Journal of Economic Perspectives—Volume 28, Number 2—Spring 2014—Pages 75–98

Privacy and Data-Based Research

Ori Heffetz and Katrina Ligett

On August 9, 2006, the "Technology" section of the *New York Times* contained a news item titled "A Face Is Exposed for AOL Searcher No. 4417749," in which reporters Michael Barbaro and Tom Zeller (2006) tell a story about big data and privacy:

> Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield. No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything." And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia." It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga. . . . Ms. Arnold, who agreed to discuss her searches with a reporter, said she was shocked to hear that AOL had saved and published three months' worth of them. "My goodness, it's my whole personal life," she said. "I had no idea somebody was looking over my shoulder." . . ."We all have a right to privacy," she said. "Nobody should have found this all out."

■ *Ori Heffetz is Assistant Professor of Economics, Samuel Curtis Johnson Graduate School of Management, Cornell University, Ithaca, New York. Katrina Ligett is Assistant Professor of Computer Science and Economics, California Institute of Technology, Pasadena, California. Their email addresses are oh33@cornell.edu and katrina@caltech.edu.*

http://dx.doi.org/10.1257/jep.28.2.75 doi=10.1257/jep.28.2.75

Empirical economists are increasingly users, and even producers, of large datasets with potentially sensitive information. Some researchers have for decades handled such data (for example, certain kinds of Census data), and routinely think and write about privacy. Many others, however, are not accustomed to think about privacy, perhaps because their research traditionally relies on already-publicly-available data, or because they gather their data through relatively small, "mostly harmless" surveys and experiments. This ignorant bliss may not last long; detailed data of unprecedented quantity and accessibility are now ubiquitous. Common examples include a private database from an Internet company, data from a field experiment on massive groups of unsuspecting subjects, and confidential administrative records in digital form from a government agency. The AOL story above is from 2006; our ability to track, store, and analyze data has since then dramatically improved. While big data become difficult to avoid, getting privacy right is far from easy—even for data scientists.

This paper aims to encourage data-based researchers to think more about issues such as privacy and anonymity. Many of us routinely promise anonymity to the subjects who participate in our studies, either directly through informed consent procedures, or indirectly through our correspondence with Institutional Review Boards. But what is the informational content of such promises? Given that our goal is, ultimately, to publish the results of our research—formally, to publish functions of the data—under what circumstances, and to what extent, can we guarantee that individuals' privacy will not be breached and their anonymity will not be compromised?

These questions may be particularly relevant in a big data context, where there may be a risk of more harm due to both the often-sensitive content and the vastly larger numbers of people affected. As we discuss below, it is also in a big data context that privacy guarantees of the sort we consider may be most effective.

Our paper proceeds in three steps. First, we retell the stories of several privacy debacles that often serve as motivating examples in work on privacy. The first three stories concern intentional releases of de-identified data for research purposes. The fourth story illustrates how individuals' privacy could be breached even when the data themselves are not released, but only a seemingly innocuous function of personal data is visible to outsiders. None of our stories involves *security* horrors such as stolen data, broken locks and passwords, or compromised secure connections. Rather, in all of them information was released that had been *thought* to have been anonymized, but, as was soon pointed out, was rather revealing.

Second, we shift gears and discuss *differential privacy*, a rigorous, portable privacy notion introduced roughly a decade ago by computer scientists aiming to enable the release of information while providing *provable* privacy guarantees. At the heart of this concept is the idea that the addition or removal of a single individual from a dataset should have nearly no effect on any publicly released functions of the data, but achieving this goal requires introducing randomness into the released outcome. We discuss simple applications, highlighting a privacy-accuracy tension: randomness leads to more privacy, but less accuracy.

Third, we offer lessons and reflections, discuss some limitations, and briefly mention additional applications. We conclude with reflections on current promises of "anonymity" to study participants—promises that, given common practices in empirical research, are not guaranteed to be kept. We invite researchers to consider either backing such promises with meaningful privacy-preserving techniques, or qualifying them. While we are not aware of major privacy debacles in economics research to date, the stakes are only getting higher.

Intuition May Not Be Enough: Cautionary Tales

Well-intentioned government or private entities in possession of a sensitive database may wish to make an anonymized version of the data public—for example, to facilitate research. We retell and discuss a few cautionary tales that illustrate how intuition-based attempts at anonymization may fail, sometimes spectacularly.[1]

When Anonymization Failed

The first story is from the mid 1990s, when William Weld, then Governor of Massachusetts, approved the release of certain medical records of state employees to researchers, assuring the public that individual anonymity would be protected by eliminating obvious identifiers from the data (Greely 2007). A few days after Weld's announcement, Latanya Sweeney—then a graduate student at MIT—re-identified Weld's personal records (including diagnoses and prescriptions) in the database; she then had his records delivered to his office.

While the medical data—officially, the Massachusetts "Group Insurance Commission" (GIC) data—had been "de-identified" by removing fields containing patients' name, address, and social security number (SSN) prior to the data release, the nearly 100 remaining fields included ZIP code, birth date, and sex. As Ohm (2010) tells the story, Sweeney

> . . . knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 residents and seven ZIP codes. For twenty dollars, she purchased the complete voter rolls from the city of Cambridge—a database containing, among other things, the name, address, ZIP code, birth date, and sex of every voter. By combining this data with the GIC records, Sweeney found Governor Weld with ease. Only six people in Cambridge shared his birth date; only three were men, and of the three, only he lived in his ZIP code.

Barth-Jones (2012) revisits and critiques this story. Perhaps in response, Sweeney, Abu, and Winn (2013) use a similar method to re-identify individuals in the publicly

[1] As hinted above, these stories are well known in the computer science community that studies privacy. The first three were revisited and discussed by Ohm (2010), a legal scholar, who provides further references and links to primary sources.

available Personal Genome Project database. Sweeney's "How Unique Are You?" interactive website invites the visitor to "Enter your ZIP code, date of birth, and gender to see how unique you are (and therefore how easy it is to identify you from these values)." Her methodology is explained on the website (http://aboutmyinfo.net, accessed on August 9, 2013):

> Birthdate . . . gender, and 5-digit postal code (ZIP) uniquely identifies most people in the United States. Surprised? . . . 365 days in a year × 100 years × 2 genders = 73,000 unique combinations, and because most postal codes have fewer people, the surprise fades. . . . [T]here are more than 32,000 5-digit ZIP codes in the United States; so 73,000 × 32,000 is more than 2 billion possible combinations but there are only 310 million people in the United States.

The next story, involving Ms. Arnold above, is from roughly a decade later. In 2006, AOL Research released detailed Internet search records of 650,000 users covering a three-month period, amounting to 20 million search queries.[2] The stated purpose of the release was expressed by then AOL Research head Abdur Chowdhury:

> AOL is embarking on a new direction for its business—making its content and products freely available to all consumers. To support those goals, AOL is also embracing the vision of an open research community, which is creating opportunities for researchers in academia and industry alike. . . . with the goal of facilitating closer collaboration between AOL and anyone with a desire to work on interesting problems.[3]

Prior to the data release, the search logs were de-identified, for example by removing usernames and IP addresses, using instead unique identifiers (such as "4417749") to link all of a single user's queries. This de-identification, however, quickly proved far from sufficient for the intended anonymization, as illustrated by the *New York Times* article on Ms. Arnold. Within days of the release, AOL apologized, removed the data website as well as a few employees, and silenced its research division. Of course, to this day, the data are widely available through a simple web search; once published, you cannot take it back.

The third story is also from 2006. About two months after the AOL debacle, Netflix announced a competition—the Netflix Prize—for improving the company's algorithm that predicts user ratings of films, using only past user ratings. To allow competitors to train their algorithms, Netflix released a database with 100 million ratings of 17,770 films by about 500,000 subscribers covering a six-year period.

[2] As Ohm (2010) notes, different numbers appear in different accounts. The 650,000 figure above was described as 500,000 in the original post, and the 20 million figure in the original post has later been reported by some as 36 million.

