A Precautionary Approach to Big Data Privacy

Arvind Narayanan arvindn@cs.princeton.edu

Joanna Huey jhuey@princeton.edu

Edward W. Felten felten@cs.princeton.edu

March 19, 2015

Once released to the public, data cannot be taken back. As time passes, data analytic techniques improve and additional datasets become public that can reveal information about the original data. It follows that released data will get increasingly vulnerable to re-identification—unless methods with provable privacy properties are used for the data release.

Due to the ad hoc de-identification methods applied to currently released datasets, the chances of re-identification depend highly on the progress of re-identification tools and the auxiliary datasets available to an adversary. The probability of a privacy violation in the future is essentially unknowable. In general, a precautionary approach deals with uncertain risk by placing the burden of proof that an action is not harmful on the person taking the action. Here, we argue for a weak version of the precautionary approach, in which the idea that the burden of proof falls on data releasers guides policies that incentivize them not to default to full, public releases of datasets using ad hoc de-identification methods.

In Section 1, we argue that privacy risks due to inference go beyond the stereotypical re-identification attack that links a de-identified record to PII. We review and draw lessons from the history of re-identification demonstrations, including both "broad" and "targeted" attacks. In Section 2, we explain why the privacy risk of data that is protected by ad hoc de-identification is not just unknown, but unknowable, and contrast this situation with provable privacy techniques like differential privacy.

Sections 3 and 4 contain our recommendations for practitioners and policy makers.[1] In Section 3, we discuss the levers that policymakers can use to influence data releases: research funding choices that incentivize collaboration between privacy theorists and practitioners, mandated transparency of re-identification risks, and innovation procurement. Meanwhile, practitioners and policymakers have numerous pragmatic options for narrower releases of data. In Section 4, we present advice for six of the most common use cases for sharing data. Our thesis is that the problem of "what to do about re-identification" unravels once we stop looking for a one-size-fits-all solution, and in each of the six cases we propose a solution that is tailored, yet principled.

[1] Though many of the examples are U.S.-centric, the policy recommendations have widespread applicability.

1 Ill-Founded Promises of Privacy: The Failures of Ad Hoc De-identification

Significant privacy risks stem from current de-identification practices. Analysis methods that allow sensitive attributes to be deduced from supposedly de-identified datasets pose a particularly strong risk, and calling data "anonymous" once certain types of personally identifiable information ("PII") have been removed from it is a recipe for confusion. The term suggests that such data cannot later be re-identified, but such assumptions are increasingly becoming obsolete.

The U.S. President's Council of Advisors on Science and Technology ("PCAST") was emphatic in recognizing these risks:

> 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.[2]

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.

1.1 Privacy-Violating Inferences Go Beyond Stereotypical Re-identification

It is important to consider the full scope of privacy violations that can stem from data releases. 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.

First, when discussing identifiers, the relevant question is not so much "can this data be linked to PII?" as "can this data be linked to a user?" Account numbers, persistent tags such as device serial numbers, or long-lived tracking identifiers—such as enduring pseudonyms[3]—can all be associated with a collection of information about a user, whether or not they are included in exist-

[2] Executive Office of the President, President's Council of Advisors on Science and Technology, *Report to the President: Big Data and Privacy: A Technological Perspective* (Washington, DC: 2014): 38-39.

[3] Ed Felten, "Are pseudonyms 'anonymous'?," *Tech@FTC*, April 30, 2012, https://techatftc.wordpress.com/2012/04/30/are-pseudonyms-anonymous/.

ing definitions of PII.[4] Nissenbaum and Barocas point out that oxymoronic “anonymous identifiers” such as Google’s AdID assigned by an organization to a user do nothing to alleviate the user’s privacy worries when interacting with that organization or the universe of applications with which the identifier is shared.[5] A recent example of such problems is Whisper, a social media app that promises anonymity but tracks users extensively and stores their data indefinitely.[6] The false distinction between defined PII and other potential identifiers allows Whisper to monitor the movements of “a sex obsessed lobbyist,” noting “[h]e’s a guy that we’ll track for the rest of his life and he’ll have no idea we’ll be watching him,” while still maintaining that “Whisper does not request or store any personally identifiable information from users, therefore there is never a breach of anonymity.”[7]

Second, re-identification affects a user’s privacy whenever an inference of a sensitive attribute can be made. Suppose an analyst can narrow down the possibilities for Alice’s record in a de-identified medical database to one of ten records.[8] If all ten 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 that there is a high likelihood of Alice having liver cancer.[9] Either way, Alice’s privacy has been impacted, even though no individual database record could be associated with her.

1.2 Re-identification Attacks May Be Broad or Targeted

Two main types of scenarios concern us as 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 have a specific person of interest (an adversary could be someone who wants to learn medical information about a potential employee).

[4] Paul Ohm, “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization,” *UCLA Law Review* 57 (2010): 1742-43, http://uclalawreview.org/pdf/57-6-3.pdf.

[5] Solon Barocas and Helen Nissenbaum, “Big Data’s End Run Around Anonymity and Consent,” in ***Privacy, Big Data, and the Public Good: Frameworks for Engagement***, ed. Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum (New York: Cambridge University Press, 2014), 52-54.

[6] Paul Lewis and Dominic Rushe, “Revealed: how Whisper app tracks ‘anonymous’ users,” *The Guardian*, October 16, 2014, http://www.theguardian.com/world/2014/oct/16/-sp-revealed-whisper-app-tracking-users.

[7] Ibid. A poster self-identified as the CTO of Whisper reiterated this point: “We just don’t have any personally identifiable information. Not name, email, phone number, etc. I can’t tell you who a user is without them posting their actual personal information, and in that case, it would be a violation of our terms of service.” rubyrescue, October 17, 2014, comment on blackRust, “How Whisper app tracks ‘anonymous’ users,” *Hacker News*, October 17, 2014, https://news.ycombinator.com/item?id=8465482.

[8] This is consistent with the database having a technical property called *k-anonymity*, with k=10. Latanya Sweeney, “*k*-anonymity: A Model for Protecting Privacy,” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10, no. 5 (2001): 557-70. Examples like this show why k-anonymity does not guarantee privacy.

[9] Heuristics such as l-diversity and t-closeness account for such privacy-violating inferences, but they nevertheless fall short of the provable privacy concept we discuss in the next section. Ashwin Machanavajjhala et al., “*l*-diversity: Privacy beyond *k*-anonymity,” *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, no. 1 (2007): 3; Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian, “*t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity,” in *IEEE 23rd International Conference on Data Engineering, 2007* (Piscataway, NJ: IEEE, 2007): 106-15.

1.2.1 Broad Attacks: Examples and Lessons

Many released datasets can be re-identified with no more than basic programming and statistics skills. But even if current techniques do not suffice, that is no guarantee of privacy — the history of re-identification has been a succession of surprising new techniques rendering earlier datasets vulnerable.

In 2000, Sweeney showed that 87% of the U.S. population can be uniquely re-identified based on five-digit ZIP code, gender, and date of birth.[10] Datasets released prior to that publication and containing such data became subject to re-identification through simple cross-referencing with voter list information. For example, through comparison with the Social Security Death Index, an undergraduate class project re-identified 35% of Chicago homicide victims in a de-identified dataset of murders between 1965 and 1995.[11] Furthermore, because research findings do not get put into practice immediately, datasets still are being released with this type of information: Sweeney showed that demographic information could be used to re-identify 43% of the 2011 medical records included in data sold by the state of Washington,[12] and Sweeney, Abu, and Winn demonstrated in 2013 that such demographic cross-referencing also could re-identify over 20% of the participants in the Personal Genome Project, attaching their names to their medical and genomic information.[13]

For years, security experts have warned about the failure of simple hash functions to anonymize data, especially when that data has an easily guessable format, such as the nine digits of a social security number.[14] Yet, simple hashing was commonly thought of as an anonymization method, and once again, continues to be used in released datasets. The 2013 dataset released by New York City's Taxi and Limousine Commission after a FOIL request[15] exposed sensitive information in part by using a simple hash function to try to anonymize drivers and cabs, allowing for easy re-identification of taxi drivers:

[10] Latanya Sweeney, "Simple Demographics Often Identify People Uniquely" (Data Privacy Working Paper 3, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2000), http://dataprivacylab.org/projects/identifiability/paper1.pdf.

[11] Salvador Ochoa et al., "Reidentification of Individuals in Chicago's Homicide Database: A Technical and Legal Study" (final project, 6.805 Ethics and Law on the Electronic Frontier, Massachusetts Institute of Technology, Cambridge, Massachusetts, May 5, 2001), http://mike.salib.com/writings/classes/6.805/reid.pdf.

