Review in Advance first posted online on December 21, 2016. (Changes may still occur before final publication online and in print.)

Exposed! A Survey of Attacks on Private Data

Cynthia Dwork,[1] Adam Smith,[2] Thomas Steinke,[3] and Jonathan Ullman[4]

[1]Microsoft Research, Mountain View, California 94043; email: dwork@microsoft.com

[2]Department of Computer Science and Engineering, Pennsylvania State University, State College, Pennsylvania 16802; email: asmith@psu.edu

[3]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138; email: tsteinke@seas.harvard.edu

[4]College of Computer and Information Science, Northeastern University, Boston, Massachusetts 02115; email: jullman@ccs.neu.edu

Annu. Rev. Stat. Appl. 2017. 4:12.1–12.24

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-060116-054123

Keywords

privacy, privacy attacks, reidentification, reconstruction attacks, tracing attacks, differential privacy

Abstract

Privacy-preserving statistical data analysis addresses the general question of protecting privacy when publicly releasing information about a sensitive dataset. A privacy attack takes seemingly innocuous released information and uses it to discern the private details of individuals, thus demonstrating that such information compromises privacy. For example, re-identification attacks have shown that it is easy to link supposedly de-identified records to the identity of the individual concerned. This survey focuses on attacking aggregate data, such as statistics about how many individuals have a certain disease, genetic trait, or combination thereof. We consider two types of attacks: reconstruction attacks, which approximately determine a sensitive feature of all the individuals covered by the dataset, and tracing attacks, which determine whether or not a target individual's data is included in the dataset. We also discuss techniques from the differential privacy literature for releasing approximate aggregate statistics while provably thwarting any privacy attack.

1. INTRODUCTION

Beginning in the mid-2000s, the field of privacy-preserving statistical analysis of data has witnessed an influx of ideas developed some two decades earlier in the cryptography community. These include the formalization of the notion of a privacy adversary, the introduction of a meaningful measure of privacy loss, the development of general and robust definitions of privacy, the development of a theory of how privacy loss compounds over repeated privacy-preserving data access (a process known as composition), the design of basic privacy-preserving computational building blocks, the development of techniques for combining these basic building blocks in creative ways to obtain privacy-preserving algorithms for sophisticated analytical tasks, and an investigation of the limits of what can be achieved while preserving privacy. Privacy attacks—that is, algorithms for the privacy adversary to execute—are central to establishing fundamental limits of what is possible; they also play a seminal role in formulating achievable privacy goals. Privacy attacks are the subject of this article.

We focus on the simple scenario in which there is a dataset x containing sensitive information, and the goal is to release statistics about the dataset to the public. These statistics may be fixed in advance, or may be chosen by the analyst, who queries the dataset. Speaking intuitively (because we have not yet defined privacy), the goal in privacy-preserving data analysis is to protect the privacy of the individual records in the dataset, even if the analyst maliciously chooses queries according to an attack strategy designed to compromise privacy. We restrict our discussion to a single analyst, as a collection of colluding analysts can be modeled by a single analyst.

To study a class of attacks, we need to characterize the notion of success for the attacker, that is, what it means for the adversary to win. This can make sense even before settling on a formal definition of privacy. For example, we can set a ridiculously low threshold for privacy, or, equivalently, a high threshold for what constitutes a privacy break, such as "The adversary can correctly guess the sickle cell status of 99.999% of the members in the dataset." Most people would agree that success of this type is inconsistent with any reasonable notion of privacy, so an attack that achieves this goal on arbitrary datasets ostensibly protected by a given technique must be viewed as a repudiation of the protection technique. By the same token, if the protection technique provably satisfies a candidate definition of privacy, then such an attack refutes the value of the definition.

The remainder of this article is organized as follows. We begin with a discussion of three adversarial goals: re-identification; reconstruction, which is precisely the kind of 99.999% correct guessing just described; and tracing, in which the adversary determines whether a given individual is, or is not, present in a given dataset. Tracing can be significant if, for example, the dataset comprises medical records of participants in a pharmaceutical trial or patient records from an abortion clinic. We also present some basic definitions that will be used throughout. Reconstruction and tracing attacks are then surveyed in Sections 2 and 3, respectively. Finally, Section 4 discusses differential privacy, a definition of privacy tailored to statistical data analysis, and highlights a variant that achieves the limits established by the attacks. (Although differential privacy is a worst-case notion of privacy, it is interesting that our attacks require no strong assumptions either about the data or the information available to the adversary.)

1.1. Adversarial Goals and Resources

Achieving different adversarial goals may require different resources, so to fully specify an attack, we must also specify the resources to which the adversary has access. Examples of resources include computational capabilities and additional, or auxiliary, information beyond what is supplied by interacting with the dataset. Examples of useful auxiliary information might be personal details about an individual, known, for example, to a sibling or coworker, such as the approximate dates

on which one has watched a few movies on Netflix (Narayanan & Shmatikov 2008), and outputs from a product recommendation system (Calandrino et al. 2011).

First, we formalize the computational model for interacting with a dataset. Roughly speaking, the raw data remain hidden from the data analyst—who is also our adversary. Information is obtained by posing a query, which is simply a function mapping datasets to a range, such as the real numbers, and receiving a response in the range. Note that the queries may be specified ahead of time, for example, when a government agency decides on a set of tables to release, or may be specified implicitly, for example, when a release of synthetic data promises to preserve certain statistics of the original data. Regardless, the algorithm that provides the response is called a mechanism. This survey focuses on linear queries such as, for example, "What fraction of the rows in the dataset satisfy property P?" (see Definition 1 below).

> **Definition 1** (Mechanism). A mechanism is a randomized algorithm, often denoted $\mathcal{M}$, mapping datasets to an arbitrary set of outputs. Let x be an arbitrary dataset. Because $\mathcal{M}$ is randomized, $\mathcal{M}(x)$ yields a probability distribution on the range of $\mathcal{M}$.

