Big Data Threats

Felix T. Wu*

The pros and cons of big data are the subject of much debate. The "pro" side points to the potential to generate unprecedented new knowledge by gathering, aggregating, and mining data, knowledge that can be used for everything from earlier detection of drug side effects to better management of electricity and traffic.[1] The "con" side says that big data raises privacy issues.

To talk about a big data privacy problem, however, is far too imprecise. In this context, the concept of "privacy" stands for a diverse set of interests. In order to evaluate those interests, weigh them against competing interests, and design appropriate regulatory responses, we need to disentangle them.

Consider the problem of online behavioral advertising, that is, the targeting of advertising based upon one's prior online activity. Perhaps the problem with behavioral advertising is that the tracking technologies that make such advertising possible cause users to feel surveilled as they go about their business, online or off. The problem could also be that stored tracking information might be revealed, to acquaintances or to the government. Alternatively, it might be the targeting itself that is the problem, and that there is something wrong with using tracking information to determine what advertisements a person sees.

Similarly, think about the story of Target, which apparently computed a pregnancy prediction score based upon its customers' purchases and used this score to determine to whom to send coupons for baby products.[2] Maybe it makes Target shoppers "queasy" to think that Target is able to predict whether they are pregnant, or even to think that Target is trying to do so.[3] Target's practices might also lead to inadvertent disclosure, as when a father supposedly learned of his daughter's pregnancy for the first time from seeing the coupons Target sent to her.[4] Perhaps it is a problem for pregnant women to get different offers than non-pregnant women. While there might be nothing wrong with targeting baby products to the people who might actually buy them, perhaps differing offers for other products, or on the basis of other predictions, might be more problematic.

In the context of big data in particular, it is helpful to think not just in terms of privacy in general, but in terms of specific privacy threats.[5] When faced with a big data practice, the key question is: "How could this go wrong?" Even for a single practice, that question has many potential answers.

---

* Associate Professor, Benjamin N. Cardozo School of Law

[1] *See* Omer Tene & Jules Polonetsky, *Privacy in the Age of Big Data: A Time for Big Decisions*, 64 STAN. L. REV. ONLINE 63 (2012).

[2] *See* Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES, Feb. 16, 2012.

[3] *Id.* ("If we send someone a catalog and say, 'Congratulations on your first child!' and they've never told us they're pregnant, that's going to make some people uncomfortable . . . . [E]ven if you're following the law, you can do things where people get queasy.").

[4] *Id.*

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

One can conceive of at least three broad categories of big data threats: surveillance, disclosure, and discrimination. By surveillance, I mean the feeling of being watched, which can result from the collection, aggregation, and/or use of one's information.[6] The feeling of being surveilled might be an intrinsic problem, akin to emotional distress. It might also be a problem because such a feeling can affect how people behave, if people start to think twice about the things they do, read, or search for.[7] On this account, one problem with pervasive web tracking is the possibility that people will avoid certain searches or certain sources of information, for fear that doing so inevitably reveals interests, medical conditions, or other personal characteristics they would rather remain hidden.

The feeling of surveillance can arise from the mere collection of information, as when visits to sensitive websites are tracked. As in the Target example, however, it can also arise from the particular form of processing being done to the data. Presumably any unease that customers feel from receiving baby products coupons comes from feeling that Target "knows" about their pregnancy, rather than from knowing that Target has recorded a list of their purchases. Thus, it can be the data mining itself or the characteristic being mined for that converts a mere collection of information into a feeling of surveillance.

Other problems can be conceived of as problems of disclosure of data outside of the context in which it was collected. One disclosure threat might be the nosy employee who looks up people he knows in a corporate database. Another might be an identity thief who successfully hacks into a database. Problems of insecurity are in this sense problems of disclosure. Less maliciously, information might be revealed to people who happen to be nearby and see the ads on another person's computer. Similarly, as in the Target example, people in the same household might see one another's mail. Disclosure to the government is a different potential threat. Government as threat is also not a monolithic one, and could encompass everything from a rogue government employee to a systematic campaign that harasses people on the basis of lawful activity.

Other big data problems are problems of discrimination, that is, treating people differently on the basis of information collected about them. Again, there are many different kinds of discrimination threats. The most obvious might be trying to predict membership in some protected class, such as race or religion, and then discriminating on that basis. Some might further object to any discrimination that is correlated with a protected characteristic, whether or not it forms the explicit basis for the targeting. Consumers, however, seem to also have a visceral reaction against certain forms of price discrimination, even when based on nothing like race or religion.[8] The ability of big data to result in highly personalized pricing seems to be a source of concern for many.

Personalized persuasion is another form of discrimination enabled by big data that might be problematic.[9] The idea here is that rather than simply altering the price or

---

[6] Ryan Calo calls this a "subjective privacy harm." *See* M. Ryan Calo, *The Boundaries of Privacy Harm*, 86 IND. L.J. 1131, 1144–47 (2011).
[7] *See* Neil M. Richards, *Intellectual Privacy*, 87 TEX. L. REV. 387 (2008).
[8] *See* Jennifer Valentino-DeVries et al., *Websites Vary Prices, Deals Based on Users' Information*, WALL ST. J., at A1, Dec. 24, 2012.
[9] *See* Ryan Calo, *Digital Market Manipulation* (draft).

product being sold, the advertiser alters the sales pitch itself so as to best exploit each individual's own cognitive biases.[10] Big data may make it more possible to identify widely shared biases, and this might already be a source of concern. Even if we are willing to tolerate the exploitation of widely shared biases, however, the exploitation of individual biases raises additional concerns about an imbalance of power between advertisers and consumers.

Lack of transparency can in some ways constitute a fourth category of threats. Without transparency, individuals may find themselves feeling stuck in a world in which consequential decisions about them are being made opaquely, thereby inducing a sense of helplessness.[11] That feeling, distinct from the feeling of being surveilled, might itself be a problem with big data.

On the other hand, transparency might also be understood as a tool to mitigate some of the other threats identified above. Appropriate transparency could, at least in theory, make it possible for individuals to choose to deal with companies that minimize disclosure risks. Transparency could also diminish the effectiveness of personalized persuasion, again at least in theory.

Even though the word "threat" implies that there is something problematic about the occurrence of the threat, in speaking about threats, I am not necessarily arguing that everything laid out above should in fact be a cognizable threat. One could, for example, hold the view that certain types of discrimination are perfectly acceptable, even desirable. Similarly, some might argue that some of the negative consequences of big data that I have described are not privacy problems at all, but problems of a different sort.[12] In this brief essay, I am not trying to delimit the boundaries of privacy versus other types of harms.

Nor does distinguishing among threats necessarily mean we need distinct regulatory responses. The threat of discrimination might be dealt with by restricting the practice, but it may be far easier to regulate the collection of the relevant information than to detect its misuse.

My goal here instead is simply to catalogue some of the different things that people mean when they say that there is a privacy problem with big data. Doing so helps to frame the big data privacy analysis better. It can help us determine when tools like de-identification can be effective at balancing privacy and utility,[13] and it can help us determine in what contexts the benefits outweigh the burdens of big data analysis.

---

[10] *Id.*

[11] Dan Solove has argued that the appropriate metaphor is to Franz Kafka's The Trial. *See* Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393 (2001).

[12] *See, e.g.*, Calo, *supra*, at 1158.

[13] *See* Wu, *supra*.