[12] Latanya Sweeney, "Matching Known Patients to Health Records in Washington State Data" (White Paper 1089-1, Data Privacy Lab, Harvard University, Cambridge, Massachusetts, June 2013), http://dataprivacylab.org/projects/wa/1089-1.pdf.

[13] Latanya Sweeney, Akua Abu, and Julia Winn, "Identifying Participants in the Personal Genome Project by Name" (White Paper 1021-1, Data Privacy Lab, Harvard University, Cambridge, Massachusetts, April 24, 2013), http://dataprivacylab.org/projects/pgp/1021-1.pdf. Sweeney and her team matched 22% of participants based on voter data and 27% based on a public records website.

[14] Ben Adida, "Don't Hash Secrets," *Benlog*, June 19, 2008, http://benlog.com/2008/06/19/dont-hash-secrets/; Ed Felten, "Does Hashing Make Data 'Anonymous'?," *Tech@FTC*, April 22, 2012, https://techatftc.wordpress.com/2012/04/22/does-hashing-make-data-anonymous/; Michael N. Gagnon, "Hashing IMEI numbers does not protect privacy," *Dasient Blog*, July 26, 2011, http://blog.dasient.com/2011/07/hashing-imei-numbers-does-not-protect.html.

[15] Chris Whong, "FOILing NYC's Taxi Trip Data," March 18, 2014, http://chriswhong.com/open-data/foil_nyc_taxi/.

> 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.[16]

Additional information in the data leaves the door open to re-identification of riders, which is discussed in the following section.

New attributes continue to be linked with identities: search queries,[17] social network data,[18] genetic information (without DNA samples from the targeted people),[19] and geolocation data[20] all can permit re-identification, and Acquisti, Gross, and Stutzman have shown that it is possible to determine some people's interests and Social Security numbers from only a photo of their faces.[21] The realm of potential identifiers will continue to expand, increasing the privacy risks of already released datasets.

Furthermore, even staunch proponents of current de-identification methods admit that they are inadequate for high-dimensional data.[22] 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[23] and genetic data can have millions.[24] We expect that datasets will continue this trend towards higher dimensionality as the costs of data storage decrease and the ability to

[16] Vijay Pandurangan, "On Taxis and Rainbows: Lessons from NYC's improperly anonymized taxi logs," *Medium*, June 21, 2014, https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1.

[17] Michael Barbaro and Tom Zeller, Jr., "A Face Is Exposed for AOL Searcher No. 4417749," *New York Times*, August 9, 2006, http://www.nytimes.com/2006/08/09/technology/09aol.html.

[18] Ratan Dey, Yuan Ding, and Keith W. Ross, "The High-School Profiling Attack: How Online Privacy Laws Can Actually Increase Minors' Risk" (paper presented at the 13th Privacy Enhancing Technologies Symposium, Bloomington, IN, July 12, 2013), https://www.petsymposium.org/2013/papers/dey-profiling.pdf; Arvind Narayanan and Vitaly Shmatikov, "De-anonymizing Social Networks," in *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy* (Washington, D.C.: IEEE Computer Society, 2009): 173-87.

[19] Melissa Gymrek et al., "Identifying Personal Genomes by Surname Inference," *Science* 339, no. 6117 (January 2013): 321-24, doi:10.1126/science.1229566.

[20] Philippe Golle and Kurt Partridge, "On the Anonymity of Home/Work Location Pairs," in *Pervasive '09 Proceedings of the 7th International Conference on Pervasive Computing* (Berlin, Heidelberg: Springer-Verlag, 2009): 390-97, https://crypto.stanford.edu/~pgolle/papers/commute.pdf.

[21] Alessandro Acquisti, Ralph Gross, and Fred Stutzman, "Faces of Facebook: Privacy in the Age of Augmented Reality" (presentation at BlackHat Las Vegas, Nevada, August 4, 2011). More information can be found in the FAQ on Acquisti's website: http://www.heinz.cmu.edu/~acquisti/face-recognition-study-FAQ/.

[22] "In the case of high-dimensional data, additional arrangements [beyond de-identification] may need to be pursued, such as making the data available to researchers only under tightly restricted legal agreements." Ann Cavoukian and Daniel Castro, *Big Data and Innovation, Setting the Record Straight: De-identification* Does *Work* (Toronto, Ontario: Information and Privacy Commissioner, June 16, 2014): 3.

[23] The median Facebook user has about a hundred friends. Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow, "The anatomy of the Facebook social graph," (arXiv Preprint, 2011): 3, http://arxiv.org/pdf/1111.4503v1.pdf.

[24] There are roughly ten million single nucleotide polymorphisms (SNPs) in the human genome; SNPs are the most common type of human genetic variation. "What are single nucleotide polymorphisms (SNPs)?," *Genetics Home Reference: Your Guide to Understanding Genetic Conditions*, published October 20, 2014, http://ghr.nlm.nih.gov/handbook/genomicresearch/snp.

track a large number of observations about a single individual increase. High dimensionality is one of the hallmarks of "big data."

Finally, we should note that re-identification of particular datasets is likely underreported. First, the re-identification of particular datasets is likely to be included in the academic literature only if it involves a novel advancement of techniques, so while the first use of a re-identification method may be published, reuses rarely are. Similarly, people who blog or otherwise report re-identification vulnerabilities are unlikely to do so unless interesting methods or notable datasets are involved. Second, those with malicious motivations for re-identification are probably unwilling to announce their successes. Thus, even if a specific dataset has not been re-identified publicly, it should not be presumed secure.

1.2.2 Targeted Attacks: Examples and Lessons

Another important—but often under-acknowledged—type of re-identification risk stems from adversaries who target specific individuals. 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 attack comes from 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.[25]

More recently, as mentioned in the previous section, the data from the New York City Taxi and Limousine Commission not only had especially poor de-identification practices that made broad re-identification of all drivers trivial, but also allowed for the re-identification of targeted passengers even though the dataset did not nominally contain any information about passengers. First, it is possible to identify trip records (with pickup and dropoff locations, date and time, taxi medallion or license number, and fare and tip amounts) if some of that information is already known: for example, stalkers who see their victims take a taxi to or from a particular place can determine the other endpoint of those trips.[26] Second, it is possible to identify people who regularly visit sensitive locations, such as a strip club or a religious center.[27] 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. Similar analysis can be done on the recently released Transport for London dataset, which includes not only the information in the New York taxi dataset, but also unique customer identifiers for users of the public bicycle system.[28] These violations of the privacy of passengers demonstrate problems that better ad hoc de-identification still would not fix.

[25] *DHS Data Privacy and Integrity Advisory Committee FY 2005 Meeting Materials* (June 15, 2005) (statement of Latanya Sweeney, Associate Professor of Computer Science, Technology and Policy and Director of the Data Privacy Laboratory, Carnegie Mellon University), http://www.dhs.gov/xlibrary/assets/privacy/privacy_advcom_06-2005_testimony_sweeney.pdf.

[26] Anthony Tockar, "Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset," *Neustar: Research*, September 15, 2014, http://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/.

[27] Ibid. Tockar goes on to explain how to apply differential privacy to this dataset.

[28] James Siddle, "I Know Where You Were Last Summer: London's public bike data is telling everyone where you've been," *The Variable Tree*, April 10, 2014, http://vartree.blogspot.com/2014/04/i-know-where-you-were-last-summer.html.

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—a targeted attack.[29] 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. Thus broad attacks may also be possible depending on the quantity and accuracy of information available to the adversary for cross-referencing.

A 2013 study by de Montjoye et al. revealed weaknesses in anonymized location data.[30] 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.[31] 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.[32] 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. More recent work by Hooley and Sweeney suggests that this type of chaining remains effective on public hospital discharge data from thirty U.S. states in 2013.[33]

[29] Arvind Narayanan and Vitaly Shmatikov, "Robust de-anonymization of large sparse datasets," in *Proceedings 2008 IEEE Symposium on Security and Privacy, Oakland, California, USA, May 18-21, 2008* (Los Alamitos, California: IEEE Computer Society, 2008): 111-25. The Netflix Prize dataset included movies and movie ratings for Netflix users.

[30] Yves-Alexandre de Montjoye, et al., "Unique in the Crowd: The privacy bounds of human mobility," *Scientific Reports* 3 (March 2013), doi:10.1038/srep01376.

[31] Other studies have confirmed that pairs of home and work locations can be used as unique identifiers. Golle and Partridge, "On the anonymity of home/work location pairs;" Hui Zang and Jean Bolot, "Anonymization of location data does not work: A large-scale measurement study," in *Proceedings of the 17th International Conference on Mobile Computing and Networking* (New York, New York: ACM, 2011): 145-156.

