FPF
FUTURE OF PRIVACY FORUM

CIS
Stanford Law School
The Center for Internet and Society

BIG DATA AND PRIVACY: MAKING ENDS MEET

REVISITING THE 2000 STANFORD SYMPOSIUM IN LIGHT OF BIG DATA

William McGeveran[1]

On February 6, 2000, mere weeks into the 21st Century, a collection of the brightest minds considering the regulation of the digital world gathered at Stanford Law School to discuss a cutting-edge question: *Cyberspace and Privacy: A New Legal Paradigm?* Soon after, I purchased a copy of the *Stanford Law Revi*ew containing the writing that emerged from that symposium.[2] (How quaint! A bound volume, made of ink and paper!) Today this remarkable collection remains one of the most consulted books in my collection, printed or digital. Even that early in the internet era, the authors of those articles had already identified the outlines of the crucial issues that continue to occupy us today. (And, indeed, continue to occupy *them*, since almost all remain among the leading scholars specializing in internet-related topics).

Thirteen years later, questions about the emergence of a "new paradigm" often relate to "Big Data" methodologies – the analysis of huge data sets to search for informative patterns that might not have been derived from traditional hypothesis-driven research. Big Data burst into general public consciousness within the last year, and so did its implications for privacy. But the core practices of Big Data go back to 2000 and earlier, albeit at scales not quite as Big. By 2000, Google had already refined its search algorithm by analyzing huge numbers of users' queries. Transportation engineers already planned road improvements by running simulations based on numerous observations of real traffic patterns. Epidemiological research already relied on mass quantities of patient data,

[1] Associate Professor, Vance Opperman Research Scholar, University of Minnesota Law School.

[2] Symposium, *Cyberspace and Privacy: A New Legal Paradigm?*, 52 STAN. L. REV. 987 (2000).

including both health and demographic information. And, as demonstrated by Michael Froomkin's inventory of "privacy-destroying technologies" in the 2000 Symposium, we were already experiencing massive data collection and inevitable subsequent processing.[3]

Today's Symposium, cosponsored by Stanford once more, asks whether Big Data represents something entirely new for privacy. Well, leafing through the pages of the 2000 Stanford Symposium, one encounters all the same debates that are arising now in the context of Big Data – perhaps with a few twists, but still quite familiar. This brief essay offers some examples.

I have now heard a number of smart people suggest that treating personal information as a species of property would address many concerns about Big Data. After all, the insights gleaned from Big Data analysis are valuable. They think propertization would require those analyzing data to internalize privacy costs generated by their processing, give individuals leverage, or ensure that resulting windfalls are shared with the people whose information contributed to the profit. We have had this argument before. At the time of the 2000 Symposium, Pamela Samuelson aptly critiqued a portion of the privacy debate as "a quasi-religious war to resolve whether a person's interest in her personal data is a fundamental civil liberty or commodity interest."[4] Up to that point many commentators had similarly suggested that conceiving of personal information as one's property would be an attractive way to secure privacy. There is an initial attraction to the idea. But at the 2000 Symposium and soon thereafter, a growing scholarly consensus joined Samuelson in expressing great skepticism about that notion. [5]

Mixing property concepts with privacy concepts brought up doctrinal complications. To begin with, IP regimes such as copyright exist to encourage broad distribution of the underlying content, the very opposite purpose of privacy rules intended to limit the audience for information.[6] Further, complex

[3] A. Michael Froomkin, *The Death of Privacy?*, 52 STAN. L. REV. 1461, 1468-1501 (2000).

[4] Pamela Samuelson, *Privacy as Intellectual Property?*, 52 STAN. L. REV. 1125, 1157-58 (2000).

[5] *See* Julie E. Cohen, *Examined Lives: Informational Privacy and the Subject as Object*, 52 STAN. L. REV. 1373 (2000); Jerry Kang & Benedikt Buchner, *Privacy in Atlantis*, 18 HARV. J. L. & TECH. 229 (2004); Mark A. Lemley, *Private Property*, 52 STAN. L. REV. 1545 (2000); Jessica Litman, *Information Privacy/Information Property*, 52 STAN. L. REV. 1283 (2000); Samuelson, supra note 4; Paul M. Schwartz, *Internet Privacy and the State*, 32 CONN. L. REV. 815 (2000); Jonathan Zittrain, *What the Publisher Can Teach the Patient: Intellectual Property and Privacy in an Age of Trusted Privication*, 52 STAN. L. REV. 1201 (2000); *see also* William McGeveran, *Programmed Privacy Promises: P3P and Web Privacy Law*, 76 N.Y.U. L. REV. 1812, 1834-45 (2001) (applying this reasoning to the once-promising privacy-enhancing technology known as P3P).

