Big Data and Consumer Privacy in the Internet Economy
79 Fed. Reg. 32714 (Jun. 6, 2014)

Comment of Solon Barocas, Edward W. Felten,
Joanna N. Huey, Joshua A. Kroll, and Arvind Narayanan

Thank you for the opportunity to comment on the issues raised by big data and consumer privacy. We focus our response on question 11, which relates to de- and re-identification of data.

Significant privacy risks stem from re-identification. Analysis methods that allow sensitive attributes to be deduced from supposedly de-identified datasets pose a particularly strong risk. Although de-identification is often used as a first step, additional technological and policy measures must be developed and deployed to reduce the risks of privacy-sensitive data.

Calling data "anonymous" once personal information has been removed from it is a recipe for confusion. The term suggests that such data cannot later be re-identified. However, as we describe here and others have described elsewhere, such assumptions are increasingly becoming obsolete.

The President's Council of Advisors on Science and Technology (PCAST) was emphatic in recognizing the risks of re-identification.

> Anonymization of a data record might seem easy to implement. Unfortunately, it is increasingly easy to defeat anonymization by the very techniques that are being developed for many legitimate applications of big data. In general, as the size and diversity of available data grows, the likelihood of being able to re-identify individuals (that is, re-associate their records with their names) grows substantially.
>
> [...]
>
> Anonymization remains somewhat useful as an added safeguard, but it is not robust against near-term future re-identification methods. PCAST does not see it as being a useful basis for policy.[1]

The PCAST report reflects the consensus of computer scientists who have studied de- and re-identification: there is little if any technical basis for believing that common de-identification methods will be effective against likely future adversaries.

[1] President's Council of Advisors on Science and Technology. *Report to the President: Big Data and Privacy: A Technological Perspective*, at 38-39 (May 2014).

1. Defining Re-identification

The stereotypical example of re-identification is when a name is reattached to a record that was previously de-identified. However, privacy violations often occur through other, less obvious forms of re-identification. In particular, 1) any identifier can affect privacy, not just typical identifiers such as name and social security number, and 2) sensitive attributes of a user can be inferred, even when that user cannot be matched directly with a database record.

When discussing identifiers, we must recognize that new types of identifiers exist that may not be included in existing definitions of personally identifiable information ("PII"). Account numbers, persistent tags such as device serial numbers, or long-lived tracking identifiers can all be associated with a collection of information about a user. If a sensitive attribute is linked to an identifier and that identifier is linked to a user, then the user's privacy is at issue, regardless of whether the identifier meets some regime's definition of PII.

When considering the types of re-identification that may cause privacy harms, we must remember that a user's privacy is affected whenever an inference of a sensitive attribute can be made. Consider a hypothetical de-identified medical database and a hypothetical user, Alice, such that a database analyst can narrow down the possibilities for Alice's record to a set of ten records.[2] If all ten of the records show a diagnosis of liver cancer, the analyst learns that Alice has liver cancer. If nine of the ten show liver cancer, then the analyst can infer a 90% probability that she has liver cancer.[3] Either way, Alice's privacy has been impacted, even though no individual database record could be associated with her.

To be sure, probabilistic inferences can be harmless when the probabilities involved are negligible. For example, if general population statistics imply that Alice would have liver cancer with 0.01% probability, and an analyst with access to data can adjust that probability estimate to 0.010003%, then the impact on Alice is likely too small to worry about. However, probabilistic inferences about a person should be considered as impacting privacy when they can affect the person's interests. Our example is an extreme case in which the news that Alice has liver cancer with 90% probability clearly could alter her chances of getting a job, insurance, or credit, but less dramatic differences in probability also may affect a person's interests.

2. Re-identification Capabilities

Two main types of re-identification scenarios concern us as current threats to privacy: 1) broad attacks on large databases and 2) attacks that target a particular individual within a dataset. Broad attacks seek to get information about as many people as possible (an adversary in this case could be someone who wants to sell comprehensive records to a third party), while targeted attacks

[2] This is consistent with the database having a technical property called *k-anonymity*, with k=10. Examples like this show why k-anonymity does not guarantee privacy.

[3] To simplify the explanation, we assume that the analyst is maximally uncertain as to which of the ten records is Alice's, so that the analyst assigns a 10% probability that each of the ten records is Alice's. The inferred probability might differ from 90% if the analyst has reason to believe that one of the ten records is more likely than the others to match Alice. Provided that the analyst is highly uncertain about which of the ten is Alice, the probability that Alice has liver cancer will be close to 90%.

have a specific person of interest (an adversary could be someone who wants to learn medical information about a potential employee).

A. Broad Re-identification Attacks

Many current datasets can be re-identified with no more than basic programming and statistics skills. The amount of possible re-identification is grossly underrepresented in academic literature because only a small fraction of re-identification attempts are of academic interest.