[32] A similar type of chaining in a different context can trace a user's web browsing history. A network eavesdropper can link the majority a user's web page visits to the same pseudonymous ID, which can often be linked to a real-world identity. Steven Englehardt et al., "Cookies that give you away: Evaluating the surveillance implications of web tracking," (paper accepted at 24th International World Wide Web Conference, Florence, May 2015).

[33] Sean Hooley and Latanya Sweeney, "Survey of Publicly Available State Health Databases" (White Paper 1075-1, Data Privacy Lab, Harvard University, Cambridge, Massachusetts, June 2013), http://dataprivacylab.org/projects/50states/1075-1.pdf.

2 Quantifiable Risks and Provable Privacy

Current de-identification methods are ad hoc, following a penetrate-and-patch mindset. Proponents ask whether a de-identification method can resist certain past attacks,[34] rather than insisting on affirmative evidence that the method cannot leak information regardless of what the attacker does.

The penetrate-and-patch approach is denounced in the field of computer security[35] because systems following that approach tend to fail repeatedly.[36] Ineffective as the penetrate-and-patch approach is for securing software, it is even worse for de-identification. End users will install patches to fix security bugs in order to protect their own systems, but data users have no incentive to replace a dataset found to have privacy vulnerabilities with a patched version that is no more useful to them. When no one applies patches, penetrate-and-patch becomes simply penetrate.

In addition, ad hoc de-identification makes it infeasible to quantify the risks of privacy violations stemming from a data release. Any such risk calculation must be based on assumptions about the knowledge and capabilities of all potential adversaries. As more data releases occur and more re-identification techniques are honed, such assumptions break down. Yet, accurate risk calculations are a prerequisite for well-informed policy choices, which must weigh the risks to privacy against the benefits of data releases.

These vulnerabilities of de-identification call for a shift in the focus of data privacy research, which currently suffers from ill-defined problems and unproven solutions. 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 sometimes troubled cryptog-

[34] "Thus, while [Sweeney's re-identification of Governor Weld] speaks to the inadequacy of certain de-identification methods employed in 1996, to cite it as evidence against current de-identification standards is highly misleading. If anything, it should be cited as evidence for the *improvement* of de-identification techniques and methods insofar as such attacks are no longer feasible under today's standards precisely because of this case." Cavoukian and Castro, *De-identification* Does *Work*: 5.

"Established, published, and peer-reviewed evidence shows that following contemporary good practices for de-identification ensures that the risk of re-identification is very small. In that systematic review (which is the gold standard methodology for summarizing evidence on a given topic) we found that there were 14 known re-identification attacks. Two of those were conducted on data sets that were de-identified with methods that would be defensible (i.e., they followed existing standards). The success rate of the re-identification for these two was very small." Khaled El Emam and Luk Arbuckle, "Why de-identification is a key solution for sharing data responsibly," *Future of Privacy Forum*, July 24, 2014, http://www.futureofprivacy.org/2014/07/24/de-identification-a-critical-debate/.

[35] Gary McGraw and John Viega, "Introduction to Software Security," *InformIT*, November 2, 2001, http://www.informit.com/articles/article.aspx?p=23950&seqNum=7.

[36] Anup K. Ghosh, Chuck Howell, and James A. Whittaker, "Building Software Securely from the Ground Up," *IEEE Software* (January/February 2002): 14-16.

raphy[37]—theorists need to develop usable constructs and practitioners need to adopt methods with provable privacy.

2.1 Ad Hoc De-identification Leads to Unknowable Risks

The prominence of ad hoc de-identification has led some authors to endorse ad hoc calculation of re-identification probabilities.[38] However, these calculations are specious and offer false hope about privacy protections because they depend on arbitrary and fragile assumptions about what auxiliary datasets and general knowledge are available to the adversary.

Consider an example recently cited by Cavoukian and Castro: Golle's re-examination of unique identification from U.S. census data.[39] Golle found that, using the census data from 2000, 63.3% of individuals were uniquely identifiable by year, five-digit ZIP code, and birthdate, 4.2% when birthdate was replaced by month and year of birth, and 0.2% when replaced by only birth year. Cavoukian and Castro conclude: "The more effectively the data is de-identified, the lower the percentage of individuals who are at risk of re-identification. The risk of re-identification for weakly de-identified data, such as datasets released with gender, ZIP code, and date of birth, is not the same as for strongly de-identified data."[40] It is true that making data more abstract affects re-identification risk, but the percentages can be misleading standing alone:

- The data will doubtless contain other attributes that the adversary could use for re-identification. A common technique of categorizing columns as useful or not useful for re-identification produces an overly optimistic view of re-identification risk because any column containing nontrivial data poses some risk.
- The focus on whether individuals are uniquely identifiable misses privacy violations through probabilistic inferences.[41]

In short, a released dataset without birth day and month will be less vulnerable to re-identification through purely demographic information, but the actual effect removal of that information has on re-identification depends highly on the goals and ever-expanding auxiliary data held by the adversary. Furthermore, with high-dimensional datasets, there are strong limits to how much the data can be generalized without destroying utility, whereas auxiliary information has the tendency to get more specific, accurate, and complete with each passing year.

A more specific example offered by Cavoukian and Castro comes from the Heritage Health Prize, released for a data-mining competition to predict future health outcomes based on past

[37] For example, the description for a 2012 conference notes that communication between researchers and practitioners is "currently perceived to be quite weak." "Is Cryptographic Theory Practically Relevant?," Isaac Newton Institute for Mathematical Sciences, http://www.newton.ac.uk/event/sasw07. In addition, "[m]odern crypto protocols are too complex to implement securely in software, at least without major leaps in developer know-how and engineering practices." Arvind Narayanan, "What Happened to the Crypto Dream?, Part 2," *IEEE Security & Privacy* 11, no. 3 (2013): 68-71.

[38] El Emam and Arbuckle, "Why de-identification is a key solution."

[39] Philippe Golle, "Revisiting the Uniqueness of Simple Demographics in the US Population," in *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society* (New York, New York: ACM, 2006): 77-80.

[40] Cavoukian and Castro, *De-identification* Does *Work*: 4.

[41] See Section 1.1 .

hospitalization (insurance claims) data. The dataset was de-identified by El Emam and his team,[42] and Cavoukian and Castro note that "it was estimated that the probability of re-identifying an individual was .0084."[43]

However, El Emam's estimates were derived based on a specific, somewhat arbitrary set of assumptions, such as that "the adversary would not know the exact order of the claims,"[44] in other words, that the adversary would not know that the heart attack occurred before the broken arm. Yet, adversaries could gain detailed timeline information by cross-referencing auxiliary information from online reviews of medical providers or by using personal knowledge of targeted subjects, or by using medical knowledge that certain pairs of conditions or treatments, when they occur together, tend to happen in a particular order.

In his report to the Heritage Health Prize organizers, Narayanan shows the arbitrariness of the re-identification probability calculation by using a different, but equally plausible, set of assumptions. In particular, he assumes that the adversary knows the year but not the month or day of each visit and derives dramatically different re-identification probabilities: up to 12.5% of members are vulnerable.[45]

Happily for the patients in this dataset, large-scale auxiliary databases of hospital visits and other medical information that could be used for re-identification did not appear to be available publicly at the time of the contest. However, some auxiliary information is available in the form of physician and hospital reviews on Yelp, Vitals, and other sites. Furthermore, in 2014 the Centers for Medicare & Medicaid Services publicly released detailed Medicare physician payment data, including physicians' names and addresses, summaries of services provided, and payments for services.[46] Although the Medicare data is for 2012, it is easy to imagine that such data could have been released for the time period spanned by the contest dataset instead and used to match particular providers with contest records. Physician and hospital reviews could then more easily be matched to those records, and more patients identified. In addition, though this Medicare dataset does not include dates, the safe harbor HIPAA de-identification standards permit inclusion of the year for admission and discharge dates;[47] it is plausible that future releases could include such information and make Narayanan's assumptions clearly more valid than El Emam's.

[42] Khaled El Emam et al., "De-identification methods for open health data: the case of the Heritage Health Prize claims dataset," *Journal of Medical Internet Research* 14, no. 1 (2012): e33, doi:10.2196/jmir.2001.

[43] Cavoukian and Castro, *De-identification* Does *Work*: 11.

[44] El Emam et al., "Heritage Health"

[45] Arvind Narayanan, "An Adversarial Analysis of the Reidentifiability of the Heritage Health Prize Dataset" (unpublished manuscript, 2011).

[46] The dataset "contains information on utilization, payment (allowed amount and Medicare payment), and submitted charges organized by National Provider Identifier (NPI), Healthcare Common Procedure Coding System (HCPCS) code, and place of service." "Medicare Provider Utilization and Payment Data: Physician and Other Supplier," Centers for Medicare & Medicaid Services, last modified April 23, 2014, http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html.