As noted above, datasets will be collections of rows, and the mechanisms will provide approximate answers to linear queries over the dataset. Each dataset row will correspond to the data of a single individual.[1] Thus, in a dataset containing information about physical attributes of a collection of individuals, a row might hold the height, weight, and age of a single individual, a query might ask, "What fraction of the members of the dataset are over six feet tall?", and the mechanism might compute the true answer to the query, and produce as output the sum of the true answer and some random noise. In this survey the attacker's access to the dataset will be exclusively through these mechanisms.

1.1.1. Re-identification/de-anonymization. The technical literature and popular press frequently speak about re-identifying data. Such references implicitly assume an approach to privacy protection in which individual data records, containing explicit identifying information, are putatively de-identified or anonymized. Re-identification refers to reversing this step, tracing an individual record back to its human source. Although re-identification may seem difficult when a dataset is considered in isolation, anyone looking at supposedly de-identified data who also knows auxiliary information about a member of the dataset may well be in a position to re-identify. By linking public, not anonymous records, such as voter registration records, with de-identified data, strangers can do this too (Sweeney 1997, Narayanan & Shmatikov 2008). Indeed, the richer the dataset, the greater the set of possibilities for useful auxiliary information, and a host of results suggest that "de-identified data isn't", meaning that it is either not de-identified or no longer can serve as data. In the words of the President's Council of Advisors on Science and Technology (2014, p. 38), "Anonymization of a data record might seem easy to implement. Unfortunately, it is increasingly easy to defeat anonymization by the very techniques that are being developed for many legitimate applications of big data." For this reason we focus herein on the privacy risks posed by the release of statistics.

1.1.2. Reconstruction. Reconstruction is most easily understood by thinking of the dataset as a collection of rows, one per individual. Imagine that each row contains a large amount of nonprivate identifying information and a secret bit, one per individual, for example, indicating

[1] In fact, the attacks we discuss are robust to a fair amount of noise in the data.

whether or not the individual has the gene for Alzheimer's disease. The goal in a reconstruction attack is to determine the secret bits for nearly all individuals in the dataset.

> **Definition 2** (Reconstruction). Consider an n-row dataset in which each row contains a unique identifier and a single bit, possibly with additional information. For example, the identifiers might be the numbers $1, 2, \ldots, n$ and the bit might be the sickle cell status. Let b be the column vector of the bits. The reconstruction goal is to produce a vector c of n bits that agrees with b in all but $o(n)$ locations.

A few remarks are in order. First, the identifier is an abstraction: Individuals could be identified, for example, by a collection of attributes. There is no need for the adversary waging a reconstruction attack to know an ordering on the rows. Rather, the adversary will learn that the individual with a given set of attributes has a given sickle cell status. At the time of the attack, the attacker might not know the identity of the individual who actually has this set of attributes. This is scant comfort, however, as the attacker might learn such information at a later date. Second, reconstruction attacks can be launched against a subset of the rows of a dataset by proper formulation of the linear query. An example of a properly formed linear query would be, on the members of an extended family, "what fraction of the rows in the dataset correspond to members of family F that are over six feet tall?"

There is by now a rich literature showing that any mechanism providing overly accurate answers to too many linear queries is blatantly nonprivate, meaning that it succumbs to a reconstruction attack. Indeed, there is a single attack strategy that succeeds against all such overly accurate answering of too many queries. Here, "too many" is quite small (e.g., only n queries) and "overly accurate" means having fractional error on the order of $o(1/\sqrt{n})$. This literature is the subject of Section 2.

1.1.3. Tracing. Reconstruction represents spectacular success on the part of the adversary, or, conversely, a spectacular failure of the putative privacy mechanisms. Tracing—that is, determining whether or not a specific individual is a member of a given dataset—is a much more modest adversarial goal. (There are settings in which tracing attacks are possible, but reconstruction attacks are provably impossible.) [2]

Tracing entered the popular consciousness when a group of researchers showed how to use 1-way marginals, specifically allele frequency statistics in a genome-wide association study, together with the DNA of a target individual and allele frequency statistics for the general population, to determine the target's presence or absence in the study (Homer et al. 2008). In response, the US National Institutes of Health and the Wellcome Trust changed the access policy to statistics of this type in the studies they fund. Section 3 surveys results in tracing.

1.1.4. The attack that isn't: correlation detection. We are interested in statistical analysis of data—for example, learning facts about a population such as "smoking causes cancer." Releasing the information that smoking and cancer are correlated may reveal sensitive medical information about an individual who is known to smoke. However, we do not view this as a privacy compromise, as facts of this type can be learned even if the given individual is not in

[2] The name has its roots in the close connection to the traitor tracing problem in cryptography (Chor et al. 1994); see Dwork et al. (2009).

the dataset. These facts about the population as a whole are precisely what we seek to learn in statistical data analysis. Several works in the literature discuss attacks that in fact consist of correlation detection. We discuss the distinction further in Section 4.

1.1.5. Differential privacy. Differential privacy ensures that even a highly informed adversary, knowing a dataset x and an additional data record r, and interacting with a dataset $y \in \{x \cup \{r\}, x\}$ through a differentially private mechanism, cannot determine whether $y = x$ or $y = x \cup \{r\}$ (see Section 4 for details). Thus, differential privacy by definition prevents tracing; it also protects against reconstruction and re-identification. At the same time, it permits the analyst to learn precisely the type of statistical correlations just discussed.

Surprisingly, this very strong guarantee comes at no extra cost in accuracy, in the following sense: The bounds shown in Section 4 for achieving differential privacy match the limits imposed by reconstruction and tracing attacks in Sections 2 and 3.

2. RECONSTRUCTION ATTACKS

Suppose we have a dataset of n individuals. For each individual i, we have information (x_i, s_i). There is one bit, denoted s_i, that is considered sensitive and unknown a priori to an attacker (perhaps it indicates i's political party affiliation, diabetes status, or lack of interest in Bayesian decision theory). The remainder of the record, denoted x_i, is public and easily available—for example, the demographic information visible on i's Facebook account.

