TECHNOLOGY POLICY INSTITUTE
Studying the Global Information Economy

The Big Data Revolution: Privacy Considerations

December 2013

Thomas M. Lenard and Paul H. Rubin

1099 NEW YORK AVE, NW ▫ SUITE 520 ▫ WASHINGTON, DC 20001
PHONE: 202.828.4405 ▫ E-MAIL: info@techpolicyinstitute.org ▫ WEB: www.techpolicyinstitute.org

THE BIG DATA REVOLUTION: PRIVACY CONSIDERATIONS

Thomas M. Lenard and Paul H. Rubin*

I. Introduction

The Information Technology revolution has produced a data revolution—now commonly referred to as "big data"—in which massive amounts of data can be collected, stored and analyzed at relatively low cost. This data revolution is based on the flow of new digital data, which has grown from an estimated 0.6 to 2.1 exabytes in 2000 to 2,700 exabytes in 2012, as shown in Figure 1. About one-third of the data collected globally is estimated to originate in the United States.[1]

While one may be skeptical of the hype surrounding the big data revolution, it clearly creates the potential for significant innovation in specific sectors as well as the overall economy. Reports by the World Economic Forum, McKinsey Global Institute and others describe the potential benefits for health care, government services, fraud protection, retailing, manufacturing and other sectors. McKinsey estimates that big data and analytics could yield benefits for health care alone of more than $300 billion annually. Gains for the overall economy could be up to $610 billion in annual productivity and cost savings.[2]

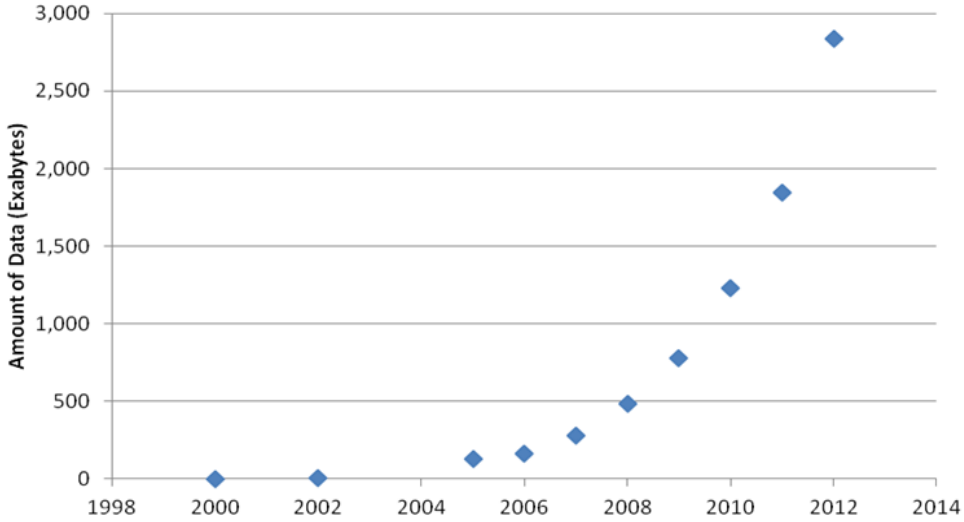
---

* Lenard is President and Senior Fellow at the Technology Policy Institute. Rubin is Samuel Candler Dobbs Professor of economics at Emory University and Senior Fellow at TPI. The authors thank Arlene Holen, Amy Smorodin and Scott Wallsten for helpful comments, and Corwin Rhyan for outstanding research assistance.

[1] IDC, sponsored by EMC, *The Digital Universe in 2020: Big Data, Bigger Data Shadows, and Biggest Growth in the Far East*, December, 2012.

[2] McKinsey Global Institute, *Game changers: Five opportunities for US growth and renewal*, July, 2013, p. 66.

Figure 1: Digital Data Created Annually Worldwide



Source: *Digital Universe Reports*, IDC

The emergence of big data also has raised privacy concerns on the part of advocates and government officials. Much of the concern relates to the collection and use of data by governments for national security purposes, an issue we do not address here. Significant concerns have also been expressed about the commercial and other non-surveillance uses of big data. Edith Ramirez devoted her first major speech on privacy as Federal Trade Commission Chairwoman to big data, arguing that "the challenges [big data] poses to privacy are familiar, even though they may be of a magnitude we have yet to see."[3] She added that "the solutions are also familiar, [a]nd, with the advent of big data, they are now more important than ever."

Chairwoman Ramirez's speech raises the question of whether big data are associated with new privacy harms and a concomitant increase in the need for government action. It also suggests that we should look to the "familiar solutions"—the Fair Information Privacy Practices (FIPPs)

[3] Edith Ramirez, "The Privacy Challenges of Big Data: A View from the Lifeguard's Chair", Speech at Technology Policy Institute's Aspen Forum, August, 2013, accessed at http://ftc.gov/speeches/ramirez.shtm.

involving notice and choice, use specification and limits, and data minimization—to solve any privacy problems brought about by big data.

To address these issues, this paper focuses on the following questions that, while not new, have become more salient in the world of big data:

- How should we think about the "reuse" of data—i.e., the use of data for purposes not initially identified or even envisioned?
- Similarly, how should we think about the combined use of data from different sources?
- What are the implications of big data for data security—data breaches and identity fraud?
- What are the implications of big data for profiling individuals and using algorithms to draw inferences for purposes ranging from marketing to credit and employment decisions?

We conclude that there is no evidence at this stage that the use of big data for commercial and other non-surveillance purposes has caused privacy harms. Moreover, the "familiar solutions" associated with the FIPPs are a potentially serious barrier to much of the innovation we hope to see from the big data revolution.

II. Defining Big Data

Although the term is in widespread use, "there is no rigorous definition of big data."[4] McKinsey defines big data as referring to "datasets whose size is beyond the ability of typical database

[4] Mayer-Schönberger and Cukier, "Big Data: a revolution that will transform how we live, work and think", Houghton Mifflin Harcourt, 2013, p. 6.

software tools to capture, store, manage and analyze."[5] Mayer-Schönberger and Cukier, in their recent book on big data, focus on what the data can produce: "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, in ways that change markets, organizations, the relationship between citizens and governments, and more."[6] They focus on the ability of large data sets to yield correlations between variables that can provide important public and private benefits.