[3] Posting of Abdur Chowdhury, cabdur@aol.com, to SIGIR-IRList, irlist-editor@acm.org, http://sifaka.cs.uiuc.edu/xshen/aol/20060803_SIG-IRListEmail.txt, as cited in Ohm (2010, accessed on August 9, 2013).

Each record contained a movie title, a rating date, and a five-point rating. As in the Massachusetts Group Insurance Commission and AOL cases, records were de-identified prior to the release, replacing user names with unique identifiers.

The illusion of protecting users' anonymity was, again, short-lived. Two weeks after the data release, Narayanan and Shmatikov (2008; first version posted in 2006) demonstrated that "an adversary who knows a little bit about some subscriber can easily identify her record if it is present in the dataset, or, at the very least, identify a small set of records which include the subscriber's record." How little is "a little bit"? In many cases, a user could be identified knowing as little as that user's approximate dates and ratings of two or three movies. In their demonstration, Narayanan and Shmatikov used ratings from the Internet Movie Database (IMDB), which are publicly available and are linked to the raters' identities, and showed how a handful of a user's IMDB ratings, even when they yield imprecise information, could uniquely identify that user in the Netflix database.

Whereas IMDB's public ratings may reveal only those movies that individuals are willing to tell the world that they have watched, Netflix ratings may reveal *all* of the movies an individual has rated, including those the rater may prefer to keep private—for example, films that may reflect a person's sexual, social, political, or religious preferences. Moreover, to be re-identified, one does not have to be on IMDB: as Ohm (2010) advises his readers, "the next time your dinner party host asks you to list your six favorite obscure movies, unless you want everybody at the table to know every movie you have ever rated on Netflix, say nothing at all."

De-identification and Beyond

De-identified data were defined by Sweeney (1997) as data in which "all explicit identifiers, such as SSN (Social Security number), name, address, and telephone number, are removed, generalized, or replaced with a made-up alternative." Her definition seems to describe accurately the released Massachusetts health insurance, AOL, and Netflix data in the stories above. Some more recent definitions (like those under federal health records privacy regulations) are stricter and would not consider the Massachusetts data released by Weld as de-identified, but these definitions still focus on removing only specific kinds of information (Greely 2007). Indeed, more than 15 years after Sweeney's powerful demonstration, her definition of de-identified data *still* describes, more or less accurately, commonplace practices among many researchers. For example, prior to publicly posting their data online (as required by some academic journals), economists often de-identify their data by merely withholding explicit identifiers such as subject names. However, as in the stories above, the stated aim of such de-identification—and what is often promised to subjects, directly or via an Institutional Review Board—is *anonymization*. In Sweeney's (1997) definition, *anonymous data* "cannot be manipulated or linked to identify an individual." Clearly, de-identification does not guarantee anonymization.

Sweeney's re-identification of people in the Massachusetts Group Insurance Commission data used birthday and five-digit ZIP code, neither of which are typically included in datasets publicly posted by economists. But it is not difficult to imagine re-identification of specific subjects based on combinations of demographics such as study major, age/class, gender, and race, which are often not considered "identifiable private information" and are routinely included in posted data.[4] Re-identification is still easier with knowledge regarding, for example, the day and time in which a classmate or a roommate participated in a specific study session. (Sweeney, 2013, applies this idea outside the lab: she uses newspaper stories that contain the word "hospitalized" to re-identify individual patients in a publicly available health dataset in Washington state.) But re-identification is possible even without such special knowledge, and it may be straightforward when targeting certain individuals who have a characteristic that is uncommon in a specific setting, such as minorities or women in certain fields and occupations.

This discussion highlights a weakness of de-identification: if one assumes no restrictions on outside information (also referred to below as auxiliary information), then, short of removing *all* data fields prior to a release, some individuals may be uniquely identified by the remaining fields. One potential response to this weakness is an approach called *k-anonymity*, which combines the assumption that there are *some* restrictions on outside information with the removal (or partial removal) of *some* fields. Specifically, assuming that outside information could only cover certain fields in the database, one could suppress these fields or, when possible, generalize them (for example, replace date of birth with year of birth) so that any combination of the values reported in these fields would correspond to at least *k* individuals in the data (Sweeney 2002). This approach has several weaknesses, and in many applications it implies either an unreasonably weak privacy guarantee or a massive suppression of data: notice that the amount of information that can be released is expected to shrink as *k* grows and as restrictions on outside information are weakened. Narayanan and Shmatikov (2008), for example, discuss the issues with such an approach in the Netflix context.

An alternative approach is to make it harder for an attacker to leverage outside information. For example, prior to making the query logs publicly available, AOL could have replaced not only user identities but also the search keywords themselves with uniquely identifying random strings. Similarly, Netflix could have replaced movie names with unique identifiers. Such an approach, known as "token-based hashing," would preserve many features of the data, hence maintaining usefulness of

[4] For example, according to Cornell's Office of Research Integrity and Assurance (at http://www.irb.cornell.edu, accessed on August 13, 2013):

> Identifiable private information is defined as: name; address; elements of dates related to an individual (e.g., birth date); email address; numbers: telephone, fax, social security, medical record, health beneficiary/health insurance, certificate or license numbers, vehicle, account numbers (e.g., credit card), device identification numbers, serial numbers, any unique identifying numbers, characteristics, or codes (e.g., Global Positioning System (GPS) readings); Web URLs; Internet Protocol (IP) addresses; biometric identifiers (e.g., voice, fingerprints); full face photographs or comparable images.

the database for some (though clearly not all) research purposes. But the preserved features of the underlying data make this type of scheme vulnerable as well.

Indeed, shortly after the disaster at AOL Research, a group at Yahoo! Research (Kumar, Novak, Pang, and Tomkins 2007) showed that an attacker with access to a "reference" query log (for example, early logs released by Excite or Altavista) could use it to extract statistical properties of tokenized words in the database, and "invert the hash function"—that is, break the coding scheme—based on co-occurrences of tokens within searches. Along similar lines, Narayanan and Shmatikov (2008) speculate that in the Netflix case, such an approach "does not appear to make de-anonymization impossible, but merely harder."

Privacy Risk without Data Release

Our fourth story, of privacy compromised on Facebook by Korolova (2011), "illustrates how a real-world system designed with an intention to protect privacy but without rigorous privacy guarantees can leak private information . . . Furthermore, it shows that user privacy may be breached not only as a result of data publishing using improper anonymization techniques, but also as a result of internal data-mining of that data."

Facebook's advertising system allows advertisers to specify characteristics of individuals to whom an ad should be shown. At the time of Korolova's (2011) attack, it was possible to specify those characteristics (for example, gender, age, location, workplace, alma mater) so finely that they would correspond to a unique Facebook user. Then, two versions of the ad campaign could be run—for example, one with those same characteristics plus "Interested in women"; the other with those characteristics plus "Interested in men." Even if this user's interests were not visible to her friends, if she had entered them in her profile, they would be used for ad targeting. Thus, if the advertiser received a report that, for example, the "Interested in women" version of her ad had been displayed, the advertiser could infer the targeted individual's private interests. Other attacks were possible too. "Using the microtargeting capability, one can estimate the frequency of a particular person's Facebook usage, determine whether they have logged in to the site on a particular day, or infer the times of day during which a user tends to browse Facebook."

Korolova (2011) quotes failed promises by Facebook executives, such as that Facebook doesn't "share your personal information with services you don't want" and doesn't "give advertisers access to your personal information." She notes: "We communicated our findings to Facebook on July 13, 2010, and received a very prompt response. On July 20, 2010, Facebook launched a change to their advertising system that made the kind of attacks we describe much more difficult to implement in practice, even though, as we discuss, they remain possible in principle."

This Facebook story helps demonstrate that if one seeks to use functions of data—be it via research findings, policy decisions, or commercial services and products—the privacy of the individuals comprising the data may be at risk without an approach providing (provable) privacy guarantees.

