Comment
Office of Science and Technology Policy
Big Data Study

March 31, 2014

James C. Cooper
George Mason University School of Law
Law & Economics Center

"Big Data," as it has come to be known broadly is the application of analytics to large datasets that come close to covering an entire population, which allows researchers to find relationships that working with small samples would not reveal.[1] By discovering these new relationships, big data stands to be transformative: Google Flu Trends and IBM's Watson are among the standard bearers of big data, but are merely the tip of the iceberg. Big data also has allowed dramatic improvements in fraud detection, and promises to make driverless cars and smart homes a reality. Working with large data sets, moreover, provides researchers with deeper understandings of social and physical phenomena that may improve living in ways we currently cannot apprehend.

At the same time, some are concerned about the risks that big data may pose. The ubiquitous collection of observations from both our online and offline lives that feeds big data has the potential to intrude on privacy. Further, some worry about the security of these immense data caches. Finally, others have expressed concern over the use of big data to sort consumers in increasing granular categories that will determine the offers and information they receive.

This brief comment makes two key points about regulatory approaches to big data. First, any regulation must be guided by an empirically grounded benefit-cost analysis, not unsupported hypotheticals. Second, because the reduction in private information improves the efficiency of markets, the ability of big data to make more granular classifications – either through firm sorting or consumer signaling – should be considered a benefit rather than a harm.

---

[1] *See* VIKTOR MAYER-SCHÖNBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW

A. Benefit-Cost Analysis

Any regulatory framework that addresses big data must be guided by empirically grounded benefit-cost analysis. The benefits of big data are tangible and can be measured objectively,[2] and regulations that retard big data will deprive consumers of these benefits. Accordingly, proponents of restrictions on big data should have the burden to demonstrate empirically that such policies are necessary to ameliorate actual or likely consumer harm, and that the avoided harm is greater than the foregone benefits.

Broadly, privacy harms can be classified as tangible or intangible. Tangible harms include the extent to which big data is likely to increase the risk of identity fraud or reputational harm from data breaches of sensitive financial or personal information. Such harms can be measured objectively with metrics like fraudulent charges, inconvenience costs, or lost marketplace opportunities due to stigma. Intangible harms include the discomfort associated with ubiquitous observation and the revelation of embarrassing information. Further, some scholars recently have written about the harm associated with receiving information flows only tailored to predicted interests.[3] These harms are suffered internally and therefore are not amendable to objective measurement.

This is not to say that intangible harm should be ignored or could never form the basis for regulatory action. Before relying on intangible harms as a justification for restrictions on big data, however, policy makers should have a firm grasp on their variance and magnitude. The harm associated with unauthorized monitoring of intimate activities or revelation of sensitive health information, for example, is probably significant for most of the population. At the same time, any discomfort associated with the collection and analysis of anonymized data streams for honing predictive algorithms, or receiving tailored information or advertisements, is likely to vary widely among consumers.

When the variance in sensitivity to a particular form of observation and analysis is low, uniform standards will approximate optimality for most of the population. Alternatively, when harm suffered varies widely, a uniform standard –

---

[2] *See id*.; Thomas M. Lenard & Paul H. Rubin, *The Big Data Revolution: Privacy Considerations*, at 4-9 (Dec. 2013), *at* https://www.techpolicyinstitute.org/files/lenard_rubin_thebigdatarevolutionprivacyconsiderations.pdf.

[3] *See*, *e.g.*, ELI PARISER, THE FILTER BUBBLE: HOW THE NEW PERSONALIZED WEB IS CHANGING WHAT WE READ AND HOW WE THINK (2012); Cynthia Dwork & Deirdre K. Mulligan, *It's Not Privacy and Its Not Fair*, 66 STANFORD L. REV. ONLINE 35, 37 (2013).

especially one geared toward to those who are most sensitive – will impose costs on wide swaths of the population. Many willingly would accept less privacy in exchange for lower prices, richer and more customized content, or superior functionality. Consequently, is it crucial that policy makers refrain from imposing a "one size fits all" solution based on "worst-case" hypotheticals that lack any empirical grounding.

B. Classification Harms

A common theme in several works concerning the potential threats of big data is that data driven algorithms increasingly will be used to sort consumers into categories that will determine the types of offers and information they receive.[4] As explained below, the reduction in private information should not be considered harm. Instead, because increased information improves market efficiency through better matching of buyers and sellers, the more granular classifications made possible by big data are likely to be beneficial.

First, businesses currently categorize consumers using the data that is available, and more data typically will allow for more, not less, accurate estimates of parameters like credit risk, health status, or interests. Further, firms have incentives to place consumers into correct categories; companies that systematically offer high interest credit cards to people with good credit or expensive auto insurance to good drivers, will see their profits suffer.

Although more accurate categorization likely means that some consumers receive worse terms, it also means that many consumers will receive better terms. Because ability to pay is negatively correlated with income, moreover, poorer individuals who were previously priced out of a market are likely benefit from differential pricing. What's more, differential pricing does not occur in a vacuum; just as big data may allow firms to charge consumers with relatively inelastic demand higher prices, the same data driven algorithms will allow their competitors to target these consumers with discounts. In this manner, big data driven classification actually can *intensify* competition.[5]

---

[4] *See*, *e.g.*, Dwork & Mulligan, *supra* note 3; Scott Peppet, *Unraveling Privacy: The Personal Prospectus and the Threat of a Full-Disclosure Future*, 105 NORTHWESTERN U. L. REV. 1116 (2012); Ryan Calo, *Digital Market Manipulation*, 82 GEO. WASH. L. REV. (forthcoming 2014).

[5] *See* Kenneth S. Corts, *Third-Degree Price Discrimination in Oligopoly: All-Out Competition and Strategic Commitment*, 29 RAND J. ECON. 306 (1998); James C. Cooper *et al.*, *Does Price Discrimination Intensify Competition? Implications for Antitrust*, 72 ANTITRUST L. J. 327 (2005). This type of competition can be seen regularly at the grocery store, when after purchasing one brand of diapers or cereal, a consumer often will receive a coupon for a competing brand.

Relatedly, some have expressed concern about the flip side of firms sorting consumers based on big data – consumer signaling.[6] In this scenario, the ubiquity of cheap monitoring technology is increasingly allowing consumers to send credible signals to firms that they possess qualities that should lead to better offers (*e.g.*, lower health risks, better driving skills). The upshot is that those who do not agree to be monitored will be presumed to possess less desirable qualities (*e.g.*, poor health or diving skills), and hence will receive worse terms (*e.g.*, higher health or car insurance rates). It is unclear how providing contracting parties with more information about their true types will harm societal welfare. Again, although some parties will receive worse terms, those with above-average qualities will receive better terms. As more information is revealed to the market, matches between firms and consumers will improve, increasing welfare. Finally, limiting classification will create a moral hazard; the inability to receive the benefits from desirable behavior (*e.g.*, healthy eating, safe driving, or good grades), or to be punished for undesirable behavior, will reduce incentives to engage in desirable behavior.

C. Conclusion

A responsible regulatory approach must rest on an empirical showing that government action is necessary to prevent substantial consumer harm. Because markets work more efficiently with greater amounts of information, regulators should refrain from restricting the ability to use big data driven algorithms to make more accurate classifications. Concerns over inequities generated by improved classification are best deal with through transfer payments to those who are negatively impacted, not by the forced concealment of private information, which will reduce the efficiency of the price system. Big data stands to be transformative, and regulators should exercise appropriate caution and humility to avoid unnecessarily depriving consumers of the substantial benefits that big data promises to offer.

[6] *See* Peppet, *supra* note 4.