Provable Privacy: Modern Methods and the FTC's Role

Joshua A. Kroll[1], Nitin Kohli[1], Will Monge[2], Nathaniel Good[1,2], and Paul Laskowski[1]

[1]School of Information, University of California, Berkeley
[2]Good Research, LLC

Respectfully submitted in response to Question 5 in the request for comments in advance of the Competition and Consumer Protection in the 21st Century Hearings, Project Number P181201.

1 Introduction

The last ten years have seen substantial development both in privacy-aware technologies and in technologies that threaten privacy. As this landscape changes, it is worth reassessing the FTC's role in guiding privacy practices as they impact consumers. Privacy technologies provide provable guarantees that can form the basis of sound privacy policymaking [15]. The FTC should guide and encourage the deployment of privacy supportive technologies by demanding honest privacy claims from businesses, by vigorously enforcing in situations where consumers are harmed, and by examining whether the proper and adequate use of privacy technologies should be recognized as a safe harbor for compliance, enforcement, or both.

When consumers navigate the modern economy, their locations, purchases, browsing habits, and other behaviors are increasingly tracked and recorded by computer systems. Meanwhile, advances in data science and artificial intelligence are empowering firms and institutions with new abilities to extract useful information from these data and from a growing variety of data sources. Data about consumers are traded on robust markets, and a constellation of companies works hard to produce ever-more-complete profiles of as many individuals as possible. In this environment, it is becoming increasingly difficult for the average person to understand their privacy profile or to assert any meaningful control over their data.

Artificial intelligence, machine learning, and the increasing sophistication of data analysis make it even harder for individuals to conceptualize privacy risks by making possible surprising and sensitive inferences from seemingly unrelated or innocuous data [3]. Location traces can reveal who goes to alcoholics anonymous [5]. Analysts can track the movements of celebrities and politicians using public transportation data [20]. Social media data can out LGBT people who would rather stay in the closet [11]. Value can be diverted away from consumers to companies that know exactly how much they are willing to pay [2]. Authoritarian governments can identify dissidents and oppress them based on their browsing history [4].

The legal cornerstone of this ecosystem is the idea that consumers consent to these activities, but modern developments make this interpretation difficult to maintain [21]: 1. consumers often do not know that data about them has been collected or exchanged. 2. consumers may regard individual pieces of data as innocuous, not recognizing how various data sources may be combined to construct a detailed account of behavior. 3. Some of the services that collect detailed data about consumers have become critical parts of modern society, and it would require drastic lifestyle changes to avoid many forms of data collection. A consumer cannot truly be said to give informed consent freely when a service is a critical part of modern society or when a product's disclosures or user interface are misleading or vague as to the extent of data collection. All too often, consumers have no way to know that data about them has been collected or exchanged.

2 Consumer Injury in Today's Landscape

Consumers little understand the privacy implications of modern technologies. Products and services, even those that promise advanced privacy protections, regularly collect detailed profiles of users. A secondary market for aggregating and cross-correlating these data means that nearly every action a consumer takes is or could be associated with their identity in perpetuity. While the FTC can and should improve enforcement around privacy policies and disclosures made to consumers [24], the surprising inferences and identifications that modern data analysis can yield require deeper engagement. While a large proportion of consumers expresses interest in protecting their data [9, 1], they may overshare information with companies that avoid sharing data or that use privacy-enhancing technologies in ways that offer little actual protection or confuse consumers as to the scope of protection offered. We consider here two specific types of harm: the risk of re-identification of consumers using rich data and the gap between the privacy claims made in marketing products and services and the actual privacy profile of those products and services.

