TECHNOLOGY POLICY INSTITUTE
Studying the Global Information Economy

March 31, 2014

Office of Science and Technology Policy
Eisenhower Executive Office Building
1650 Pennsylvania Ave., NW
Washington, DC 20502
Attn: Big Data Study

Re: Big Data Request for Information

These comments are in response to the Office of Science and Technology's March 4, 2014 Notice of Request for Information. OSTP is requesting input on the Administration's "comprehensive review of the ways in which 'big data' will affect how Americans live and work, and the implications of collecting, analyzing and using such data for privacy, the economy, and public policy."

These comments are largely based on a 2013 Technology Policy Institute paper[1] and address Question 1 of the RFI:

(1) What are the public policy implications of the collection, storage, analysis, and use of big data? For example, do the current U.S. policy framework and privacy proposals for protecting consumer privacy and government use of data adequately address issues raised by big data analytics?

The emergence of big data and the Internet of Things, which generates a growing supply of objects from which data can be collected, has raised the question of whether big data are associated with new privacy harms and a concomitant increase in the need for government action. If so, should policy makers look to the standard solutions involving notice and choice, use specification and limits, and data minimization to solve any privacy problems brought about by big data?

In the study referenced above, we conclude that there is no evidence at present that big data used for commercial and other non-surveillance purposes have caused privacy harms. Moreover, the standard solutions associated with the U.S. policy framework and privacy proposals—Privacy by Design, the Fair Information Practice Principles (FIPPs) and the Organization of Economic Cooperation and Development (OECD) principles—represent a potentially serious barrier to much of the innovation we anticipate from the big data revolution.

---

[1] Thomas M. Lenard and Paul H. Rubin, "The Big Data Revolution: Privacy Considerations," December 2013, available at http://www.techpolicyinstitute.org/files/lenard_rubin_thebigdatarevolutionprivacyconsiderations.pdf.

The Promise of Big Data[2]

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."[3] Data are now available in real time, at larger scale, with less structure, and on different types of variables than previously.[4]

Because big data analysis involves finding previously unobserved correlations and patterns, it almost necessarily involves uses of data that were not anticipated at the time the data were collected. Examples of the serendipitous uses of data are numerous and include health studies, economic studies, marketing, and the development of new products that help consumers gain access to credit and search for lower prices. Big data are also used to protect against adverse events ranging from credit card fraud to terrorism. Many of the innovations described above use multiple sources of data, which involves transferring data to third parties.

Potential Privacy Threats

Advocates have highlighted a number of potential privacy threats from big data, but as of now there is no evidence that any of these threats has materialized. I discuss them in turn.

Big data increase the risks associated with identity fraud and data breaches.[5]

In theory, big data could increase or decrease identity fraud and data breaches. On the one hand, there are more data at risk. On the other hand, the data themselves are useful in preventing fraud. Moreover, countervailing forces provide strong incentives for data holders (e.g., credit card companies) to protect their data. So, it is useful to examine what the data on identity fraud and data breaches show.

In fact, the proliferation of big data in recent years does not appear to have increased identity fraud and/or data breaches. Since 2005, the overall incidence of identity fraud has been relatively flat and the total dollar amount of fraud has fallen—from an average of $29.1 billion for 2005-2009 to $19.2 billion for 2010-2013. While there has been a slight increase in the number of U.S. data breaches, the trend in records breached since 2005 is relatively constant or even declining slightly. All these measures show more of a decline when considered relative to the growth of the economy and e-commerce.

---

[2] See Lenard and Rubin, pp. 1-10.

[3] 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.

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

[5] See Lenard and Rubin, pp. 10-15.

The use of big data to develop predictive models is harmful to consumers.[6]

The assertion that predictive models harm consumers, if valid, would apply to quantitative analysis used for decision-making throughout the economy. Much of the concern seems to be that the predictive models may not be totally accurate. However, big data improve accuracy.

Use of credentials and test scores, from credit scores to class rankings, is ubiquitous in American life. 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 reduce errors. The data and the models have limited value unless they improve accuracy. Thus, big data should lead to fewer consumers being mis-categorized and less arbitrariness in decision-making.