[47] "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," U.S. Department of Health & Human Services, http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html.

The later release of publicly available auxiliary information like the Medicare data could enable a broad attack unaccounted for in the initial re-identification probability estimates. The possibility of such future releases can never be ruled out. Even without such a data release, the contest data is vulnerable to targeted attacks by adversaries with specific knowledge about people in the dataset.

It is very tempting to look for assurances about the probability of privacy violations from an ad hoc de-identified dataset, but there is simply no scientific basis for interpreting ad hoc re-identification probability estimates of ad hoc de-identified high-dimensional datasets as anything more than (weak) lower bounds. Ad hoc estimates tend to be based on many assumptions, so that the probability claims must be accompanied by multiple caveats. In practice, the caveats likely will be lost, as they were when Cavoukian and Castro cited El Emam's 0.0084 probability without noting any of the assumptions that El Emam details in his paper. Rigorously quantified privacy risks are only possible when using methods designed to allow for such calculations.

2.2 The Promise of Provable Privacy

As noted earlier, data releases are permanent and re-identification capabilities are improving, making protocols and systems with proven privacy properties an urgent need. The foundation for such protocols and systems are methods of handling data that preserve a rigorously defined privacy, even in the face of unpredicted advances in data analysis, while also permitting useful analysis. At present, algorithms that yield differential privacy are the only well-developed methodology that satisfies these requirements.

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 this type of 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 algorithms[48] typically add "noise" — small, quantified error — to the outputs of 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.

Differential privacy is a criterion for privacy. Different algorithms can satisfy this criterion in different ways, and the approach to achieving differential privacy might differ from case to case, although the privacy criterion stays the same.

[48] The following sources contain introductions to differential privacy. Cynthia Dwork et al., "Differential Privacy - A Primer for the Perplexed" (paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Tarragona, Spain, October 2011); Erica Klarreich, "Privacy by the Numbers: A New Approach to Safeguarding Data," *Quanta Magazine* (December 10, 2012); Christine Task, "An Illustrated Primer in Differential Privacy," *XRDS* 20, no. 1 (2013): 53-57.

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 must be.[49] 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 do not depend on artificial assumptions about the adversary's capabilities or access to auxiliary information. 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 due to that person's inclusion in the dataset remains strictly limited.

Given these advantages, differential privacy is a valuable tool for data privacy. Further research is needed on 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.

3 Practical Steps towards Improved Data Privacy

Given the weaknesses of ad hoc de-identification and the nascent state of provable privacy research, we turn to the difficult policy question of how to handle current datasets: 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 combining the dataset with others, which may escalate both the losses and the gains.

In this Section, we explain why releasing datasets to the public using ad hoc de-identification methods should not be the default policy. Then, we consider methods by which policymakers can push the default to be access using provable privacy methods or restricted access to a narrow audience. The individualized nature of each dataset access means that one-size-fits-all solutions must be either incomplete or incorrect—certain broad policies may be useful, but no single rule for dealing with all data access will give good results in every case. We offer policy recommendations below that promote a more cautious and more tailored approach to releasing data: 1) incentivize the development and use of provable privacy methods and 2) encourage narrower data accesses that still permit analysis and innovation. Finally, we argue for increased transparency around re-identification risks to raise public awareness and to bolster the other recommendations.

3.1 Defining a Precautionary Approach

The precautionary principle deals with decision-making and risk regulation in the face of scientific uncertainty. It has many, much-debated formulations, ranging from very weak (for example, that regulation should be permitted when risks are uncertain) to very strong (for example, that any action with an uncertain risk should be barred completely until the actor can prove that the risks are acceptable). We do not wish to engage in the debate over the general formulation of the principle and the breadth of its applicability. Instead, we focus on the specific problem of how to react to the unknowable risks of ad hoc de-identification. Precautionary approaches often shift where the burden of proof for the decision about an action falls when risks are uncertain, and we

[49] Ohm, "Broken Promises of Privacy": 1752-55.

argue that placing the burden more heavily on data providers will yield better results than the status quo.

The difficulty at the heart of this issue is weighing uncertain privacy risks against uncertain data access benefits. The loss of these benefits—such as potential medical advances or research progress from wider data sharing—is also legitimately characterized as an uncertain risk. The impossibility of completely avoiding both uncertain risks has led to Sunstein's criticism of strong versions of the precautionary principle for creating paralysis by "forbid[ding] all courses of action, including inaction."[50] However, like most proponents of precautionary approaches, we do "not impose a burden on any party to prove zero risk, nor…state that all activities that pose a possible risk must be prohibited."[51] Instead, we see a way forward by altering default behaviors and incentives.

Currently, there is a presumption that data releases to the public are acceptable as long as they use ad hoc de-identification and strip out classes of information deemed to be PII. This presumption draws a line and the burden of proof shifts when it is crossed: if data providers have used ad hoc de-identification and removed PII, then the burden of proof falls on privacy advocates to show that the particular datasets are re-identifiable or could cause other harms; if data providers have not done so, then they are obliged to demonstrate why data releases that do not conform to standard practices are permissible.

We argue that this line—and the attendant standard practices—should shift. A spectrum of choices for the line exist, with the endpoints completely prioritizing data access or privacy, and current standards lean too far towards data access. Ad hoc de-identification has unknowable risks, and the continued release of ad hoc de-identified data presents the threat of unacceptable widespread re-identification of past datasets. In addition, data providers have the power to limit their data releases and reduce those risks. As such, release of ad hoc de-identified data to the entire public should require justification; it should not be the default behavior. Parties releasing data using ad hoc de-identification methods should have the responsibility, at a minimum, to limit that release to the narrowest possible scope likely to yield the intended benefit.[52]

Ad hoc de-identification is useful to practitioners as an additional layer of defense. However, we join PCAST in urging policymakers to stop relying on it and to stop treating it as a sufficient privacy protection on its own.

Alternatively, data providers could avoid the uncertainty of ad hoc de-identification and the need to take precautionary measures by using provable privacy methods instead. Because provable privacy methods have precisely calculable risks, they allow for traditional risk-benefit analyses and remove the possibility of snowballing re-identification risk that comes with continued unfettered release of data using ad hoc de-identification.

[50] Cass R. Sunstein, "The Paralyzing Principle," *Regulation* 25, no. 4 (2002): 33-35.

[51] Noah M. Sachs, "Rescuing the Strong Precautionary Principle from Its Critics," *Illinois Law Review* 2011 no.4 (2011): 1313.

[52] Alternatively, a data provider could show that the expected benefit outweighs the privacy cost of complete re-identification of the entire dataset. In other words, the data provider would need to show that there still would be a net benefit from releasing the data even if the names of all individuals involved were attached to their records in the dataset. This standard would be, in most cases, significantly more restrictive.

3.2 Researching and Implementing Provable Privacy

Additional funding for provable privacy research is the clearest way to encourage development of provable privacy methods. However, such methods are necessary, but not sufficient, for responsible data practices because once they exist, they still need to be deployed widely. Achieving broad adoption of those methods is as much a social and policy problem as a technical one.

We emphasize two main goals to help propagate these methods and create more real-world applications of provable privacy like the U.S. Census Bureau's OnTheMap[53] and Google's RAPPOR[54]. First, privacy researchers must communicate with data scientists so that the theoretical privacy work is developed with practical uses in mind. Second, data scientists must accept and use these new methodologies.

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 scientists 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.

Privacy research funding can encourage collaborations with or feedback from practitioners. Data science funding can favor projects that implement provable privacy methods instead of ad hoc 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 involving unproven and risky data privacy practices.

Governments can also encourage development of provable privacy by entering the market for such technologies as a consumer or by making data available under a provably private interface. Innovation procurement—using government demand to drive the development and diffusion of new products or processes—has gained support,[55] particularly in Europe.[56] Provable privacy technologies appear to be a good candidate for this kind of stimulus, as purchasing systems based on these technologies can fulfill both innovation goals and the core goals of obtaining high-quality, useful products for the public sector.[57] Similarly, providing government data through a differential privacy-based interface would serve both innovation and privacy goals by incentivizing data users to learn how to use such interfaces and protecting the people included in the datasets.

[53] "OnTheMap," U.S. Census Bureau, http://onthemap.ces.census.gov/; Klarreich, "Privacy by the Numbers."

[54] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova, "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (Scottsdale, Arizona: ACM, 2014): 1054-67.

[55] Jakob Edler and Luke Georghiou, "Public procurement and innovation—Resurrecting the demand side," *Research Policy* 36, no. 7 (September 2007): 949-63.