A well-intentioned curator might want to release various statistics about how the secret vector $s = (s_1, \ldots, s_n)$ relates to the public variables x. For example, if each x_i is a list of d binary attributes, $x_i = (x_i(1), \ldots, x_i(d)) \in \{0,1\}^d$, then one might want to release, for example:

1. The joint marginal distribution of s_i with each of the attributes—that is, for each j, a 2×2 contingency table indicating how many records have each of the four possible combinations for the pair $s_i, x_i(j)$
2. The joint marginal of s_i together with every subset of k public attributes, for some integer $k > 1$—that is, a collection of $\binom{d}{k}$ contingency tables, each with 2^{k+1} entries
3. The coefficients of a logistic regression model fit to predict s_i from x_i

Under what conditions do such releases allow an attacker, who knows the x_is and the released statistics, to reconstruct all or most of the vector s? What if the curator releases only approximate statistics? These questions are addressed by reconstruction attacks, as introduced by Dinur & Nissim (2003) and developed by a large body of subsequent work.

Let $\mathcal{M}$ the denote the mechanism used by the curator to generate a vector of released statistics $\hat{q} = \mathcal{M}(x, s)$.

> **Definition 3** (Blatant nonprivacy). A mechanism $\mathcal{M}$ is blatantly nonprivate for a public dataset x if there is an attack A such that for every vector s, we have $\mathrm{Ham}(s, \hat{s}) \leq \frac{n}{10}$, where $\hat{s} = A(x, \mathcal{M}(x, s))$. Here $\mathrm{Ham}(s, \hat{s})$ denotes the Hamming distance between two vectors (that is, the number of positions in which they differ).

Thus, as discussed informally in the Introduction, a mechanism is blatantly nonprivate if it allows an attacker to reconstruct almost all secret bits s_i. There is nothing special about reconstruction of 9/10 of the entries of s; we could have used any other constant close to 1. When we

say that a class (or set or collection) of mechanisms is blatantly nonprivate, we mean that there exists a single attack algorithm A that works against every mechanism in the class.

Note that because the attack works for every secret vector s, it cannot rely on detecting underlying statistical correlations between x and s. It is not an instance of correlation detection (see Section 1.1.4),[3] but rather an attack learning information highly specific to this particular dataset.

2.1. Reconstruction from Linear Statistics

The releases discussed above all share a particular structure, namely that the exact statistics are linear functions of the secret vector s: For each one, we can interpret the released statistics $\hat{q}$ as an approximation to Bs for some matrix B, whose rows correspond to queries. For example, releasing the pairwise marginals of s with each public column (the setting in Example 1) reveals, in particular, Bs where $B = x^\intercal$ and x is the matrix of public values. Releasing $(k+1)$-way contingency tables (item 2 in the list above) reveals Bs, where the rows of B consist of all $\binom{d}{k}$ entry-wise products of subsets of k columns of x. Less obviously, releasing the logistic regression coefficients also corresponds to a linear release, by viewing the released coefficients as a vector where the gradient of the logistic loss function, summed over all pairs (x_i, s_i), is 0^d.

In all these settings, the task of the attacker A is to solve a noisy system of linear equations. To allow the comparison of results, we normalize the matrix B so that entries lie in $[0,1]$, and divide the result by n to obtain an answer in $[0,1]$.

> **Definition 4** (Fractional linear query). A fractional linear query is specified by a vector $b \in [0,1]^n$; the exact answer is $q_b(s) = \frac{1}{n} b^\intercal s$ (which lies in $[0,1]$ as long as s is binary). An answer $\hat{q}_b$ is α-accurate if $|\hat{q}_b - q_b(s)| \leq \alpha$.

If a collection of fractional linear query statistics, given by the rows of a matrix B, is answered to within some error α, we get the following problem:

> **Definition 5** (B-reconstruction problem). Given a matrix B and a vector $\hat{q} = \frac{1}{n} Bs + e$, where $\|e\|_\infty \leq \alpha$ and $s \in \{0,1\}^n$, find $\hat{s}$ with $\mathrm{Ham}(\hat{s}, s) \leq \frac{n}{10}$. The reconstruction error is the fraction $\frac{\mathrm{Ham}(\hat{s}, s)}{n}$.

Understanding reconstruction attacks based on linear statistics thus boils down to understanding when the B-reconstruction problem can be solved, and how efficiently. The theory of noisy linear systems is deep and well developed, with extensive connections to numerical analysis, geometry, compressed sensing, and the theory of streaming algorithms. In the remainder of this section, we give a taste of how it applies to reconstruction.

2.2. An Exponential Attack

An important class of linear statistics are sums of subsets of the bits of s, which correspond to matrices B with entries in $\{0,1\}$. As a warm-up, consider what happens when approximations to all possible subset sums are released, that is, when B has 2^n rows, one for every vector in $\{0,1\}^n$.

[3] In particular, this means that reconstruction attacks are fundamentally different from the statistical literature on constructing good predictors from aggregate statistics and "ecological" correlations, as in the problem of learning from label proportions (see, e.g., Quadrianto et al. 2009 and following work).

Because the normalized subset sums lie in [0, 1], the accuracy parameter must be less than $\frac{1}{2}$ for the answers to convey any information at all about s (if $\alpha \geq \frac{1}{2}$, simply releasing $\frac{1}{2}$ as the approximation to each normalized subset sum, regardless of s, satisfies the accuracy requirement).

In this case, we can get nontrivial reconstruction attacks whenever the accuracy parameter α goes to 0 (that is, when the error in answering each query is a vanishing fraction of each query's maximum possible value).

Theorem 1 (Dinur & Nissim 2003). When $B \in \{0,1\}^{2^n \times n}$ has all possible rows in $\{0,1\}^n$, there is an attack A that solves the B-reconstruction problem with reconstruction error at most 4α (given α-accurate query answers), for every $\alpha > 0$. In particular, every mechanism that releases such statistics is blatantly nonprivate when $\alpha < 1/40$.

Proof. This brute-force attack simply enumerates all vectors $\tilde{s} \in \{0,1\}^n$ and picks one that agrees, within α, with all entries of $\hat{q}$, meaning $\|\hat{q} - \frac{1}{n}B\tilde{s}\|_\infty \leq \alpha$. We know such an $\tilde{s}$ exists because s is a solution. Let us call it $\hat{s}$.