The use of big data in marketing decisions favors the rich.[7]

The argument that data collection favors the rich over the poor is presented without evidence. Likely the concern relates to price discrimination, which involves charging different prices to different consumers for the same product based on their willingness to pay.[8] 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.[9]

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, other things the same, 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, the use of big data, to the extent it facilitates price discrimination, should usually work to the advantage of lower-income consumers.

Moreover, big data are being used to develop products that specifically benefit lower-income consumers. For example, 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.[10] This provides a better alternative to people who otherwise

---

[6] See Lenard and Rubin, pp. 15-18.

[7] See Lenard and Rubin, pp. 20-22.

[8] See Tene, ¶ 4,6.

[9] 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.

[10] See http://www.zestfinance.com/how-we-do-it.html.

might rely on payday lenders or even loan sharks. LendUp, BillFloat, and ThinkFinance are companies following similar models that can provide better loan options for lower-income consumers, while Kabbage and On Deck Capital provide lending services to very small businesses.

Big data have the potential to provide cost savings to consumers through the analysis and comparison of the prices of goods sold over the internet. Two successful startups, Farecast and Decide.com, use big data to help consumers find the lowest prices.[11] 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.

Firms use big data to manipulate consumers.[12]

Some writers argue that firms use big data to manipulate consumers to behave irrationally and purchase things they don't really want.[13] Drawing a boundary between what is called "manipulation" and the provision of information that helps a consumer make purchasing decisions is difficult. In general, it is not possible to determine whether any given purchase is "rational" or not, because consumers' utility functions are not directly observable.

In a market economy, firms are rewarded for giving consumers what they want. The economist's criterion of performance is how close the economy comes to maximizing "total surplus." Firms want to capture as much of that surplus as possible, and may use data to more precisely target times when they can charge consumers a higher price. However, the transaction will still be beneficial to the consumer, or the transaction would not occur; she may just capture less consumer surplus. Moreover, this is also a way that firms can efficiently price discriminate. Others may get a lower price. Importantly, such price discrimination may be necessary to cover costs of production and for the product to be available at all.

An implicit assumption in discussions of manipulation is that firms lack competition. Even assuming firms can manipulate consumers and thereby earn super-competitive profits, unless there are barriers to entry, other firms will be induced to enter and compete away those profits. This is a check on whatever manipulation might be possible.

---

[11] Mayer-Schönberger and Cukier, "Big Data: a revolution that will transform how we live, work and think", Houghton Mifflin Harcourt, 2013, p. 124.

[12] See Lenard and Rubin, pp. 22-24.

[13] 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.

Individuals will be forced to reveal data about themselves, thereby eroding privacy.

With advanced information technologies, individuals will increasingly be able to voluntarily make available a range of *verified* (because it is linked directly to the source) information about themselves, including health information, employment records, court records, driving behavior and credit history. For example, your health data may be generated by wearable monitors, your driving behavior by sensors in your car, etc.

The data made available could determine eligibility and the terms for many economically important items, including jobs, insurance and admission to schools. Those with the most favorable data will find it in their interest to reveal it. Others will then be "forced" to reveal their data because failure to do so will reflect negatively on those who do not. Some are concerned that this contains within it the threat of unraveling privacy altogether.[14]

Generally, this phenomenon is considered efficient, because it provides information to the market and helps solve asymmetric information problems. In the absence of information, markets may "unravel" in another way. This is the well-known "lemons" problem.[15]

In the same way that prohibiting producers from hiding defects in their products leads to better products, there are positive incentive effects when individuals are unable to conceal adverse personal information. If individuals were able to conceal their credit histories, we would find more delinquent payments, which would raise the costs of borrowing generally. The fact that automobile insurance rates are lower for young males with a better grade point average is an incentive to study harder, or, at least, for the parents to make sure the student studies harder.

A simple example illustrates the potential costs of restricting this type of information sharing. It is now possible to monitor driving behavior for a variety of purposes. Mapping programs do this in order to direct drivers to the fastest route at any given time. But data can also be used by insurance companies to set rates. A driver can install a device in her car to monitor the time of day she drives and the distance traveled and have the data automatically delivered to her insurance company.[16] Presumably, safe drivers will want to do this so they can get lower rates. Insurance companies might rationally assume that drivers who failed to install such devices were less safe than those who did and charge them a higher rate. This would likely result in at least a partial "unraveling" as more and more drivers installed the monitoring devices.