[56] Charles Edquist and Jon Mikel Zabala-Iturriagagoitia, "Public Procurement for Innovation as mission-oriented innovation policy," *Research Policy* 41, no. 10 (December 2012): 1757-69.

[57] Elvira Uyarra and Kieron Flanagan, "Understanding the Innovation Impacts of Public Procurement," *European Planning Studies* 18, no. 1 (2010): 123-43.

3.3 Flexible Options for Narrower Releases of Data

Although we argue that data providers should justify public releases of datasets that use ad hoc de-identification methods, we do not recommend hardening that burden of proof into a single legal or regulatory requirement. Because dataset releases are highly individualized, a universal one-size-fits-all requirement would lead to sub-optimal results in many cases. Instead, the burden-of-proof concept can be considered a guiding principle for an array of more flexible policy choices that can be tailored to particular circumstances, as the case studies in the next part demonstrate. Here we list, for both data custodians and policymakers, some of the considerations—not mutually exclusive—that may help in determining the appropriate scope for the release of datasets:

- Is it possible to use a provable privacy method and thus get an accurate calculation of the privacy risks to weigh against the expected benefit?
- Is it possible to host data on the custodian's system and allow researchers to query it, instead of releasing the dataset?
- Can all or most of the intended benefit of data release be achieved by computing and releasing aggregate statistics instead of raw micro-data?
- Is a limited release similarly useful? Are the people most likely to use the data beneficially a subset of the general public: researchers, affiliates of educational institutions, data analysts with past successes?
- Can multiple forms of the dataset be released so that only those who have demonstrated effectiveness or a need for more vulnerable datasets receive them?
- Can data recipients be required to sign legal contracts restricting their use and transfer of the dataset?
- Can data recipients be required to undergo ethics training?
- Can data recipients be required to provide certain information: identification, a statement of purpose for obtaining the data?

These questions can help determine whether a narrower release of a dataset is wise, and we think that it almost never will be the case that an unlimited release of a dataset to the entire public will be the optimal choice.

3.4 Enabling Transparency of Re-identification Risks

Privacy is, at least in part,[58] an individual right, and as such, transparency about data usage and data flows is a natural response to big data privacy concerns. Such transparency has appeared as a central tenet in governmental pronouncements on big data: for example, the U.K.'s Information Commissioner's Office includes transparency among the "practical aspects to consider when us-

[58] Solove, among others, has discussed how privacy is traditionally viewed as an individual right but also has social value. Daniel J. Solove, "'I've Got Nothing to Hide' and Other Misunderstandings of Privacy," *San Diego Law Review* 44 (2007): 760-64.

ing personal data in big data analytics,"[59] and the U.S. White House makes transparency one of the seven rights in its Consumer Privacy Bill of Rights.[60]

This transparency should include informing people about re-identification risks stemming from data collected about them. Knowledge about the possibility of re-identification is necessary "to enabl[e] consumers to gain a meaningful understanding of privacy risks and the ability to exercise Individual Control."[61] We propose that, wherever notice about data collection can be given, a short statement should be included that briefly describes what steps will be taken to protect privacy and notes whether records may be re-identified despite those steps. Users also should be able to access further details about the privacy protection measures easily, perhaps through a link in the notice. Among the available details should be a justification for the protective steps taken, describing why the provider has confidence that re-identification will not occur.

Giving users information about privacy protection measures and re-identification risks helps to even the information asymmetry between them and data collectors.[62] It would allow users to make more informed decisions and could motivate more conscientious privacy practices, including the implementation of provable privacy methods. It is also possible that data collectors could give users options about the privacy protection measures to be applied to their information. Such segmentation would permit personal assessments of the risks and benefits of the data collection: people who have strong desires for privacy could choose heavier protections or non-participation; people who do not care about being identified or who strongly support the potential research could choose lighter, or no, protections.[63] This segmentation is a helpful complement to narrowed releases of data: instead of restricting access to the people who can create the most benefit, segmentation restricts participation to the people who feel the least risk.

4 Specific Advice for Six Common Cases

Now we turn to six of the most common cases in which we believe it is particularly important for data custodians to look beyond ad hoc de-identification for privacy protection. In each case, we present recommendations for data custodians and policymakers, providing real-world applications of the risk-benefit assessments and policy tools described in Section 3.

[59] Information Commissioner's Office, *Big data and data protection* (July 28, 2014): 5-6, 33-37.

[60] The White House, Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy (Washington, D.C.: February 2012): 47.

[61] Ibid.

[62] Of course, simply providing information can be insufficient to protect users. It may not "be information that consumers can use, presented in a way they can use it," and so it may be ignored or misunderstood. Lawrence Lessig, "Against Transparency," *New Republic*, October 9, 2009. Alternatively, a user may be informed effectively but the barriers to opting out may be so high as to render the choice illusory. Janet Vertesi, "My Experiment Opting Out of Big Data Made Me Look Like a Criminal," *Time*, May 1, 2014. Still, we believe that concise, clear descriptions of privacy protecting measures and re-identification risks can aid users in many circumstances and should be included in the options considered by policymakers.

[63] For example, patients in clinical trials or with rare diseases might wish to have their data included for analysis, even if the risk of re-identification is high or if no privacy protecting measures are taken at all. Kerstin Forsberg, "De-identification and Informed Consent in Clinical Trials," Linked Data for Enterprises, November 17, 2013, http://kerfors.blogspot.com/2013/11/de-identification-and-informed-consent.html.

Case 0: “No PII” as a putative justification for data collection.

Companies that track user activities—often without notice or choice—frequently proffer the argument that they do not collect PII in response to privacy concerns. Third-party online tracking is a prime example—U.S. online advertising self-regulation treats PII as the primary dividing line between acceptable and unacceptable tracking.[64] Mobile apps and mall tracking based on WiFi signals are others.

Of course, we should expect that such datasets can be re-identified, and even *accidental* leaks of identity to tracking companies are rampant online.[65] As such, we recommend that policymakers and regulators not consider the absence of deliberate PII collection to be an adequate privacy safeguard. Additional privacy measures include aggregation[66] and data minimization. Requiring affirmative consent for tracking, encouraging the development of easy-to-use opt-out mechanisms, and funding the development of technical defense mechanisms are fruitful policy directions as well.

Online privacy is often a proxy for other worries such as targeting of protected groups and data-driven discrimination.[67] These worries are just as serious whether or not PII is involved or re-identification takes place. In recent years a combination of press reporting,[68] empirical research,[69] and theory[70] has helped clarify the nature of these dangers. As a result, policy makers' attention has gradually shifted to data use in addition to data collection. While restrictions on collection continue to be important, we encourage the trend toward monitoring data use and developing norms and rules.

Case 1: Companies selling data to one another.

When privacy laws place use limits on customer information, there is typically a carve-out for “anonymized” records. For example, both the EU Data Protection Directive and the proposed General Data Protection Regulation place more stringent restrictions on “personal data”: the former defines “personal data” as “information relating to an identified or identifiable natural

[64] For example, the Network Advertising Initiative’s self-regulatory Code “provides disincentives to the use of PII for Interest-Based Advertising. As a result, NAI member companies generally use only information that is not PII for Interest Based Advertising and do not merge the non-PII they collect for Interest-Based Advertising with users’ PII.” “Understanding Online Advertising: Frequently Asked Questions,” Network Advertising Initiative, http://www.networkadvertising.org/faq.

[65] Balachander Krishnamurthy and Craig E. Wills, “On the Leakage of Personally Identifiable Information Via Online Social Networks,” in *Proceedings of the 2nd ACM Workshop on Online Social Networks* (New York, New York: ACM, 2009): 7-12, http://www2.research.att.com/~bala/papers/wosn09.pdf.

[66] Data aggregation replaces individual data elements by statistical summaries.

[67] Cynthia Dwork and Deirdre K. Mulligan, “It's not privacy, and it's not fair,” *Stanford Law Review Online* 66 (2013): 35.

[68] Julia Angwin, “The web’s new gold mine: Your secrets,” *Wall Street Journal*, July 30, 2010.

[69] Aniko Hannak et al., “Measuring Price Discrimination and Steering on E-commerce Web Sites,” in *Proceedings of the 2014 Conference on Internet Measurement Conference* (Vancouver: ACM, 2014): 305-318.

[70] Solon Barocas and Andrew D. Selbst, “Big Data's Disparate Impact,” *California Law Review* 104 (forthcoming); Ryan Calo, “Digital Market Manipulation,” *George Washington Law Review* 82 (2014): 995.

person"[71]; the latter defines it as "any information relating to a data subject," who is someone who "can be identified, directly or indirectly, by means reasonably likely to be used."[72] These definitions were constructed to provide safe harbors for anonymized data.[73] However, they are only as strong as the anonymization method used. In the case of ad hoc anonymization, re-identification science has shown that such exceptions are not well-founded. It is unclear whether the EU rules will be interpreted to create loopholes or to apply stringent requirements to all data collection and release; other statutes and regulations have more explicit carve-outs for data that omits specific PII, and these rules will create more loopholes.

