

National Privacy Research Strategy
79 Fed. Reg. 56091 (Sep. 18, 2014)

Comment of Edward W. Felten, Joanna N. Huey, and Arvind Narayanan

Thank you for the opportunity to comment on the development of a National Privacy Research Strategy to guide federally-funded privacy research and provide a framework for coordinating research and development in privacy-enhancing technologies.

Currently, privacy research suffers from ill-defined problems and unproven solutions. A National Privacy Research Strategy presents an opportunity to refocus research on developing well-founded theories of privacy and to encourage implementations of those theories in practice.

The field of privacy can learn from the successes and struggles in cryptography research. The concept of provable security can be translated to this area: “privacy” can be defined rigorously and data practices can be designed to have provable levels of privacy. In addition, privacy researchers should be careful to avoid the disconnect between theorists and practitioners that has troubled cryptography—theorists need to develop usable constructs and practitioners need to adopt methods with provable privacy.

Research into differential privacy methods follows this principle of provable privacy. We propose that the National Privacy Research Strategy 1) should promote further research into how differential privacy methods, or other well-founded theories of privacy, can be used in practice and 2) should incentivize scientists to adopt these methods.

1. Privacy Problems: The Trouble with Ad Hoc De-Identification

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 certain specified information has been removed from it is a recipe for confusion. The term suggests that such data cannot later be re-identified or used to infer sensitive attributes of a person. 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.

A few illustrative examples of problems stemming from de-identification in various domains are listed below:

- In 1997, Sweeney demonstrated that she could re-identify the medical record of then-governor William Weld using only his date of birth, gender, and ZIP code.²
- The 2013 dataset released by New York City's Taxi and Limousine Commission after a FOIL request³ exposed sensitive information about both drivers and passengers. The re-identification of driver information stems from especially poor de-identification practices,⁴ but the re-identification of passenger information demonstrates privacy problems that better ad hoc de-identification still would not fix. First, it is possible to identify trip records (with pickup and dropoff locations, date and time, medallion or license number, and fare and tip amounts) if you know some of that information: for example, stalkers who see their victims take a taxi to or from a particular place can

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

² 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.

³ Chris Whong, FOILING NYC's Taxi Trip Data, Mar. 18, 2014, http://chriswhong.com/open-data/foil_nyc_taxi/.

⁴ Vijay Pandurangan, On Taxis and Rainbows: Lessons from NYC's improperly anonymized taxi logs, Jun. 21, 2014, <https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1> ("Security researchers have been warning for a while that simply using hash functions is an ineffective way to anonymize data. In this case, it's substantially worse because of the structured format of the input data. This anonymization is so poor that anyone could, with less than 2 hours work, figure which driver drove every single trip in this entire dataset. It would even be easy to calculate drivers' gross income, or infer where they live.").

determine the other endpoint of those trips.⁵ Second, it is possible to identify people who regularly visit sensitive locations, such as a strip club or a religious center.⁶ The data includes specific GPS coordinates. If multiple trips have the same endpoints, it is likely that the other endpoint is the person's residence or workplace, and searching the internet for information on that address may reveal the person's identity.

- 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, which included movies and movie ratings for Netflix users.⁷ 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. Although some movie ratings are not always considered highly private information, identifying full viewing and rating histories can reveal political preferences, religious affiliations, and other tastes that users may have preferred not to share.
- A 2013 study by de Montjoye et al. revealed weaknesses in anonymized location data.⁸ 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.⁹ If an adversary knows four random data points, which a user easily could reveal through social media, 95% of mobility traces are uniquely identifiable.

Two conclusions can be drawn from these examples. First, 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.

⁵ Anthony Tockar, *Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset*, Sep. 15, 2014, <http://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/>.

⁶ Id. Tockar goes on to explain how to apply differential privacy to this dataset.

⁷ Arvind Narayanan & Vitaly Shmatikov, *Robust de-anonymization of large sparse datasets*, in Proc. 2008 IEEE Symp. on Security and Privacy 111-125 (2008).

⁸ Yves-Alexandre de Montjoye et al., *Unique in the Crowd: The privacy bounds of human mobility*, *Scientific Reports* 3 (2013).

⁹ 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).

Second, 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 dimensions¹⁰ 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.

