Big Data and the "New" Privacy Tradeoff

Robert H. Sloan[*] and Richard Warner[**]

Predictions of transformative change surround Big Data. It is routine to read, for example, that "with the coming of Big Data, we are going to be operating very much out of our old, familiar ballpark."[1] But, as both Niels Bohr and Yogi Berra are reputed to have observed, "Prediction is difficult, especially about the future." And, they might have added, especially regarding the effects of major technological change. In the Railroad Mania of nineteenth century England, for example, some made the typical prediction that a new communication network meant the end of an old one: namely, that that face-to-face communication over the emerging railroad network would entail a drastic drop in postal mail. In fact, mail volume increased.[2] Given the difficulty of forecasting transformative change, we opt for a "prediction" about the present: Big Data *already* presents a "new" and important privacy challenge. As the scare quotes indicate, the challenge is not truly new. What Big Data does is compel confrontation with a difficult trade-off problem that has been glossed over or even ignored up to now. It does so because both the potential benefits and risks from Big Data analysis are so much larger than anything we have seen before.

We confine our inquiry to the private sector. Governmental concerns are critically important, but they require separate treatment.

---

[*] Professor and Head, Department of Computer Science, University of Illinois at Chicago.
Partially supported by National Science Foundation Grant No. DGE-1069311.
[**] Professor of Law, Chicago-Kent College of Law, Visiting Foreign Professor, University of Gdańsk, Poland.

[1] Alex (Sandy) Pentland, *Reinventing Society in the Wake of Big Data*, EDGE, 2012, http://www.edge.org/conversation/reinventing-society-in-the-wake-of-big-data.
[2] Andrew Odlyzko, *The Volume and Value of Information*, 6 INT. J. COMMUN. 920, 925 (2012).

The Tradeoff Problem and Big Data

We claim Big Data greatly exacerbates a now decades old problem about how to balance the benefits of data collection and analysis against the relevant privacy risks. In the 1990s and early 2000s, before the current Big-Data era, commentators typically identified the following benefits of data collection: increased economic efficiency, improved security, better personalization of services, increased availability of relevant information, and innovative platforms for communication.[3] The tradeoff task was to balance that relatively short list of benefits against the loss of informational privacy. (By informational privacy, we mean the ability to control who collects information about you and what they do with it, and data collection and analysis reduces one's control.) Unfortunately, while privacy advocates and policy makers acknowledge tradeoff issues, they typically pay little attention to them.[4] Instead, they concentrate on the—also crucial—task of ensuring free and informed consent to businesses' data collection and use practices. Big Data compels a change: it involves such large and important risks *and* benefits that there is no longer any excuse for setting tradeoff issues aside.

"Big Data" refers to the acquisition and analysis of massive collections of information, collections so large that until recently the technology needed to analyze them did not exist.[5] The analysis can reveal patterns that would otherwise go unnoticed, and this has already yielded an astonishing array of benefits from

[3] *See, e.g.*, Jerry Kang, *Information Privacy in Cyberspace Transactions*, 50 STAN. L. REV. 1193–1294 (1998) (emphasizing availability of relevant information, increased economic efficiency, improved security).
[4] *See* Fred Cate, *The Failure of Fair Information Practice Principles*, *in* THE FAILURE OF FAIR INFORMATION PRACTICE PRINCIPLES 342, 361–367 (Jane Winn ed., 2006).
[5] Omer Tene & Jules Polonetsky, *Privacy In The Age Of Big Data: A Time For Big Decision*, 64 STAN. L. REV. ONLINE 63 (2012).

detecting drug interactions to improving access to social services in India by creating digital IDs for citizens.[6] The risks are equally serious. The risk of a massive loss of informational privacy has become much larger, and there are other risks as well. Consider improving access to social services in India. A significant improvement will increase the demand for the services. Meeting that demand may require an increased investment in those services, thereby creating at least two risks: social discontent if the demand is not met; and, the diversion of scarce resources from other critical areas if it is. An acceptable level of information flow into Big Data analysis is one that yields acceptable tradeoffs between risks and benefits. The problem is to find a level of information flow that does that. The current mechanisms for determining the proper level are woefully inadequate.

Mid-20th Century Information Processing

To see why, it helps to turn back the clock to the mid-twentieth century. Data collection was in its infancy, with only the beginnings of credit reporting practices. Direct marketing was not widely used until the 1970s because prior to that time it was too difficult to differentiate among consumers (the change came when the government began selling census data on magnetic tapes).[7] People did disclose information to businesses, governmental and private licensing agencies, and so on, but the information was typically stored in paper records and geographically scattered. There was no convenient way to search all of it or to retrieve readily storable, reusable information. You could by and large regulate the

[6] RICK SMOLAN & JENNIFER ERWITT, THE HUMAN FACE OF BIG DATA 34 (2012).
[7] DANIEL J. SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 18 (2004).

flow of your information to private businesses in the way you thought best. The sum of the individual decisions about data collection provided the answer to how much information should flow to businesses for analysis.

Did this yield an acceptable level of information flow? The answer did not matter much because mid-twentieth century information processing did not generate significant risks and benefits compared to today, but, in general, summing individual decisions is not a good way to answer "acceptable level" tradeoff questions, as the following example illustrates.[8] Imagine that in a community that does not have a telephone book, everyone would like to have one. However, each person prefers not to have his or her phone number listed and so refuses to consent to listing. No phone book is the result—a result each regards as much worse than having one.

