized by the White House and the President's Council of Advisors on Science and Technology (PCAST) have culminated in reports [20], [21] that stress the need for privacy-preserving technologies and a multidisciplinary investigation of privacy in the context of Big Data. For a comprehensive discussion that addresses questions arising at various stages of the privacy pipeline such as how consumers' data should be collected, who should own it, the transparency of the data trail, and recommendations for corporations and policy makers, see the Federal Trade Commission's (FTC) report [10]. For a more recent comprehensive coverage of privacy, especially in the context of surveillance, see Schneier's book *Data and Goliath* [18].

A. Importance of Context in the Pipeline

While understanding information flow through several stages of this privacy pipeline, it is important to

consider the context in which data are gathered and shared. Often, in debates and discourses on privacy, the purpose of sharing data for research gets conflated with using them for commercial purposes, overlooking the importance of context. This is especially the case when companies share data. However, individuals and policy makers make a sharp distinction between the two, recognizing the inherent "social good" in making data available for noncommercial and investigative purposes, particularly those that seek to improve our lives as a society.

Nissenbaum [16] argues that when sharing information and considering privacy, it is important to consider the context in which data are shared. She argues for "context-specific substantive norms" that should determine who collects data, who it could be shared with, and under what circumstances. Nissenbaum expounds that information flow, *per se*, is not problematic, rather, the problem lies in the departure of information flow processes from "entrenched norms." To illustrate, she presents the example of how patients entrust their medical information to their doctors and how selling this information to other parties would be a disruption of an established norm. She proposes "contextual integrity" as a framework to reason about privacy, which tells us that the appropriate methodologies to share information are those that are rooted in "norms" entrenched in distinct spheres of our social life.

If we view the decision of companies sharing data with researchers through the lens of "contextual integrity," can we develop frameworks that enable sharing of these data within well-accepted norms? This begs the question: What are the norms of information flow through each of the three stages of the privacy pipeline? Formulating norms of information flow through stages of the privacy pipeline that correspond to collection and use of data by companies is a complex issue, very much in flux (see the FTC report [10], for example). However, restricting ourselves to stage 3) of the privacy pipeline that corresponds to sharing of data, can we formulate frameworks for companies to share data with researchers? Can we do so in a manner that respects context-specific "substantive norms" of information flow? Can we do so, while not being besieged by important, complex ongoing discussions involving stages 1) and 2), corresponding to gathering, storing, and internally using data? In the following sections, I argue that we can; we do not have to wait for consensus of norms in earlier stages of the pipeline and that it is possible for companies to share data with researchers in what I call a "forward-extensible" manner that is cognizant of current discourses in privacy.

Forward extensibility is a software system design principle "where the implementation takes future growth into consideration."[1] Forward-extensible data-sharing mechanisms should 1) be capable of working with any norms, legal or otherwise, that as a society we may formulate in the future for the data-gathering and data-usage stages of the pipeline; and 2) be informed by both current opinions of the public and current scholarly work regarding privacy, incorporating essential elements of these in the implementation. It is crucial that these sharing practices do not disrupt, but rather support evolving norms.

The argument builds on two facts: 1) accounting for people's preferences about sharing their data in a specific context—to researchers, for the public good; and 2) assessing the success and failure of models of data sharing that companies have employed in the past.

IV. SHARING DATA FOR RESEARCH

People are more willing to share data when they see a tangible social good in such sharing and when norms of privacy are respected. Such norms are often informed by an underlying trust in processes and institutions—including academic institutions—that serve the public good.

Researchers supported by the Robert Wood Johnson Foundation [17] surveyed individuals who track their own personal health using personal health data (PHD) devices, such as FitBits. PHD devices that self-track everyday behavior and lifestyle can help fill the gap in traditional public health that relies on electronic health records, periodic surveys, and clinical trials. When asked if they would be willing to share personal health and activity data with researchers, 78% of respondents answered either "probably would" or "definitely would"; 57% said the most important condition for making their personal health data available for research was an assurance of privacy of their data; and 90% of respondents said that it was important that the data be anonymous.

Another recent NPR poll indicates that most Americans are willing to share their data for research—overall 53% of participants said that they would be comfortable sharing their data for research as long as it is anonymous [12].

On surveying personally controlled health records (PCHRs) users, Weitzman *et al.* [23] found that 91% were willing to share medical information for health research with 59% favoring an opt-in sharing model. Further, willingness to share was "conditioned by anonymity, research use, engagement with a trusted intermediary, transparency around PCHR access and use, and payment."

