64 STAN. L. REV. ONLINE 63
February 2, 2012

SYMPOSIUM ISSUE†

PRIVACY IN THE AGE OF BIG DATA: A TIME FOR BIG DECISIONS

Omer Tene* & Jules Polonetsky**

We live in an age of "big data." Data has become the raw material of production, a new source of immense economic and social value. Advances in data mining and analytics and the massive increase in computing power and data storage capacity have expanded, by orders of magnitude, the scope of information available to businesses, government, and individuals.[1] In addition, the increasing number of people, devices, and sensors that are now connected by digital networks has revolutionized the ability to generate, communicate, share, and access data.[2] Data create enormous value for the global economy, driving innovation, productivity, efficiency, and growth. At the same time, the "data deluge" presents privacy concerns that could stir a regulatory backlash, dampening the data economy and stifling innovation.[3] In order to craft a balance be-

† The Privacy Paradox: Privacy and Its Conflicting Values.

* Associate Professor, College of Management Haim Striks School of Law, Israel; Senior Fellow, Future of Privacy Forum; Visiting Researcher, Berkeley Center for Law and Technology; Affiliate Scholar, Stanford Center for Internet and Society. I would like to thank the College of Management Haim Striks School of Law research fund and the College of Management Academic Studies research grant for supporting research for this Essay.

** Co-Chair and Director, Future of Privacy Forum.

1. *See, e.g.*, Kenneth Cukier, *Data, Data Everywhere*, ECONOMIST, Feb. 27, 2010, at 3-5, *available at* http://www.economist.com/node/15557443.

2. *See, e.g.*, Omer Tene, *Privacy: The New Generations*, 1 INT'L DATA PRIVACY LAW 15 (2011), *available at* http://idpl.oxfordjournals.org/content/1/1/15.full.

3. Consider, for example, the draft Regulation proposed on January 25, 2012, by the European Commission to replace the 1995 Data Protection Directive. It is poised to significantly increase sanctions, expand the geographical scope of the law, tighten requirements for explicit consent, and introduce a new "right to be forgotten." *See Commission Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data*

tween beneficial uses of data and the protection of individual privacy, policymakers must address some of the most fundamental concepts of privacy law, including the definition of "personally identifiable information," the role of consent, and the principles of purpose limitation and data minimization.

BIG DATA: BIG BENEFITS

The uses of big data can be transformative, and the possible uses of the data can be difficult to anticipate at the time of initial collection. For example, the discovery of Vioxx's adverse effects, which led to its withdrawal from the market, was made possible by the analysis of clinical and cost data collected by Kaiser Permanente, a California-based managed-care consortium. Had Kaiser Permanente not connected these clinical and cost data, researchers might not have been able to attribute 27,000 cardiac arrest deaths occurring between 1999 and 2003 to use of Vioxx.[4] Another oft-cited example is Google Flu Trends, a service that predicts and locates outbreaks of the flu by making use of information—aggregate search queries—not originally collected with this innovative application in mind.[5] Of course, early detection of disease, when followed by rapid response, can reduce the impact of both seasonal and pandemic influenza.

While a significant driver for research and innovation, the health sector is by no means the only arena for transformative data use. Another example is the "smart grid," which refers to the modernization of the current electrical grid to achieve a bidirectional flow of information and electricity. The smart grid is designed to allow electricity service providers, users, and other third parties to monitor and control electricity use. Some of the benefits accrue directly to consumers, who are able to reduce energy consumption by learning which devices and appliances consume the most energy, or which times of the day put the highest or lowest overall demand on the grid. Other benefits, such as accurately predicting energy demands to optimize renewable sources, are reaped by society at large.

Traffic management and control is another field witnessing significant data-driven environmental innovation. Governments around the world are establishing electronic toll pricing systems, which set forth differentiated payments based on mobility and congestion charges. Users pay depending on their

---

(General Data Protection Regulation), COM (2012) 11 final (Jan. 25, 2012), *available at* http://ec.europa.eu/justice/data-rotection/document/review2012/com_2012_11_en.pdf.

