Big Data: A Pretty Good Privacy Solution

Ira S. Rubinstein[*]

Introduction

Big data—by which I mean the use of machine learning, statistical analysis, and other data mining techniques to extract hidden information and surprising correlations from very large and diverse data sets—raises numerous privacy concerns. A growing number of privacy scholars (myself included) have argued that big data casts doubt on the Fair Information Practices ('FIPs'), which form the basis of all modern privacy law.[1] With the advent of big data, the FIPs seem increasingly anachronistic for three reasons. First, big data heralds the shift from data actively collected with user awareness and participation to machine-to-machine transactions (think of electronic toll-collection systems) and passive collection (data collected as a by-product of other activities like searching or browsing the web).[2] Thus, big data nullifies informed choice, undermining the FIPs at their core. Second, big data thrives on comingling and sharing large data sets to create economic value and innovation from new and unexpected uses, making it inimical to collection, purpose, use or retention limitations, without which the FIPs are toothless. Finally, big data seems to make anonymization impossible. Why? The amount of data available for analysis has increased exponentially and while much of it seems non-personal, researchers have shown that almost any attribute, when combined with publicly available background information, can be linked back to an individual.[3] There is a large and growing literature on whether anonymization is no longer an effective strategy for protecting privacy[4] and to what extent this failure makes it impossible to publicly release data that is both private and useful.[5]

This indictment of the FIPs paints big data with a broad brush. And yet a moment's thought suggests that not every big data scenario is necessarily alike or poses the same risk to privacy. Having reviewed dozens of big-data analyses culled from the lay literature, I want to explore whether they have distinguishing characteristics that would allow us to categorize them as having a low, medium, or high risk of privacy violations.[6] In what follows, I offer a tentative and preliminary categorization of big data scenarios and their varying levels of risks. And I emphasize two supplemental FIPs that may help address some (but not all) of the riskier scenarios: first, a default prohibition on the transfer

[*] Senior Fellow and Adjunct Professor of Law, Information Law Institute, New York University School of Law.

[1] *See* Omer Tene & Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics*, 11 NW. J. TECH. & INTELL. PROP. 239, 257-63 (2013); Ira S. Rubinstein*, Big Data*: *The End of Privacy or a New Beginning?* 3 INT'L DATA PRIV. L. 74, 78 (2012).

[2] *See* World Economic Forum, *Unlocking the Value of Personal Data: From Collection to Usage* 7-8 (2013), http://www.weforum.org/reports/unlocking-value-personal-data-collection-usage.

[3] *See* Arvind Narayanan & Vitaly Shmatikov, Robust De-anonymization of Large Sparse Datasets, 2008 PROC. 29TH IEEE SYMP. ON SECURITY & PRIVACY 111.

[4] *Compare* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010) *with* Jane Yakowitz, *Tragedy of the Data Commons* 25 HARV. J. L. & TECH. 1 (2011).

[5] *See* Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. (forthcoming 2013).

[6] This paper confines itself to consumer and research scenarios and does not address government data mining.

of large data sets to third parties for secondary uses without the explicit, opt-in consent of the data subject; and, second, a broad prohibition on the re-identification of anonymized data, with violators subject to civil and/or criminal sanctions. This approach is partial and imperfect at best but perhaps offers a pretty good privacy solution for the moment.

Discussion

In a recent book explaining big data for the lay reader, Viktor Mayer-Schönberger and Kenneth Cukier describe dozens of scenarios in which big data analytics extract new insights.[7] Several of these scenarios are low-risk and raise no or minimal privacy alarms. As they observe, "Sensor data from refineries does not [contain personal information], nor does machine data from factory floors or data on manhole explosions or airport weather."[8] What about services using billions of flight-price records to predict the direction of prices on specific airline routes or popular web services using billions of text or voice samples and "machine learning" algorithms to develop highly accurate spam filters, grammar and spell checkers, and translation and voice recognition tools? These scenarios are low risk for several reasons: they mainly involve first-party collection and analysis of non-personal or de-identified data, they seek to improve or enhance devices or systems that affect consumers rather than specific individuals, and they involve either very limited or pre-defined data sets that are not shared with others. And the services have little incentive to re-identify individuals; indeed, they may have made binding promises to safeguard data security.

If other risk factors are present, however, first party collection and analysis of limited data sets may be more troubling. Medium-risk scenarios occur when (1) the data is personal and/or the first party contemplates (2) sharing the data with a third party for secondary uses or (3) a broad or public data release. And yet it is possible to reduce the privacy risks in each of these cases.

A good example of (1) is Google Flu Trends, which uses search engine query data and complex models for the early detection of flu epidemics. Although search queries are IP-based and therefore identifiable, Google safeguards privacy by aggregating historical search logs and discarding information about the identity of every user.[9]

A good example of (2) is any smart meter system subject to California's SB 1476, a recently-enacted privacy law that "requires aggregators of energy consumption data to obtain consumer consent before sharing customer information with third parties; mandates that third parties may only have access to such data when they are contracting with the utility to provide energy management-related services; stipulates that data be kept secure from unauthorized parties; and mandates that electricity ratepayers opt in to authorize any sharing of their energy consumption data for any secondary commercial purpose[s]."[10] Absent such protections, utilities might be tempted to sell consumption data for analysis and secondary use by third parties for marketing purposes or to determine insurance risk. SB 1476 permits first party data analysis for operational purposes that

[7] *See* VIKTOR MAYER-SCHÖNBERGER AND KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK (2013). All of the big data scenarios discussed below are drawn from this book unless otherwise noted.

[8] *Id.* at 152.

