66 STAN. L. REV. ONLINE 81
September 3, 2013

BIG DATA IN SMALL HANDS

Woodrow Hartzog & Evan Selinger*

"Big data" can be defined as a problem-solving philosophy that leverages massive datasets and algorithmic analysis to extract "hidden information and surprising correlations."[1] Not only does big data pose a threat to traditional notions of privacy, but it also compromises socially shared information. This point remains underappreciated because our so-called public disclosures are not nearly as public as courts and policymakers have argued—at least, not yet. That is subject to change once big data becomes user friendly.

Most social disclosures and details of our everyday lives are meant to be known only to a select group of people.[2] Until now, technological constraints have favored that norm, limiting the circle of communication by imposing transaction costs—which can range from effort to money—onto prying eyes. Unfortunately, big data threatens to erode these structural protections, and the common law, which is the traditional legal regime for helping individuals seek redress for privacy harms, has some catching up to do.[3]

---

* Woodrow Hartzog is Assistant Professor, Cumberland School of Law, Samford University; Affiliate Scholar, Stanford Center for Internet and Society. Evan Selinger is Associate Professor of Philosophy, Rochester Institute of Technology; Fellow, Institute for Ethics and Emerging Technology.

1. Ira Rubinstein, *Big Data: The End of Privacy or a New Beginning?*, 3 INT'L DATA PRIVACY L. 65, 74 (2013). The term "big data" has no broadly accepted definition and has been defined many different ways. *See* VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 6 (2013) ("There is no rigorous definition of big data . . . . One way to think about the issue today . . . is this: big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value . . . .").

2. *See, e.g.*, Lior Jacob Strahilevitz, *A Social Networks Theory of Privacy*, 72 U. CHI. L. REV. 919 (2005).

3. *See* Patricia Sánchez Abril, *Recasting Privacy Torts in a Spaceless World*, 21 HARV. J.L. & TECH. 1, 19-20 (2007); Danielle Keats Citron, *Mainstreaming Privacy Torts*, 98 CALIF. L. REV. 1805, 1827 (2010); Andrew Jay McClurg, *Bringing Privacy Law Out of the Closet: A Tort Theory of Liability for Intrusions in Public Places*, 73 N.C. L. Rev. 989, 1057 (1995); Neil M. Richards, *The Limits of Tort Privacy*, 9 J. TELECOMM. & HIGH TECH. L. 357, 383 (2011); Neil M. Richards & Daniel J. Solove, *Prossers Privacy Law: A Mixed Legacy*, 98 CALIF. L. REV. 1887, 1889 (2010); Harry Surden, *Structural Rights in Privacy*, 60 SMU L. REV. 1605 (2007).

To make our case that the legal community is under-theorizing the effect big data will have on an individual's socialization and day-to-day activities, we will proceed in four steps.[4] First, we explain why big data presents a bigger threat to social relationships than privacy advocates acknowledge, and construct a vivid hypothetical case that illustrates how democratized big data can turn seemingly harmless disclosures into potent privacy problems. Second, we argue that the harm democratized big data can inflict is exacerbated by decreasing privacy protections of a special kind—ever-diminishing "obscurity." Third, we show how central common law concepts might be threatened by eroding obscurity and the resulting difficulty individuals have gauging whether social disclosures in a big data context will sow the seeds of forthcoming injury. Finally, we suggest that one way to stop big data from causing big, unredressed privacy problems is to update the common law with obscurity-sensitive considerations.

I. BIG, SOCIAL DATA

For good reason, the threat big data poses to social interaction has not been given its due. Privacy debates have primarily focused on the scale of big data and concentrations of power—what big corporations and big governments can do with large amounts of finely analyzed information. There are legitimate and pressing concerns here, which is why scholars and policymakers focus on Fair Information Practice Principles (FIPPs), deidentification techniques, sectoral legislation protecting particular datasets, and regulatory efforts to improve data security and safe international data transfers.[5]

---

4. A notable exception is Paul M. Schwartz and Daniel J. Solove's *Reworking Information Privacy Law: A Memorandum Regarding Future ALI Projects About Information Privacy Law* (Aug. 2012), http://law.duke.edu/sites/default/files/images/centers/judicialstudies/Reworking_Info_Privacy_Law.pdf. They write:

> People also expect "privacy by obscurity," that is, the ability to blend into a crowd or find other ways to be anonymous by default. This condition is rapidly disappearing, however, with new technologies that can capture images and audio nearly everywhere. As an example, facial recognition technology is constantly improving. Already, Facebook and Apple use technologies that permit the automatic tagging of photographs. One day devices, such as Google Glasses, may permit the identification of passing pedestrians on the street. In short, if the privacy torts are to be rethought, more guidance must be provided as to the underlying concept of privacy.