2.1 Traditional anonymization does not adequately protect consumers

Traditional anonymization methods, such as removing sensitive fields from a database, hashing identifiers, or replacing identities with pseudonyms, fail to adequately protect consumers. These methods have been shown not to provide useful limits on data analysis and inference on individuals.[1] This happens because these methods do not speak directly to the expressivity of the data being treated, and so do not provide a good analytic framework for evaluating the risk associated with identification [17, 19, 18].[2] Further, these methods are often analyzed and deployed in a too-weak adversary model, where it is assumed that the analyst attempting to identify a person in the data set has limited additional information beyond what is contained in the data set (in which case, removing or redacting information in the data set would seem to be a good solution). In realistic application scenarios, it is generally the case that someone has access to relevant additional information or that such information can be readily collected [14]. Given the state of the science and knowledge of modern methods for defeating traditional anonymization approaches, we offer that claiming robust privacy based on these methods is deceptive.

2.1.1 The concept of PII is outdated

Because traditional anonymization methods fail to protect consumers from re-identification, the utility of giving different levels of sensitivity to data which are "identifiable" and data that are "not identifiable" is questionable at best. Rehearsing a thought experiment from Narayanan [18], we think purely along the lines of how to *assign* identities to individuals, we discover that any feature which is not common to all people provides at least one "bit" of distinguishing information. With the current world population somewhere around 7.4-7.5 billion people, it would take only 33 bits to uniquely identify every person on earth (that is, there exists a set of 33 yes/no questions for which each person in the world has a different set of answers—the answers uniquely identify each person in the world). For this reason, we offer that considering some data as "PII" and other data as "non-PII" speaks less to the identifiability of the data and more to the extent to which the data are overtly personal (that is, descriptive of a natural person).

2.1.2 Deidentification of rich data is impossible

Following on the idea that the distinction between PII and non-PII may not be relevant given modern data analysis methods, we offer that deidentifying any sufficiently rich data set is impossible, meaning that only privacy-aware technologies including provable privacy methods (described below) can improve

[1] Sometimes, such methods may even be used to explicitly enable resolution of identities across disparate data sources in such a way that no single entity appears to have both sensitive data and an identifier for the consumer to whom those data pertain. Several online tracking businesses, such as LiveRamp, RocketFuel, and Adobe's "Audience Manager", offer such resolution as a service.

[2] Despite this, it is all-too-common to attempt ad-hoc analysis of the safety of such techniques, sometimes even going as far as to estimate the "probability" that an individual in a data set can be re-identified [7]. However, rigorous scientific

consumer privacy in large data sets. Specifically, in any data set, as the number of variables increases, the number of ways to distinguish individuals increases. This means that the chance that new approaches to distinguishing people will uniquely identify people increases. At the least, such methods will define small groups of individuals within the data set which have homogenous features. If an analyst wishes to know whether Alice has liver cancer, the analyst will be satisfied if she can place Alice in a group of ten people who all have liver cancer, even if the group is not identified by this feature (for example, the analyst may know that Alice is represented in a certain data set and also that Alice is in a certain subgroup of that data set distinguished by being 52 years old, female, and living in ZIP code 44115; if every member of this group is also identified as having liver cancer, then Alice's diagnosis becomes known incidentally).

A common example of rich data is location data. For example, identifying just two data points in a location history as the individual's home and work locations can uniquely identify over 50% of people. Knowing four locations increases the accuracy of location-based identification to over 95% [5, 8, 25]. For this reason, data sets that contain location information reveal much about the identities and habits of the people whose data comprise them. It is for this reason that issues of location privacy rightfully receive heightened scrutiny from all sides, such as in the recent *Carpenter* Supreme Court decision and during the more recent debate around the privacy implications of location tracking controls in Android mobile devices.

Worse yet, there are fundamental mathematical reasons that data points are easier to separate and identify uniquely as the number of dimensions defining them increases, a phenomenon known as the "Curse of Dimensionality."