Differential Privacy

A common theme in the examples above has been the crucial role played by *auxiliary information*, that is, knowledge from sources outside the dataset under consideration. In the examples above, attackers consulted various outside sources not foreseen by the database owners, including public records such as voter rolls, complementary databases such as IMDB, or, simply, personal familiarity with an individual in the database. To identify individuals, the attackers then carried out a variant of a so-called "linkage attack": they matched fields that overlap across the auxiliary data and the attacked database.

More generally, one may invite trouble when making specific assumptions regarding what information a potential attacker might have and how the attacker might use it. If such assumptions are ever violated—even in the future, as new technology and information become available—privacy may be compromised. One approach to addressing the auxiliary-information concern would be to seek to provide privacy guarantees free from such assumptions. The approach we discuss here, *differential privacy*, seeks to do just that. It emerged from work in computer science theory by Dinur and Nissim (2003), Dwork and Nissim (2004), and Dwork, McSherry, Nissim, and Smith (2006). Our discussion and examples draw on a number of surveys, including Dwork (2006), Dwork and Smith (2010), Dwork (2011b, a), and Dwork, McSherry, Nissim, and Smith (2011). These surveys additionally present historical aspects of the development of the differential privacy definition, more examples, and a much broader range of applications than we discuss here. Our working paper, Heffetz and Ligett (2013), contains slightly more technical detail than presented here as well as more references to recent work on differential privacy. We also recommend a recent popular article on differential privacy research by Klarreich (2012).

The Differential Privacy Definition

To fix ideas, consider the released outcome of some function of a database: for example, the released number of Facebook users to whom an ad was displayed, or some published table of statistics in an empirical research paper, or even a released version of the entire database. Consider a potential participant in the database: for example, someone who considers joining Facebook, or someone who considers participating in a research study. Compare two possible scenarios: in one, this person joins and is added to the database; in the other, the person does not join and hence is not in the database.

Informally, differential privacy seeks to guarantee to the potential participant that, irrespective of the decision whether to participate, *almost* the same things can be learned from the released outcome—regardless of outside information, of the data already in the database, or of the participant's own personal data. Differential privacy hence gives participants (and nonparticipants) in the database a form of plausible deniability: they could always deny that their data took specific values or even that they participated (or did not participate), and an observer would have almost no evidence either way.

Here is an often-used example: one could conduct a differentially-private analysis that revealed a correlation between smoking and cancer, so long as that correlation depended only negligibly on the participation of any one individual in the study. Revealing (that is, publishing) this correlation might allow observers to draw inferences about an individual smoker, and that person might then feel that his or her privacy has been harmed. But since essentially the same conclusions would have been drawn regardless of whether that smoker participated in the study, the *differential* privacy of that person has been respected.

In a more formal sense, consider pairs of databases that are identical except that one of the databases has one additional row (or record) over the other. We refer to such a pair as *neighboring* databases, and think of each row as corresponding to one individual. Thus, two neighboring databases differ by only the participation of one individual. Now consider some computation that is carried out on such databases, and consider the space of possible outcomes of the computation. A differentially-private computation (or function, or mechanism) selects its output using a degree of randomness, such that the probability of any given outcome is similar under any two neighboring databases.

How similar? A common differential privacy definition, *ϵ-differential privacy* (Dwork, McSherry, Nissim, and Smith 2006), requires that the probability of any given outcome under any two neighboring databases cannot differ by more than a multiplicative constant, e^{ϵ}, where e is Euler's number and the parameter ϵ is a positive number that quantifies the amount of privacy.[5] The smaller ϵ is, the stronger is the privacy guarantee, but the less useful is the computation's output: in the limiting case of $\epsilon = 0$, we would replace the word "similar" above with "identical" since in that limiting case, e^{ϵ} would equal 1, requiring that the differentially-private mechanism be indistinguishable on any two input databases. In other words, maximum differential privacy means useless published output. More generally, the definition makes precise an intuitive tradeoff between privacy and usefulness.

The output of a differentially-private mechanism is readily publishable. It could, for example, be a single statistic (or a collection of statistics) to which a sufficient amount of random noise was added so that the inclusion of an additional record in, or exclusion of an existing record from, the database would have

[5] Here is a formal definition:

A randomized function K provides ϵ-differential privacy if for every $\mathcal{S} \in \text{Range}(K)$ and for all neighboring databases D and D',

$$\text{Prob}[K(D) = \mathcal{S}] \leq e^{\epsilon} \cdot \text{Prob}[K(D') = \mathcal{S}]$$

for $\epsilon \geq 0$ and where the probability space in each case is over the randomness of K.

Note that in particular, for any neighboring pair (D, D'), the definition must hold with the larger quantity (that is, $\max\{\text{Prob}[K(D) = \mathcal{S}],\ \text{Prob}[K(D') = \mathcal{S}]\}$) on the left, constraining it to be larger by at most a multiplicative e^{ϵ}. There are other variants on this definition, which we do not emphasize here. A common generalization of differential privacy allows an *additive* δ difference in the probabilities, in addition to the mulitiplicative difference e^{ϵ} (for example, Dwork, Kenthapadi, McSherry, Mironov, and Naor 2006). Such generalization provides a weaker privacy guarantee, but may allow for more accurate outcomes.

almost no effect on the distribution of the statistic (or statistics). Or it could be an entire *synthetic* database—a database consisting of artificial records, created with a degree of randomness from the original records in a way that preserves certain statistical properties of the original database but does not give away the inclusion of any individual record. The following subsections will have more to say about ways of using randomness and how much randomness is necessary. As discussed above, notice again that when we write here "would have almost no effect" and "does not give away" we imply that ϵ is small. How small should it be? The definition of differential privacy does not prescribe an answer to this normative question, a point we return to below.

Observations Regarding the Definition

The concept of differential privacy readily extends to provide a privacy guarantee to a group of individuals of a certain size: an ϵ-differentially-private mechanism is $k\epsilon$-differentially private from the point of view of a group of k individuals (or one individual whose data comprise k rows in the database). Intuitively, the inclusion in or exclusion from a database of a *group* of rows could have larger cumulative effect on outcomes of computations, weakening the privacy guarantee. In the smoking-and-cancer example above, it is more difficult to guarantee that adding an entire group of people to the study—say, all the residents of a specific city—would have almost no effect on outcomes.

The differential privacy definition also immediately yields an elegant composition property: running ℓ ϵ-differentially-private mechanisms—for example, publishing ℓ statistics based on a database—gives a guarantee of $\ell\epsilon$-differential privacy.[6] Equivalently, one may split a fixed total "privacy budget" of ϵ across a set of desired computations.

This composition property is particularly important in the context of potential real-world applications—including academic research and public- and private-sector implementations—where individuals may participate in more than one database, and where on each database typically more than one analysis is conducted. Differential privacy hence provides a tool for understanding the cumulative privacy harm incurred by an individual whose data appear in multiple databases, potentially used by different entities for different purposes and at different points in time. One could discuss assessments of individuals' cumulative, lifelong privacy loss, and use them as an input into the discussion of how small ϵ should be in each specific computation. Moreover, some socially desired cap on such cumulative privacy loss could be thought of as an individual's lifetime privacy budget. That budget is then to be carefully allocated, and prudently spent, across computations over one's lifetime to guarantee a desired amount of lifetime privacy.

Finally, it can be shown that differential privacy guarantees hold up under post-processing of their outputs: if one conducts an ϵ-differentially-private

[6] More generally, running any ℓ differentially-private mechanisms with guarantees $\epsilon_1,\ldots,\epsilon_\ell$ gives $\left(\sum_{i=1}^{\ell}\epsilon_i\right)$-differential privacy.

computation, one is then free to perform any subsequent computation on the output of that computation, and the result will still be ϵ-differentially private. In other words, once one has produced differentially-private statistics on a dataset, those statistics can be made public for all eternity, without concern that at some later date a clever hacker will find some new privacy-revealing weakness.