This point is echoed by Einav and Levin in a recent paper discussing the potentially revolutionary effects of data on economic analysis and policymaking. Big data's potential comes from "the identification of novel patterns in behavior or activity, and the development of predictive models, that would have been hard or impossible with smaller samples, fewer variables, or more aggregation."[7] Data are now available in real time, at larger scale, with less structure, and on different types of variables than previously.[8]

III. Examples: The Benefits of Unanticipated Uses of Big Data

One of the "familiar solutions" long promoted by privacy advocates is that data should only be collected for an identified purpose. This is reflected in the FIPPs dating back to the 1970s, the OECD Privacy Principles of 1980, current European Union regulations, and the

---

[5] McKinsey Global Institute, *Big Data: The Next Frontier for Innovation, Competition and Productivity*, June, 2011 p. 1.

[6] Mayer-Schönberger and Cukier, p. 6.

[7] Liran Einav and Jonathan Levin, "The Data Revolution and Economic Analysis", Prepared for the NBER Innovation Policy and the Economy Conference, April, 2013, p. 2.

[8] Einav and Levin, pp. 5-6.

recommendations of the FTC's 2012 Privacy Report.[9] Indeed, according to Chairwoman Ramirez, the First Commandment of data hygiene is: "Thou shall not collect and hold onto personal information unnecessary to an identified purpose."[10] Similarly, Commissioner Julie Brill laments the fact that firms, "without our knowledge or consent, can amass large amounts of private information about people to use for purposes we don't expect or understand."[11]

Chairwoman Ramirez's First Commandment is particularly ill-suited to the world of big data and, in fact, is inconsistent with other parts of her speech where she points out beneficial uses of big data, such as: improving the quality of health care while cutting costs, making more precise weather forecasts, forecasting peak electricity consumption, and delivering better products and services to consumers at lower costs.[12] These beneficial uses often involve using medical data, utility billing records and other data for purposes other than those for which they were initially collected.

Moreover, the government itself routinely violates the data-hygiene First Commandment. When people paid their taxes, for example, they did not know that data from their returns would later be used to determine their eligibility for health insurance subsidies. Individuals could not have been informed of that potential use, which was only recently envisioned.

---

[9] An excellent summary of the evolution of the FIPPs comes from Robert Gellman, "FAIR INFORMATION PRACTICES: A Basic History", last updated October 8, 2013, available at http://www.bobgellman.com/rg-docs/rg-FIPShistory.pdf, and the current FTC FIPPs are posted at http://www.ftc.gov/reports/privacy3/fairinfo.shtm.

[10] Ramirez, p. 4.

[11] Julie Brill, "Demanding transparency from data brokers", *Washington Post Opinions*, August 15, 2013.

[12] Ramirez, p.1.

Because big data analysis involves finding correlations and patterns that might otherwise not be observable, it almost necessarily involves uses of data that were not anticipated at the time the data were collected. Mayer-Schönberger and Cukier emphasize that "in a big-data age, most innovative secondary uses haven't been imagined when the data is first collected." They add, "[w]ith big data, the value of information no longer resides solely in its primary purpose. As we've argued, it is now in secondary uses."[13]

The poster child for big data is Google Flu. Testing 450 million models, researchers identified 45 search terms that could predict the spread of flu more rapidly than the Centers for Disease Control, which relies on physicians' reports.[14] By tracking the rate at which the public searched for terms like "flu" and "cough medicine" using Google, an outbreak of influenza could be spotted a week or two ahead of CDC reports.[15] Using data from internet searches for a service such as Google Flu was not and could not be envisioned when these data were collected.

The serendipitous use of data is not, however, a new phenomenon or confined to the digital era. Mayer-Schönberger and Cukier give the example of Commander Mathew Maury, who, in the middle of the 19th century, used data from logbooks of past voyages to devise more efficient routes and mapped out the shipping lanes that are still in use today. His data were also used to

---

[13] Mayer-Schönberger and Cukier, p. 153.

[14] Mayer-Schönberger and Cukier, pp. 2-3. While Google Flu has generally been very accurate, there have been glitches. Google Flu seems to have overestimated the incidence of flu early in the 2013 season, because widespread press reports of a particularly severe outbreak may have induced more searches by people who didn't actually have the flu. See http://www.nature.com/news/when-google-got-flu-wrong-1.12413.

[15] Jeremy Ginsburg et al., "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature*, Vol. 457, February 2009, pp. 1012-14, http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html.

lay the first transatlantic telegraph cable.[16] Commander Maury "took information generated for one purpose and converted it into something else."[17]

The examples of the serendipitous use of data are numerous. In the health care area, for example, the Danish Cancer Society combined Denmark's national registry of cancer patients with cell phone subscriber data to study whether cell phone use increased the risk of cancer.[18] The FDA used Kaiser Permanente's database of 1.4 million patients to show that the arthritis drug Vioxx increased the risk of heart attacks and strokes.[19] The Centers for Disease Control combine airline records, disease reports, and demographic data to track health risks.[20]

Einav and Levin survey new research by economists using large-scale, real-time data to better track and forecast economic activity using measures that supplement official government statistics. The Billion Prices Project, for example, uses data on retail transactions from hundreds of online retail websites to produce alternative price indices that are made available in real time, before the official Consumer Price Indexes.[21] In the same vein, Choi and Varian have used

---

[16] Mayer-Schönberger and Cukier, pp. 73-76.

[17] Mayer-Schönberger and Cukier, p. 76.

[18] See the Danish study by Cardis et al, "The INTERPHONE study: design, epidemiological methods, and description of the study population", *European Journal of Epidemiology,* Vol. 22, No. 9, 2007, pp. 647-664.

[19] See the original study by Graham et al, "Risk of acute myocardial infarction and sudden cardiac death in patients treated with cyclo-oxygenase 2 selective and non-selective non-steroidal anti-inflammatory drugs: nested case-control study", *The Lancet,* Vol. 365, No. 9458, 2005, pp. 475-481.