We call for a move away from such exceptions in future lawmaking and rulemaking, except in cases where strong provable privacy methods are used. Meanwhile, we make two recommendations to minimize privacy risks in domains in which such loopholes do or may exist. First, data custodians must use legal agreements to restrict the flow and use of data—in particular, to prohibit resale of such datasets and specify acceptable uses including limits on retention periods. Second, policymakers should increase the transparency of the data economy by requiring disclosures of "anonymized" data sharing in privacy policies. This change will fix the current information asymmetry between firms and consumers and allow the market to price privacy more efficiently.

Case 2: Scientific research on data collected by companies.

From telephone call graphs to medical records, customer data collected by private companies has always been tremendously valuable for scientific research. The burgeoning field of computational social science has made great strides in adapting online self-reported data, such as information on social networks, for drawing statistically sound conclusions.[74] Such data were previously considered less useful for research but this thinking is being overturned.

Privacy and re-identification risks are again a vexing concern if these companies are to open their datasets to external researchers. The silver lining is that the largest companies with the most interesting research datasets usually have in-house research teams—AT&T, Microsoft, and more

[71] Directive 95/46/EC, of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, Art. 2(a), 1995 O.J. (C 93).

[72] Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, Art. 4(1)-(2) (January 25, 2012).

[73] Ohm, "Broken Promises of Privacy": 1704, 1738-41.

[74] Pablo Barberá, "How Social Medial Reduces Mass Political Polarization: Evidence from Germany, Spain, and the U.S." (unpublished manuscript, October 18, 2014), https://files.nyu.edu/pba220/public/barbera-polarization-social-media.pdf; Amaney Jamal et al., "Anti-Americanism and Anti-Interventionism in Arabic Twitter Discourses" (unpublished manuscript, October 20, 2014), http://scholar.harvard.edu/files/dtingley/files/aatext.pdf; Margaret E. Roberts, "Fear or Friction? How Censorship Slows the Spread of Information in the Digital Age" (unpublished manuscript, September 26, 2014), http://scholar.harvard.edu/files/mroberts/files/fearfriction_1.pdf.

Computational social scientists can also generate their own self-reported data online. Matthew J. Salganik and Karen E.C. Levy, "Wiki surveys: Open and quantifiable social data collection" (unpublished manuscript, October 2, 2014), http://arxiv.org/abs/1202.0500.

recently, Facebook are good examples. However, there are two problems with relying on in-house research; we now discuss these problems and potential solutions.

First, benefits from published research have large positive externalities, often far exceeding the benefits to the firm, which include improved reputation or increased knowledge about users. So, economic theory would predict that these research teams will be smaller than the public would want them to be. Rather than dealing with this externality by encouraging public release of company data, governments should seek ways to incentivize research publications of this type with fewer privacy implications, such as by sponsoring programs for academic researchers in visiting positions at companies.

Second, in-house research may not be reproducible. However, much of the interesting user research at companies seems to involve interventional experiments on users. For such experiments, publishing data will not enable reproducibility, and the best option for verifying results is for the company to permit outside researchers to visit and re-run experiments on new batches of users. When access to the data would help with reproducibility, we would follow the recommendations laid out below in Case 4 for scientific research in general.

Case 3: Data mining contests.

The ease of data collection means that even small companies that cannot afford in-house research teams often have interesting datasets for scientific research or knowledge discovery—colloquially termed data mining. Data mining contests, such as the Netflix prize discussed above, have recently gained popularity as a way for such companies to incentivize research that utilizes their data.

Such contests are spurs to innovation, and the most effective scope for data release depends on balancing two factors: having more contestants reduces their motivation because they become less likely to win, but it also increases the chance of having a contestant put forth a rare solution.[75] As such, Boudreau, Lacetera, and Lakhani have concluded that expansive competitions are most useful for problems where the solutions are highly uncertain, including multi-domain problems where it is less clear who would solve them best and how.[76] Jeppesen and Lakhani also suggest that broadening the scope of contestants can bring in people on the margins of the technical fields and social groups primarily associated with the contest problem and that those marginal people are more likely to succeed in these contests.[77]

We make three recommendations for data custodians running contests:

- Consider whether the group of contestants can be narrowed. If the solution desired is less uncertain, perhaps because it lies in a single domain or known methodologies are expected to work, research suggests that a contest between few participants can be more ef-

[75] Kevin J. Boudreau, Nicola Lacetera, and Karim R. Lakhani, “Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis,” *Management Science* 57, no. 5 (2014): 843-63, doi: 10.1287/mnsc.1110.1322.
[76] Ibid., 860-61.
[77] Lars Bo Jeppesen and Karim R. Lakhani, “Marginality and Problem Solving Effectiveness in Broadcast Search,” *Organization Science* 21, no. 5 (2010): 1016-33.

fective. It may also be possible to invite participants with diverse backgrounds and views to provide the advantage from marginal contestants, though we recognize that identifying such people may be difficult because they are on the margins.

- Whenever possible, switch to a model in which data is made available under provable privacy guarantees. We expect that the expense and development effort involved in applying the appropriate data transformations and carrying out privacy analyses will be similar to the current process of data pre-processing and evaluating de-identification methods. Contest organizers are in a good position to effect a behavior change among data scientists because of the financial incentives.
- If de-identified data is released, use a multi-stage process. Early stages can limit the amount or type of data released by releasing data on only a subset of users, minimizing the quantitative risk, or by releasing a synthetic dataset created to mimic the characteristics of the real data.[78] Later stages can permit access to a broader dataset but add some combination of the following restrictions: requiring contestants to sign a data-use agreement; restricting the contest to a shortlist of best performers from the first stage; and switching to an "online computation model" where participants upload code to the data custodian's server (or make database queries over its network) and obtain results, rather than download data.

Case 4: Scientific research, in general.

Nearly all scientific research on human subjects would be improved if data could be shared more freely among researchers, enhancing efficiency and reproducibility. These advantages have led to calls for open data, which can be interpreted as advocating the public release of datasets used in research. However, the gains come predominantly from scientists having the data, and so restricted access to a data-sharing system is a good solution in this area.[79] Such a system should implement various gatekeeping functions, such as demanding proof of academic or peer-reviewed standing, requiring ethical training, and designing and overseeing the security of the system.[80] In addition, government research funding can incentivize scientists to use provable privacy methods.

A good example of gatekeeping is the U.S. State Inpatient Databases (SIDs) developed for the Healthcare Cost and Utilization Project sponsored by the Agency for Healthcare Research and Quality (AHRQ). AHRQ wishes this data to be used more broadly than just among scientific researchers, but it is cognizant of the very serious re-identification risk presented by the datasets. Obtaining them involves a number of steps:[81] completing an online Data Use Agreement Train-

[78] Researchers already have developed methods for creating such synthetic data. Avrim Blum, Katrina Ligett, and Aaron Roth, "A Learning Theory Approach to Non-Interactive Database Privacy," in *Proceedings of the 40th ACM SIGACT Symposium on Theory of Computing* (Victoria, British Columbia: ACM, 2008).

[79] "If there are privacy concerns I can imagine ensuring we can share the data in a 'walled garden' within which other researchers, but not the public, will be able to access the data and verify results." Victoria Stodden, "Data access going the way of journal article access? Insist on open data," *Victoria's Blog*, December 24, 2012, http://blog.stodden.net/2012/12/24/data-access-going-the-way-of-journal-article-access/.

[80] Genomics researchers have proposed one such system. Bartha Maria Knoppers, et al., "Towards a data sharing Code of Conduct for international genomic research," *Genome Medicine* 3 (2011): 46.

[81] HCUP, SID/SASD/SEDD Application Kit (October 15, 2014), http://www.hcup-us.ahrq.gov/db/state/SIDSASDSEDD_Final.pdf.

ing Course; paying a fee; providing information including name, address, and type of organization; describing the intended project, areas of investigation, potential uses of any products created, and reasons for requesting the data; and physically signing a data-use agreement that prohibits the use of the data "to identify any person"—this last requirement could be further strengthened by defining identification to include any use of the data "to infer information about, or otherwise link the data to, a particular person, computer, or other device."[82]

Case 5: Open government data.