From a Bayesian point of view, differential privacy can be given the following interpretation: an observer with access to the output of a differentially-private function should draw almost the same conclusions whether or not one individual's data are included in the analyzed database, regardless of the observer's prior. This interpretation highlights that differential privacy is a property of the function (the mapping from databases into outcomes), not of the output (a particular outcome). Kasiviswanathan and Smith (2008) credit Cynthia Dwork and Frank McSherry with the first formulation of this interpretation, which can be formalized and proven equivalent to differential privacy.

The Bayesian "observer" may of course refer to anyone with access to the output of the function, including malicious attackers, (legitimate) advertisers on Facebook, or the readers of a research paper that reports some statistic. Notice that this Bayesian interpretation does not rule out performing analyses and reporting outcomes that vastly alter the observer's posterior view of the world, so long as the outcomes are not very sensitive to the presence or absence of any one individual in the original database. Our example above, regarding a differentially-private analysis that revealed a correlation between smoking and cancer, illustrates this point.

From Definition to Application: Noise and Sensitivity

With the concept of differential privacy in hand, consider the computation (and subsequent release) of the mean income of individuals in a database. While the mean might seem like a fairly innocuous statistic, all statistics reveal *something* about the data, and in certain worst-case situations, the mean might be quite revealing. For example, if the mean salary in a certain economics department prior to hiring a new faculty member is known to an observer (for instance, due to a previous release), then releasing the new mean after the hire reveals the new hire's salary. This is a variant of the so-called "differencing attack."

A simple technique for guaranteeing differential privacy is to add randomly generated noise to the true mean prior to its release. How much noise? Since differential privacy is a worst-case guarantee over all possible pairs of neighboring databases and over all possible outcomes, if the distribution of incomes is not a priori bounded, the noise would have to be unboundedly large (to guarantee that even the addition of an extreme outlier to the database would have little effect on the differentially-private statistic). With a limit on the range of incomes, however, one could add a limited amount of noise to the true mean in order to guarantee differential privacy.

More formally, when a function that we wish to compute on a database returns a real number, we say that the *sensitivity* of that function, denoted Δf, is the largest possible difference (in absolute value) between the two outputs one might get when applying that function to two neighboring databases.[7] The definition makes it clear that sensitivity is a property of the function, given a universe of possible databases, and is independent of the actual input database. Intuitively, this maximum difference between the values that the function could take on any two neighboring databases must be hidden in order to preserve differential privacy. We next focus on a technique that hides this maximum difference by adding noise, in the context of a concrete example.

A Single-Statistic Example: Mean Salary

To illustrate some of the delicate issues involved in actually carrying out a differentially-private computation, consider the release of mean salary among (all or some of) the faculty in an economics department. For concreteness, consider the following scenario: each faculty member is asked to voluntarily and confidentially agree to have their individual salary included in some database; statistics from the database are to be released in a differentially-private manner.

Notice that the details of this scenario, including details of the differentially-private mechanism to be used, can be publicly announced. What one aims to hide is only the confidential participation decision by any individual faculty. Our example will illustrate that this individual decision is easier to hide the larger the database is, the fewer statistics are to be published, and the less sensitive these statistics are given the considered universe of possible databases.

Dwork, McSherry, Nissim, and Smith (2006) show that one way to get an ϵ-differentially-private release of a statistic is to add "Laplace noise" to the (true) statistic prior to its release: that is, the noise is drawn from a Laplace distribution with mean equal to zero and standard deviation $= \sqrt{2}\,\Delta f/\epsilon$, where Δf is the sensitivity of the statistic—that is, the maximum difference in the statistic between any two neighboring databases—and ϵ is the parameter that quantifies the level of privacy we choose to guarantee.[8]

To apply this technique one therefore needs, first, to choose a value for ϵ (the smaller it is, the stronger is the privacy guarantee), and second, to calculate Δf. Remember that because these quantities do not depend on the underlying data, they are not in themselves private.

Choosing a value for ϵ: Recall that ϵ quantifies the maximum multiplicative difference possible between a differentially-private computation's outcome probabilities across two neighboring databases. In choosing a value for ϵ, one therefore chooses the maximum such difference that one is willing to allow under the differential privacy guarantee. But what should this maximum difference be? For the

[7] Formally, the sensitivity of a function f is $\Delta f = \max_{D,\,D'} |f(D) - f(D')|$, for $(D,\, D')$ neighboring databases.

[8] With scale parameter $b = \Delta f/\epsilon$, the probability density function of this distribution is $\frac{1}{2b}\, e^{-\frac{|x|}{b}}$ and its standard deviation is $\sqrt{2}\, b$. This distribution is a natural choice because its exponential form satisfies the multiplicative e^{ϵ} constraint in the differential privacy definition.

most part, the differential privacy literature is silent on this question. Developing the reasoning and intuition necessary for determining a socially desired value may take time. Concrete proposals may eventually emerge from a combination of philosophical and ethical inquiry, and social, political, and legislative processes, and could depend on context; further research is clearly needed.[9] For illustrative purposes only, in our mean salary example we will consider the values $\epsilon = 0.1$ and $\epsilon = 1$.

Calculating Δf*:* As mentioned above, to yield practical results our technique requires Δf to be bounded. Our example involves salary rather than total income, because salary is bounded from below (in the worst case, at zero). One still needs an upper bound, which cannot be naively calculated from the data, but should be a property of the known universe of possible salaries. For simplicity, we assume that it is known to be some $\bar{y}$. With these bounds, and with mean salary as our function of interest, the absolute value of the difference between the function applied to two neighboring databases will be less than or equal to the highest possible salary divided by the number (denoted by n) of individuals in the larger database: $|f(D) - f(D')| \leq \bar{y}/n$, for any two neighboring databases (D, D').

Since the number of faculty members participating in the database is not publicly known, the universe of possible databases includes the case $n = 1$, and therefore $\Delta f = \bar{y}$.[10] With such high sensitivity, a naive application of the Laplace noise technique yields a uselessly uninformative outcome at any n: the noise added to the true mean has standard deviation $\sqrt{2}\,\bar{y}/\epsilon$, which, even with $\epsilon = 1$, is larger than the upper bound on salaries.

An easy modification of the technique, however, yields noise that shrinks with n. The idea is to think of the mean as the function sum/n, that is, as a function of two statistics—the sum of salaries, and the sample size n—to be calculated and released in a differentially-private manner. One then divides the privacy budget ϵ between the two statistics: $\epsilon_{sum} + \epsilon_n = \epsilon$ (recall that the composition property allows such a division). The sensitivity of the sum of salaries is $\bar{y}$ because the maximum difference between the sum of salaries across two databases that differ only in the inclusion versus exclusion of one record is the upper bound on one additional salary. The sensitivity of n is 1, because by the definition of neighboring

[9] As Dwork et al. (2011) note in a defense of differential privacy:

> Yes, this research is incomplete. Yes, theorems of the following form seem frighteningly restrictive:
>
> > If an individual participates in 10,000 adversarially chosen databases, and if we wish to ensure that her cumulative privacy loss will, with probability at least $1 - e^{-32}$, be bounded by e^1, then it is sufficient that each of these databases will be $\epsilon = 1/801$-differentially private.
>
> But how else can we find a starting point for understanding how to relax our worst-case adversary protection? How else can we measure the effect of doing so? And what other technology permits one to prove such a claim?

[10] For simplicity (and conservativeness), we define mean salary in a database with 0 individuals to be at the lower bound 0.

databases, the difference between the number of records across any two neighboring databases is 1. The noise added to the two statistics would therefore have standard deviations $\sqrt{2}\,\bar{y}/\epsilon_{sum}$ and $\sqrt{2}/\epsilon_n$, respectively. Because the two statistics increase with n, the noise-to-true-statistic ratio of each vanishes asymptotically. With $\epsilon = 1$ and a favorable setting—a large department with high rate of voluntary participation in the database, and with mean salary not much below $\bar{y}$—the differentially-private release may convey some usable information about the true mean; but generally, the promise of the approach is more apparent on bigger data.