[20] A discussion of the new CDC tool, Biomosiac, can be found in Amy O'Leary, "In New Tools to Combat Epidemics, the Key is Context", *The New York Times Bits Blog,* June 19, 2013, accessible at http://bits.blogs.nytimes.com/2013/06/19/in-new-tools-to-combat-epidemics-the-key-is-context/?smid=tw-share.

[21] See http://bpp.mit.edu/usa/.

Google search engine data to provide accurate measures of unemployment and consumer confidence.[22] Wu and Brynjolfsson have used search data to predict housing market trends.[23]

In the private sector, big data are being used to develop products that create value for firms and consumers. ZestFinance, using many more variables than traditional credit scoring, helps lenders determine whether or not to offer small, short-term loans to people who are otherwise poor credit risks.[24] This provides a better alternative to people who otherwise might rely on payday lenders or even loan sharks.

Two successful startups, Farecast, purchased by Microsoft, and Decide.com, recently purchased by eBay, use big data to help consumers find the lowest prices.[25] Farecast uses billions of flight-price records to predict the movement of airfares, saving purchasers an average of $50 per ticket. Decide.com predicts price movements for millions of products with potential savings for consumers of around $100 per item.

Another new company, Factual, collects data on over 65 million user locations and combines them with other data to help provide location-specific services, content and advertising.[26]

---

[22] Hyunyoung Choi and Hal Varian, Predicting the Present with Google Trends, Economic Record, Vol. 88, pp. 2-9.

[23] Lynn Wu and Erik Brynjolfsson, "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities", *ICIS 2009 Proceedings*, Paper 147, 2009, http://aisel.aisnet.org/icis2009/147.

[24] See an explanation of ZestFinance's techniques at http://www.zestfinance.com/how-we-do-it.html and a discussion of early success for the strategy at http://techcrunch.com/2013/07/31/data-focused-underwriting-and-credit-analysis-platform-zestfinance-raises-20m-from-peter-thiel-and-others/. Also, see Mayer-Schönberger and Cukier, p. 47.

[25] Mayer-Schönberger and Cukier, p. 124.

[26] See discussions of Factual's growth and methods in interviews with its CEO at http://streetfightmag.com/2013/06/05/with-disparate-data-factual-founder-sees-opportunity/ and http://www.adexchanger.com/mobile/factual-eyes-new-opportunities-in-location-data/.

Big data are also used to protect against adverse events ranging from credit card fraud to terrorism. As Mayer-Schönberger and Cukier note, "The detection of credit card fraud works by looking for anomalies, and the best way to find them is to crunch all the data rather than a sample."[27] Einav and Levin cite a "Palo Alto company, Palantir, [which] has become a multi-billion dollar business by developing algorithms that can be used to identify terrorist threats using communications and other data, and to detect fraudulent behavior in health care and financial services."[28] They also cite work from a group at Dartmouth using large samples of Medicare claims to demonstrate substantial unexplained variation in Medicare spending per enrollee that could be due to inefficiencies or fraud.[29]

Many of the innovations described above use multiple sources of data, which involves transferring data to third parties This practice is suspect among privacy advocates and making such transfers more difficult is one of the "familiar solutions" referred to in the FTC privacy report.[30] However, precluding the exchange of data greatly diminishes its value for purposes ranging from epidemiology studies (e.g., the Danish study of cellphone use and cancer risk) to marketing. A recent study from the Direct Marketing Association found that individual-level consumer data were an integral component in producing over $150 billion in marketing services and that over 70 percent of these services required the ability to exchange data between firms.[31]

---

[27] Mayer-Schönberger and Cukier, p. 27.

[28] Einav and Levin, p. 7.

[29] Einav and Levin, pp. 10-11.

[30] Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers,* March, 2012, pp. 44, available at: http://ftc.gov/os/2012/03/120326privacyreport.pdf.

[31] John Deighton and Peter A. Johnson, "The Value of Data: Consequences for Insight, Innovation & Efficiency in the US Economy", *The Data Driven Marketing Institute,* October, 2013, available at http://ddminstitute.thedma.org/#valueofdata.

These marketing services reduce the cost of matching producers with potential consumers in a marketplace, and are particularly valuable to smaller firms and new entrants.

IV. Big Data Has Not Increased Identity Fraud and Data Breaches

In theory, big data could increase or decrease identity fraud and data breaches. These security issues might indicate a market failure because of the difficulty of imposing costs on the perpetrators, who may be able to remain anonymous. Countervailing forces, however, provide strong incentives for data holders (e.g., credit card companies) to protect their data, while the data themselves are useful in preventing fraud, as discussed above.

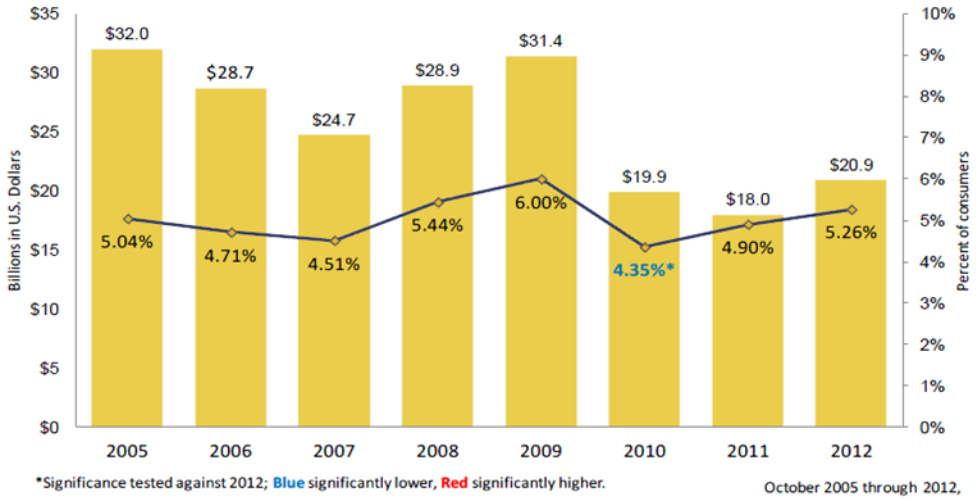
In her speech, Chairwoman Ramirez suggests that big data increase the risks associated with identity fraud and data breaches.[32] It is useful, therefore, to examine whether the proliferation of data in recent years has shown up in greater incidence of identity fraud and/or data breaches.