2.1.3 Machine learning, data science, and re-identification

Researchers are advancing the field of re-identification. As machine learning technologies become more widely available and more widely deployed in the work flow of designing real products, the pace of this advance will only increase and the skill required of analysts to successfully perform re-identification will decrease. This is because machine learning works by automatically discovering patterns in data in order to identify data points with each other and with labels into "classes." Identification is simply a class with only one data point in it—namely, an individual natural person. Further, machine learning can also be used to identify the features which best separate data points into individuated classes. Especially in situations where the object of analysis is unstructured data such as text or images, machine learning techniques can discover the proper "features" by which to analyze data [16]. Certain technologies, such as deep learning, might even re-identify consumers or associate them with adverse inferences (such as inferences of their race, religious preferences, or health status) in the service of advancing some other goal.

Beyond this, side information–that is, information which is not in the database under analysis but which an analyst has access to—gets more abundant with time simply because of the proliferation of recorded data about individuals. As a result, we are being forced to contemplate a world in which rich multidimensional data is available for the vast majority of citizens, one in which the technology to automatically re-identify anonymized data is within the grasp of even non-experts.

2.2 Stated privacy guarantees may be illusory

When companies make privacy claims, either in marketing materials or in their privacy policies, consumers may rightly be curious whether those claims correspond to the company's actual practices. And yet, at best, consumers must rely on these assertions and the community's best view of whether the company is complying with its own promises. And yet, there are many approaches that companies could take to provide verifiable evidence directly convincing to consumers that they are satisfying their own stated claims [13, 12].

3 Privacy-aware computing

Against the backdrop of increasing information gathering, researchers have been studying techniques for protecting individual privacy. The advances of the last 10 years have yielded new technologies, designed to

protect individuals while deriving value from data. At the same time, we have gained a clearer understanding of the limits of privacy protection. Today, a number of prominent companies advertise their use of the latest privacy-enhancing technologies, highlighting the value that consumers place on protecting their data. Unfortunately, the extent to which this technology actually improves the privacy landscape for consumers remains unclear. Nonetheless, when properly deployed, these modern privacy technologies can and will improve consumer privacy and the FTC should acknowledge and support this transition.

Prior to the last decade, efforts to protect individuals represented in datasets centered around controlling access to the data and anonymizing the data. Anonymization, or de-identification, refers to a collection of practices focused on the types of variables in a dataset, and how frequently values appear. For example, sensitive fields, such as names and social security numbers, may be removed from a dataset or replaced with pseudonyms. Rare values may be combined together into larger categories [22]. Such practices often emerge organically in response to privacy violations and come with no formal guarantees of protection, nor with any way to evaluate the protection afforded by these techniques [15]. Anonymization is often codified in law, with anonymized data treated as less privacy sensitive [19] and data sets broken out into sensitive data containing *personally identifiable information* and other, non-sensitive data. However, modern data analysis technology can (and regularly does) render "anonymized" data identifiable [17, 15].

A major obstacle to anonymization lies in the rich structure of today's datasets. Data about consumers is often high-dimensional, making each person's data profile unique and specific to them [15]. This is precisely what makes these data valuable—being able to infer a consumer's life situation from their purchasing or browsing habits makes it possible to understand what products a consumer is likely to want, forming the basis of the online behavioral advertising ecosystem. Unfortunately, this richness also makes each person's personal record highly recognizable, even without explicit "identifiers" like names or social security numbers. As a result, there is a risk that anonymized datasets are *re-identified* [17, 19, 20].

In the last decade, an important alternative to traditional anonymization has emerged through the study of *provable privacy*. Researchers in this field use mathematics to quantify the amount of information that a computer system can leak about an individual in the worst possible case. They focus in parallel on designing new systems while generating rigorous proofs that the information that can be learned about individuals is within a fixed threshold. This is usually achieved by injecting a specific amount of noise or randomness into a computation.