Id. at 11 (citations omitted).

5. *See, e.g.*, Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. Rev. 1701, 1776 (2010) ("Easy reidentification represents a sea change not only in technology but in our understanding of privacy."); Rubinstein, *supra* note 1, at 74; Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 256-57 (2013); Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data*, 64 STAN. L. REV. ONLINE 63 (2012); Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117 (2013); danah boyd, Address at the WWW2010 Conference: "Privacy and Publicity in the Context of Big Data" (Apr. 29, 2010), http://www.danah.org/papers/talks/2010/WWW2010.html. *But see* Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH. 1 (2011).

This trajectory fails to address the full scope of big data as a disruptive force in nearly every sector of the patchwork approach to privacy protection in the United States. Individuals eventually will be able to harness big datasets, tools, and techniques to expand dramatically the number and magnitude of privacy harms to themselves and others, perhaps without even realizing it.[6] This is problematic in an age when so many aspects of our social relationships with others are turned into data.

Consider web-scraping companies that dig up old mugshots and showcase them online, hoping embarrassed or anxious citizens will pay to have their images taken down. It isn't hard to imagine that the next generation of this business will cast a wider net, capitalizing on stockpiles of aggregated and filtered data derived from diverse public disclosures. Besides presenting new, unsettling detail about behavior and proclivities, they might even display predictive inferences couched within litigation-buttressing weasel wording—e.g., "correlations between X and Y have been known to indicate Z." Everyone, then, will be at greater risk of unintentionally leaking sensitive personal details. Everyone will be more susceptible to providing information that gets taken out of its original context, becomes integrated into a new profile, and subsequently harms a friend, family member, or colleague.

Inevitably, those extracting personal details from big data will argue that the information was always apparent and the law should not protect information that exists in plain sight.[7] The law has struggled with protecting privacy in public long before big data. However, we envision a tipping point occurring whereby some pro-publicity precedent appears more old than wise.

II. MORE DATA, LESS OBSCURITY

Socialization and related daily public disclosures have always been protected by varying layers of obscurity, a concept that we previously defined as follows:

> Obscurity is the idea that when information is hard to obtain or understand, it is, to some degree, safe. Safety, here, doesn't mean inaccessible. Competent and determined data hunters armed with the right tools can always find a way to get it. Less committed folks, however, experience great effort as a deterrent.
>
> Online, obscurity is created through a combination of factors. Being invisible to search engines increases obscurity. So does using privacy settings

6. Although we're focusing on how the law should respond to the dark side of big data, some see mastering quantitative legal prediction as essential to the future of entrepreneurial law firms and the law schools that train students to work in them. *See, e.g.*, Daniel Martin Katz, *Quantitative Legal Prediction—or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, 62 EMORY L.J. 909 (2013).

7. *See* MAYER-SCHÖNBERGER & CUKIER, *supra* note 1, at 29 (describing one instance of information discovered from big data analysis as "always apparent [as] [i]t existed in plain sight").

> and pseudonyms. Disclosing information in coded ways that only a limited audience will grasp enhances obscurity, too. Since few online disclosures are truly confidential or highly publicized, the lion's share of communication on the social web falls along the expansive continuum of obscurity: a range that runs from completely hidden to totally obvious.[8]

In the past, individuals have been able to roughly gauge whether aspects of their daily routines and personal disclosures of information would be safeguarded at any appropriate level of privacy protection by (sometimes implicitly) guessing the likelihood their information would be discovered or understood by third parties who have exploitative or undesirable interests. In the age of big data, however, the confidence level associated with privacy prognostication has decreased considerably, even when conscientious people exhibit due diligence.

Increasingly powerful and often secretive (proprietary and governmental) algorithms combined with numerous and massive datasets are eroding the structural and contextual protections that imposed high transactional costs on finding, understanding, and aggregating that information. Consumers got a taste of both the ease and power in which these processes can occur when Facebook rolled out Graph Search, denied it had privacy implications, then also revealed how readily what we "like" gets translated into who we are.