For illustration, consider $\epsilon = 1$, mean salary $\bar{y}$ (this is the unrealistically favorable case of all salaries in the department equal, at the upper limit), and $n = 30$. Then the standard deviation on the noise added by the Laplace technique would be $\left(\sqrt{2}/0.5\right)/30 = 9.4$ percent of each of the two (true) statistics, assuming for simplicity we divide the privacy budget equally between the two statistics.

For comparison, consider mean salary among the American Economic Association (AEA) membership in 2012, reported at 18,061 members (Rousseau 2013). Pick a tenfold stronger privacy guarantee, that is, $\epsilon = 0.1$, and assume a more realistic relation between the upper bound and the true mean, say, mean salary $= \bar{y}/10$. Assuming that all members volunteer to participate in the database, the much larger n means that in spite of these significantly more conservative conditions, the standard deviation on the noise added by the Laplace technique would be a much more tolerable 1.6 percent of the true sum of salaries and 0.16 percent of the true n (that is, a standard deviation of 28 members), if the privacy budget is again divided equally—rather than optimally—between the two statistics. Of course, things look still better with still bigger data and cleverer techniques.

Mean Salary Revisited: When the Database Size is Known

Dwork (2011a) suggests that "[s]ometimes, for example, in the census, an individual's participation is known, so hiding presence or absence makes no sense; instead we wish to hide the values in an individual's row." Our examples above could be modified to match such settings. Instead of a scenario where each faculty member is asked to voluntarily join a database, consider a different scenario where some administrative database with everyone's salaries is already known to exist. As above, statistics from the database are to be released in a differentially-private manner. Under this modified scenario, the databases from our examples above are now known to include all the faculty in an economics department and all AEA members.

In such settings, where participation is publicly known, it may make sense to modify our above definition of neighboring databases, from pairs "that are identical except that one of the databases has one additional row," to pairs of known size n, that differ in the content of exactly one row. In this form, differential privacy guarantees participants that if their true salary y were replaced with some fake salary $y' \in [0, \bar{y}]$, the probability of any given outcome would not change by much. With this modification, differentially-private release of mean salary requires only the sum of salaries to be computed and released in a differentially-private manner.

Historically, this alternate definition (with databases of fixed and publicly known *n*) was used in the first papers that sparked the differential privacy literature, and it is still used in much of the work on differential privacy and statistics—a body of work that has grown quickly over the past few years. Work in this area has repeatedly established the feasibility of achieving common statistical goals while maintaining differential privacy. Differentially-private versions have been developed for large classes of estimators—including those used routinely by empirical economists—often with little effective cost in terms of accuracy of the released results.[11]

Multiple Statistics

Of course, researchers wish to publish more than one statistic per database. In our example above, the privacy budget ϵ was divided between two statistics, *sum* and *n*, and each was then independently computed in a differential-privacy-preserving way. An alternative approach is to compute the two (or more) statistics jointly, which in some cases may significantly reduce the amount of added noise, as demonstrated by the case of histograms (Dwork, McSherry, Nissim, and Smith 2006).

Consider the release of a frequency histogram of salaries in some database. Treating each bin as a separate statistic (for example, "the count of rows with salary $0–10,000" is one statistic) would require dividing the privacy budget ϵ between the bins. The sensitivity (that is, Δf) of each such bin statistic is 1. It turns out that a generalized sensitivity concept applied jointly to the entire histogram is also 1, since adding an individual to a database always adds 1 to the count of exactly one of the bins and 0 to all others. In this example, calculating all the bins of the histogram jointly reduces the added noise because it saves the need to first divide the privacy budget between the statistics—a division whose cost in added noise increases with the number of bins. More generally, consider the maximum possible effect on a statistic of adding one individual to the database; if such a worst-case effect cannot occur on each of a group of statistics at the same time, considering them jointly may improve results.

One of the main focuses of research in differential privacy in recent years has been to develop algorithms that can handle very large numbers of queries jointly with far less noise than simple noise addition would permit. This large literature, which begins with Blum, Ligett, and Roth (2013) and continues with Hardt and Rothblum (2010) and Hardt, Ligett, and McSherry (2012), develops techniques for generating "synthetic data"—a set of valid database rows—that approximate the correct answers to

[11] Here we provide a few examples; see Heffetz and Ligett (2013) for a fuller reference list. Dwork and Lei (2009) demonstrate differentially-private algorithms for interquartile distance, median, and linear regression. Lei (2011) and Nekipelov and Yakovlev (2011) study differentially-private M-estimators. Smith (2008, 2011) finds that for almost any estimator that is asymptotically normal on independent and identically distributed samples from the underlying distribution (including linear regression, logistic regression, and parametric maximum likelihood estimators, under regularity conditions), there are differentially-private versions with asymptotically no additional perturbation. Along with these and other theoretical papers, a number of papers empirically investigate the performance of differentially-private estimators; useful starting points include Vu and Slavkovic (2009), Chaudhuri, Monteleoni and Sarwate (2011), and Abowd, Schneider and Vilhuber (2013).

all of a large, fixed set of queries. These techniques go far beyond just perturbing the data. Using ideas from geometry and computational learning theory, they generate synthetic data consisting of artificial records that cannot be connected with a single or small number of records in the original data. These approaches have started to show practicality, in the form of simple implementations that achieve good accuracy when tested on common statistical tasks using standard benchmark data (Hardt, Ligett and McSherry 2012), but much remains to be done.[12]

From Intuitions to Provable Guarantees

What insights can the differential privacy literature offer regarding the cautionary tales above? What tools could it provide for researchers working with data? We offer some thoughts, and highlight how different approaches respond differently to the inherent, unavoidable tradeoff between privacy and accuracy. We then discuss some of the limitations, as well as additional applications, of differential privacy.

Lessons and Reflections

In the Massachusetts Group Insurance Commission case—and, more generally, regarding the "anonymization" of complex datasets—lessons from differential privacy suggest considering two alternatives. First, one could release a differentially-private, synthetic (that is, artificial) version of the original database, after removing or coarsening complex fields such as text (which, without coarsening, would have made the data too high-dimensional for a synthetic version to be feasible in practice). The synthetic data would only be useful for a predetermined (though potentially quite large) set of statistics.[13] Second, one could withhold the full data but provide a differentially-private interface to allow researchers (or possibly the general public) to issue queries against the database.

Both approaches—providing a sanitized database, and providing sanitized answers to individual queries—face the inescapable tradeoff between privacy and usefulness (or accuracy). To achieve privacy, they limit usefulness in different ways: while the first approach limits in advance the type of queries (and hence of analysis)

[12] Another growing literature considers large sets of queries of a particular type, and aims to get a better understanding of the privacy–accuracy tradeoffs for a specific combined task. Beginning with Barak et al. (2007), one application that has received substantial attention is contingency tables, which are computed from sets of *k-way marginal* queries; see Heffetz and Ligett (2013) for references to more recent work.

[13] Kinney et al. (2011) provide evidence of the promise of synthetic data, describing the generation of an initial version of the SynLBD, a synthetic version of the US Census Bureau's Longitudinal Business Database (https://www.census.gov/ces/dataproducts/synlbd/). Their synthetic database is designed to preserve aggregate means and correlations from the underlying, confidential data. While their algorithm for generating synthetic data is not explicitly designed to preserve any particular level of differential privacy, they present an interesting illustrative assessment—inspired by differential privacy—of privacy risk. See Machanavajjhala et al. (2008) for an earlier exploration of the challenges of generating privacy-preserving synthetic data from other Census datasets.

possible, the second maintains flexibility but might more severely limit the overall *number* of queries, since the system has to manage a privacy budget dynamically (and hence potentially less efficiently) to answer arbitrary queries as they arrive and would eventually run out of its ϵ privacy budget and then would have to refuse new queries. This idea of an overall limit—a privacy budget that places a quantifiable constraint on any approach—is a useful metaphor that highlights one of the costs of preserving privacy: it imposes fundamental limits on how much information can be revealed about the data.