A hallmark of provable privacy is the use of strong adversary models. This can be seen in the most popular privacy definition in the field, which is called *differential privacy*. In a differentially private system, if you change a single row of a database, the behavior of the system cannot change by very much [6]. To be a little more precise, for two databases, D_1 and D_2 that differ in just one row (e.g., where one contains a data point about an individual and the other does not), the probability that the system produces any output is almost the same in a world where the real data is D_1 and in a world where the real data is D_2. A tunable parameter defines how similar these probabilities must be. The tighter the limit on the difference in behavior between the two scenarios, the less the system's output reveals about the underlying data.

An example may help clarify how differential privacy protects consumers. Suppose that a differentially private system tells you that there are 304 taxis in Washington D.C. Because there is randomness in the output, you don't know if the real number is 304 or 305 or perhaps some other number. If we demand that the system use a strong privacy bound, the probability that the system outputs 304 is almost the same if the real number is 304 and if the real number is 305 or 303. This means that even if a powerful attacker knows the position of every single taxi in the local fleet except for one, they still cannot tell whether that remaining taxi is in in D.C. or in a neighboring city.

Because provable privacy technologies such as differential privacy limit the amount of information about the underlying data that can be learned from queries, these technologies improve privacy while limiting what can be learned from data. Seemingly, this reduces the useful insights the data revolution promises. However, this tradeoff is often illusory in practice: by carefully considering the goals of analysis, it is generally possible to implement a provable privacy method that limits disclosure and inference of sensitive information without meaningfully damaging the value of the data.

3.1 Provable Privacy Must be Deployed Responsibly

We have argued that companies that continue to employ traditional anonymization techniques pose risks to consumers. Unfortunately, even when a company employs modern techniques from the field of provable privacy, the strength of privacy protection depends on many factors. Since these methods can be difficult for consumers to understand, there is an even greater risk that consumers are misled into thinking that their information is safer than it really is. There are several common reasons that a provably private system may not offer meaningful protection in practice.

First, provable privacy guarantees are not simply on-or-off, but rather expressions of the *degree* of protection. Take, for example, the most popular provable privacy definition, differential privacy [6]. The level of protection (that is, the maximum amount of detail which can be learned about an individual in the protected database) afforded by a differentially private system is controlled by a tunable parameter. From the human perspective, this parameter measures the largest possible risk any individual faces by having their data included as an input to the system by expressing the maximum amount of information an adversarial analyst could extract through any arbitrary query. It might seem that it would be beneficial to tune this parameter all the way towards "more privacy", but there are no free lunches to be had—this protection comes at a cost. Typically differential privacy is achieved by introducing randomness or noise into computation process. As such, privacy trades off with the accuracy of queries against the protected database—too much noise degrades the accuracy of query results, whereas too little noise provides overly accurate results to queries, thereby degrading individual privacy. The amount of noise that must be added depends precisely on the level of protection demanded. For differential privacy to have any meaningful guarantees, the maximum amount of allowed inference needs to be small. But in real-world implementations, of which there are now several at major tech companies, it is unclear whether this key parameter is set small enough to provide any meaningful protection for individuals. For example, a team of researchers inferred that Apple's differential privacy implementation had a daily privacy budget that was unlikely to provide meaningful guarantees [23] and which was orders of magnitude larger than what is proposed in the literature. These tradeoffs can generally be managed by careful engineering of queries so that noisy answers can be tolerated.

A second reason that a provably private system may not offer strong protection is that it will continue to leak information if it is used over and over. If a system is designed to provide a differential privacy guarantee at a certain level of protection, and an attacker queries it twice, the total possible amount of information disclosure will be twice as much. Designers of privacy-aware systems therefore consider setting a "privacy budget", which is a limit on how much information can be leaked in total, over the entire operation of the system. Each time a differentially private system computes an answer to a query use a data element, we have to pay a price from the privacy budget to compute an answer. Once the total privacy budget is utilized, the data is, theoretically, never supposed to be used again.