We now argue that $\frac{\mathrm{Ham}(\hat{s},s)}{n} \leq 4\alpha$. Let $b_0 = s$, and let b_1 denote the bit-wise complement of s (that is, the n-bit vector with zeros in positions where s has ones, and ones where s has zeros). Because $\hat{s}$ agreed with $\hat{q}$ in the position corresponding to b_0, we have $|\frac{\langle b_0,\hat{s}\rangle}{n} - \hat{q}_{b_0}| \leq \alpha$. Because by assumption $|\frac{\langle b_0,s\rangle}{n} - \hat{q}_{b_0}| \leq \alpha$, we have that s and $\hat{s}$ disagree on at most $2\alpha n$ locations in which s is zero. An analogous argument shows they disagree on at most $2\alpha n$ locations in which s is one (based on their mutual agreement with $\hat{q}_{b_1}$).

Theorem 1 has important implications: there is no way to construct a noisy table that will permit highly accurate answers to be derived for computations that are not specified at the outset, even if only a relatively small number of linear queries will ever be of interest. Because we do not know in advance which queries will be of interest, the table must permit the analyst to learn accurate answers to all queries. Theorem 1 tells us that any such table providing answers to all 2^n queries described in the theorem will succumb to a reconstruction attack. As a result, when releasing information about sensitive data, we must make choices: Because no method can accurately and privately provide answers to everything, thought must be given to the use of the resource.

2.3. Attacks Requiring only Polynomially Many Queries

The attack of the previous section runs in exponential time and requires a release of exponentially many statistics. What can we do when the number of released statistics and the time available to the attacker are more limited? Before giving a general answer to this question, we consider a few special cases.

Theorem 2 (Dwork & Yekhanin 2008). There exists a matrix $B \in \{0,1\}^{2n \times n}$ and an attack A running in time $O(n \log n)$ that solves the B-reconstruction problem with reconstruction error at most $16\alpha^2 n$ when the answers are α-accurate. In particular, every mechanism that releases such statistics is blatantly nonprivate when $\alpha < \frac{1}{13\sqrt{n}}$.

A similar result is known to hold when the entries of B are chosen uniformly at random, though the number of rows must then be larger than n by a constant factor and the attack takes longer (about the time required to multiply two $n \times n$ matrices), and even if a certain constant number of responses have unbounded error (Dinur & Nissim 2003, Dwork et al. 2007). Furthermore, one

can interpolate smoothly through Theorems 1 and 2. The following slightly generalizes a result of Dinur & Nissim (2003):

Theorem 3. There exists an attack A such that, if B is chosen uniformly at random in $\{0,1\}^{m\times n}$ and $1.1n \le m \le 2^n$ then, with high probability over the choice of B, $A(B,\hat{q})$, given any α-accurate answers $\hat{q}$, solves B-reconstruction with error $\beta = o(1)$ as long as $\alpha = o(\sqrt{\frac{\log(m/n)}{n}})$. In particular, there is a $c > 0$ such that every mechanism for answering the queries in B with error $\alpha \le c\sqrt{\log(\frac{m}{n})/n}$ is blatantly nonprivate.

The constant 1.1 in the theorem is somewhat arbitrary. It suffices that $\log(m/n)$ be bounded below by a positive constant. We omit the proof of Theorem 3, though we outline below a general connection to discrepancy theory on which the proof is based.

2.4. Reconstruction, Spectral Bounds and Discrepancy

Understanding linear reconstruction attacks boils down to understanding the geometric properties of the query matrix B. We start by describing a very efficient attack, first presented by Dwork & Yekhanin (2008), who provide a proof of Theorem 2. The attack relies on bounding the eigenvalues of B.

Proof of Theorem 2. Suppose for now that $n = 2^\ell$ is an integer power of 2. To simplify computations, we allow the coefficients of the query matrix to lie in $\{-1,1\}$ instead of $\{0,1\}$. (One can always simulate a query with $\{-1,1\}$ coefficients using two queries with $\{0,1\}$ coefficients, at the cost of doubling the allowed error α).

We take B to be the Hadamard matrix H_ℓ, defined recursively by the formula $H_0 = (1)$ and $H_{i+1} = \begin{pmatrix} H_i & H_i \\ H_i & -H_i \end{pmatrix}$. H_ℓ is a $n \times n$ matrix (because $n = 2^\ell$) with entries in $\{\pm 1\}$ with the property that $H_\ell^2 = nI$ where I is the identity matrix. This means that the inverse of H_ℓ is $\left(\frac{1}{n}H_\ell\right)$, and that the eigenvalues of H_ℓ are all $\pm\sqrt{n}$.

Given $\hat{q} = \frac{1}{n}H_\ell s + e$, the attacker first multiplies by $H_\ell = (\frac{1}{n}H_\ell)^{-1}$ to obtain

$$r = H_\ell \hat{q} = H_\ell\left(\frac{1}{n}H_\ell\right)s + H_\ell e = s + H_\ell e\,.$$

Now the ℓ_2 norm of e is at most $\alpha\sqrt{n}$ because each of its entries has absolute value at most α. Because the eigenvalues of H_ℓ are $\pm\sqrt{n}$, the ℓ_2 norm of $H_\ell e$ is at most αn. Thus, we have $\|r - s\|_2^2 \le \alpha^2 n^2$.

In the second step, the attacker rounds each entry of r to the nearer of $\{0,1\}$ to obtain a candidate dataset $\hat{s} \in \{0,1\}^n$. We will use the following claim (proved below).

Claim 1. Let $s \in \{0,1\}^n$ and $r \in \mathbb{R}^n$ be arbitrary, and let $\hat{s}$ be obtained by rounding the entries of r to $\{0,1\}$. Then $\mathrm{Ham}(s,\hat{s}) \le 4\|r - s\|_2^2$.

The attacker's reconstruction error β is thus $\mathrm{Ham}(\hat{s},s) \le 4\alpha^2 n$. The constant claimed in the theorem statement is slightly higher than 4, because we must take into account the conversion from ± 1 to $\{0,1\}$ in the query coefficients, and also the padding required to get n to the next largest power of 2.