Unfortunately, society has not yet—in the opening decades of the twenty-first century—changed its ways. Summing individual decisions still plays a key role in answering the "acceptable level" question. Summing individual decisions works extremely well for setting prices in highly competitive markets with no externalities, but can work very poorly indeed when results of individual decisions come with significant externalities. For Big Data today, there are tremendous externalities: Decisions by individual consumers to withhold data may have large negative externalities for society's overall ability to reap the benefits of Big Data, and decisions by individual businesses may have large negative externalities for citizens' privacy.

[8] Amartya Sen, *Social Choice*, *in* THE NEW PALGRAVE DICTIONARY OF ECONOMICS (2nd ed. 2008), http://www.dictionaryofeconomics.com/dictionary.

The Current Mechanism for Summing Individual Decisions

Outside the health and finance sectors, private businesses are relatively unconstrained in their data collection and analysis practices, and summing individual decisions still plays a key role in determining the level of information that flows to private businesses. We focus on the online context, but similar remarks hold for offline situations. Online, the current summing mechanism is Notice and Choice (sometimes called Notice and Consent). The "notice" is a presentation of terms. The "choice" is an action signifying acceptance of the terms (typically using a website or clicking on an "I agree" button). Implementations of Notice and Choice lie along a spectrum. One extreme is home to implementations that place few restrictions on Notices (how they are presented and what they may or must say) and few restrictions on what counts as choice (using the site, clicking on an "I agree" button); the other extreme is occupied by restrictive implementations requiring conformity to some or all of the Fair Information Practice Principles of transparency, error correction, restriction of use of data to purposes stated at the time of collection, deletion of data when it is no longer used for that purpose, and data security.

Proponents of Notice and Choice make two claims. First: when adequately implemented, (the appropriate version of) Notice and Choice ensures that website visitors can give free and informed consent to businesses' data collection and use practices. For purposes of this essay, we grant the first claim.[9] Our concern is with the second claim: namely, that the sum of the individual consent decisions determines an acceptable level of information flowing to businesses. We see little

[9] We criticize and reject the claim in Robert H. Sloan & Richard Warner, *Beyond notice and Choice: Privacy, Norms, and Consent*, ___ SUFFOLK UNIV. J. HIGH TECHNOL. LAW ___ (2014).

reason to think it is true. As the telephone book example illustrates, summing individual decisions can lead to information flows that are inconsistent with what the individuals making those decisions would collectively agree is good overall. We believe Notice and Choice will not yield results good for society as a whole. In all its versions, Notice and Choice leaves tradeoff issues largely to the discretion of private business.[10] The Notices under which they collect consumers' information leave the subsequent uses of that information largely up to the businesses. By way of illustration, consider one well-known example. Microsoft allowed Dr. Russ Altman to analyze Bing searches for search terms correlated with dangerously high blood sugar levels. This was a key step in Altman's confirming that the antidepressant Paxil together with the anti-cholesterol drug Pravachol could result in diabetic blood sugar levels.[11] Our point is that the decision about how to use the Bing searches was *Microsoft's*. The Altman result is a life-saving one, but not all uses of Big Data are so uncontroversially good. Target, for example, infamously uses Big Data analysis to predict which of their customers are pregnant,[12] and it would be remarkable if decisions by businesses about data use reliably yielded acceptable society-wide balances of risks and benefits. Each business will balance in ways that serve its business goals, and there is no reason to think that summing up business decisions will yield an acceptable balance of risks and benefits from the point of view of society as a whole. This is just the "summing" problem over again with

[10] The point is widely accepted. We give our reasons for it in Richard Warner & Robert H Sloan, *Behavioral Advertising: From One-Sided Chicken to Informational Norms*, VANDERBILT ENTERTAIN. TECHNOL. LAW J. 15 (2012).
[11] See Peter Jaret, *Mining Electronic Records for Revealing Health Data*, NEW YORK TIMES, January 14, 2013, http://www.nytimes.com/2013/01/15/health/mining-electronic-records-for-revealing-health-data.html?pagewanted=all.
[12] ERIC SIEGEL, PREDICTIVE ANALYTICS: THE POWER TO PREDICT WHO WILL CLICK, BUY, LIE, OR DIE Kindle Locations 1368–1376 (Kindle Edition ed. 2013).

businesses making the decisions instead of consumers. Since the businesses do not suffer any of the negative effects on consumers of the loss of informational privacy, they will undervalue consumers' interests and reach an unacceptably biased overall tradeoff.

The Not-New-But-Now-More-Difficult-and-Important Problem

Is there a way to balance risks and benefits that reliably yields acceptable results? We will not answer that question here.[13] Our point is that this problem is not new, but that Big Data does make it both considerably more difficult and considerably more important. We can certainly no longer reasonably rely on an approach that was acceptable in the mid-twentieth century only because back then information processing created relatively small benefits and risks.

[13] We offer a partial answer in ROBERT H. SLOAN & RICHARD WARNER, UNAUTHORIZED ACCESS: THE CRISIS IN ONLINE PRIVACY AND INFORMATION SECURITY (2013).