Unfortunately, a common practice in today's industry is to reset the privacy budget at regular intervals. This means that users may suffer additional privacy loss with each time period that passes and with no limit on total privacy loss. An example of this is Apple, which was observed to reset their privacy budget daily [23], allowing for a worst-case privacy loss to be incurred every day that an individual used their Apple product. Implementations that reset their privacy budgets overstate the true level of privacy protection provided, and marketing claims based on the protections provided by such systems run the risk of being false at a technical level.

A third reason that a provably private system may not provide meaningful protection derives from how the unit of analysis is defined. Differential privacy applies to rows of a database, but a single row may not be equivalent to a single person. For example, in the system used by Apple, differential privacy is applied to individual uses of emoji This suggests that no attacker can determine which emoji was used at a specific time. However, there is no guarantee for repeated uses of emoji over the course of a day. In fact, this is likely on purpose; the company may want to learn whether an individual used a lot of happy emojis or sad emojis on a particular day.

4 FTC Authority and Provable Privacy

The FTC's authority extends to regulating deceptive practices and unfair practices. The uses of privacy we describe implicate both dimensions, though here we focus on deception claims.

Under FTC guidelines,[3] three key factors in determining whether a practice constitutes deception. Specifically, there must be:

(i) a representation, omission, or practice that is likely to mislead a consumer; (ii) the interpretation of that act or practice is considered from the perspective of a reasonable consumer; (iii) the representation must be material (i.e. a practice or act that is likely to alter a consumers choice or conduct regarding a practice).

That is, a deception case involves a material representation, omission, or practice that is likely to mislead consumers acting reasonably under the circumstances [10]. In this section, we argue that the act of misrepresenting or ommiting the guarantees afforded by a provable privacy technology can be subject to a deception claim.

Consumers expect a reasonable level of protection when they see any claim of protecting privacy. And as we have seen, it is possible for claims of privacy protection to outstrip actual privacy protection in several ways, even when provable privacy technologies are in use. Consider, for example, an organization that misrepresents the parameter chocies used in their differentially private system in a way that overstates the level of privacy afforded (i.e. they disclose a smaller maximum inference limit than is actually provided).

Requirements (1) and (2) of deception follow immediately. In the language of economics, privacy can be classified as a credence good. Unlike search goods (where quality can be addressed pre-purchase) and experience goods (where quality can be addressed post-purchase), credence goods are defined by the peculiar characteristic that the quality cannot be addressed even post-purchase / consumption, without extraordinary means. Privacy falls into the credence classification as consumers have no real means to deduce, infer, or abduct the quality of protection they are being given. Since privacy is a credence good, there is no possible way for a consumer to know that the reported value of the privacy parameters is true just from such a claim alone, or even through ordinary use of a product. As such, any reasonable consumer can only view take the disclosed parameters as given. By overstating the extent of privacy protection, a reasonable consumer is likely to be misled by the false representation of the reported parameters.

The issue of materiality present in Requirement (3) is a bit subtler. For a claim to be material, it must be the case that some consumers decision towards a product has been altered based on the misrepresentation. In general, a published choice of privacy parameters will not be interpretable to the general public. For such consumers, misrepresenting these parameters is unlikely to alter their decision-making process. However, for those with a basic understanding of differential privacy and the role of these parameters play in determining the afforded level of protection, consumer choice outcomes can vary based on such misrepresentations. After all, privacy protection is something that consumers value and that guides their economic decisions, if perhaps irrationally [9, 1].

As noted above, different choices of parameters in a differentially private system indicate different levels of protection. Suppose that a consumer is deciding between two products, A and B, both of which offer protection via differential privacy. For concreteness, suppose that A and B both collect similar information about the consumer and perform similar analyses as well. The only difference between the two products is that A uses a weak choice of parameters and B uses stronger parameters. All things, equal, for consumers with strong privacy preferences, option B is more desirable than option A. However, suppose that the producers of A misrepresent the true parameters as stronger than they really are. Such a deviation changes the decision of this consumer from B to A now, making the misrepresentation material.