These examples indicate that people are willing to share personal data for the public good, when certain substantive norms (such as a trust or a guarantee that the researcher will not compromise the privacy of the individual) are respected. However, they are largely unwilling to do so for commercial purposes. The Pew Research Center conducted a study of cloud services users in 2008 [13]: 90% said that they "would be very concerned" if the company at which their data were stored sold them to

[1]https://en.wikipedia.org/wiki/Extensibility

another party; 68% said that they would be "very concerned if companies who provided these services analyzed their information and then displayed ads to them based on their actions."

A recent Pew study [4] found that 80% of social networking site users are "concerned about third parties like advertisers or businesses accessing the data they share on these sites." Further, 91% of adults in the survey "agree" or "strongly agree" "that consumers have lost control over how personal information is collected and used by companies."

Researchers, on their end, can harness the unique opportunities these data present to serve the public good. For example, a study used cell phone data in Kenya [25] to show that human movements contribute to the spread of the malarial parasite—and hence ultimately transmission of malaria—on spatial scales that extend beyond that of mosquito travel. These data also helped construct models that identify the sources of such travel-initiated infection and pinpoint likely sites of transmission, potentially improving malaria control programs. Similar studies harness mobile phone usage data and ferry traffic data between Zanzibar and Tanzania [14] to quantify malaria importation rates and identify potentially high-risk travelers.

In the context of the recent Ebola outbreak crisis, Wesolowski *et al.* argue for the utility of CDRs in understanding human mobility in the spread and possible containment of Ebola [24].

As another example, Google searches seeking information about major mental health disorders have been observed to follow seasonal patterns [1], suggesting a strong correlation between mental health illness and seasons. Such passive monitoring can help fill in the gaps in traditional health monitoring programs such as in-person or over-the-phone surveys. One could posit that the surveys indicate that most people would be willing to share their data with researchers for uses such as the above, as long as their expectations of privacy are upheld and the data are not "free for all."

I will discuss two models of data sharing that have been used by companies in the past, contrasting them in the ways they are or are not forward extensible. Note that I do not discuss the specifics of how these data were collected and how they were used by the collecting entity. Nevertheless, one can still argue about the data-sharing mechanisms themselves.

A. "Release-and-Forget" Model

One extreme is the so-called "release-and-forget" model where data are "sanitized" to remove any obvious identifiers (such as names or social security numbers) and then publicly released, affording flexibility to a wide array of researchers, whether academically affiliated or not. However, a history of well-intentioned but ultimately failed attempts cautions against such an approach. As an example, in 2004, AOL released an anonymized log of 20 million search queries to embrace "the vision of an open research community, which is creating opportunities for researchers in academia and industry alike."[2] Shortly after, two *New York Times* journalists [2] used these data to identify a person, user number 4417749 identified as Thelma Arnold of Lilburn, Georgia, in this data set based on their searches. Arnold's thoughts on associating the search terms—that "We all have a right to privacy. Nobody should have found this all out"—indicate that a norm had been disrupted, making the resulting flow of information problematic.

In another high-profile case in 2008, researchers deanonymized the Netflix Prize data set, an anonymized data set of millions of movie reviews by Netflix customers, by cross referencing it with the IMDB database of movie reviews [15]. Consequently, Netflix had to cancel the second part of this competition, which was intended to (and did) stimulate algorithm design for improving recommendation algorithms.

Owing to the fallout of these highly publicized failures, companies are less likely to share data with researchers, though, as is evident from both the FTC report [10] and Schenier's [18] recent book, internally, they continue to use them.

As we saw earlier, mobile phone data present unique opportunities to deploy data for social good, especially as cell phones are ubiquitous in developing countries where traditional methods of gathering data are either too impractical or expensive. While studying the spread of epidemics such as malaria and Ebola using CDRs is important, we cannot advocate using a release-and-forget approach to share CDRs. Recent studies have shown that such spatiotemporal data are highly unique: four spatiotemporal points with a temporal aggregation of one hour and a spatial resolution specified by the carrier's antennas are sufficient to uniquely identify an individual in a data set consisting of months of CDRs of millions of people [6]. Such uniqueness makes data, even if anonymized, particularly vulnerable to cross-correlational attacks.

A recent report by the U.S. PCAST [21] stresses that the anonymized release-and-forget approach is ineffective in providing privacy, with difficulties exacerbated by advances in data mining and Big Data. The report recommends employing privacy policy to ensure that use of data does not compromise privacy. In all the examples above of the "release-and-forget" model, we see that the data-release mechanisms lack any semblance of control by the companies: data once released cannot be reclaimed, and no future evolution of norms can reverse the damage. Further, such sharing mechanisms do not seem to be fully aware of privacy concerns of scholars from various disciplines, including computer science. Therefore,

[2] http://sifaka.cs.uiuc.edu/xshen/aol/20060803_SIG-IRListEmail.txt

"release-and-forget" is not a forward-extensible model. Supplementing data sharing by policies that mandate how data can be shared, with whom, for how long, and for what purposes makes forward extensibility possible.