Identity Fraud

Javelin Strategy and Research compiles the only statistically representative series on identity fraud of which we are aware. These data are presented in Figure 2. Despite concerns voiced by the FTC and others, the overall incidence of identity fraud has been flat since 2005. During the same period, the total dollar amount of fraud has fallen—from an average of $29.1 billion for 2005-2009 to $19.6 billion for 2010-2012.

[32] Ramirez, p. 6.

Figure 2: Overall Identity Fraud Incidence Rate and Total Fraud Amount by Year



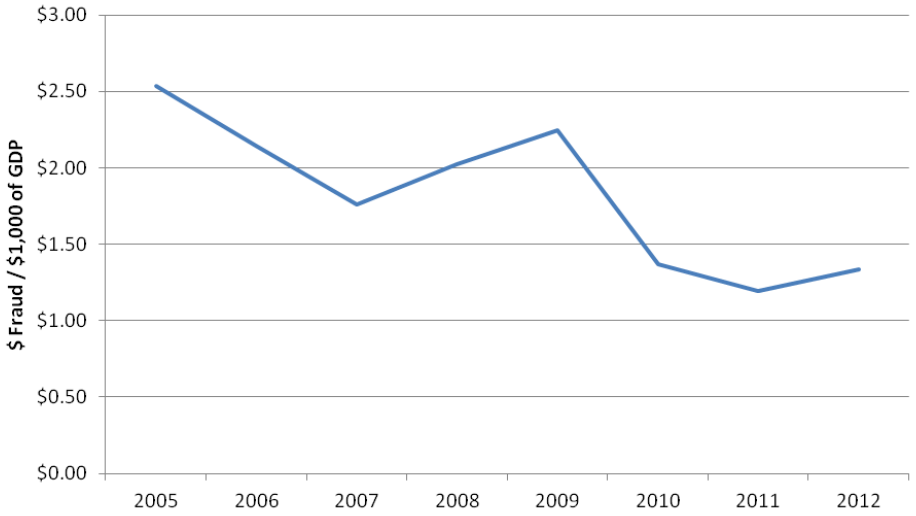
Source: *2013 Identity Fraud Report Sample*, Javelin Strategy and Research

To obtain a clearer picture of what has happened to the "risk" of identity fraud, we need to normalize the data on identity fraud by some measure of exposure.[33] Figure 3 shows that the cost of identity fraud per $1,000 of US GDP has been declining since 2005. If the identity fraud cost data were deflated by ecommerce retail sales the downward trend would be steeper, because ecommerce has grown more rapidly than GDP. However, GDP is probably a more appropriate deflator, since the great majority of identity fraud is due to offline behavior.[34]

[33] This is the same thing analysts do when examining, for example, the risks associated with driving. They don't simply look at the number of accidents. They look at the number of accidents per mile driven.

[34] Only about 15 percent is associated with data breaches and online causes. The remainder is due to offline causes, including a stolen wallet or purse, auto burglary, home burglary and signature forgery. See Travelers Insurance, "73% of identity fraud cases resulted from stolen personal items", November 2012, Online, available at: http://investor.travelers.com/phoenix.zhtml?c=177842&p=irol-newsArticle&ID=1761670.

Figure 3: Annual Cost of Identity Fraud (in dollars) Deflated by US GDP



Sources: *How Consumers Can Protect Against Identity Fraudsters in 2013*, Javelin Strategy and Research; Federal Reserve Economic Data, Federal Reserve Bank of St. Louis

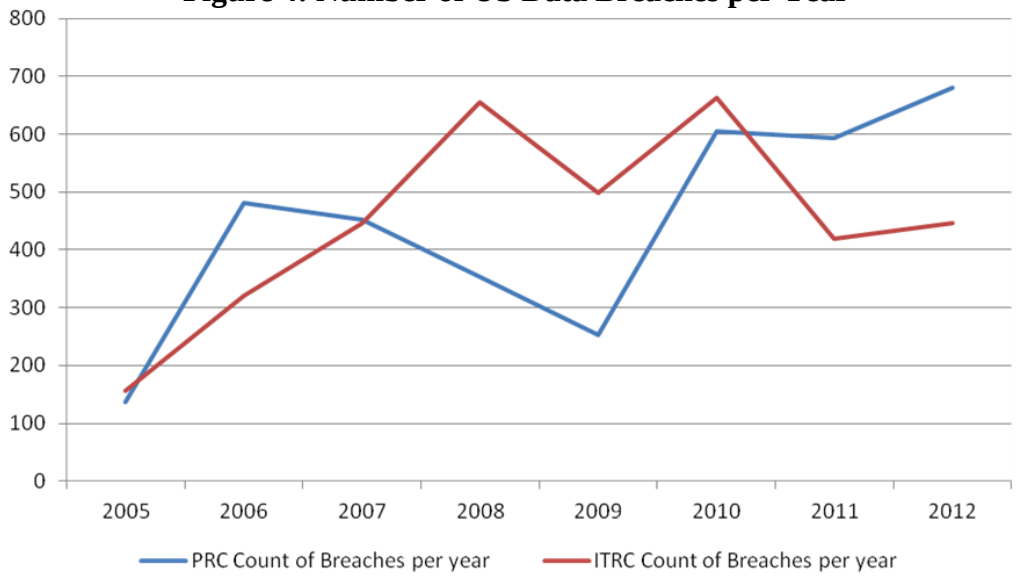
Data Breaches

There are two sources of data on data breaches—the Privacy Rights Clearinghouse and the Identity Theft Resource Center. Both of these sources aggregate information on data breaches from the media, public databases, and news releases from state governments; however, the annual totals vary slightly based on methodology and their individual definitions of a data breach.[35] Figure 4, which shows both series, suggests that the trend is slightly up since 2005.

Data breaches are purely an online phenomenon, so it is appropriate to deflate them by a measure of online activity. When deflated by the volume of ecommerce, the risk of a data breach has been relatively constant, as shown in Figure 5.