The damage done by these attacks is exacerbated by the failure to de-identify datasets properly. The recently released New York City taxi log data used a simple hash function in an attempt to anonymize drivers and cabs. That method is known to be flawed[4] and allowed for easy re-identification of taxi drivers.[5] Additional information in the data leaves the door open to possible re-identification of riders.

However, while better de-identification can prevent certain casual attacks, even data de-identified with the standard techniques permit a certain level of re-identification. For example, the Heritage Health Prize was a contest asking people to infer—based on a database of de-identified historical health insurance claims—which patients were most likely to be admitted to a hospital in the year after scoring. In his report to the contest organizers, Narayanan calculated that up to 12.5% of members could be re-identified by an adversary assumed to know certain information about the targets, such as the year during which visits to medical practitioners were made.[6] With large enough datasets, even a small percentage of re-identification becomes a significant concern; if the database covers the entire United States population, a 1% rate of re-identification means that over three million identities will be compromised.

Furthermore, current de-identification is inadequate for high-dimensional data. These high-dimensional datasets, which contain many data points for each individual's record, have become the norm: social network data has at least a hundred dimensions[7] and genetic data at least a million. We expect that datasets will continue this trend towards higher dimensionality as the costs of data storage decrease and the ability to track a large number of observations about a single individual increase.

B. Targeted Re-identification Attacks

An important type of re-identification risk stems from adversaries with specific targets. If someone has knowledge about a particular person, identifying him or her within a dataset becomes much easier. The canonical example of this type of re-identification comes from

---

[4] Edward Felten, *Does Hashing Make Data "Anonymous"?*, Tech@FTC blog (2012), available at http://techatftc.wordpress.com/2012/04/22/does-hashing-make-data-anonymous/.

[5] Vijay Pandurangan, *On Taxis and Rainbows: Lessons from NYC's improperly anonymized taxi logs* (2014), available at https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1.

[6] Arvind Narayanan, *An Adversarial Analysis of the Reidentifiability of the Heritage Health Prize Dataset*, Manuscript (2011).

[7] Johan Ugander, Brian Karrer, Lars Backstrom & Cameron Marlow, *The anatomy of the Facebook social graph*, arXiv Preprint, arXiv:1111.4503 (2011) (noting that the median Facebook user has about a hundred freinds).

Sweeney's 1997 demonstration that she could re-identify the medical record of then-governor William Weld using only his date of birth, gender, and ZIP code.[8]

More recently, research by Narayanan and Shmatikov revealed that with minimal knowledge about a user's movie preferences, there is an over 80% chance of identifying that user's record in the Netflix Prize dataset.[9] In addition, they showed as a proof-of-concept demonstration that it is possible to identify Netflix users by cross-referencing the public ratings on IMDb.

In addition, a 2013 study by de Montjoye et al. revealed weaknesses in anonymized location data.[10] Analyzing a mobile phone dataset that recorded the location of the connecting antenna each time the user called or texted, they evaluated the uniqueness of individual mobility traces (i.e., the recorded data for a particular user, where each data point has a timestamp and an antenna location). Over 50% of users are uniquely identifiable from just two randomly chosen data points. As most people spend the majority of their time at either their home or workplace, an adversary who knows those two locations for a user is likely to be able to identify the trace for that user—and to confirm it based on the patterns of movement.[11] If an adversary knows four random data points, which a user easily could reveal through social media, 95% of mobility traces are uniquely identifiable.

Many de-identified datasets are vulnerable to re-identification by adversaries who have specific knowledge about their targets. A political rival, an ex-spouse, a neighbor, or an investigator could have or gather sufficient information to make re-identification possible.

As more datasets become publicly available or accessible by (or through) data brokers, the problems with targeted attacks can spread to become broad attacks. One could chain together multiple datasets to a non-anonymous dataset and re-identify individuals present in those combinations of datasets.[12] Sweeney's re-identification of then-Governor Weld's medical record used a basic form of this chaining: she found his gender, date of birth, and ZIP code through a public dataset of registered voters and then used that information to identify him within the de-identified medical database.

---

[8] Latanya Sweeney, Statement before the Privacy and Integrity Advisory Committee of the Department of Homeland Security, Jun. 15, 2005, http://www.dhs.gov/xlibrary/assets/privacy/privacy_advcom_06-2005_testimony_sweeney.pdf.

[9] Arvind Narayanan & Vitaly Shmatikov, *Robust de-anonymization of large sparse datasets*, in Proc. 2008 IEEE Symp. on Security and Privacy 111-125 (2008). The Netflix Prize dataset included movies and movie ratings for Netflix users.

[10] Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The privacy bounds of human mobility*, Scientific Reports 3 (2013).

[11] Other studies have confirmed that pairs of home and work locations can be used as unique identifiers. *See* Hui Zang & Jean Bolot, *Anonymization of location data does not work: A large-scale measurement study*, in Proc. 17th Int'l. Conf. on Mobile Computing and Networking 145-156 (2011); Philippe Golle & Kurt Partridge, *On the anonymity of home/work location pairs*, Pervasive Computing 390-397 (2009).