The running time of the attack, perhaps surprisingly, is less than the time it takes to write down the matrix H_ℓ. Because of the recursive form of H_ℓ, we can use a

divide-and-conquer algorithm similar to the fast Fourier transform to multiply any n-entry vector by H_ℓ in time $O(n \log n)$. (Multiplication by H_ℓ is in fact a Fourier transform over an appropriate group.)

Finally, we prove Claim 1. Notice that $\hat{s}$ and s differ only in positions j where $|r(j) - s(j)| \geq \frac{1}{2}$ (and hence $(r(j) - s(j))^2 \geq 1/4$). The average of $(r(j) - s(j))^2$ over all entries is $\|r - s\|_2^2/n$. By Markov's inequality, the fraction of squared entries over $1/4$ is at most $4\|r - s\|_2^2/n$, which proves the claim.

A careful inspection of the proof of the previous theorem shows that B-reconstruction is possible roughly whenever $B \in [0,1]^{m \times n}$ has at least n rows, and its least singular value is bounded below by a known quantity $\sigma_{\min}$. The attack simply multiplies $\hat{q}$ by $nB^\dagger$, where $B^\dagger$ is the left pseudoinverse of B, and rounds the result to $\{0,1\}^n$. The reconstruction error is then at most $4\alpha^2 nm/\sigma_{\min}^2$, because the maximum singular value of the pseudoinverse is $1/\sigma_{\min}$. This general connection was used by Kasiviswanathan et al. (2010, 2013) to get results for k-way marginal releases and several more general kinds of releases, including the logistic regression example at the beginning of this section.

For k-way marginal releases, the number of released statistics is $2^k\binom{d}{k-1}$ (because each released contingency table gives marginal statistics for a set of $k-1$ public attributes and the secret attribute). The question raised by Kasiviswanathan et al. (2010) was, how large does d have to be to carry out meaningful reconstruction? So far, the attacks we have considered use matrices B with $m > n$ rows, but the rows were selected independently (as in Theorem 3) or to be far apart from each other (as in the proof of Theorem 2). Such query matrices can arise with 2-way statistics (setting item 1 from the list at the beginning of Section 2), but require high-dimensional data: d must be at least n. To get matrices with n independent rows in the setting of k-way statistics would also require $d > n$, effectively wasting many of the released statistics. Building on results from random matrix theory, Kasiviswanathan et al. (2010) showed that in fact, it suffices that the total number of released statistics $2^k\binom{d}{k-1}$ be at least $c_k n$ for a constant $c_k > 1$ (that depends on k but not d or n); in particular, the dimension of the data d need only grow as $n^{1/k}$.

Though the spectral argument is very useful, it has limitations. A more sophisticated argument, based on restricted isometry properties, was used by Dwork et al. (2007) (and later generalized by De 2012) to handle releases where most statistics are answered α-accurately, but the error on some fraction (bounded by a parameter η) of statistics is arbitrarily high. For various classes of random matrices, the attack runs in polynomial time as long as η is a sufficiently small constant. In fact, the attack works as long as $\eta < \frac{1}{2} - \Omega(1)$ (that is, as long as a strict majority of statistics are α-accurate), though it takes exponential time in general.

2.4.1. Discrepancy-based bounds. We can in fact characterize when reconstruction from linear statistics is possible, using a combinatorial analogue of the spectral argument. The framework we use here was formulated by Muthukrishnan & Nikolov (2012), abstracting the idea in the proof of Theorem 1. Consider an attacker who knows B and an α-accurate answer vector $\hat{q}$, and wants to decide if a particular dataset $\hat{s}$ is a plausible candidate for the true dataset s.

A natural approach is simply to check if all the entries of $\frac{1}{n}B\hat{s}$ are within α of the entries of $\hat{q}$ (that is, if $\|\hat{q} - \frac{1}{n}B\hat{s}\|_\infty \leq \alpha$), and accept $\hat{s}$ as plausible if that is the case. This procedure will always accept the true vector s; under what conditions will it accept an incorrect vector $\hat{s}$? By the

triangle inequality, if this procedure accepts $\hat{s}$ then every entry of $\frac{1}{n}B(s-\hat{s})$ must be at most 2α:

$$\left\|\frac{1}{n}B(s-\hat{s})\right\|_\infty \le \underbrace{\left\|\frac{1}{n}Bs-\hat{q}\right\|_\infty}_{\le\alpha \text{ since } \hat{q} \text{ accurate}} + \underbrace{\left\|\frac{1}{n}B\hat{s}-\hat{q}\right\|_\infty}_{\le\alpha \text{ since } \hat{s} \text{ accepted}} \le 2\alpha.$$

If $\hat{s}$ differs from s in more than βn positions, then $(s-\hat{s})$ is a vector with entries in $\{-1,0,1\}$, of which at least βn entries are not zero. We can therefore ensure that no vector $\hat{s}$ that is at Hamming distance βn from s gets accepted if we ensure that Bz is large for all appropriate z.

Definition 6. The β-partial discrepancy of a matrix $B \in \mathbb{R}^{m\times n}$, denoted $\mathrm{disc}_{\infty,\beta}(B)$, is

$$\mathrm{disc}_{\infty,\beta}(B) \stackrel{\mathrm{def}}{=} \min_{\substack{z\in\{-1,0,1\}^n,\\ \|z\|_1\ge\beta n}} \|Bz\|_\infty.$$

If the partial discrepancy is at least $2\alpha n$, then no candidate $\hat{s}$ which is at Hamming distance more than βn from s will be accepted by the procedure above, that is, the reconstruction attack succeeds. We can thus define an (exponential-time) attack which has reconstruction error at most β whenever the accuracy α satisfies

$$\alpha \le \mathrm{disc}_{\infty,\beta}(B)/2n\,.$$

This idea underlies, among other results, the proof of Theorem 3.

Conversely, if the partial discrepancy is less than $2\alpha n$, then we can find vectors $s, \hat{s}$, and $\hat{q}$ such that s and $\hat{s}$ are far apart (at Hamming distance at least βn), but $\hat{q}$ is α-accurate for both data sets. Thus, no reconstruction attack based on $\hat{q}$ can reliably have reconstruction error less than $\frac{\beta}{2}$; reconstruction attacks work if and only if the partial discrepancy of the query matrix exceeds the error parameter α of the mechanism releasing the statistics.