Next, we turn our attention to the case of omission. Consider an organization that claims to be using differential privacy but does not provide the value of the parameters used. Furthermore, suppose that this organization is using a very weak set of parameter choices. While the organization is technically using differential privacy, it is doing so in name only. Weakly chosen differential privacy parameters provide little-to-no meaningful privacy protection. When presented with a claim of protection via differential privacy, a

[3] FTC Policy Statement on Deception, October 14, 1983, 103 F.T.C. 110, 174 (1984). `https://www.ftc.gov/public-statements/1983/10/ftc-policy-statement-deception`

reasonable consumer would indeed believe that they are being offered some meaningful amount of protection. As such, Requirements (1) and (2) are satisfied. Furthermore, Requirement (3) of materiality is also satisfied, as the presence of information about the strength of the parameters could alter the decision-calculus of a consumer. That is, if the organization offered privacy protection that didn't perform privacy protection in material ways, then a consumer may have opted for the different service had the value of the parameters been known.

References

[1] Alessandro Acquisti and Jens Grossklags. What can behavioral economics teach us about privacy. *Digital privacy: theory, technologies and practices*, 18:363–377, 2007.

[2] Le Chen and Christo Wilson. Observing Algorithmic Marketplaces In-the-Wild. *ACM SIGecom Exchanges*, 15(2):34–39, January 2017.

[3] Kate Crawford and Jason Schultz. Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College L. Rev.*, 55:93, 2014.

[4] Rogier Creemers. China's social credit system: An evolving practice of control. *SSRN*, 2018. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3175792`.

[5] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3:1376, 2013.

[6] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[7] Khaled El Emam, 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(1), 2012.

[8] Philippe Golle and Kurt Partridge. On the anonymity of home/work location pairs. In *International Conference on Pervasive Computing*, pages 390–397. Springer, 2009.

[9] Jens Grossklags and Alessandro Acquisti. When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. In *WEIS*, 2007.

[10] Chris Jay Hoofnagle. *Federal Trade Commission privacy law and policy*. Cambridge University Press, 2016.

[11] Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, page 201218772, 2013.

[12] Joshua A. Kroll. The fallacy of inscrutability. *Philosophical Transactions of the Royal Society—A*, 2018.

[13] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. Accountable algorithms. *University of Pennsylvania Law Review*, 165(3), 2017.

[14] Arvind Narayanan and Edward W Felten. No silver bullet: De-identification still doesnt work. *White Paper*, pages 1–8, 2014.

[15] Arvind Narayanan, Joanna Huey, and Edward W. Felten. A precautionary approach to big data privacy. *Computers, Privacy, and Data Protection*, 2015.

[16] Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. On the feasibility of internet-scale author identification. In *2012 IEEE Symposium on Security and Privacy (S&P),*, pages 300–314. IEEE, 2012.

[17] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy, 2008. S&P 2008*, pages 111–125, 2008.

[18] Arvind Narayanan and Vitaly Shmatikov. Myths and fallacies of personally identifiable information. *Communications of the ACM*, 53(6):24–26, 2010.

[19] Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *Ucla L. Rev.*, 57:1701, 2009.

[20] Vijay Pandurangan. On taxis and rainbows: Lessons from nycs improperly anonymized taxi logs. *Medium*, 2014. `https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1`.

[21] Daniel J Solove. Privacy self-management and the consent dilemma. *Harvard L. Rev.*, 126:1880, 2012.

[22] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[23] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. Privacy loss in apple's implementation of differential privacy on macos 10.12. *arXiv preprint arXiv:1709.02753*, 2017.

[24] Joseph Turow, Chris Jay Hoofnagle, Deirdre K Mulligan, and Nathaniel Good. The federal trade commission and consumer privacy in the coming decade. *ISJLP*, 3:723, 2007.

[25] Hui Zang and Jean Bolot. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, pages 145–156. ACM, 2011.