In the case of the AOL debacle, the data to be released were so high-dimensional (the space of rows being all possible search histories) that they clearly could not be handled with differential privacy without some initial dimension reduction. This point in itself is worth observing—free text and other high-dimensional data (for example, genetic information) are potentially extraordinarily revealing, and deserve careful attention. Korolova, Kenthapadi, Mishra, and Ntoulas (2009), in response to AOL's data release, propose releasing an alternate data structure called a *query click graph*, and demonstrate on real search log data that a differentially-private query click graph can be used to perform some research tasks that one might typically run on search logs.[14] As the authors note, it remains to be seen how broadly useful such sanitized data are, but such findings "offer a glimmer of hope" on reconciling research usability with privacy concerns.

Regarding the Netflix challenge, the *manner* in which it was carried out—releasing a large, very high-dimensional dataset—is difficult to implement in a differentially-private way. However, the *goals* of the challenge—namely, producing recommendations from collective user behavior—could be achievable while guaranteeing differential privacy. To explore this possibility, McSherry and Mironov (2009) evaluate several of the algorithmic approaches used in the challenge, showing that they could have been implemented in a differentially-private manner (via privacy-preserving queries issued against the database) without significant effect on their accuracy.

The Facebook goal—giving advertisers a count of the number of times their ad was shown—at first sounds as if it might be well-suited to differential privacy: one could simply add an appropriate level of noise to the true count. However, charging advertisers based on noisy counts may be considered objectionable, and regardless, privacy would then degrade as the number of ad campaigns increased (or, alternatively, Facebook would have to discontinue the service once they ran out of a certain ϵ budget to which they had committed). Even if we assume that advertisers do not share the statistics Facebook reports to them (and so perhaps each advertiser can be apportioned a separate privacy budget rather than sharing a single budget among

[14] The differentially-private query click graph the authors propose to publish is a noisy version of a "graph where the vertices correspond to both queries and URLs and there is an edge from a query to a URL with weight equal to the number of users who click on that URL given they posed the query. Each query node is labeled by the number of times this query was posed in the log. Similarly, there is an edge from one query to another query with weight equal to the number of users who posed one query and reformulated to another."

them all), large advertisers likely run so many campaigns that the noise necessary in order to ensure any reasonable level of privacy would swamp any signal in the data. Korolova (2011) suggests that an approach like differential privacy would provide the most robust starting point for privately addressing Facebook's goal, and discusses these and other challenges that leave the targeted-ads application an intriguing open problem.

More generally, what tools and other thoughts could differential privacy potentially offer to researchers who work with data?

While no standardized implementations yet exist, and while conventions (for example regarding setting ϵ) have not yet been established, a rich set of theoretical results already provides the foundations for a useful toolbox for the data-based researcher.

If one would like to publish a single statistic (or a small set of statistics), differentially-private estimator versions might already exist. As discussed above, the accuracy cost imposed by the added noise may be negligible when the sample size n is sufficiently large.

Regardless of whether the statistic of interest has received attention in the differential privacy literature, the study of differential privacy suggests that it may be helpful to understand the *sensitivity* of the statistic to changes in one person's information—how much can varying one entry in the database affect the statistic? Such understanding not only helps assess how much noise one could add to achieve differential privacy in the simplest manner; it is also helpful for getting an intuitive understanding of how and why a statistic might be revealing. There are also differentially-private techniques that can provide good accuracy even on high-sensitivity statistics, so long as the statistics are "well-behaved" on the data of interest (Nissim, Raskhodnikova, and Smith 2007; Dwork and Lei 2009). Finally, if one wishes to publish a large set of statistics or produce sanitized data, as we discussed, general purpose techniques for doing so already exist, but it is possible that a researcher's particular properties of interest would be even better served by a specialized differentially-private mechanism.

The centrality of the notion of sensitivity to the ongoing research on differential privacy highlights an old truth from a new perspective: it underscores the importance of thinking about the robustness of the statistics we report. If reporting a statistic while preserving privacy requires introducing an unacceptable level of randomness, this may indicate that one's dataset is too small for one's desired levels of privacy and accuracy, but it may also suggest that worst-case scenarios exist under which the statistic is simply not robust—that is, it may be quite sensitive to potential individual outliers.

Finally, the concept of differential privacy offers one way to quantify the often loosely used notions of privacy and anonymity. Researchers may find such quantification helpful in thinking about whether study participants should be given a different, more qualified, promise of privacy/anonymity than is typically given—especially in settings where implementing a specific guarantee (not necessarily the one offered by differential privacy) is not practical.

Limitations

Like any other rigorous approach, the differential privacy approach makes some assumptions that may be questioned. For example, it assumes that an individual's private data are conveniently represented as a row in a database (an assumption violated by, for example, social network data), and it implicitly assumes that a particular definition—involving a bound on the ratio of outcome probabilities—captures what we mean by privacy.

Strong privacy guarantees necessarily obscure information. The intentional introduction of randomness into published outcomes may require adjustments to specific implementations of scientific replication. More generally, for some applications the very idea of deliberately introducing randomness is problematic: preventable mistakes such as allocating the wrong resources to the wrong groups or making the wrong policy decisions could have grave consequences.

As hinted above, a potential limitation of differentially-private mechanisms producing synthetic data is that they require the data analyst to specify the query set in advance. In many research settings, one may not know in advance exactly which statistics one wishes to compute or what properties of a dataset must be preserved in order for the data to be useful. There is a natural tension between an analyst's desire to "look at the data" before deciding what to do with them and a privacy researcher's desire that all computations that touch the original data be made formal and privacy-preserving.

As a practical response to this limitation, rather than attempting to define the query set a priori, one could consider using some of the privacy budget for *interactive queries* where the analyst poses queries one at a time and receives privacy-preserving answers, and could then base the choice of future queries on the answers previously received. The analyst thus establishes via this sequence of interactive queries what properties of the original database to preserve in the sanitized version, and can then use the rest of the privacy budget to produce sanitized data.

More generally, with the growth of big data, the "look at the data" approach is destined to change: in practical terms, "looking" at enormous datasets means running analyses on them. As soon as "looking at the data" has a technical meaning, one can try to enable it in a privacy-preserving manner.

Finally, for particular applications, differentially-private mechanisms may not yet have been developed, or the existing technology may not enable a satisfying privacy–accuracy tradeoff. Such limitations may merely suggest that more research is needed. Even when a satisfying privacy–accuracy tradeoff is formally proved impossible, in many cases such impossibility results are not specific to differential privacy, but rather reflect that certain tasks are inherently revealing and hence may be fundamentally incompatible with privacy.

Differential Privacy and Mechanism Design

The last few years have seen a growth of interest in a number of topics at the intersection of differential privacy and economics, in particular, privacy and mechanism design; see Pai and Roth (2013) for a survey. Some of the key questions under consideration include how one might incorporate privacy considerations into utility functions and how one might model the value of privacy. Work in this area includes Ghosh and Roth (2011), Nissim, Orlandi, and Smorodinsky (2012), Fleischer and Lyu (2012), Roth and Schoenebeck (2012), Ligett and Roth (2012), Xiao (2013), Chen et al. (2013), and Ghosh and Ligett (2013).

From a mechanism design point of view, the differential privacy guarantee—that a participant's inclusion or removal from the database would have almost no effect on the outcome—could be viewed as a valuable guarantee even in the absence of privacy concerns. In particular, consider settings where participants in a database can misrepresent their individual data, and have preferences over the possible outcomes of a function to be computed from the data. A differentially-private computation implies that such participants have only limited incentive to lie, because lying would have only a limited effect on the outcome. McSherry and Talwar (2007) were the first to observe that differential privacy implies asymptotic (or approximate) "strategyproofness" (or truthfulness). Of course, under differential privacy, not only do individuals have almost no incentive to lie; they also have almost no incentive to tell the truth (Nissim, Smorodinsky and Tennenholtz 2012; Xiao, 2013); however, a small psychological cost of lying could strictly incentivize truth-telling.