The partial discrepancy generalizes the spectral arguments, because the smallest singular value of B gives a lower bound on its partial discrepancy: $\mathrm{disc}_{\infty,\beta}(B) \ge \frac{\sigma_{min}\sqrt{\beta n}}{\sqrt{m}}$ (because every z in $\{-1,0,1\}^n$ with at least βn nonzero entries has Euclidean norm at least $\sqrt{\beta n}$).

At the end of this survey, we describe differential privacy, a class of algorithms that resist reconstruction attacks (and enjoy other important properties). Perhaps surprisingly, when the size n of the dataset is very large, one can in fact answer a batch of linear queries with error roughly comparable to a slight generalization of the partial discrepancy, called the hereditary discrepancy (Nikolov et al. 2013). We omit the exact statement here.

3. TRACING ATTACKS

Reconstruction attacks are devastating when they occur, but simply avoiding reconstruction is not a satisfactory guarantee of privacy on its own. For example, consider the subsampling algorithm: subsample a random τ fraction of the dataset and release those samples. The reconstruction attacks from Section 2 will fail when given only this subsample, because the released subsample is completely independent of the data of a $(1-\tau)$ fraction of the rows, and yet a large number of individuals lose all privacy. Note that the subsampling algorithm introduces error proportional to $\frac{1}{\sqrt{\tau n}} > \frac{1}{\sqrt{n}}$, so there is no contradiction with the results in Section 2.

We remark that subsampling ensures that very few people—only those unlucky enough to be in the subsample—can possibly have their privacy compromised. We have often heard objections of the form "only a few people will be hurt" in defense of weak privacy protections. When such

risks are acceptable, subsampling provides a crisp privacy solution, accompanied by a plethora of utility results.

Tracing is a more subtle privacy breach than reconstruction. In a tracing attack, the attacker has (possibly noisy) statistics about the dataset and the data of a target individual, and wants to determine if that target individual is present in the dataset or not. As we have discussed, mere presence in the dataset can be highly sensitive information. By weakening the attacker's goal to tracing, we can reason about the privacy risks of very simple statistics, perturbed with a large amount of noise, even when the attacker has very limited auxiliary information.

As we will demonstrate, in a rather general model of a tracing attack, not only does the subsampling algorithm allow an attacker to trace a τ fraction of the individuals in the dataset, but so does any algorithm of comparable utility!

3.1. Tracing from Exact Statistics

Sankararaman et al. (2009) presented a formal model of tracing attacks based on hypothesis testing. Their model is well suited to capturing the attack of Homer et al. (2008) on genome-wide association data, with one difference noted below. The dramatis personae of their model are as follows.

1. We model our population by a distribution P over $\{\pm 1\}^d$. We call d the dimension of the data. We assume that P is a product distribution, and we let
$$p = \mathop{\mathbb{E}}_{x\sim P}[x]$$
be the population mean. We call p_j the jth marginal of the population. Note that P is entirely described by p.
2. There is a dataset $x = \{x_1, \ldots, x_n\}$ consisting of n independent and identically distributed (i.i.d.) samples from P.
3. The sample mean
$$q = \frac{1}{n}\sum_{i=1}^{n} x_i = \mathop{\mathbb{E}}_{x_i\sim x}[x_i]$$
is released. We use standard notation from the computation science literature and write $x_i \sim x$ to mean that x_i is sampled from the uniform distribution over the elements of x. We call q_j the jth marginal of the sample.
4. The attacker has the data $y \in \{\pm 1\}^d$ of a target individual. That individual y is either IN the dataset, meaning y is a uniformly random element of x, or OUT of the dataset, meaning y is an independent random sample from P. The attacker's goal is to distinguish these two cases.
5. The population P is unknown to the attacker. However, the attacker has a collection of m i.i.d. reference samples $z = \{z_1, \ldots, z_m\}$ from P.

The attack is a function $A(y, q, z)$ that takes the data of the target individual, the released marginals, and the reference samples and outputs a value in $\{\mathsf{IN}, \mathsf{OUT}\}$.

Homer et al. (2008) and Sankararaman et al. (2009) (see also the analysis of Yu 2015) design an attack based on hypothesis testing. Consider the null hypothesis H_0 corresponding to the case where y is OUT of the dataset, meaning that y is a random sample from the population P, that is, y is sampled from a product distribution with marginals p. Also consider the alternative

hypothesis H_1 corresponding to the case where y is IN the dataset.[4] In this case, y is a random sample from the dataset x, which can be approximated by a product distribution with mean q. If the attacker has p and q, then the optimal way to determine if y is sampled from the population or the dataset is to perform a log-likelihood test. In our model, the vector q is released, but the vector p is unknown. However, if the attacker has sufficiently many reference samples $z_1, \ldots, z_m$, then the average $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} z_i$ will be a suitable approximation to p for the attack. All of the attacks discussed in this section can (directly or with minor changes) be carried out using no information about the reference set z except its mean. We can summarize what is known about this attack in the following theorem.

Theorem 4 (Sankararaman et al. 2009). There is an attack $A(y, q, z)$ that takes the data $y \in \{\pm 1\}^d$ of a targeted individual, the exact sample mean q of a dataset x of dimension $d = O(n \log(1/\delta))$, and $m = O(n)$ reference samples $z = \{z_1, \ldots, z_m\} \subseteq \{\pm 1\}^d$ such that for every nontrivial product distribution P,

1. If y is IN the dataset x, then $\mathbb{P}[A(y, q, z) = \mathsf{IN}] \geq 1 - \delta$.
2. If y is OUT of the dataset x, then $\mathbb{P}[A(y, q, z) = \mathsf{OUT}] \geq 1 - \delta$.

In both of these statements, the probabilities are taken over the random choices of the dataset $x \sim P^n$, the reference samples $z \sim P^m$, and the choice of y according to either $y \sim x$ (IN) or $y \sim P$ (OUT). In other words, the probabilities of both type I and type II errors are at most δ.