4. Rita Rubin, *How Did Vioxx Debacle Happen?,* USA TODAY, Oct. 12, 2004, at D1, *available at* http://www.usatoday.com/news/health/2004-10-12-vioxx-cover_x.htm.

5. *See Google Flu Trends: How Does This Work?*, GOOGLE, http://www.google.org/flutrends/about/how.html (last visited Jan. 25, 2012). Also consider Google Translate, which provides a free and highly useful statistical machine translation service capable of translating between roughly sixty languages by relying on algorithms and data freely available on the Web. *See Inside Google Translate*, GOOGLE, http://translate.google.com/about/intl/en_ALL/ (last visited Jan. 25, 2012).

use of vehicles and roads. These and other uses of data for traffic control enable governments to "potentially cut congestion and the emission of pollutants."[6]

Big data is also transforming the retail market. Indeed, Wal-Mart's inventory-management system, called Retail Link, pioneered the age of big data by enabling suppliers to see the exact number of their products on every shelf of every store at each precise moment in time. Many of us use Amazon's "Customers Who Bought This Also Bought" feature, prompting users to consider buying additional items selected by a collaborative filtering tool. Analytics can likewise be used in the offline environment to study customers' in-store behavior in order to improve store layout, product mix, and shelf positioning.

BIG DATA: BIG CONCERNS

The harvesting of large data sets and the use of analytics clearly implicate privacy concerns. The tasks of ensuring data security and protecting privacy become harder as information is multiplied and shared ever more widely around the world. Information regarding individuals' health, location, electricity use, and online activity is exposed to scrutiny, raising concerns about profiling, discrimination, exclusion, and loss of control. Traditionally, organizations used various methods of de-identification (anonymization, pseudonymization, encryption, key-coding, data sharding) to distance data from real identities and allow analysis to proceed while at the same time containing privacy concerns. Over the past few years, however, computer scientists have repeatedly shown that even anonymized data can often be re-identified and attributed to specific individuals.[7] In an influential law review article, Paul Ohm observed that "[r]eidentification science disrupts the privacy policy landscape by undermining the faith that we have placed in anonymization."[8] The implications for government and businesses can be stark, given that de-identification has become a key component of numerous business models, most notably in the contexts of health data (regarding clinical trials, for example), online behavioral advertising, and cloud computing.

---

6. MCKINSEY GLOBAL INST., BIG DATA: THE NEXT FRONTIER FOR INNOVATION, COMPETITION, AND PRODUCTIVITY 91-92 (2011), *available at* http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation.

7. This line of research was pioneered by Latanya Sweeney and made accessible to lawyers by Paul Ohm. *See* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010); Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, 2008 PROC. OF IEEE SYMP. ON SECURITY & PRIVACY 111; Latanya Sweeney, *Simple Demographics Often Identify People Uniquely* 2 (Carnegie Mellon Univ., Data Privacy Working Paper No. 3, 2000).

8. Ohm, *supra* note 7, at 1704.

We urge caution, however, when drawing conclusions from the re-identification debate. One possible conclusion, apparently supported by Ohm himself, is that all data should be treated as personally identifiable and subjected to the regulatory framework.[9] Yet such a result would create perverse incentives for organizations to abandon de-identification and therefore increase, rather than alleviate, privacy and data security risks.[10] A further pitfall is that with a vastly expanded definition of personally identifiable information, the privacy and data protection framework would become all but unworkable. The current framework, which is difficult enough to comply with and enforce in its existing scope, may well become unmanageable if it extends to any piece of information. Moreover, as Betsy Masiello and Alma Whitten have noted, while