Prohibiting this practice, as some privacy advocates suggest, would mean there is no payoff to voluntarily providing your monitoring data to the insurance company. This would penalize safe drivers to the benefit of average and less-safe drivers. The prohibition would increase accidents,

---

[14] See Scott Peppet, *Unraveling Privacy: The Personal Prospectus and the Threat of a Full-Disclosure Future,* Northwestern University Law Review, Vol. 105, No. 3, p. 1166.

[15] 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.

[16] See for example: http://www.progressive.com/auto/snapshot-how-it-works/.

because even the safest drivers will drive more carefully when they know they are being monitored. There may be a significant increase in safety from drivers further down the spectrum, who would be induced to install the monitor.

Policy Considerations

There is no obvious reason to approach privacy policy issues arising from big data differently than we approach issues involving smaller amounts of data. The same questions are relevant:

- *Is there evidence of a market failure or 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. Discussions of harm are largely speculative and hypothetical. Moreover, the available data do not indicate that there has been 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?* Since the harms are largely hypothetical, so are the benefits.

The threshold question is whether there are harms that can be reduced by the adoption of privacy policies. Otherwise, there are no benefits to privacy regulations. The absence of identified harms implies that privacy policies cannot be expected to yield significant benefits, even in the absence of costs.

The privacy remedies typically discussed are, however, likely to impose costs. A standard solution long promoted by privacy advocates is that data should only be collected for a specific identified purpose. This is reflected in the FIPPs dating back to the 1970s, the OECD Privacy Principles of 1980, the current European Union regulations, and the recommendations of the FTC's 2012 Privacy Report.[17]

Requiring that data only be collected for an identified purpose is particularly ill-suited to the world of big data. Using data in unanticipated ways has been a hallmark of the big data revolution, for commercial, research and even public sector uses. Therefore, the standard solutions that would limit the reuse or sharing of data would be particularly harmful if applied to big data because they are inconsistent with the innovative ways in which data are being used.

This also calls into question the principles of notice and choice, which become almost meaningless when data may be used in unpredictable ways. Even absent questions concerning big data, these principles have become increasingly irrelevant. As Beales and Muris note, "The reality that decisions about information sharing are not worth thinking about for the vast majority

[17] The Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers,* March, 2012.

of consumers contradicts the fundamental premise of the notice approach to privacy." They continue, "The FIPs principle of choice fares no better." [18]

Concern about the use of data for predictive scoring and the possibility that algorithms may mis-categorize individuals sometimes leads to recommendations for greater transparency. The notion that consumers should understand who is collecting their data and how they are being used is an appealing one, but it is largely meaningless, especially in the big data era where scores may be based on hundreds of data points and very complex calculations. For example, it is not clear that a person rejected for credit by a complex algorithm 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.[19]

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 made difficult to do so, for good reason. 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.

The purpose of collecting information that affects decisions about individuals—e.g., credit decisions, insurance decisions, or employment decisions—is to ameliorate asymmetric information problems. As Beales and Muris point out, "In our economy, there are vital uses of information sharing [such as credit reporting] that depend on the fact that consumers cannot choose whether to participate."[20]

Moreover, if we make it easier for individuals to access their data then we also make it easier for those bent on fraud to access the same data. If fraudsters have access to large amounts of data about a person, they can more easily defraud that individual. Thus, ease of consumer monitoring is at best a two-edged sword.

Conclusion

Any attempt to limit "harmful" uses of information will limit beneficial uses as well. Although many have suggested "meaningful oversight" as a remedy for what they perceive as harms to

---

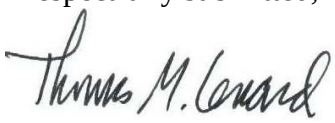
[18] J. Howard Beales and Timothy Muris, *Choice or Consequences: Protecting Consumer Privacy in Commercial Information*, University of Chicago Law Review, 2008, pp. 113-118.

[19] 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/.

[20] Beales and Muris, p. 115.

consumers, there has been no evidence of any concrete privacy harms. 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 uninformed regulation will stifle innovation rather than provide net benefits. The "familiar solutions"—such as those 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.

Respectfully submitted,



Thomas M. Lenard
President and Senior Fellow
Technology Policy Institute