[9] *See* Jeremy Ginsberg, et al., *Detecting Influenza Epidemics Using Search Engine Query Dat*a 457 NATURE 1012 (2009).

[10] *See* John R. Forbush, *Regulating the Use and Sharing of Energy Consumption Data: Assessing California's SB 1476 Smart Meter Privacy Statute*, 75 ALB. L. REV. 341, 343 (2012).

benefit both consumers and society while also addressing the risks associated with third party sharing for secondary uses.

A good example of (3) is using anonymized geo-location data derived from GPS-equipped devices to optimize public transit systems. The analysis relied on a research challenge dubbed "Data for Development" in which the French telecom Orange "released 2.5 billion call records from five million cell-phone users in Ivory Coast. … The data release is the largest of its kind ever done. The records were cleaned to prevent anyone identifying the users, but they still include useful information about these users' movements."[11] Locational data is highly sensitive and it has proven very difficult to achieve anonymization by removing identifiers from mobility datasets.[12] However, the researchers who gained access to the Orange data set had to be affiliated with a public or private research institution, submit a research proposal for approval, and sign a data-sharing agreement.[13] These agreements typically prohibit re-identification of the data subject and impose additional security and privacy safeguards such as audits, privacy impact assessments, and data destruction upon completion of the research.[14] This contractual approach seems to finesse the "de-identification dilemma"[15] by avoiding both Ohm's Scylla (that anonymized data sets lack either privacy or utility) and Yakowitz's Charybdis (that all useful research requires the public release of anonymized data sets).[16]

High-risk scenarios occur whenever big data analytics result in actions taken regarding groups with sensitive attributes or affecting specific individuals. Mayer-Schönberger and Cukier provide several relevant examples such as startups that would determine a consumer's credit rating based on "behavioral scoring" using rich social media data sets not regulated by fair credit reporting laws; insurance firms that would identify health risks by combining credit scores with various lifestyle data not regulated by any privacy laws; and the notorious Target incident, in which the firm used big data analytics to predict whether female shoppers were newly pregnant and then marketed baby-related products to them, even though they may have delayed sharing this news with family members.[17] Why are these high-risk scenarios? First, the data sets are large and heterogeneous, increasing the likelihood that analysis will reveal sensitive or intimate attributes, even though we think of the underlying data as non-personal. Second, the data comes from multiple sources, so individuals are unaware of how third parties collect, store or use it and therefore lack any ability to access their

[11] *See* David Talbot, *African Bus Routes Redrawn Using Cell-Phone Data*, MIT TECH. REV. (Apr. 30, 2013).

[12] *See* Y.-A. de Montjoye, et al., *Unique in the Crowd: The Privacy Bounds of Human Mobility*, SCIENTIFIC REPORTS 3 (March 25, 2013), http://www.readcube.com/articles/10.1038%2Fsrep01376. However, efforts are underway to make large-scale mobility models provably private without unduly sacrificing data accuracy using new techniques based on differential privacy; *see* Darakshan J. Mir, et al., *Differentially Private Modeling of Human Mobility at Metropolitan Scale* (2013) (unpublished paper on file with the author).

[13] *See* D4D Challenge, Learn More, http://www.d4d.orange.com/learn-more (last visited June 25, 2013).

[14] *See* Khaled El Emam, *Risk-Based De-Identification of Health Data,* IEEE SEC & PRIV, May-June 2010, at 66, 64-67. Both Ohm, *supra* note 4 at 1770, and Yakowitz, *supra* note 4 at 48-49, endorse penalizing improper re-identification.

[15] *See* Robert Gellman, *The Deidentification Dilemma: A Legislative and Contractual Proposal*, 21 FORDHAM INTELL. PROP. MEDIA & ENT. L. J. 33 (2010).

[16] Dozens of papers describing presumably valuable research results from the D4D Challenge were presented at the 2013 NetMob conference at MIT, *available at* http://perso.uclouvain.be/vincent.blondel/netmob/2013/.

[17] For a fascinating and detailed account, *see* Charles Duhigg, *How Companies Learn Your Secrets*, NY TIMES (Feb. 16, 2012), *available at* http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all.

data or control detrimental uses of inferred attributes. Third, when firms rely on big data analytics to infer sensitive attributes (creditworthiness, insurability, pregnancy), they often skirt regulations limited to the collection and use of specific types of personal data. Another problem is that these analytic techniques are imperfect and may result in erroneous or unfair decisions.[18] In any case, the underlying privacy issues in high-risk scenarios are far more difficult to address: at a minimum, they require stronger default rules and perhaps a major shift in business models and new and innovative data frameworks.[19]

Conclusion

This short essay seeks to characterize big data scenarios according to their level of privacy risks and to identify supplemental FIPs that might help mitigate these risks. Whether this approach is worthwhile requires further study of many more scenarios and development of a more comprehensive set of risk criteria and supplemental privacy principles. A risk-based approach is at best a compromise. Yet is has the virtue of acknowledging that while the anonymous release of useful data is no silver bullet for privacy, neither is big data in all cases a poison pill.

[18] *See* Kate Crawford & Jason Schultz, *The Due Process Dilemma: Big Data and Predictive Privacy Harms* (2013) (unpublished paper on file with the author).

[19] The World Economic Forum has published several reports championing a new business model based on personal data stores (PDS); *see* http://www.weforum.org/issues/rethinking-personal-data (last visited June 25, 2013). For a privacy-protective implementation of PDS, *see* Y.-A. de Montjoye, et al., *On the Trusted Use of Large-Personal Data*, 35 IEEE DATA ENG. BULL. 5 (2012).