> [a]nonym[ized] information will always carry some risk of re-identification . . . . [m]any of the most pressing privacy risks . . . exist only if there is certainty in re-identification, that is if the information can be authenticated. As uncertainty is introduced into the re-identification equation, we cannot know that the information truly corresponds to a particular individual; it becomes more anonymous as larger amounts of uncertainty are introduced.[11]

Most importantly, if information that is not ostensibly about individuals comes under full remit of privacy laws based on a possibility of it being linked to an individual at some point in time through some conceivable method, no matter how unlikely to be used, many beneficial uses of data would be severely curtailed. Such an approach presumes that a value judgment has been made in favor of individual control over highly beneficial uses of data, but it is doubtful that such a value choice *has* consciously been made. Thus, the seemingly technical discussion concerning the scope of information viewed as personally identifiable masks a fundamental normative question. Policymakers should engage with this normative question, consider which activities are socially acceptable, and spell out the default norms accordingly. In doing so, they should assess the value of data uses against potential privacy risks, examine the practicability of obtaining true and informed consent, and keep in mind the enforceability of restrictions on data flows.

---

9. *See id*. at 1742-43.

10. Ann Cavoukian & Khaled El Emam, Info. & Privacy Comm'r of Ont., Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy 7 (2011), *available at* http://www.ipc.on.ca/images/Resources/anonymization.pdf.

11. Betsy Masiello & Alma Whitten, *Engineering Privacy in an Age of Information Abundance*, 2010 AAAI Spring Symp. Series 119, 122, *available at* http://www.aaai.org/ocs/index.php/SSS/SSS10/paper/viewFile/1188/1497; *see also* Cynthia Dwork, *Differential Privacy*, 2006 INT'L COLLOQUIUM ON AUTOMATA, LANGUAGES AND PROGRAMMING pt. II, at 8, *available at* http://www.dbis.informatik.hu-berlin.de/fileadmin/lectures/SS2011/VL_Privacy/Differential_Privacy.pdf (introducing the concept of a privacy-preserving statistical database, which enables users to learn properties of a population while protecting the privacy of individuals).

OPT-IN OR OPT-OUT?

Privacy and data protection laws are premised on individual control over information and on principles such as data minimization and purpose limitation. Yet it is not clear that minimizing information collection is always a practical approach to privacy in the age of big data. The principles of privacy and data protection must be balanced against additional societal values such as public health, national security and law enforcement, environmental protection, and economic efficiency. A coherent framework would be based on a risk matrix, taking into account the value of different uses of data against the potential risks to individual autonomy and privacy. Where the benefits of prospective data use clearly outweigh privacy risks, the legitimacy of processing should be assumed even if individuals decline to consent. For example, web analytics—the measurement, collection, analysis, and reporting of internet data for purposes of understanding and optimizing web usage—creates rich value by ensuring that products and services can be improved to better serve consumers. Privacy risks are minimal, since analytics, if properly implemented, deals with statistical data, typically in de-identified form. Yet requiring online users to opt into analytics would no doubt severely curtail its application and use.

Policymakers must also address the role of consent in the privacy framework.[12] Currently, too many processing activities are premised on individual consent. Yet individuals are ill-placed to make responsible decisions about their personal data given, on the one hand, well-documented cognitive biases, and on the other hand the increasing complexity of the information ecosystem. For example, Alessandro Acquisti and his colleagues have shown that, simply by providing users a *feeling* of control, businesses encourage the sharing of data, regardless of whether or not a user has actually gained control.[13] Joseph Turow and others have shown that "[w]hen consumers see the term 'privacy policy,' they believe that their personal information will be protected in specific ways; in particular, they assume that a website that advertises a privacy policy will