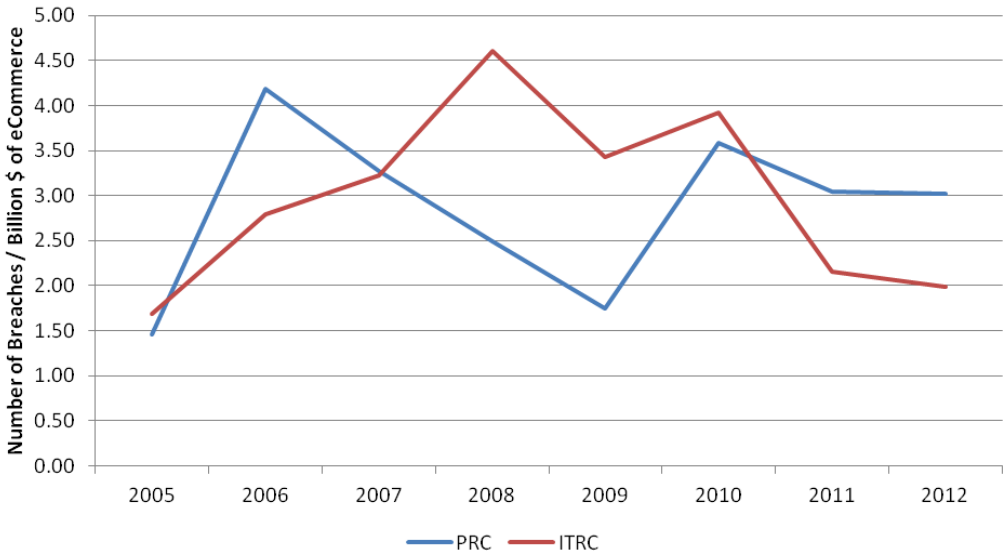
[35] For more information on the ITRC and PRC databases see http://www.idtheftcenter.org/id-theft/data-breaches.html and https://www.privacyrights.org/data-breach-FAQ.

Figure 4: Number of US Data Breaches per Year



Sources: *Data Breach Reports*, Identity Theft Resource Center, *Chronology of Data Breaches*, Privacy Rights Clearinghouse

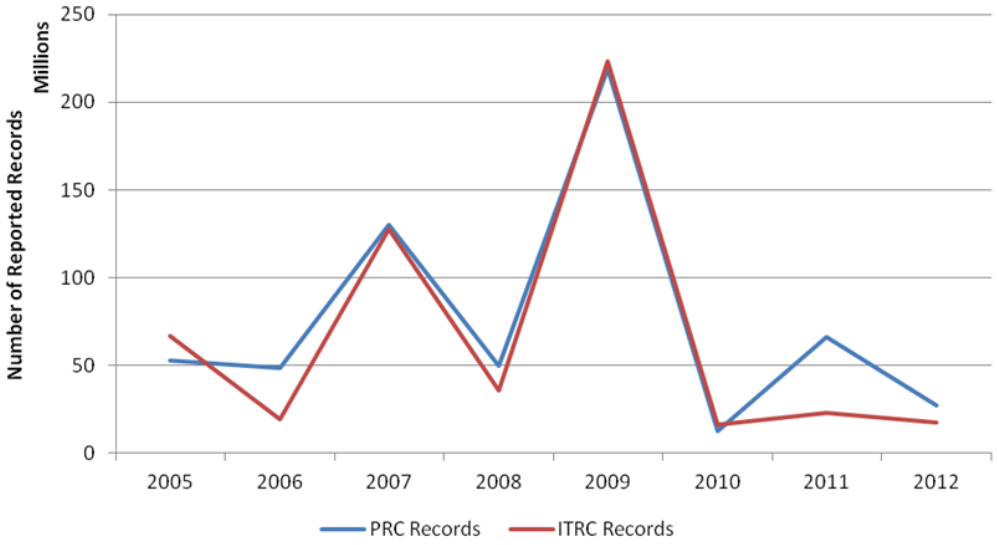
Figure 5: Number of US Data Breaches per Year Deflated by US Ecommerce



Sources: *Data Breach Reports*, Identity Theft Resource Center, *Chronology of Data Breaches*, Privacy Rights Clearinghouse; *Quarterly E-Commerce Reports*, US Census Bureau

More important than the number of breaches is the number of records compromised and the number of records compromised deflated by some measure of exposure such as ecommerce

Figure 6: Number of Reported Individual Records Compromised by Data Breaches



Sources: *Data Breach Reports,* Identity Theft Resource Center, *Chronology of Data Breaches,* Privacy Rights Clearinghouse

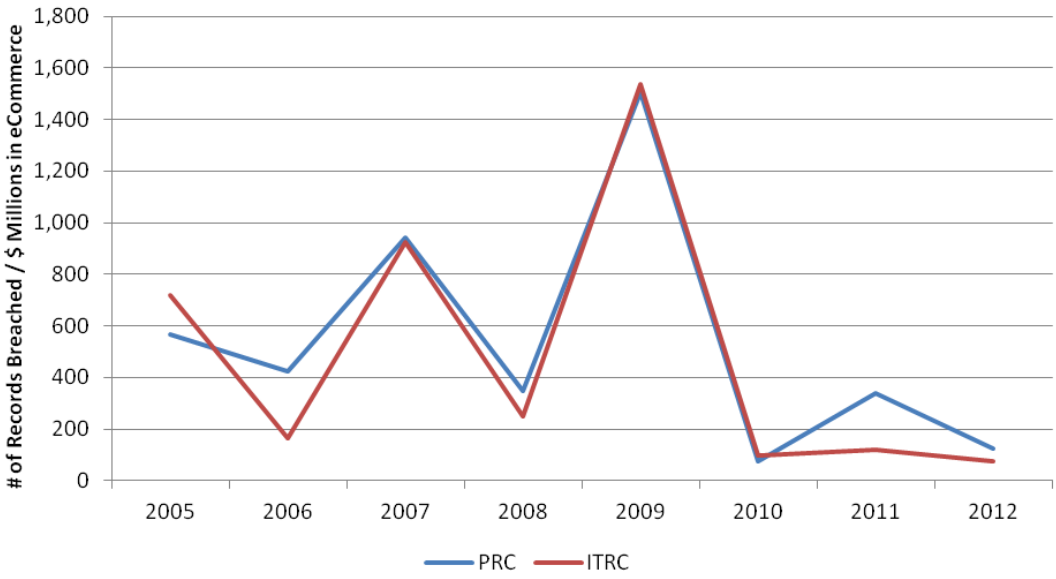
dollars. These are shown in Figures 6 and 7, respectively.[36]