This implication of approximate truthfulness may be of particular interest to researchers who wish to gather survey data in settings where participation is voluntary and the accuracy of responses cannot be easily verified. More generally, the asymptotic strategyproofness implied by differential privacy inherits some of the latter's useful additional properties. For example, because of the way differential privacy extends to groups of *k* individuals, this strategyproofness extends to the case of *k* colluding individuals (a collusion resistance that deteriorates with the coalition size *k*). The strategyproofness also holds under repeated application of the mechanism (again, with a deterioration as the number of repetitions rises). Finally, this asymptotic truthfulness has inspired further work on privacy-preserving mechanism design (Huang and Kannan 2012; Kearns, Pai, Roth, and Ullman 2014) and has enabled differential privacy to be used as a tool in the design of truly strategyproof mechanisms (for example, Nissim, Smorodinsky, and Tennenholtz 2012).

Concluding Thoughts

Privacy concerns in the face of unprecedented access to big data are nothing new. More than 35 years ago, Dalenius (1977) was discussing "the proliferation of computerized information system[s]" and "the present era of public concern about 'invasion of privacy.'" But as big data get bigger, so do the concerns. Greely (2007) discusses genomic databases, concluding:

> The size, the cost, the breadth, the desired broad researcher access, and the likely high public profile of genomic databases will make these issues especially

> important to them. Dealing with these issues will be both intellectually and politically difficult, time-consuming, inconvenient, and possibly expensive. But it is not a solution to say that "anonymity" means only "not terribly easy to identify," . . . or that "informed consent" is satisfied by largely ignorant blanket permission.

Replacing "genomic databases" with "big data" in general, our overall conclusion may be similar.

The stories in the first part of this paper demonstrate that relying on intuition when attempting to protect subject privacy may not be enough. Moreover, privacy failures may occur even when the raw data are never publicly released and only some seemingly innocuous *function* of the data, such as a statistic, is published.

The differential privacy literature provides a framework for thinking more precisely about privacy–accuracy tradeoffs. With computer scientists using phrases such as "the amount of privacy loss" and "the privacy budget," the time seems ripe for more economists to join the conversation. Is privacy a term in the utility function that can in principle be compared against the utility from access to accurate data? Should individuals be entitled to privacy—or to a certain lifelong privacy budget—as a basic right, or as a property right? Should a certain privacy budget be allocated across interested users of publicly owned data, like Census data, and if so, how? If a budget were allocated to individuals, should fungible, transferable ϵ be allowed to be sold in markets from private individuals to potential data users, and if so, what would its price be?

When big data means large n, an increasing number of common computations can be achieved in a differentially-private manner with little cost to precision. It is not inconceivable that within a few years, many of the computations that have been—and those that are yet to be—proven achievable in theory will be applied in practice. Dwork and Smith (2010) write that they "would like to see a library of differentially-private versions of the algorithms in R and SAS." In a similar spirit, we would be happy to have a differentially-private option in estimation commands in STATA. But ready-to-use, commercial-grade applications will not be developed without sufficient demand from potential users. We hope that the incorporation of privacy considerations into the vocabulary of empirical researchers will help raise demand, and stimulate further discussion and research—including, we hope, regarding additional approaches to privacy.

Until such applications are available, it might be wise to pause and reconsider researchers' promises and, more generally, obligations to subjects. When researchers (and Institutional Review Boards!) are confident that the data pose only negligible privacy risks—as in the case of some innocuous small surveys and lab experiments—it may be preferable to replace promises of complete anonymity with promises for "not terribly easy" identification or, indeed, with no promises at all. In particular, researchers could explicitly inform subjects that a determined attacker may be able to identify them in posted data, or even learn things about them merely by looking at the empirical results of a research paper. We caution against taking the naive alternate route of simply refraining from making seemingly harmless data publicly available; freedom of information, access to data, transparency, and scientific replication are all dear to us.[15] Of course, the tradeoffs, and in particular the question of what privacy risks are negligible and what data are harmless, should be considered and discussed; a useful question to ask ourselves may resemble the old "*New York Times* test": Would our subjects mind if their data were identified and published in the *New York Times*?

■ *For useful comments on an early draft, we thank Dan Benjamin, Avrim Blum, Kamalika Chaudhuri, Hank Greely, Aleksandra Korolova, Frank McSherry, Ilya Mironov, Denis Nekipelov, Kobbi Nissim, Ted O'Donoghue, Grant Schoenebeck, Moses Shayo, Adam Smith, Latanya Sweeney, Kunal Talwar, and Jonathan Ullman. We are also grateful to the editors—David Autor, Chang-Tai Hsieh, Ulrike Malmendier, and Timothy Taylor—for their encouragement and comments. Ligett's work was supported in part by an NSF CAREER award (CNS-1254169), the US–Israel Binational Science Foundation (grant 2012348), the Charles Lee Powell Foundation, a Google Faculty Research Award, and a Microsoft Faculty Fellowship.*

References

Abowd, John M., Matthew J. Schneider, and Lars Vilhuber. 2013. "Differential Privacy Applications to Bayesian and Linear Mixed Model Estimation." *Journal of Privacy and Confidentiality* 5(1).

Barak, Boaz, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. 2007. "Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release." In *PODS '07: Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems,* pp. 273–82. ACM Digital Library.

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

Barth-Jones, Daniel C. 2012. "The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now." Available at SSRN: http://ssrn.com/abstract=2076397.

Blum, Avrim, Katrina Ligett, and Aaron Roth. 2013. "A Learning Theory Approach to Noninteractive Database Privacy." *Journal of the ACM* 60(2): Article 12.

Chaudhuri, Kamalika, Claire Monteleoni, and Anand D. Sarwate. 2011. "Differentially Private Empirical Risk Minimization." *Journal of Machine Learning Research* 12(March): 1069–1109.

Chen, Yiling, Stephen Chong, Ian A. Kash, Tal Moran, and Salil P. Vadhan. 2013. "Truthful Mechanisms for Agents that Value Privacy." In *EC '13: Proceedings of Fourteenth ACM Conference on Electronic Commerce*, pp. 215–32. ACM Digital Library.

[15] Flood, Katz, Ong, and Smith (2013) provide a comprehensive discussion of such a transparency–confidentiality tradeoff in a context that is very different from ours, yet of great interest to economists—that of financial supervision and regulation.

Dalenius, Tore. 1977. "Towards a Methodology for Statistical Disclosure Control." *Statistisk tidskrift*, 15: 429–44.

Dinur, Irit, and Kobbi Nissim. 2003. "Revealing Information While Preserving Privacy." In *PODS '03: Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems,* pp. 202–210. ACM Digital Library.

Dwork, Cynthia. 2006. "Differential Privacy." ICALP '06: Proceedings of the 33rd *International Conference on Automata, Languages and Programming,* pp. 1–12. ACM Digital Library.

Dwork, Cynthia. 2011a. "A Firm Foundation for Private Data Analysis." *Communications of the ACM* 54(1): 86–95.

Dwork, Cynthia. 2011b. "The Promise of Differential Privacy: A Tutorial on Algorithmic Techniques." In *FOCS '11: Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science*, pp. 1–2. ACM Digital Library.

Dwork, Cynthia, and Jing Lei. 2009. "Differential Privacy and Robust Statistics." In *STOC '09: Procceedings of the Forty-first Annual ACM Symposium on Theory of Computing,* pp. 371–80. ACM Digital Library.

Dwork, Cynthia, and Kobbi Nissim. 2004. "Privacy-Preserving Datamining on Vertically Partitioned Databases." In *Advances in Cryptology—CRYPTO* 2004, *24th Annual International Cryptology Conference,* pp. 528–44.