Maintaining obscurity will be even more difficult once big data tools, techniques, and datasets become further democratized and made available to the non-data-scientist masses for free or at low cost. Given recent technological trends, this outcome seems to be gradually approaching inevitability. At the touch of a button, Google's search engine can already unearth an immense amount of information that not too long ago took considerable effort to locate. Looking ahead, companies like Intel are not shy about letting the public know they believe "data democratization is a good bet."[9]

Decreasing confidence in our ability to judge the privacy value of disclosures puts us on a collision course for deepening the problem of "bounded rationality" and, relatedly, what Daniel Solove recognized as the problems of

---

8. Woodrow Hartzog & Evan Selinger, *Obscurity: A Better Way to Think About Your Data than 'Privacy*,' ATLANTIC (Jan. 17, 2013), http://www.theatlantic.com/technology/archive/2013/01/obscurity-a-better-way-to-think-about-your-data-than-privacy/267283 (explaining how obscurity is the proper conceptual framework for analyzing the privacy implications that follow from the introduction of Graph to Facebook's interface and analytics); *see also* Woodrow Hartzog & Frederic Stutzman, *The Case for Online Obscurity*, 101 CALIF. L. REV. 1 (2013) (identifying four key factors that define an obscurity continuum); Woodrow Hartzog & Frederic Stutzman, *Obscurity by Design*, 88 WASH. L. REV. 385 (2013) (explaining how obscurity considerations can enhance privacy by design efforts); Fred Stutzman & Woodrow Hartzog, *Boundary Regulation in Social Media*, *in* PROCEEDINGS OF THE ACM 2012 CONFERENCE ON COMPUTER SUPPORTED COOPERATIVE WORK 769 (2012), *available at* http://dl.acm.org/citation.cfm?id= 2145320&bnc=1 (observing that the creation of obscurity is part of the boundary regulation process of social media users).

9. *See* Jordan Novet, *Why Intel Thinks Data Democratization is a Good Bet*, GIGAOM (May 30, 2013), http://gigaom.com/2013/05/30/why-intel-thinks-data-democratization-is-a-good-bet.

scale, aggregation, and assessing harm.[10] It appears that the courts will need to grapple with a new wave of allegations of harms arising from behavior that yielded unintended and unforeseeable consequences.

As a thought experiment that crystalizes our guiding intuitions, consider a big data update to the problems that occurred when college students were revealed to be gay to their disapproving parents after a third party added them as members to Facebook's Queer Chorus group.[11] In the original instance, the salient tension was between how Facebook described its privacy settings and what users expected when utilizing the service. But what if someday a parent, teacher, or other authority figure wanted to take active steps to determine if their child, student, or employee was gay? Using democratized big data, a range of individually trivial, but collectively potent, information could be canvassed. Geolocation data conveyed when the child, or, crucially, his or her friends, used services like Foursquare combined with increasingly sophisticated analytical tools could lead to a quick transition from checking in to being outed. People-search services like Spokeo are well positioned to offer such user-friendly big data services.

III. THE COMMON LAW PRIVACY IMPLICATIONS OF BIG DATA FOR EVERYONE

Once big data is democratized and obscurity protections are further minimized, peer-to-peer interactions are poised to challenge many traditional common law concepts. Because the courts already make inconsistent rulings on matters pertaining to what reasonable expectations of privacy are, tort law is especially vulnerable.[12]

Here are a few of the fundamental questions we expect the courts will struggle to answer:

What Constitutes a Privacy Interest? A crucial question for both the tort of public disclosure of private facts and the tort of intrusion upon seclusion is whether the plaintiff had a privacy interest in a certain piece of information or

---

10. Daniel J. Solove, *Introduction: Privacy Self-Management and the Consent Dilemma*, 126 HARV. L. REV. 1879, 1888-93 (2013) ("The point is that it is virtually impossible for a person to make meaningful judgments about the costs and benefits of revealing certain data."); *see, e.g.*, Alessandro Acquisti & Jens Grossklags, *Privacy and Rationality: A Survey*, *in* PRIVACY AND TECHNOLOGIES OF IDENTITY: A CROSS-DISCIPLINARY CONVERSATION 15, 16 (Katherine R. Strandburg & Daniela Stan Raicu eds., 2006); Danielle Keats Citron, *Reservoirs of Danger: The Evolution of Public and Private Law at the Dawn of the Information Age*, 80 S. CAL. L. REV. 241 (2007); Paul M. Schwartz, *Privacy and Democracy in Cyberspace*, 52 VAND. L. REV. 1609, 1661 (1999) ("The difficulty with privacy-control in the Information Age is that individual self-determination is itself shaped by the processing of personal data.").