The spikes in records breached in 2007 and 2009 are due to three major breaches—TJ Maxx in 2007 (100 million records) and Heartland Payment Systems (130 million records) and a military veterans database (76 million records) in 2009. Overall, the trend in records breached since 2005 is relatively constant or even declining slightly, and the trend in records breached deflated by ecommerce volume is somewhat more negative.

The data on identity fraud and breaches are far from complete. Nevertheless, there is no indication that either has gone up with the rise of big data.

[36] Note these values should be viewed with some caution as the number of records compromised is not known for every reported breach. In fact, the percentage of reports with a known number of records has varied from 30% in some years to 87% in others.

Figure 7: Number of Reported Individual Records Compromised by Data Breaches Deflated by US Ecommerce



Sources: *Data Breach Reports*, Identity Theft Resource Center, *Chronology of Data Breaches*, Privacy Rights Clearinghouse; *Quarterly E-Commerce Reports*, US Census Bureau

Indeed, one would expect that the use of big data would reduce identity fraud. This is because credit card companies, who bear most of the costs, have strong incentives to police misuse of their cards. Although these companies understandably do not publicize their procedures, one obvious method is monitoring purchases and notifying consumers when purchases seem to be outside of normal behavior, as determined by analysis of big data. Note that this policing involves use of data for purposes other than for which they were initially collected.

V. "Data Determinism": The Benefits of Algorithms

The systematic use of individuals' data for a wide range of purposes is not new. The direct marketing industry, for example, has for decades assembled mailing lists of consumers interested in specific products and services. Credit bureaus use formulas that determine individuals'

eligibility for loans and the rates they may be offered. Similarly, the insurance industry uses key variables that indicate risk to determine whether and at what rates to offer insurance policies.

A theme running through the privacy-centric big data literature is that the use of data to develop predictive models incorporating seemingly unrelated variables is harmful to consumers. As FTC Chairwoman Ramirez said, "[t]here is another risk that is a by-product of big data analytics, namely, that big data will be used to make determinations about individuals, not based on concrete facts, but on inferences or correlations that may be unwarranted."[37] She notes that "[i]ndividuals may be judged not because of what they've done, or what they will do in the future, but because inferences or correlations drawn by algorithms suggest they may behave in ways that make them poor credit or insurance risks, unsuitable candidates for employment or admission to schools or other institutions, or unlikely to carry out certain functions." She further points out, "[a]n error rate of one-in-ten, or one-in-a-hundred, may be tolerable to the company. To the consumer who has been miscategorized, however, that categorization may feel like arbitrariness-by-algorithm."[38]

This point is also made by Commissioner Brill: "They [data brokers] load all this data into sophisticated algorithms that spew out alarmingly personal predictions about our health, financial status, interests, sexual orientation, religious beliefs, politics and habits…. [I]ncreasingly our

[37] Ramirez, p.7.

[38] Ramirez, p.8.

data fuel more than just what ads we are served. They may also determine what offers we receive, what rates we pay, even what jobs we get."[39]

Of course, such distinctions are commonly made. Indeed, the educational testing industry is based on this type of correlation.[40] For example, the Federal Government, including the FTC, uses class rank in hiring lawyers. Use of credentials and test scores is universal in American life. All of these decisions are based on "small data"—sometimes, one test score or one data point. Big data can only improve this process. If more data points are used in making decisions, then it is less likely that any single data point will be determinative, and more likely that a correct decision will be reached.

Companies that devote resources to gathering data and undertaking complex analysis do so because it is in their interest to make more accurate decisions. Thus, the use of big data should lead to fewer consumers being mis-categorized and less arbitrariness in decision-making.

It is unclear what "inferences or correlations may be unwarranted." Insurance companies typically give a discount on auto insurance to students with good grades, for example. They also differentiate on the basis of the gender of young drivers. This is presumably because the data show that there is a correlation between these variables—school performance and gender—and accident costs.[41]

---

[39] Brill, ¶ 3,4.

[40] See, for example, William A. Mehrens, "Using Test Scores for Decision Making", *Test Policy and Test Performance: Education, Language, and Culture*, 1989, pp. 93-113.

[41] See for example: http://www.cnbc.com/id/100863117.

The use of more variables made possible by big data should lead to more accurate decisions that also might be "fairer." For example, ZestFinance, described above, uses its big data analysis to help underwrite loans to individuals who would otherwise not qualify. Another example is the greater use of data by state parole boards to help inform parole decisions.[42] Whether this is fairer is unclear, but proponents believe the use of big data in this manner provides more accurate predictions of the risk of recidivism and therefore can help determine which prisoners should be released and thereby increase public safety and perhaps also reduce prison costs.

Greater Transparency is a Questionable Remedy

Concern about the impact of mis-categorizing individuals sometimes leads to the recommendation that the algorithms should be more transparent and there should be "procedures to remediate decisions that adversely affect individuals who have been wrongly categorized by correlation."[43] This is the thrust of Commissioner Brill's "Reclaim Your Name" initiative. One major data broker, Acxiom, has taken a step in that direction with its aboutthedata.com web site, which allows individuals to view and potentially correct some of the data in Acxiom's file.[44]

The maxim that "…consumers [should] understand who is collecting their data and what it is being used for…" is often repeated by regulators and privacy advocates, but it is largely meaningless. For example, it is not clear that a person rejected for credit by a complex algorithm

---

[42] See Joseph Walker, "State Patrol Boards Use Software to Decide Which Inmates to Release", *The Wall Street Journal Online*, October 12, 2013, available at: http://online.wsj.com/news/article_email/SB10001424052702304626104579121251595240852-lMyQjAxMTAzMDEwMDExNDAyWj.