[6] *See, e.g.*, Litman, supra note 5, at 1295-96.

adjustments to preserve speech interests and the public domain overwhelmed the simplicity of the property model.[7]

At a deeper theoretical level, it wasn't terribly clear what a property rationale really accomplished. The "quasi-religious" dispute often turned on framing without affecting substance. Certainly, as Julie Cohen pointed out in the 2000 Symposium and in much of her later work, the rhetoric of ownership has an effect. If we talk about Big Data organizations "buying" personal information from the willing sellers depicted by that information, we will enshrine assumptions about consent, knowledge, and utility that merit closer inspection.[8] But as a matter of legal design, merely calling an entitlement "property" does not make it any stronger. If the data subject can bargain the right away, all that really matters is the structure of that interaction – default rules, disclosure obligations, imputed duties. Regimes such as the European Union's data protection directive or the HIPAA privacy rules impose significant privacy obligations on data processing without calling the resulting individual rights "property." If I own my data but can sell it to a data miner (Big or Small) by clicking an "I agree" button at site registration, then what difference does that ownership make on the ground? I encourage those who would turn to ownership as the silver-bullet response to Big Data to read those 2000 Symposium articles first.

Another renewed debate that was already in full cry at the 2000 Symposium relates to technological protections. Big Data is made possible by rapid advances in computational power and digital storage capacity. Why not, smart people now ask, use these same features to ensure that downstream Big Data entities respect individuals' preferences about the use of their data? Ideas like persistent tagging of data with expiration dates or use restrictions are in vogue. Internet scholars such as Viktor Mayer-Schönberger and Jonathan Zittrain emphasize the importance of curtailing data permanence through a variety of measures including technological ones.[9] And developments like the European Union's deliberation over a "right to be forgotten" and California's "shine the light" law might create incentives to design Big Data mechanisms that allow individuals to inspect the personal data entities hold about them, and to delete it if they withdraw their consent for processing.

Unlike the propertization strategy, I think this approach has some potential merit, if it is backed by legal rules ensuring adoption and compliance. But nothing about Big Data makes any of these new concepts. Zittrain certainly

[7] *See, e.g.*, Lemley, supra note 5, at 1548-50.

[8] *See, e.g.*, JULIE E. COHEN, CONFIGURING THE NETWORKED SELF: LAW, CODE, AND THE PLAY OF EVERYDAY PRACTICE (2012); Cohen, supra note 5.

[9] *See* VIKTOR MAYER-SCHÖNBERGER, DELETE: THE VIRTUE OF FORGETTING IN THE DIGITAL AGE (2011); JONATHAN ZITTRAIN, THE FUTURE OF THE INTERNET--AND HOW TO STOP IT (2009).

recognizes this, because he was one of several speakers at the Symposium debating the potential of "trusted systems" to embed privacy protection in the architecture of data systems.[10] And Lawrence Lessig's notion that "code is law" was a centerpiece of the debate by 2000.[11] Proposals for trusted intermediaries or data brokers who handled information with a duty to protect the data subject's privacy interests were already in wide circulation by 2000 as well. These types of techno-architectural responses should be guided by history, such as the failure of P3P and the very slow uptake for other privacy-enhancing technologies, all discussed in the 2000 Symposium. As we already knew in 2000, technology can contribute greatly to addressing privacy problems, but cannot solve them on its own.

A third argument that has flared up with renewed vigor, fueled by Big Data, asks how much speech-related protection might apply to processing of data.[12] This discussion relates to new regulatory proposals, particularly those that advocate increased control at the processing and storage phases of data handling. These rules, it is said, contrast with the collection-focused rules that now dominate privacy law, especially in the US.

Once again, the seminal work was already happening in the 2000 Symposium. In his contribution, Eugene Volokh memorably characterized much of privacy law as "a right to stop people from speaking about you."[13] Others in the Symposium took up both sides of the argument.[14] The speech aspects of Big Data activities resemble very much the speech aspects of past data mining activities. While downstream regulation may be more attractive, there is still no real sea change in the dissemination of personal information. Neither its larger scale nor its lack of hypothesis should influence application of First Amendment

[10] *See* Zittrain, supra note 5; *see also* Henry T. Greely, *Trusted Systems and Medical Records: Lowering Expectations*, 52 STAN. L. REV. 1585 (2000); Jonathan Weinberg, *Hardware-Based ID, Rights Management, and Trusted Systems*, 52 STAN. L. REV. 1251 (2000).