In one sense, open government data may be the most difficult case because most of our earlier prescriptions do not apply. First, in most cases there is no ability to opt out of data collection. Second, while some research could be done in-house by government agencies, it is not possible to anticipate all beneficial uses of the data by external researchers, and the data is not collected for a specific research purpose. Finally, restricting access runs contrary to the transparency goals of improving government by shedding light on its practices.

However, in another sense, re-identification worries are minimal because the vast majority of open government datasets do not consist of longitudinal observations of individuals. Interestingly, for a variety of datasets ranging from consumer complaints to broadband performance measurement, the data is not *intended* to track users longitudinally, but it might *accidentally* enable such tracking if there is enough information about the user in each measurement data point. To prevent such accidental linkability, de-identification is indeed a valuable approach.

Certain aggregate or low-dimensional government data, such as many of the datasets published by the U.S. Census Bureau, seem to avoid privacy violations fairly well by using statistical disclosure control methodologies. However, high-dimensional data is problematic, and there is no reason to expect it cannot be de-anonymized. For these datasets, it seems that the best solution is to implement provable privacy techniques, as the Census Bureau did with its OnTheMap data, or to wait to release such data until provable privacy techniques can be implemented satisfactorily.

These cases illustrate how our various policy recommendations can be applied to practical situations, and the variation among the recommendations demonstrates the importance of a flexible policy response. Data custodians and policymakers will need to make granular decisions about the risks and benefits of releasing specific datasets, and we hope that the factors and examples in this paper will serve as a guide.

[82] Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers* (Washington, DC: March 2012) 21, http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf.

Bibliography

Acquisti, Alessandro, Ralph Gross, and Fred Stutzman. “Faces of Facebook: Privacy in the Age of Augmented Reality.” Presentation at BlackHat Las Vegas, Nevada, August 4, 2011.

Adida, Ben. “Don’t Hash Secrets.” *Benlog*, June 19, 2008. http://benlog.com/2008/06/19/dont-hash-secrets/.

Angwin, Julia. “The web’s new gold mine: Your secrets.” *Wall Street Journal*, July 30, 2010.

Barbaro, Michael and Tom Zeller, Jr. “A Face Is Exposed for AOL Searcher No. 4417749.” *New York Times*, August 9, 2006. http://www.nytimes.com/2006/08/09/technology/09aol.html.

Barberá, Pablo. “How Social Medial Reduces Mass Political Polarization: Evidence from Germany, Spain, and the U.S.” Unpublished manuscript, October 18, 2014. https://files.nyu.edu/pba220/public/barbera-polarization-social-media.pdf.

Barocas, Solon and Andrew D. Selbst. “Big Data's Disparate Impact,” *California Law Review* 104 (forthcoming).

Barocas, Solon and Helen Nissenbaum. “Big Data’s End Run Around Anonymity and Consent.” In *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, edited by Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nissenbaum, 44-75. New York: Cambridge University Press, 2014.

Blum, Avrim, Katrina Ligett, and Aaron Roth. “A Learning Theory Approach to Non-Interactive Database Privacy.” In *Proceedings of the 40th ACM SIGACT Symposium on Theory of Computing*. Victoria, British Columbia: ACM, 2008.

Boudreau, Kevin J., Nicola Lacetera, and Karim R. Lakhani. “Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis.” *Management Science* 57, no. 5 (2014): 843-63. doi: 10.1287/mnsc.1110.1322.

Calo, Ryan. “Digital Market Manipulation,” *George Washington Law Review* 82 (2014): 995-1051.

Cavoukian, Ann and Daniel Castro. *Big Data and Innovation, Setting the Record Straight: De-identification* Does *Work*. Toronto, Ontario: Information and Privacy Commissioner, June 16, 2014.

de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. “Unique in the Crowd: The privacy bounds of human mobility.” *Scientific Reports* 3 (March 2013). doi:10.1038/srep01376.

Dey, Ratan,Yuan Ding, and Keith W. Ross. “The High-School Profiling Attack: How Online Privacy Laws Can Actually Increase Minors’ Risk.” Paper presented at the 13th Privacy Enhancing Technologies Symposium, Bloomington, IN, July 12, 2013. https://www.petsymposium.org/2013/papers/dey-profiling.pdf.

DHS Data Privacy and Integrity Advisory Committee FY 2005 Meeting Materials (June 15, 2005). Statement of Latanya Sweeney, Associate Professor of Computer Science, Technology and Policy and Director of the Data Privacy Laboratory, Carnegie Mellon University). http://www.dhs.gov/xlibrary/assets/privacy/privacy_advcom_06-2005_testimony_sweeney.pdf.

Directive 95/46/EC, of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data, Art. 2(a), 1995 O.J. (C 93).

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. “Differential Privacy - A Primer for the Perplexed.” Paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Tarragona, Spain, October 2011.

Dwork, Cynthia and Deirdre K. Mulligan. “It's not privacy, and it's not fair.” *Stanford Law Review Online* 66 (2013): 35-40.

Edler, Jakob and Luke Georghiou. “Public procurement and innovation—Resurrecting the demand side.” *Research Policy* 36, no. 7 (September 2007): 949-63.

Edquist, Charles and Jon Mikel Zabala-Iturriagagoitia. “Public Procurement for Innovation as mission-oriented innovation policy.” *Research Policy* 41, no. 10 (December 2012): 1757-69.

El Emam, Khaled, Luk Arbuckle, Gunes Koru, Benjamin Eze, Lisa Gaudette, Emilio Neri, Sean Rose, Jeremy Howard, and Jonathan Gluck. “De-identification methods for open health data: the case of the Heritage Health Prize claims dataset.” *Journal of Medical Internet Research* 14, no. 1 (2012): e33. doi:10.2196/jmir.2001.

El Emam, Khaled and Luk Arbuckle. “Why de-identification is a key solution for sharing data responsibly.” *Future of Privacy Forum*, July 24, 2014. http://www.futureofprivacy.org/2014/07/24/de-identification-a-critical-debate/.

Englehardt, Steven, Dillon Reisman, Christian Eubank, Peter Zimmerman, Jonathan Mayer, Arvind Narayanan, and Edward W. Felten. “Cookies that give you away: Evaluating the surveillance implications of web tracking.” Paper accepted at 24th International World Wide Web Conference, Florence, May 2015.

Erlingsson, Úlfar, Vasyl Pihur, and Aleksandra Korolova. “RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response.” In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 1054-67. Scottsdale, Arizona: ACM, 2014.

Executive Office of the President, President's Council of Advisors on Science and Technology, *Report to the President: Big Data and Privacy: A Technological Perspective* (Washington, DC: 2014).

Federal Trade Commission, *Protecting Consumer Privacy in an Era of Rapid Change: Recommendations for Businesses and Policymakers* (Washington, DC: March 2012). http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf.

Felten, Ed. "Are pseudonyms 'anonymous'?" *Tech@FTC*, April 30, 2012. https://techatftc.wordpress.com/2012/04/30/are-pseudonyms-anonymous/.

Felten, Ed. "Does Hashing Make Data 'Anonymous'?" *Tech@FTC*, April 22, 2012. https://techatftc.wordpress.com/2012/04/22/does-hashing-make-data-anonymous/.

Forsberg, Kerstin. "De-identification and Informed Consent in Clinical Trials." Linked Data for Enterprises, November 17, 2013. http://kerfors.blogspot.com/2013/11/de-identification-and-informed-consent.html.

Gagnon, Michael N. "Hashing IMEI numbers does not protect privacy." *Dasient Blog*, July 26, 2011. http://blog.dasient.com/2011/07/hashing-imei-numbers-does-not-protect.html.

Ghosh, Anup K., Chuck Howell, and James A. Whittaker. "Building Software Securely from the Ground Up." *IEEE Software* (January/February 2002).

Golle, Philippe and Kurt Partridge. "On the Anonymity of Home/Work Location Pairs." In *Pervasive '09 Proceedings of the 7th International Conference on Pervasive Computing*, 390-97. Berlin, Heidelberg: Springer-Verlag, 2009. https://crypto.stanford.edu/~pgolle/papers/commute.pdf.

Golle, Philippe. "Revisiting the Uniqueness of Simple Demographics in the US Population." In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society*, 77-80. New York, New York: ACM, 2006.

"Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," U.S. Department of Health & Human Services. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html.

Gymrek, Melissa, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. "Identifying Personal Genomes by Surname Inference." *Science* 339, no. 6117 (January 2013): 321-24. doi:10.1126/science.1229566.

Hannak, Aniko, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. "Measuring Price Discrimination and Steering on E-commerce Web Sites." In *Proceedings of the 2014 Conference on Internet Measurement Conference*, 305-318. Vancouver: ACM, 2014.

HCUP, SID/SASD/SEDD Application Kit (October 15, 2014). http://www.hcup-us.ahrq.gov/db/state/SIDSASDSEDD_Final.pdf.