[43] Ramirez, p. 8.

[44] This effort, however, has been criticized by privacy advocates as being too limited. See Natasha Singer, "Acxiom Lets Consumers See Data It Collects", *The New York Times*, September 4, 2012, available at: http://www.nytimes.com/2013/09/05/technology/acxiom-lets-consumers-see-data-it-collects.html

would particularly benefit by being shown the equation used. The FICO score, an early example of a calculation based on a complex algorithm, is virtually impossible to explain to even an informed consumer because of interactions and nonlinearities in the way that elements enter into the score.[45]

Moreover, electronic information is frequently used in complex ways that are difficult or impossible to explain. For example, a Wall Street Journal series titled "What They Know" consisted of several lengthy articles explaining uses of data. It would not be feasible for websites to meaningfully convey this information through a notice, and consumers would not devote the hours required to understand it. Indeed, from the Wall Street Journal series, it appears that many practitioners do not themselves understand the ways in which they are using data.[46] In Rubin and Lenard there is a complex schematic showing the uses of data as of 2001.[47] That schematic is very difficult to follow and since then the system has become much more complex.

Giving consumers the ability to correct their information may be more complicated than it might appear, even aside from the administrative complexities. Consumers do have the right to correct information used in deriving their credit scores, but it is difficult to do so, for good reason. The purpose of collecting information that affects decisions about individuals—e.g., credit decisions, insurance decisions, or employment decisions—is to ameliorate an asymmetric information

---

[45] The major inputs to a credit score are well known; however, the calculation of credit scores from credit report data is proprietary and exceedingly complex. See for example FDIC, *Credit Card Activities Manual*, Ch. 8 – Scoring and Modeling, 2007, available at: http://www.fdic.gov/regulations/examinations/credit_card/.

[46] Julia Angwin, "The Web's New Gold Mine: Your Secrets", *The Wall Street Journal*, July 30, 2010, Online, available at: http://online.wsj.com/news/articles/SB10001424052748703940904575395073512989404.

[47] See Paul Rubin and Thomas Lenard, "Privacy and the Commercial Use of Personal Information", Kluwer Academic Publishers and Progress and Freedom Foundation, 2001, p. 26.

problem. Individuals have much more information about themselves than lenders, insurance companies or prospective employers. Asymmetric information is a feature of some markets that potentially can lead to market breakdown.[48]

An individual who thinks she has been wrongly categorized clearly has an interest in correcting erroneous information if that information has a negative effect. But she might also have an interest in "correcting" valid information that would adversely affect the decision, or inserting incorrect information that would have a positive effect. Distinguishing between these various "corrections" may be quite difficult.

Moreover, if we make it easier for individuals to access their data then we also make it easier for fraudsters to access the same data. If fraudsters have access to large amounts of data about an individual, they can more easily defraud that individual (perhaps by making purchases that are consistent with the individual's behavior in order to trick the credit card companies' monitoring efforts). Thus, ease of monitoring is at best a two-edged sword.

Big Data Do Not Discriminate Against the Poor

Some writers argue that the use of big data in marketing decisions favors the rich over the poor. A few particularly inflammatory quotes from critics include: "Ever-increasing data collection and analysis have the potential to exacerbate class disparities;"[49] and "[b]ig data—discrimination, profiling, tracking, exclusion—threaten the self-determination and personal

[48] See George A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism", *The Quarterly Journal of Economics,* Vol. 84, No. 3, August, 1970, pp. 488-500.

[49] Joseph Jerome, "Buying and Selling Privacy: Big Data's Different Burdens and Benefits", 66 Stanford Law Review Online 47, 2013, p. 50.

autonomy of the poor more than any other class."[50] One writer theorized, "[t]o woo the high value shoppers, they offer attractive discounts and promotions—use your loyalty card to buy Beluga caviar; get a free bottle of Champagne. Yet obviously the retailers can't take a loss for their marketing efforts. Who then pays the price of the rich shoppers' luxury goods? You guessed it, the rest of us—with price hikes on products like bread and butter."[51]

The argument that data collection favors the rich over the poor is usually presented without evidence. The example of consumption of caviar and Champagne by rich people being subsidized by price increases on bread and butter is, as far as we can tell, hypothetical.

Likely the concern expressed by these writers relates to price discrimination, which involves charging different prices to different consumers for the same product based on their willingness to pay.[52] Online data collection can yield information that can be used to infer a consumer's willingness to pay for a good and in that way facilitates price discrimination.[53]

Price discrimination transfers some (or even all in the case of perfect price discrimination) surplus from consumers to producers. However, price discrimination can be economically efficient (i.e., increase welfare overall) if it increases total output in a market. Particularly in the case of products with high fixed and low marginal costs—such as airline tickets—price discrimination may be necessary for the good to be produced at all. There would be fewer flights

---

[50] Jerome, p. 51.

[51] Omer Tene, "Privacy: For the Rich or for the Poor", Concurring Opinions Blog Post, June, 2012, available at: http://www.concurringopinions.com/archives/2012/07/privacy-for-the-rich-or-for-the-poor.html

[52] See Tene, ¶ 4,6.

[53] Hal R. Varian, "Differential Pricing and Efficiency", *First Monday* Vol. 1, No. 5, August, 1996, http://www.firstmonday.dk/ojs/index.php/fm/article/view/473/394.

if airlines were not able to charge varying prices. Many virtual goods, such as apps and software, also have high fixed and low or even zero marginal costs, and price discrimination may be essential to the production of these goods.

Price discrimination involves charging prices based on a consumer's willingness to pay, which in general is positively related to a consumer's ability to pay. This implies that a price discriminating firm will usually charge lower prices to lower-income consumers. Indeed, in the absence of price discrimination, some lower-income consumers would be unable or unwilling to purchase some products at all. So, contrary to arguments above, the use of big data, to the extent it facilitates price discrimination, should usually work to the advantage of lower-income consumers.