[11] LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (1999). For another important rendition of this argument, see Joel R. Reidenberg, *Lex Informatica: The Formulation of Information Policy Rules Through Technology*, 76 TEX. L. REV. 553, 570 (1998). *See also* Jay P. Kesan and Rajiv C. Shah, *Shaping Code*, 18 HARV. J. L. & TECH. 319 (2005) (reviewing history of code-backed restrictions).

[12] *See, e.g.*, Jane Bambauer, *Is Data Speech?*, 66 STAN. L. REV. __ (forthcoming 2013); Neil M. Richards, *Data Privacy and the Right to be Forgotten after Sorrell* (working paper, June 6, 2013, on file with author).

[13] Eugene Volokh, *Freedom of Speech and Information Privacy: the Troubling Implications of a Right to Stop People From Speaking About You*, 52 STAN. L. REV. 1049 (2000).

[14] Compare Richard Epstein, *Privacy, Publication, and the First Amendment: The Dangers of First Amendment Exceptionalism*, 52 STAN. L. REV. 1003 (2000) with Cohen, supra note 5; Paul M. Schwartz, *Free Speech vs. Information Privacy: Eugene Volokh's First Amendment Jurisprudence*, 52 STAN. L. REV. 1559 (2000).

principles to Big Data. There is no more *speaking* in Big Data than there was in Medium-Sized Data, circa 2000.

Finally, some discussion of Big Data emphasizes that, by its nature, the subsequent processing of information is unpredictable. Smart people wonder what this means for the consent that was offered at the time of initial collection. If the purposes for which data would be used later could not be specified then, could there be true consent from the data subject? In the European Union, the answer to this question has long been: no. But for a long time now, the U.S. has embraced an increasingly farcical legal fiction that detailed disclosures to data subjects generated true informed consent. The empirical silliness of this notion was brought home by a recent study calculating that it would take the average person 76 work days to read every privacy policy that applied to her.[15]

Yet again, however, the 2000 Symposium already understood the disconnection between the complexities of data collection and processing and the cognitive abilities of an individual site user to offer meaningful consent.[16] Froomkin explained the economics of "privacy myopia," under which a consumer is unable to perceive the slow aggregation of information in a profile, and therefore its true privacy costs.[17] If Big Data processing might be even more remote, then it might induce even more myopia, but we would have the tools to analyze it from the 2000 Symposium.[18]

Each of these four debates – propertization, technological measures, speech protection, and privacy myopia – takes on new salience because of Big Data. But they are not fundamentally different from the brilliant deliberations at the 2000 Symposium. To see how they apply today one must substitute the names of some companies and update some technological assumptions. But these cosmetic changes don't compromise their theoretical core.

[15] *See* Aleecia M. McDonald & Lorrie Faith Cranor, *The Costs of Reading Privacy Policies*, 4 I/S J. OF L. & POLICY 540 (2008); Alexis C. Madrigal, *Reading the Privacy Policies You Encounter in a Year Would Take 76 Work Days*, THEATLANTIC.COM (March 1, 2012) at http://www.theatlantic.com/technology/archive/2012/03/reading-the-privacy-policies-you-encounter-in-a-year-would-take-76-work-days/253851/.

[16] *See* Cohen, supra note 5; Froomkin, supra note 3.

[17] Froomkin, supra note 3, at 1501-05; *see also* Neil Weinstock Netanel, *Cyberspace Self-Governance: A Skeptical View From Liberal Democratic Theory*, 88 CAL. L. REV. 395, 476-77 (2000) (discussing ignorance of data aggregation).

[18] Granted, Big Data may result in more decisions and assumptions about an individual to her detriment – such as price discrimination or insurance underwriting. If so, then most likely those decision processes ought to be regulated in themselves, through mechanisms modeled on the Fair Credit Reporting Act, 15 U.S.C. § 1681 et seq., or the Genetic Information Nondiscrimination Act, Pub. L. 110–233 (2008).

In the end, what is different about Big Data? Basically, that it is Big. The scale of information collected and processed is considerably greater. In addition, the ability to draw inferences from data has become steadily more sophisticated. So there is more data and it is more useful. But by 2000 we already surrendered vast quantities of personal information in our everyday life. It was already mined assiduously in search of insights both aggregate and personalized. We were already worried about all that, and already considering how to respond. I don't mean to suggest that the development of Big Data isn't important. I only emphasize that the ways to think about it, and the policy debates that it generates, have been around for a long time. The 2000 Symposium remains highly relevant today – and that kind of longevity itself proves the enduring value of the best privacy scholarship.