B. "Forward-Extensible" Models

Recently, several companies have successfully developed data-sharing models that employ policy to share data with outside researchers, presenting a good middle approach. These models have succeeded because they are cognizant of social norms of information sharing.

In 2013, Orange, a major mobile telephony company, made about 2.5 billion anonymized call detail records and text messages in Ivory Coast available to researchers worldwide.[3] The main purpose of the data for development (D4D) challenge was to foster research for social development of Ivory Coast. Orange termed it as "data philanthropy" and envisioned academic institutions across the world using these data to work on problems of "data for social good." One of the main purposes was to encourage cross-correlational analyses by correlating data from various sources. However, as we saw earlier, correlation is also an important source of the privacy problem.

The challenge culminated in a conference[4] with over 80 research contributions across academic institutions in the world.[5] These spanned a diverse set of studies, such as predicting epidemics, proposing solutions for optimizing local infrastructure such as public transportation based on mobility patterns, studying social and ethnic divisions in society, and detecting weather-driven or other social or political events that may facilitate early intervention. To avail of this opportunity, researchers had to be affiliated with an academic institution, register for the challenge, submit a short blurb on their problem, and if approved, agree to a set of terms and conditions specified by Orange.

The challenge also furthered the understanding of privacy implications of sharing such data. Orange made a preliminary data set available to a team who studied how these CDRs could be deanonymized [19]. They also advised Orange to modify the data to make deidentification more difficult prior to the wider release. However, Orange still relied on a nondisclosure agreement that bound participants to abide by a "code of conduct."

Owing to the success of this venture, Orange repeated the D4D challenge[6] in 2014, this time making cell phone data from Senegal available to researchers.[7]

Similarly, Telecom Italia organized a Big Data Challenge in 2014, making data from a wide variety of resources in Italy available to researchers. To get access to the data, participants had to register and abide by the terms and conditions, some of which were: "The Data cannot be used, distributed or transferred in any form outside of the Contest, neither in its original form nor in processed or aggregate form, except for the production of scientific publications or publications for the dissemination of the Proposals prepared using the Data provided."

V. HOW IT FITS INTO THE BIGGER PICTURE

Both mechanisms used by Orange and Telecom Italia relied on public trust in academic institutions and people's willingness to contribute their data for social good. Further, by highlighting the role of a researcher as a trusted recipient of these data and as someone acting for the public good and by using precise terms of use for the data, companies can provide forward-extensible solutions that may be revised and modified to keep pace with evolving norms. One crucial difference between the "release-and-forget" and "forward-extensible" models is the lack of control over data by the releasing entity in case of the former. The ease of correlation attacks on the "release-and-forget" model facilitated by advances in technology might make it possible to violate norms of information flow that have been deemed acceptable in earlier stages of the privacy pipeline. For example, if a company collected location data from people providing them with a guarantee that unauthorized people will not be able to learn something personal about them, and then released it widely by simply anonymizing it, it would be violating that contract. However, if the company while collecting such data provided people with the guarantee that their data will only be used by researchers in a specific context bound by certain terms and conditions and by the ethical considerations of their profession, the "forward-extensible" model could easily enable stipulation of these guarantees during the sharing process.

In fact, Daries *et al.* [5] argue that data should be made available to researchers by binding them to an ethical and legal framework, thereby, ensuring concerned parties that there is an underlying trustworthy framework. This banks on a socially entrenched norm, and there are penalties if researchers depart from the norm. Both Orange and Telecom Italia leveraged this underlying model of trust.

This article does not discuss the ethical and legal questions that are raised by the methods and contexts in which data are collected and used by companies. However, it proposes a solution that is forward extensible—if the methods used for collections and analyses would be cognizant and respectful of evolving norms on privacy, the decision of sharing those data with researchers would not cause any

[3]http://www.unglobalpulse.org/D4D-NetMob

[4]http://perso.uclouvain.be/vincent.blondel/netmob/2013/

[5]http://www.unglobalpulse.org/D4D-Winning-Research

[6]http://d4d.orange.com/en/home

[7]http://www.d4d.orange.com/en/content/download/29438/273168/version/12/file/D4DSonatel_06062014Engl.pdf

disruption of these norms. While we as a society debate and discuss evolving norms of privacy for each stage of the privacy pipeline, research, particularly in areas as important as public health, should not be held hostage to such discussions. Willing companies can adopt data-sharing models such as those successfully used by Orange and Telecom Italia to share data with researchers. This in addition to making critical data available to researchers also advances research in privacy. For example, such data sharing would also provide authentic scenarios for implementations of notions such as differential privacy [8] that employ a rigorous, mathematical characterization of privacy. A combination of technology and policy that is informed by current discourses in privacy in a range of disciplines could make such data-sharing mechanisms receptive to periodic change and revision.