VI. Big Data and Consumer Choice: Do Consumers Get What They Want?

Two additional themes running through some of the recent privacy literature suggest that the use of data and algorithms may produce "harms" quite different from what we normally think of as privacy and security harms (i.e., harms which involve the exposure of individuals' data to people who shouldn't see them). Some writers argue that big data will facilitate manipulating consumers to purchase things they don't "really" want. Others are concerned that consumers will get too much of what they want—that they will live in a "filter bubble" determined by big data.

The literature on misuse of algorithms does not present any evidence that is inconsistent with our conclusion that there is little demonstrable harm from the legal use of commercial information.[54] For example, Calo's examples of "objective privacy harms" include: use of blood test data for drunk driving; data used for a no-fly list; police use of information from a psychologist.[55] None of these involve commercial information. The only example he uses of a commercial use is from Google gmail ads. But in this case, the consumer voluntarily uses the service in full knowledge that he will receive targeted ads. Moreover, the "harm" identified is speculative and quite indirect—consumers using the service are not typically aware of any harm.

More generally, it is difficult to draw a boundary between what is called "manipulation" and the provision of information that helps a consumer in making purchases. For example, in a separate paper, Calo discusses profitable opportunities for firms to capitalize on irrational behavior.[56] As an example, he suggests that a harmful use of information would be to send an obese consumer a text message from a donut shop when the consumer is trying to avoid snacking. But of course the consumer might want a donut, even though Calo thinks he should not have one. Moreover, given the rate of evolution of apps, there will soon be one (if there is not now) that a diet conscious consumer with weak willpower could program to ignore all messages with certain keywords, including "donut", or to remind him of the caloric content of the donut and his current weight-loss goal.

---

[54] See Thomas Lenard and Paul Rubin, "In Defense of Data: Information and the Costs of Privacy", *Policy & Internet*, Vol. 2, Issue. 1, April, 2010, pp. 149-183, available at http://onlinelibrary.wiley.com/doi/10.2202/1944-2866.1035/abstract.

[55] Ryan Calo, "The Boundaries of Privacy Harm", *Indiana Law Journal*, Vol. 86, No. 3, 2011, pp. 1131-1162.

[56] Ryan Calo, "Digital Market Manipulation", Legal Studies Research Paper and George Washington Law Review (forthcoming), 2013, University of Washington School of Law, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2309703.

As Calo acknowledges, profiting from irrational behavior would be difficult (perhaps impossible) since it would be extremely difficult to determine what is rational for a given consumer.[57] Moreover, he does not explain why firms would want to do this. Using large data sets, firms might simply determine when they can sell products, and most of the time that would be to consumers who want the product, and that would generally be to rational consumers. Moreover, while some firms might try to sell products that the consumer does not "really" want, others would be trying to sell products that the consumer does want, and those firms can be expected to win out.

The fundamental problem with this line of analysis is that many of the privacy advocates and writers on the subject do not trust the consumers for whom they purport to advocate. This is also apparent in writers who express concern about consumers living in a "filter bubble." For example, Pariser laments that "[t]he statistical models that make up the filter bubble write off the outliers. But in human life it's the outliers who make things interesting and give us inspiration."[58] Dwork and Mulligan are concerned that "filter bubbles" will take away "the tumult of traditional public forums—sidewalks, public parks, and street corners—where a measure of randomness and unpredictability yields a mix of discoveries and encounters that contribute to a more informed populace."[59]

---

[57] Calo, p. 15.

[58] Eli Pariser, "The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think", Penguin Books, April, 2012, p. 134.

[59] Cynthia Dwork and Deirdre K. Mulligan, "It's Not Privacy and It's Not Fair", 66 Stanford Law Review Online 47, 2013, p. 39.

If consumers want variety, big data and algorithms, particularly as they get more sophisticated, should be helpful in providing that to them. However, the notion that algorithms will give consumers "too much" of what they want at the expense of what is good for them is a more radical idea with unclear policy implications. Does it mean we should limit the collection and use of data to purposely produce less accurate algorithms? That doesn't seem to make much sense.

VII. Conclusion

The potential of big data may be different from "small data" in terms of their transformative effects on the economy and specific sectors. That remains to be seen. However, there is no obvious reason to approach privacy policy questions arising from big data differently than we approach questions involving smaller amounts of data. The same questions are relevant:

- Is there a market failure and evidence of harm to consumers? The recent literature on big data does not provide such evidence, at least as far as the legal use of data for commercial purposes is concerned. Moreover, we have found no evidence of an increase in harm to consumers from identity fraud or data breaches.
- If evidence of market failure or harm is found, is there an available remedy (or remedies) that can reasonably be expected to yield benefits greater than costs and therefore yield net benefits to consumers? As we've illustrated above, the standard solutions referenced by regulators would likely not yield net benefits.

Any attempt to limit "harmful" uses of information will limit beneficial uses as well. The FTC is viewed by writers such as Calo as a white knight, and Chairwoman Ramirez also suggests

"meaningful oversight" as a remedy for what they perceive as harms to consumers. However, the FTC has shown itself to be an overprotective steward, and has often reduced consumer welfare by excessive regulation of information.[60] Although the first task of a regulator such as the FTC should be to perform a cost-benefit analysis, in producing its recent privacy guidelines the FTC has explicitly avoided doing so.[61] Given this lack of data and analysis, particularly in a new market such as the electronic use of information, it is much more likely that an uninformed regulator will stifle innovation rather than provide net benefits. The "familiar solutions"—the FIPPs that would limit the reuse or sharing of data—would seem to be particularly harmful because they are inconsistent with the new ways in which big data are being used.

[60] Paul H. Rubin, "Regulation of Information and Advertising," Competition Policy International, Spring 2008, v. 4, No. 1, pp. 169-192.

[61] See Exec. Order 13563, Improving Regulation and Regulatory Review (Jan. 17, 2011), available at http://www.gpo.gov/fdsys/pkg/FR-2011-01-21/pdf/2011-1385.pdf. Also, see Thomas M. Lenard and Paul H. Rubin, "The FTC and Privacy: We Don't Need No Stinking Data," Antitrust Source, October, 2012, available at http://www.americanbar.org/content/dam/aba/publishing/antitrust_source/oct12_lenard_10_22 f.authcheckdam.pdf.