In this theorem, a nontrivial product distribution is one whose marginals are bounded away from -1 and 1. This condition is very mild and serves to rule out pathological cases, such as a population in which every member of the population has the same data.

3.2. Robust Tracing for Noisy Statistics

What happens if the attacker does not get the exact sample mean q, but instead only a noisy sample mean $\hat{q}$? Is it still possible to trace individuals in the dataset? As Dwork et al. (2015d) showed, the answer is resoundingly "yes!" However, it takes much more care to formalize tracing attacks when the statistics can be noisy. To see why tracing attacks are more subtle with only noisy statistics, we give a few examples of ways that the sample mean can be perturbed to make tracing difficult.

Example 1. Subsampling is one way to introduce noise into statistics. That is, we can take a dataset $x = \{x_1, \ldots, x_n\}$ of n samples obtain a dataset $\hat{x} = \{\hat{x}_1, \ldots, \hat{x}_s\} \subseteq x$ of $s \approx \tau n$ samples for some $\tau > 0$. Now we can release the exact mean $\hat{q} = \frac{1}{s}\sum_{i=1}^{s} \hat{x}_i$ of the subsample, which is a noisy sample mean of x. Sampling theory tells us that the average over coordinates j of $|q_j - \hat{q}_j| \approx \frac{1}{\sqrt{\tau n}}$. However, when we run a tracing attack on the released mean $\hat{q}$, and a random target individual x_i that is IN the dataset x, with probability $1 - \tau$, x_i is independent of $\hat{q}$. Thus, we have

$$\mathop{\mathbb{P}}_{\mathsf{IN}}[A(y, \hat{q}, z) = \mathsf{IN}] - \mathop{\mathbb{P}}_{\mathsf{OUT}}[A(y, \hat{q}, z) = \mathsf{IN}] \leq \tau \ll 1,$$

[4] In Homer et al. (2008), the attacker sees only the mean of the reference set, and the null hypothesis is that y is drawn from that set (see the discussion in Sankararaman et al. 2009).

in contrast to the case of exact statistics where Theorem 4 yields

$$\mathop{\mathbb{P}}_{\mathsf{IN}}[\mathcal{A}(y,\hat{q},z)=\mathsf{IN}]-\mathop{\mathbb{P}}_{\mathsf{OUT}}[\mathcal{A}(y,\hat{q},z)=\mathsf{IN}]\geq 1-2\delta\approx 1,$$

where $\mathop{\mathbb{P}}_{\mathsf{IN}}[E]$ is the probability of the event E when y is IN the dataset, and $\mathop{\mathbb{P}}_{\mathsf{OUT}}[E]$ is defined analogously.

The subsampling example shows that, in general, we must give up on tracing every individual who is in the dataset (but see Theorem 7 for a case in which all individuals in the dataset can be traced). Thus, we will set our sights on tracing at least one individual in the dataset.

As our next example shows, the role played by the population P is much more delicate when the statistics can be noisy.

Example 2. Suppose the population P and its mean p are fixed and known. Then instead of releasing the exact sample mean q of the dataset, the algorithm could release the noisy population mean $\hat{q}=p$. By sampling theory, we will have that $|q_j-\hat{q}_j|\approx\frac{1}{\sqrt{n}}$ on average over coordinates j. Note that, although the population mean is likely of more interest than the sample mean, they are noisy for the purposes of this example because they are not equal to the sample mean. If we are only given the population mean $\hat{q}$, then we cannot hope to trace because $\hat{q}$ is independent of the dataset. Thus,

$$\mathop{\mathbb{P}}_{\mathsf{IN}}[\mathcal{A}(y,\hat{q},z)=\mathsf{IN}]=\mathop{\mathbb{P}}_{\mathsf{OUT}}[\mathcal{A}(y,\hat{q},z)=\mathsf{IN}].$$

This example does not give an actual algorithm for releasing an approximately correct sample mean, because we cannot assume that the holder of the dataset knows the population mean. However, it demonstrates the need to model the data holder's uncertainty about the population. Thus, we assume that each of the population marginals p_j is itself random and chosen i.i.d. from some probability distribution $\mathcal{P}$ over $[-1,1]$. To simplify notation, we will assume that every marginal is chosen from the same distribution $\mathcal{P}$. (It would not affect our results if each marginal were chosen from a different distribution $\mathcal{P}_j$.)

Clearly $\mathcal{P}$ must be nontrivially random, or else the population's marginals are not really uncertain. Our final example considers two types of nontrivial yet degenerate distributions $\mathcal{P}$ that will make it impossible to trace from noisy statistics.

Example 3. Suppose that $\mathcal{P}$ is uniform on some interval $[a,b]$ for some $0<a<b<1$. Then, the noisy sample mean $\hat{q}=(\frac{b-a}{2},\frac{b-a}{2},\ldots,\frac{b-a}{2})$ will satisfy $|q_j-\hat{q}_j|\lesssim\frac{b-a}{2}$ for every coordinate j. As in the previous example, $\hat{q}$ is independent of the dataset, so tracing cannot succeed against an algorithm that outputs $\hat{q}$. This example shows that $\mathcal{P}$ must be well-spread in the sense that it must place significant mass on values that are farther apart than the amount of noise we are willing to add to the mean.

Now, suppose that $\mathcal{P}$ is either $-\frac{1}{2}$ or $+\frac{1}{2}$ with equal probability. For this choice of $\mathcal{P}$, every population marginal p_j is in $\{\pm\frac{1}{2}\}$. Consider the noisy mean $\hat{q}$ such that $\hat{q}_j=\frac{\mathrm{sign}(q_j)}{2}\in\{\pm\frac{1}{2}\}$. By sampling theory, we will have $|q_j-\hat{q}_j|\approx\frac{1}{\sqrt{n}}$ on average over coordinates j. However, changing the data of one person x_i will, with high probability, not change $\hat{q}$ at all. Therefore, $\hat{q}$ is nearly independent of any specific person in the dataset, making tracing impossible. This example shows that $\mathcal{P}$ must be smooth in addition to being well spread.