11. Geoffrey A. Fowler, *When the Most Personal Secrets Get Outed on Facebook*, WALL ST. J. (Oct. 13, 2012), http://online.wsj.com/article/SB10000872396390444165804578008740578200224.html.

12. *See, e.g.*, Strahilevitz, *supra* note 2, at 921.

context. This determination has varied wildly among the courts, and it is unclear how ubiquitous big data will alter this. For example, some courts have found that a privacy interest exists in involuntary exposure in public.[13] Other courts have found that overzealous surveillance in public that reveals confidential data can be seen to violate a privacy interest.[14] Will invasive "dataveillance" trigger the same protections?[15] Finally, courts have found, albeit inconsistently, a privacy interest in information known only to, and likely to stay within, a certain social group.[16] Does an increased likelihood that such information might be ascertained by outsiders destroy the privacy interest in information shared discreetly in small groups?[17]

What Actions Are Highly Offensive? Directly revealing or gaining access to certain kinds of information has been found to be highly offensive for purposes of the disclosure, intrusion, and false light torts.[18] In an age of predictions based upon data, would indirect disclosures of private information also be considered highly offensive? If not, does the law need to better articulate these limits? Does it matter if the eventual revelation of certain kinds of information that is highly offensive was predictable? Regarding the intrusion tort, can information gleaned from "public" big datasets ever be considered "secluded" and, if so,

13. *See, e.g.*, Daily Time Democrat v. Graham, 162 So. 2d 474, 478 (Ala. 1964).

14. *See, e.g.*, Nader v. Gen. Motors Corp., 255 N.E.2d 765, 771 (N.Y. 1970) ("[I]t is manifest that the mere observation of the plaintiff in a public place does not amount to an invasion of his privacy. But, under certain circumstances, surveillance may be so 'overzealous' as to render it actionable . . . . A person does not automatically make public everything he does merely by being in a public place."); Kramer v. Downey, 680 S.W.2d 524, 525 (Tex. App. 1984).

15. *See* Roger Clarke, ROGER CLARKE'S DATAVEILLANCE AND INFORMATION PRIVACY HOME-PAGE, http://www.rogerclarke.com/DV (last updated Jan. 6, 2013) (defining dataveillance as "the systematic use of personal data systems in the investigation or monitoring of the actions or communications of one or more persons"); *see also* Jerry Kang, *Information Privacy in Cyberspace Transactions*, 50 STAN. L. REV. 1193, 1261 (1998) (arguing that "information collection in cyberspace is more like surveillance than like casual observation"); Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393, 1417 (2001) ("Dataveillance is thus a new form of surveillance, a method of watching not through the eye or the camera, but by collecting facts and data."); Katherine J. Strandburg, *Freedom of Association in a Networked World: First Amendment Regulation of Relational Surveillance*, 49 B.C. L. REV. 741, 761 (2008) (observing that "[j]ust as "dataveillance" can chill an individual's experimentation with particular ideas or pastimes, relational surveillance can chill tentative associations and experimentation with various group identities").

16. *See, e.g.*, Y.G. v. Jewish Hosp., 795 S.W.2d 488 (Mo. Ct. App. 1990).

17. *See* Strahilevitz, *supra* note 2, at 922.