2. A Principled Approach to Data Privacy Architecture

Once a dataset is released to the public, it cannot be taken back. Re-identification techniques will continue to improve and will be bolstered as additional datasets become public. These facts make protocols and systems with proven privacy properties an urgent need.

A. Provable Privacy Research

The foundation for such protocols and systems are methods of handling data that preserve a rigorously defined privacy while also permitting useful analysis. At present, algorithms that yield differential privacy are the only well-developed methodology that satisfies these requirements. Development of additional models and methods is a useful avenue for research.

Current ad hoc de-identification methods do not provide rigorous justification for claims 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. Ad hoc de-identification techniques are best seen not as a way to prevent re-identification, but at best as a way to delay re-identification by raising the bar a bit for adversaries.

One lesson from cryptography research is the importance of getting central definitions correct. Finding a definition of security or privacy that is sound, provable, and consistent with intuitive notions of those terms can be a research contribution in itself. Such a definition enables evaluation of existing and proposed algorithms against a consistent standard.

Differential privacy is based on a formal definition: including a particular user's data in a dataset (as opposed to omitting it) must have a strictly limited effect on the output of any differentially private analysis of the data. Differential privacy algorithms¹¹ typically add noise to the outputs of

¹⁰ 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 friends).

¹¹ For introductions to differential privacy, see Cynthia Dwork, Frank McSherry, Kobbi Nissim & Adam Smith, Differential Privacy - A Primer for the Perplexed, Joint UNECE/Eurostat work session on statistical data confidentiality (2011); Christine Task, An Illustrated Primer in Differential Privacy, XRDS 20.1, 53-57 (2013); Erica Klarreich, Privacy by the Numbers: A New Approach to Safeguarding Data, Quanta Magazine, Dec. 10, 2012.

analysis and release those blurred outputs, rather than releasing the original input data or unaltered outputs. The effect of including a particular user's data in the dataset can be made arbitrarily small through variations in the type and amount of noise.

Like all protective measures, differential privacy algorithms involve a tradeoff between privacy and utility, as the stronger the privacy guarantees are made, the less accurate the estimated statistics from the data become. Increased noise both improves privacy and reduces the usefulness of the blurred outputs. However, unlike ad hoc de-identification, algorithms implementing differential privacy can quantify the tradeoff between privacy and utility, and they do not depend on artificial assumptions about the adversary's capabilities. Their guarantees do not become weaker as adversaries become more capable. No matter how much is known about the targeted person, the information learnable by the adversary because that person is included in the dataset remains strictly limited.

Given these advantages, we encourage further research investment in the development and application of differential privacy methods, as well as in the development of other computer science and mathematical techniques aimed at provable privacy.

B. Implementation of Provable Privacy

Provable privacy methods are necessary, but not sufficient, for responsible data practices. The full implementation of those methods is as much a policy problem as a technical one, and it requires multi-disciplinary cooperation.

We emphasize two main goals to help propagate these methods and create more real-world applications of provable privacy, such as the Census Bureau's OnTheMap.¹² First, privacy researchers must communicate with scientists using data so that the theoretical privacy work is developed with practical uses in mind. Second, data scientists and other data providers must accept and use these new methodologies—this sort of shift in data user behavior fits into the responsible use framework recommended by the Big Data report.¹³

Although many levers may be used to influence researchers, funding choices are an essential and practical tool. Much of the work done both by privacy researchers and by data users and providers is dependent upon governmental funding streams, so altering allocations to advance provable privacy would be a highly effective motivation to improve practices. It is also a quicker and more flexible path to behavioral change than legislative or regulatory privacy requirements.

¹² U.S. Census Bureau, OnTheMap, <http://onthemap.ces.census.gov/>; see also Erica Klarreich, Privacy by the Numbers: A New Approach to Safeguarding Data, *Quanta Magazine*, Dec. 10, 2012.

¹³ "Big Data: Seizing Opportunities, Preserving Values," May 2014, http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

Privacy research funding can encourage collaborations with or feedback from practitioners. Data science funding can favor projects that implement provable privacy methods instead of de-identification or no privacy measures. Making the development and application of provable privacy a factor in funding decisions will push practitioners to overcome the inertia that keeps them using existing ad hoc methods, which incorporate unproven and risky data privacy practices.

About the Commenters

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.

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.