[12] A similar type of chaining in a different context can trace a user's web browsing history. A network eavesdropper can link 90% of a user's web page visits to the same pseudonymous ID, which can often be linked to a real-world identity. Dillon Reisman, *Cookies that give you away: The surveillance implications of web tracking*, Apr. 4, 2014, available at http://freedom-to-tinker.com/blog/dreisman/cookies-that-give-you-away-the-surveillance-implications-of-web-tracking/. Both types of chaining are examples of data fusion, referenced in question 10.

3. Policy Responses to Re-identification

Re-identification raises a difficult policy question: how to balance privacy threats with the benefits fostered by wider access to data. Each dataset has its own risk-benefit tradeoff, in which the expected damage done by leaked information must be weighed against the expected benefit from improved analysis. Both assessments are complicated by the unpredictable effects of data fusion, since combining the dataset with others may escalate either the losses or the gains.

The current toolkit for addressing re-identification includes de-identification methods, emerging technologies like differential privacy, prophylactic restrictions of access to data, and legal backstops. Each of these tools has its strengths and weaknesses.

We should remember that once a dataset is released to the public, it cannot be taken back. Re-identification techniques will continue to improve, and additional datasets will become public and available to be chained together as described above. Current de-identification methods do not provide affirmative evidence that they cannot leak information regardless of what an adversary does. As such, a dataset that is de-identified upon its release today becomes increasingly vulnerable as adversaries get more skilled and possess more information. De-identification techniques are best seen not as a way to prevent re-identification, but as a way to delay re-identification by raising the bar a bit for adversaries.

Unlike de-identification, differential privacy does not depend on artificial assumptions about the adversary's capabilities and its guarantees do not become weaker as adversaries become more capable. Like all protective measures, differential privacy involves a tradeoff between privacy and utility, as the stronger the privacy guarantees are made, the less accurate the estimated statistics from the data become.

Restricting access to data enhances privacy and cabins data science in the expected ways. Such limitations allow data custodians to vet researchers for trustworthiness and to hold leakers accountable more easily, but they prevent crowdsourcing and reduce the chances for novel uses of the data.

Finally, legal measures can deter adversaries or make data custodians more cautious, but they depend on ease of enforcement and are cold comfort to privacy victims if re-identification has already occurred. Provisions in the terms of use or the work contracts of hired analysts can make re-identification or other privacy-impacting uses of data a contractual violation. In addition, civil or criminal penalties could be imposed on adversaries who perform re-identification or other make other privacy-harming uses of data, and on data custodians who fail to take reasonable steps to prevent these uses.

The decreasing efficacy of de-identification should lead policymakers to rely less on de-identification and more on other safeguards. Deciding which combination of these other tools should be used for a particular dataset requires an understanding of their limitations and an individualized risk-benefit calculation. We have no one-size-fits-all solution, but we believe that one helpful initiative would be to create a best practices guide. Such a guide would include general information on the tools available and at the very least would prevent naive attempts at

re-identification, such as happened with the New York City taxi data. It should also lead the data custodian through the issues outlined above and could provide case studies of how to approach particular data in specific situations. And, of course, further research into privacy technologies will strengthen and expand the toolkit and improve the tradeoffs.

About the Commenters

Solon Barocas is a Postdoctoral Research Associate at the Center for Information Technology Policy at Princeton University. He completed his doctorate in the Department of Media, Culture, and Communication at New York University, during which time he was also a Student Fellow at the Information Law Institute at the School of Law.

Edward W. Felten is the Robert E. Kahn Professor of Computer Science and Public Affairs, and the Director of the Center for Information Technology Policy, at Princeton University. In 2011-12 he served as the first Chief Technologist at the Federal Trade Commission. He is a member of the National Academy of Engineering and the American Academy of Arts and Sciences. He is Chair of ACM's U.S. Public Policy Council, and an ACM Fellow.

Joanna N. Huey is the Associate Director of the Center for Information Technology Policy at Princeton University. She holds an M.P.P. in science and technology policy from the Harvard Kennedy School and a J.D. from Harvard Law School, where she was president of the Harvard Law Review.

Joshua A. Kroll is a Ph.D. Candidate in Computer Science at Princeton University. His research spans computer security, privacy, and the interplay between technology and public policy, with a particular interest in building systems to make algorithms accountable. He received the National Science Foundation Graduate Research Fellowship in 2011.

Arvind Narayanan is an Assistant Professor in Computer Science at Princeton, and an affiliated faculty member at the Center for Information Technology Policy. He was previously a post-doctoral fellow at the Stanford Computer Science department and a Junior Affiliate Scholar at the Stanford Law School Center for Internet and Society. He studies privacy from a multidisciplinary perspective, focusing on the intersection between technology, law and policy. His research has shown that data anonymization is broken in fundamental ways, for which he jointly received the 2008 Privacy Enhancing Technologies Award. He is one of the researchers behind the "Do Not Track" proposal.