Perhaps one could envision a scenario where internal entities within companies much like Institutional Review Boards, would oversee such policies, their implementation, and revision. The data-sharing model employed by Orange and Telecom Italia is forward extensible in this sense as well. ■

REFERENCES

[1] J. W. Ayers, B. M. Althouse, J.-P. Allem, J. N. Rosenquist, and D. E. Ford, "Seasonality in seeking mental health information on Google," *Amer. J. Preventive Med.*, vol. 44, no. 5, pp. 520–525, 2013.

[2] M. Barbaro and T. Zeller, Jr., "A face is exposed for AOL searcher no. 4417749," *New York Times*, Aug. 9, 2006. [Online]. Available: http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all&r=0

[3] A. Bleicher, "Hacking the human OS," *IEEE Spectrum*, vol. 52, no. 6, p. 31, 2015, DOI: 10.1109/MSPEC.2015.7115560.

[4] Pew Research Center "Public perceptions of privacy and security in the post-Snowden era." Nov. 2014. [Online]. Available: http://www.pewinternet.org/2014/11/12/public-privacy-perceptions/

[5] J. P. Daries *et al.* "Privacy, anonymity, big data in the social sciences," *Commun. ACM*, vol. 57, no. 9, pp. 56–63, Sep. 2014.

[6] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Sci. Rep.*, vol. 32013, DOI:10.1038/srep01376.

[7] C. Duhigg, "How companies learn your secrets," *New York Times*, Feb. 2012. [Online]. Available: http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. 3rd Conf. Theory Cryptogr.*, 2006, pp. 265–284.

[9] Facebook "Advertising and our third-party partners." Apr. 2013. [Online]. Available: https://www.facebook.com/notes/facebook-and-privacy/advertising-and-our-third-party-partners/532721576777729

[10] Federal Trade Commission "Protecting consumer privacy in an era of rapid change," Mar. 2012. [Online]. Available: http://www.ftc.gov/os/2012/03/120326privacyreport.pdf

[11] S. Hassler, "Homing in on health care's sweet spots [spectral lines]," *IEEE Spectrum*, vol. 52, no. 6, pp. 12–12, Jun. 2015.

[12] S. Hensley, "Poll: Most Americans would share health data for research," NPR, Jan. 2015. [Online]. Available: http://www.npr.org/blogs/health/2015/01/09/375621393/poll-most-americans-would-share-health-data-for-research

[13] J. Horrigan, "Use of cloud computing applications and services," Pew Research Center, Sep. 2008. [Online]. Available: http://www.pewinternet.org/2008/09/12/use-of-cloud-computing-applications-and-services/

[14] A. Le Menach *et al.* "Travel risk, malaria importation and malaria transmission in Zanzibar," *Sci. Rep.*, vol. 12011, DOI: 10.1038/srep00093.

[15] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proc. IEEE Symp. Security Privacy*, Washington, DC, USA, 2008, pp. 111–125.

[16] H. Nissenbaum, *Privacy in Context: Technology, Policy, the Integrity of Social Life*. Stanford, CA, USA: Stanford Univ. Press, 2009.

[17] University of California San Diego "Personal data for the public good: New opportunities to enrich understanding of individual and population health," Robert Wood Johnson Foundation, Mar. 2014. [Online]. Available: http://www.rwjf.org/en/library/research/2014/03/personal-data-for-the-public-good.html

[18] B. Schneier, *Data and Goliath: The Hidden Battles to Capture Your Data and Control Your World*, 1st ed. New York, NY, USA: Norton, 2015.

[19] K. Sharad and G. Danezis, "De-anonymizing d4d datasets," presented at the *Workshop Hot Topics Privacy Enhancing Technol.*, Bloomington, IN, USA, 2013.

[20] U.S. President's Council of Advisers on Science and Technology (PCAST) "Big Data and privacy: A technological perspective," May 2014. [Online]. Available: http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf

[21] U.S. President's Council of Advisers on Science and Technology (PCAST) "Big Data: Seizing opportunities, preserving values," May 2014. [Online]. Available: http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

[22] E. Waltz, "The quantified olympian," *IEEE Spectrum*, vol. 52, no. 6, pp. 44–45, Jun. 2015.

[23] E. R. Weitzman, L. Kaci, and K. D. Mandl, "Sharing medical data for health research: The early personal health record experience," *J. Med. Internet Res*, vol. 12, 2010, Art. ID. e14.

[24] A. Wesolowski *et al.* "Commentary: Containing the Ebola outbreak-the potential and challenge of mobile network data," *PLOS Currents Outbreaks*, 2014, DOI: 10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e