Hooley, Sean and Latanya Sweeney. "Survey of Publicly Available State Health Databases." White Paper 1075-1, Data Privacy Lab, Harvard University, Cambridge, Massachusetts, June 2013. http://dataprivacylab.org/projects/50states/1075-1.pdf.

Information Commissioner's Office. *Big data and data protection* (July 28, 2014).

"Is Cryptographic Theory Practically Relevant?" Isaac Newton Institute for Mathematical Sciences. http://www.newton.ac.uk/event/sasw07.

Jamal, Amaney, Robert O. Keohane, David Romney, and Dustin Tingley. "Anti-Americanism and Anti-Interventionism in Arabic Twitter Discourses." Unpublished manuscript, October 20, 2014. http://scholar.harvard.edu/files/dtingley/files/aatext.pdf.

Jeppesen, Lars Bo and Karim R. Lakhani. "Marginality and Problem Solving Effectiveness in Broadcast Search." *Organization Science* 21, no. 5 (2010): 1016-33.

Klarreich, Erica. "Privacy by the Numbers: A New Approach to Safeguarding Data." *Quanta Magazine* (December 10, 2012).

Knoppers, Bartha Maria, Jennifer R. Harris, Anne Marie Tassé, Isabelle Budin-Ljøsne, Jane Kaye, Mylène Deschênes, and Ma'n H Zawati. "Towards a data sharing Code of Conduct for international genomic research," *Genome Medicine* 3 (2011): 46.

Krishnamurthy, Balachander and Craig E. Wills. "On the Leakage of Personally Identifiable Information Via Online Social Networks." In *Proceedings of the 2nd ACM Workshop on Online Social Networks*. New York, New York: ACM, 2009. http://www2.research.att.com/~bala/papers/wosn09.pdf.

Lessig, Lawrence. "Against Transparency," *New Republic*, October 9, 2009.

Lewis, Paul and Dominic Rushe, "Revealed: how Whisper app tracks 'anonymous' users." *The Guardian*, October 16, 2014. http://www.theguardian.com/world/2014/oct/16/-sp-revealed-whisper-app-tracking-users.

Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian, "*t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity." In *IEEE 23rd International Conference on Data Engineering, 2007*, 106-15. Piscataway, NJ: IEEE, 2007.

Machanavajjhala, Ashwin, Johannes Gehrke, Daniel Kifer, and Muthuramakrishnan Venkitasubramanian. “*l*-diversity: Privacy beyond *k*-anonymity.” *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, no. 1 (2007): 3.

McGraw, Gary and John Viega. “Introduction to Software Security.” *InformIT*, November 2, 2001. http://www.informit.com/articles/article.aspx?p=23950&seqNum=7.

“Medicare Provider Utilization and Payment Data: Physician and Other Supplier.” Centers for Medicare & Medicaid Services. Last modified April 23, 2014. http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html.

Narayanan, Arvind. “An Adversarial Analysis of the Reidentifiability of the Heritage Health Prize Dataset.” Unpublished manuscript, 2011.

Narayanan, Arvind and Vitaly Shmatikov, “De-anonymizing Social Networks.” In *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, 173-87. Washington, D.C.: IEEE Computer Society, 2009).

Narayanan, Arvind and Vitaly Shmatikov. “Robust de-anonymization of large sparse datasets.” In *Proceedings 2008 IEEE Symposium on Security and Privacy, Oakland, California, USA, May 18-21, 2008*, 111-25. Los Alamitos, California: IEEE Computer Society, 2008.

Narayanan, Arvind. “What Happened to the Crypto Dream?, Part 2.” *IEEE Security & Privacy* 11, no. 3 (2013): 68-71.

Ochoa, Salvador, Jamie Rasmussen, Christine Robson, and Michael Salib. “Reidentification of Individuals in Chicago’s Homicide Database: A Technical and Legal Study.” Final project, 6.805 Ethics and Law on the Electronic Frontier, Massachusetts Institute of Technology, Cambridge, Massachusetts, May 5, 2001. http://mike.salib.com/writings/classes/6.805/reid.pdf.

Ohm, Paul. “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization.” *UCLA Law Review* 57 (2010): 1742-43. http://uclalawreview.org/pdf/57-6-3.pdf.

“OnTheMap.” U.S. Census Bureau. http://onthemap.ces.census.gov/.

Pandurangan, Vijay. “On Taxis and Rainbows: Lessons from NYC’s improperly anonymized taxi logs.” *Medium*, June 21, 2014. https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1.

Proposal for a Regulation of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data. Art. 4(1)-(2) (January 25, 2012).

Roberts, Margaret E. “Fear or Friction? How Censorship Slows the Spread of Information in the Digital Age.” Unpublished manuscript, September 26, 2014. http://scholar.harvard.edu/files/mroberts/files/fearfriction_1.pdf.

rubyrescue, October 17, 2014. Comment on blackRust, “How Whisper app tracks ‘anonymous’ users.” *Hacker News*, October 17, 2014. https://news.ycombinator.com/item?id=8465482.

Sachs, Noah M. “Rescuing the Strong Precautionary Principle from Its Critics.” *Illinois Law Review* 2011 no.4 (2011): 1313.

Salganik, Matthew J. and Karen E.C. Levy. “Wiki surveys: Open and quantifiable social data collection.” Unpublished manuscript, October 2, 2014. http://arxiv.org/abs/1202.0500.

Siddle, James. “I Know Where You Were Last Summer: London’s public bike data is telling everyone where you’ve been.” *The Variable Tree,* April 10, 2014. http://vartree.blogspot.com/2014/04/i-know-where-you-were-last-summer.html.

Solove, Daniel J. “‘I’ve Got Nothing to Hide’ and Other Misunderstandings of Privacy.” *San Diego Law Review* 44 (2007): 760-64.

Stodden, Victoria. “Data access going the way of journal article access? Insist on open data.” *Victoria’s Blog*, December 24, 2012. http://blog.stodden.net/2012/12/24/data-access-going-the-way-of-journal-article-access/.

Sunstein, Cass R. “The Paralyzing Principle.” *Regulation* 25, no. 4 (2002): 33-35.

Sweeney, Latanya, Akua Abu, and Julia Winn, “Identifying Participants in the Personal Genome Project by Name.” White Paper 1021-1, Data Privacy Lab, Harvard University, Cambridge, Massachusetts, April 24, 2013. http://dataprivacylab.org/projects/pgp/1021-1.pdf.

Sweeney, Latanya. “*k*-anonymity: A Model for Protecting Privacy.” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10, no. 5 (2001): 557-70.

Sweeney, Latanya. “Matching Known Patients to Health Records in Washington State Data.” White Paper 1089-1, Data Privacy Lab, Harvard University, Cambridge, Massachusetts, June 2013. http://dataprivacylab.org/projects/wa/1089-1.pdf.

Sweeney, Latanya. “Simple Demographics Often Identify People Uniquely.” Data Privacy Working Paper 3, Carnegie Mellon University, Pittsburgh, Pennsylvania, 2000. http://dataprivacylab.org/projects/identifiability/paper1.pdf.

Task, Christine. “An Illustrated Primer in Differential Privacy.” *XRDS* 20, no. 1 (2013): 53-57.

Tockar, Anthony. “Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset.” *Neustar: Research*, September 15, 2014. http://research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset/.

Ugander, Johan, Brian Karrer, Lars Backstrom, and Cameron Marlow. “The anatomy of the Facebook social graph.” arXiv Preprint, 2011. http://arxiv.org/pdf/1111.4503v1.pdf.

“Understanding Online Advertising: Frequently Asked Questions.” Network Advertising Initiative. http://www.networkadvertising.org/faq.

Uyarra, Elvira and Kieron Flanagan. “Understanding the Innovation Impacts of Public Procurement.” *European Planning Studies* 18, no. 1 (2010): 123-43.

Vertesi, Janet. “My Experiment Opting Out of Big Data Made Me Look Like a Criminal.” *Time*, May 1, 2014.

“What are single nucleotide polymorphisms (SNPs)?” *Genetics Home Reference: Your Guide to Understanding Genetic Conditions*. Published October 20, 2014. http://ghr.nlm.nih.gov/handbook/genomicresearch/snp.

The White House, Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy (Washington, D.C.: February 2012).

Whong, Chris. “FOILing NYC’s Taxi Trip Data.” March 18, 2014. http://chriswhong.com/open-data/foil_nyc_taxi/.

Zang, Hui and Jean Bolot. “Anonymization of location data does not work: A large-scale measurement study.” In *Proceedings of the 17th International Conference on Mobile Computing and Networking*, 145-56. New York, New York: ACM, 2011.