12. See, in a different context, Omer Tene & Jules Polonetsky, *To Track or 'Do Not Track': Advancing Transparency and Individual Control in Online Behavioral Advertising,* 13 MINN. J. L. SCI. & TECH. (forthcoming 2012) (arguing that "[i]n the context of online privacy . . . emphasis should be placed less on notice and choice and more on implementing policy decisions with respect to the utility of given business practices and on organizational compliance with fair information principles"). *See also* Nicklas Lundblad & Betsy Masiello, *Opt-in Dystopias*, 7 SCRIPTed 155, 155 (2010), http://www.law.ed.ac.uk/ahrc/scripted/vol7-1/lundblad.asp (contending that while "[o]pt-in appears to be the optimal solution for anyone who believes consumers should have choice and control over their personal data collection… upon closer examination, it becomes clear that opt-in is a rhetorical straw-man that cannot really be implemented by regulatory policies without creating a number of unintended side effects, many of which are suboptimal for individual privacy").

13. Laura Brandimarte, Alessandro Acquisti & George Loewenstein, *Misplaced Confidences: Privacy and the Control Paradox,* (Sept. 2010) (unpublished manuscript), *available at* http://www.futureofprivacy.org/wp-content/uploads/2010/09/Misplaced-Confidences-acquisti-FPF.pdf.

not share their personal information."[14] In reality, however, "this is not the case."[15] Privacy policies often serve more as liability disclaimers for businesses than as assurances of privacy for consumers.

At the same time, collective action problems may generate a suboptimal equilibrium where individuals fail to opt into societally beneficial data processing in the hope of free riding on the goodwill of their peers. Consider, for example, internet browser crash reports, which very few users opt into, not so much because of real privacy concerns but rather due to a (misplaced) belief that others will do so instead. This phenomenon is evident in other contexts where the difference between opt-in and opt-out regimes is unambiguous, such as organ donation rates. In countries where organ donation is opt-in, donation rates tend to be very low compared to the rates in countries that are culturally similar but have an opt-out regime.[16] Finally, a consent-based regulatory model tends to be regressive, since individuals' expectations are based on existing perceptions. For example, if Facebook had not proactively launched its News Feed feature in 2006 and had instead waited for users to opt-in, we might not have benefitted from Facebook as we know it today. It was only after data started flowing that users became accustomed to the change.

We do not argue that individuals should *never* be asked to expressly authorize the use of their information or offered an option to opt out. Certainly, for many types of data collection and use, such as in the contexts of direct marketing, behavioral advertising, third-party data brokering, or location-based services, consent should be solicited or opt-out granted. But an increasing focus on express consent and data minimization, with little appreciation for the value of uses for data, could jeopardize innovation and beneficial societal advances. The question of the legitimacy of data use has always been intended to take into account additional values beyond privacy, as seen in the example of law enforcement, which has traditionally been allotted a degree of freedom to override privacy restrictions.

CONCLUSION

Privacy advocates and data regulators increasingly decry the era of big data as they observe the growing ubiquity of data collection and the increasingly ro-

14. Joseph Turow, Chris Jay Hoofnagle, Deirdre K. Mulligan, Nathaniel Good & Jens Grossklags, *The Federal Trade Commission and Consumer Privacy in the Coming Decade*, 3(3) I/S: J. L. & POL'Y FOR INFO. SOC'Y 723, 724 (2007).

15. *Id*.

16. Consider, for example, the donation rates in Sweden (opt-out) and Denmark (opt-in), 85.9% and 4.25% respectively; or in Austria (opt-out) and Germany (opt-in), 99.9% and 12% respectively. Notice, however, that additional factors besides individuals' willingness to participate, such as funding and regional organization, affect the ultimate conversion rate for organ transplants. Eric J. Johnson & Daniel Goldstein, *Do Defaults Save Lives?*, 302 SCIENCE 1338 (2003).

bust uses of data enabled by powerful processors and unlimited storage. Researchers, businesses, and entrepreneurs vehemently point to concrete or anticipated innovations that may be dependent on the default collection of large data sets. We call for the development of a model where the benefits of data for businesses and researchers are balanced against individual privacy rights. Such a model would help determine whether processing can be justified based on legitimate business interest or only subject to individual consent, and whether consent must be structured as opt-in or opt-out.