Dwork, Cynthia, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. "Our Data, Ourselves: Privacy via Distributed Noise Generation." In *EUROCRYTO '06: Proceedings of the 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. ACM Digital Library.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis." In *TTC '06: Proceedings of the Third Conference on Theory of Cryptography,* pp. 265–84. ACM Digital Library.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2011. "Differential Privacy: A Primer for the Perplexed." Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, WP. 26.

Dwork, Cynthia, and Adam Smith. 2010. "Differential Privacy for Statistics: What We Know and What We Want to Learn." *Journal of Privacy and Confidentiality* 1(2): 135–54.

Fleischer, Lisa, and Yu-Han Lyu. 2012. "Approximately Optimal Auctions for Selling Privacy When Costs Are Correlated with Data." In *EC '12: Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 568–85. ACM Digital Library.

Flood, Mark, Jonathan Katz, Stephen J. Ong, and Adam Smith. 2013. "Cryptography and the Economics of Supervisory Information: Balancing Transparency and Confidentiality." Federal Reserve Bank of Cleveland Working Paper 1312.

Ghosh, Arpita, and Katrina Ligett. 2013. "Privacy and Coordination: Computing on Databases with Endogenous Participation." In *EC '13: Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, pp. 543–60. ACM Digital Library.

Ghosh, Arpita, and Aaron Roth. 2011. "Selling Privacy at Auction." *EC '11: Proceedings of the 12th ACM Conference on Electronic Commerce*, pp. 199–208. ACM Digital Library.

Greely, Henry T. 2007. "The Uneasy Ethical and Legal Underpinnings of Large-Scale Genomic Biobanks." *Annual Review of Genomics and Human Genetics* 8: 343–64.

Hardt, Moritz, Katrina Ligett, and Frank McSherry. 2012. "A Simple and Practical Algorithm for Differentially Private Data Release." In *Advances in Neural Information Processing Systems* 25: 2348–56.

Hardt, Moritz, and Guy N. Rothblum. 2010. "A Multiplicative Weights Mechanism for Privacy-Preserving Data Analysis." In *FOCS '10: Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, pp. 61–70. ACM Digital Library.

Heffetz, Ori, and Katrina Ligett. 2013. "Privacy and Data-Based Research." NBER Working Paper 19433.

Huang, Zhiyi, and Sampath Kannan. 2012. "The Exponential Mechanism for Social Welfare: Private, Truthful, and Nearly Optimal." In *FOCS '12: Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, pp. 140–49. ACM Digital Library.

Kasiviswanathan, Shiva Prasad, and Adam Smith. 2008. "A Note on Differential Privacy: Defining Resistance to Arbitrary Side Information." Unpublished paper.

Kearns, Michael, Mallesh Pai, Aaron Roth, and Jon Ullman. 2014. "Mechanism Design in Large Games: Incentives and Privacy." In *ITCS '14: Proceedings of the 5th Conference on Innovations in Theoretical Computer Science*, pp. 403–410. ACM Digital Library.

Kinney, Satkartar K., Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M. Abowd. 2011. "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database." *International Statistical Review* 79(3): 362–84.

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

Korolova, Aleksandra. 2011. "Privacy Violations Using Microtargeted Ads: A Case Study." *Journal of Privacy and Confidentiality* 3(1).

Korolova, Aleksandra, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. 2009. "Releasing Search Queries and Clicks Privately." In *WWW '09: Proceedings of the 18th International Conference on the World Wide Web,* pp. 171–180. ACM Digital Library.

Kumar, Ravi, Jasmine Novak, Bo Pang, and Andrew Tomkins. 2007. "On Anonymizing Query Logs via Token-based Hashing." In *WWW '07: Proceedings of the 16th International Conference on the World Wide Web,* pp. 629–38. ACM Digital Library.

Lei, Jing. 2011. "Differentially Private M-Estimators." *Advances in Neural Information Processing Systems* 24: 361–69.

Ligett, Katrina, and Aaron Roth. 2012. "Take it or Leave it: Running a Survey When Privacy Comes at a Cost." *Internet and Network Economics* (WINE '12: Proceedings of the 8th international conference on Internet and Network Economics), pp. 378–391.

Machanavajjhala, Ashwin, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. 2008. "Privacy: Theory Meets Practice on the Map." In *ICDE '08: Proceedings of the 24th IEEE International Conference on Data Engineering,* pp. 277–86. ACM Digital Library.

McSherry, Frank, and Ilya Mironov. 2009. "Differentially Private Recommender Systems: Building Privacy into the Net." In *KDD' 09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* pp. 627–36. ACM Digital Library.

McSherry, Frank, and Kunal Talwar. 2007. "Mechanism Design via Differential Privacy." In *FOCS '07: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science,* pp. 94–103. ACM Digital Library.

Narayanan, Arvind, and Vitaly Shmatikov. 2008. "Robust De-anonymization of Large Sparse Datasets." In *SP '08: 2008 IEEE Symposium on Security and Privacy,* pp. 111–25. ACM Digital Library.

Nekipelov, Denis, and Evegeny Yakovlev. 2011. "Private Extremum Estimation." Unpublished paper.

Nissim, Kobbi, Claudio Orlandi, and Rann Smorodinsky. 2012. "Privacy-aware Mechanism Design." In *EC '12: Proceedings of the 13th ACM Conference on Electronic Commerce,* pp. 774–89. ACM Digital Library.

Nissim, Kobbi, Sofya Raskhodnikova, and Adam Smith. 2007. "Smooth Sensitivity and Sampling in Private Data Analysis." In *STOC '07: Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing,* pp. 75–84.

Nissim, Kobbi, Rann Smorodinsky, and Moshe Tennenholtz. 2012. "Approximately Optimal Mechanism Design via Differential Privacy." In *ITCS '12: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference,* pp. 203–213. ACM Digital Library.

Ohm, Paul. 2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." *UCLA Law Review* 57(6): 1701–77.

Pai, Mallesh, and Aaron Roth. 2013. "Privacy and Mechanism Design." *Sigecom Exchanges* 12(1).

Roth, Aaron, and Grant Schoenebeck. 2012. "Conducting Truthful Surveys, Cheaply." In *EC '12: Proceedings of the 13th ACM Conference on Electronic Commerce,* pp. 826–43. ACM Digital Library.

Rousseau, Peter L. 2013. "Report of the Secretary." *American Economic Review* 103(3): 669–72.

Smith, Adam. 2008. "Efficient, Differentially Private Point Estimators." arXiv:0809.4794.

Smith, Adam. 2011. "Privacy-Preserving Statistical Estimation with Optimal Convergence Rates." In *STOC '11: Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing,* pp. 813–22. ACM Digital Library.

Sweeney, Latanya. 1997. "Weaving Technology and Policy Together to Maintain Confidentiality." *Journal of Law, Medicine & Ethics* 25(2–3): 98–110.

Sweeney, Latanya. 2002. "Achieving *k*-anonymity Privacy Protection Using Generalization and Suppression." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10(5): 571–88.

Sweeney, Latanya. 2013. "Matching Known Patients to Health Records in Washington State Data." http://thedatamap.org/1089-1.pdf.

Sweeney, Latanya, Akua Abu, and Julia Winn. 2013. "Identifying Participants in the Personal Genome Project by Name." http://dataprivacylab.org/projects/pgp/1021-1.pdf.

Vu, Duy, and Aleksandra Slavkovic. 2009. "Differential Privacy for Clinical Trial Data: Preliminary Evaluations." In *ICDMW '09: Proceedings of the 2009 IEEE International Conference on Data Mining Workshops,* pp. 138–43. ACM Digital Library.

Xiao, David. 2013. "Is Privacy Compatible with Truthfulness?" In *ITCS '13: Proceedings of the 4th Conference on Innovations in Theoretical Computer Science,* pp. 67–86. ACM Digital Library.