18. *See, e.g.*, Nappier v. Jefferson Standard Life Ins. Co., 322 F.2d 502 (4th Cir. 1963) (identity of a rape victim); Crippen v. Charter Southland Hosp. Inc., 534 So. 2d 286 (Ala. 1988) (confidential medical data); Taylor v. K.T.V.B., Inc., 525 P.2d 984 (Idaho 1974) (nude photos); Brents v. Morgan, 299 S.W. 967 (Ky. 1927) (debts); Reid v. Pierce Cnty., 961 P.2d 333 (Wash. 1998) (autopsy photos).

would using tools to unearth such data ever be considered highly offensive to a reasonable person?[19]

What Kinds of Disclosures Breach a Confidence? When has a confidant disclosed enough indirect information effectively to breach a confidence? If revealing a friend's location more than once a week allows others to determine that he is visiting a doctor for treatment of a communicable disease—a secret you promised to keep confidential—have you breached your promise? Courts would likely be hesitant to find a breach if the link between the disclosure and revealed confidential information were speculative, though inevitably some indirect disclosures will be so likely to compromise the confidentiality of other pieces of information so as to result in a *de facto* disclosure of the information itself. Should contracts with privacy-protective terms between individuals and small groups contemplate potential uses in big data? What lengths must confidants go to protect facts from being uncovered via big data techniques?

IV. REGULATING THE BIG IMPACT OF SMALL DECISIONS

Given the powerful debate over large-scale regulation of big data, safeguarding smaller, peer-to-peer interaction may prove to be the most feasible and significant privacy-related protection against big data.[20] The concept of

---

19. *See* Citron, *Mainstreaming Privacy Torts*, *supra* note 3, at 1827 ("[P]laintiffs probably cannot sue database operators for intrusion on seclusion under current case law. To prevail in an intrusion suit, a plaintiff must show that a defendant invaded his physical solitude or seclusion, such as by entering his home, in a manner that would be highly offensive to the reasonable person. Database operators and data brokers, however, never intrude upon a plaintiff's private space. They do not gather information directly from individuals and, to the extent that they do, the privacy problem involves the failure to secure personal information, not its collection.") (citations omitted). *But see* Jane Yakowitz Bambauer, *The New Intrusion*, 88 NOTRE DAME L. REV. 205, 207 (2012) ("Intrusion has great, untapped potential to address privacy harms created by advances in information technology. Though the tort is associated with conduct in real space, its principles apply just as well to operations in the era of Big Data."); Lyrissa Barnett Lidsky, *Prying, Spying, and Lying: Intrusive Newsgathering and What the Law Should Do About It*, 73 TUL. L. REV. 173, 227 (1998) ("[S]everal recent examples indicate that the average citizen's privacy is protected from media intrusions primarily by media disinterest, a tenuous basis at best for privacy protection."); McClurg, *supra* note 3, at 1057 ("The tort of intrusion can be redefined in a way that would allow recovery in suitable cases of public intrusion while also accommodating the competing interests of free social interaction and free speech."); Richards, *supra* note 3, at 383 ("[I]f we are interested in protecting against what we colloquially call 'invasions of privacy,' the intrusion model is a better fit with our intuitive linguistic understandings of that metaphor.").

20. For a skeptical view on the likelihood of significant regulation limiting how businesses mine data, see Lior Jacob Strahilevitz, *Toward a Positive Theory of Privacy Law*, 126 HARV. L. REV. 2010, 2033 (2013) ("The deck is stacked against restrictions on data mining."). *Cf.* Citron, *Reservoirs of Danger*, *supra* note 10, at 296 (asserting that, as a private law response to privacy harms, "[t]he contours of a negligence regime are simply too uncertain, and inherent problems with its enforcement undermines optimal deterrence," and proposing a strict-liability response instead); Sarah Ludington, *Reigning in the Data Traders: A Tort for the Misuse of Personal Information*, 66 MD. L. REV. 140, 146 (2006) (proposing a

obscurity might be useful in guiding the common law's evolution. If embraced as part of the disclosure and intrusion privacy torts, obscurity would allow socially shared information to fall within the ambit of "private facts" and "secluded" contexts. Contracts could also be used to protect the obscurity of individuals by targeting big data analysis designed to reveal socially shared but largely hidden information. Those charged with interpreting broad privacy-related terms should keep in mind structural and contextual protections that might have been relied upon by those whose privacy was to be protected.

Those forming the common law can now choose one of two paths. They can cling to increasingly ineffective and strained doctrines that were created when structural and contextual protections were sufficient for most of our socialization and obscure activities in public. Or they can recognize the debilitating effect big data has on an individual's ability to gauge whether social disclosures and public activity will later harm themselves and others, and evolve the common law to keep small acts of socialization and our day-to-day activities from becoming big problems.

---

tort to target "insecure data practices" and "the use of personal information data for purposes extraneous to the original transaction").