$A(y, \hat{q}, z)$:
Input: target y, noisy marginals $\hat{q}$, reference sample z.
Let $T = O(\sqrt{d \log(1/\delta)})$ be a carefully chosen threshold. If

$$\langle y, \hat{q} \rangle - \langle z, \hat{q} \rangle \geq T,$$

output **IN**, otherwise output **OUT**.

Figure 1

A description of our robust tracing attack.

The distributions in the example above show that for tracing to succeed, $\mathcal{P}$ must avoid certain pathologies. The specific technical conditions we need to impose are beyond the scope of this article but can be found in Dwork et al. (2015d). In the following, we refer to a distribution $\mathcal{P}$ satisfying these unspecified conditions as a strong distribution. For example, the uniform distribution on $[-1, 1]$ is strong, as is the uniform distribution on an interval of sufficient length or a Beta distribution with reasonable parameters.

3.2.1. The model. To address these pathologies, we need to modify some features of the model we described in Section 3.1. These modifications only affect the way the population P is chosen and the way the released vector $\hat{q}$ are chosen, so we only give the modifications to these two parts of the model.

Modified 1. We model our population P by a distribution P over $\{\pm 1\}^d$. We call d the dimension of the data. We assume that P is a product distribution and we let

$$p = \mathop{\mathbb{E}}_{x \sim P}[x]$$

so that p_j is the mean of the jth attribute in the population. Each coordinate p_j will be random and chosen i.i.d. from a strong distribution $\mathcal{P}$.

Modified 3. There is an algorithm $\mathcal{M}$ that takes the dataset x and outputs a noisy sample mean $\hat{q} = \mathcal{M}(x)$. This vector satisfies $\frac{1}{d}\|q - \hat{q}\|_1 \leq \alpha$ (i.e., $|q_j - \hat{q}_j| \leq \alpha$ on average over coordinates j). Recall that $q = \frac{1}{n}\sum_{i=1}^{n} x_i$. If $\mathcal{M}$ is randomized, then we require that accuracy holds with probability at least $2/3$ over this randomization. We call an algorithm that outputs such a vector α-accurate.

3.2.2. The attack. Dwork et al. (2015d) showed how to trace in this model using the very simple attack in **Figure 1**. The attack requires no knowledge of $\mathcal{M}$, and remarkably requires only a reference sample from P.

We can interpret the quantities $\langle y, \hat{q} \rangle$ and $\langle z, \hat{q} \rangle$ as the correlation of the noisy sample mean with the target individual and a random member of the population, respectively. Thus, the attack is testing whether the target individual's data is significantly more correlated with the released statistics than a random member of the population. The attack itself is quite similar to the one used by Homer et al. (2008), but the analysis is necessarily quite different because we can no longer view the attack as testing one of two simple hypotheses. (By "simple hypotheses," we mean that each hypothesis stipulates a fixed distribution.) That is, because the distribution $\mathcal{P}$ and the algorithm $\mathcal{M}$ are unknown to the attacker, the cases of IN and OUT no longer give rise to two specific distributions on the triple $(y, \hat{q}, z)$, and instead each case now corresponds to a whole family of distributions (a composite hypothesis).

We can summarize the properties of this attack in the following theorem.

Theorem 5 (Dwork et al. 2015d). There is an attack $A(y, q, z)$ that takes a noisy sample mean $\hat{q}$ of a dataset x of dimension $d = O(n^2 \log(1/\delta))$, the data $y \in \{\pm 1\}^d$ of a targeted individual, and a single reference sample $z \in \{\pm 1\}^d$, such that if the population P's mean is chosen from any strong distribution $\mathcal{P}$, and $\hat{q} = \mathcal{M}(x)$ for any 1/2-accurate $\mathcal{M}$,

1. If y is IN the dataset x, then $\mathbb{P}[A(y, q, z) = \mathsf{IN}] \geq \Omega(1/n)$.
2. If y is OUT of the dataset x, then $\mathbb{P}[A(y, q, z) = \mathsf{OUT}] \geq 1 - \delta$.

In both of these statements, the probabilities are taken over the random choices of the population P's marginals $p \sim \mathcal{P}^d$, the dataset $x \sim P^n$, the possibly random choice of noisy marginals $\hat{q} \sim \mathcal{M}(x)$, the reference sample $z \sim P$, and the choice of y according to either $y \sim x$ (IN) or $y \sim P$ (OUT).

Some comments are in order. First, note that this theorem is nontrivial when $\delta \ll 1/n$. In this case, when given a random member of the dataset, the attack will say IN with probability $\Omega(1/n)$, but when given a random member of the population the attack says IN with probability at most $\delta \ll 1/n$.

Second, note that the condition of 1/2-accuracy is very weak—much less accurate than applications would require—and it is rather surprising that we can trace in the presence of so much noise. In exchange for requiring such a weak notion of accuracy, the attack is only guaranteed to trace when $d \gtrsim n^2$, whereas for exact marginals the attack in Theorem 4 was able to trace with dimension $d \approx n$. As we will see in Section 4, such high dimension is necessary to guarantee tracing, because when $d = o(n^2)$ we can simultaneously achieve nontrivial accuracy and a strong guarantee of privacy (see Theorem 8 for a precise statement). However, for certain algorithms M, this attack may succeed even when d is much smaller.

3.2.3. Analysis of the robust tracing attack. A full proof of Theorem 5 is beyond the scope of this article. Instead, we give some basic intuition for why the attack works and how to analyze it.

First, suppose y is OUT of the dataset x. Then y and z are independent samples from the population P. Moreover, because $x = \{x_1, \ldots, x_n\}$ is independent of y, z, and $\hat{q} = \mathcal{M}(x)$, y and z are distributed as two independent samples from P even when conditioned on any fixed value of $\hat{q}$ (even one that is not accurate). For any $\hat{q}$, we have $\mathbb{E}[\langle y, \hat{q}\rangle - \langle z, \hat{q}\rangle] = 0$. Furthermore, because P is a product distribution, the coordinates of y and z are independent and $\langle y, \hat{q}\rangle - \langle z, \hat{q}\rangle$ can be written as the sum of d bounded independent random variables. Applying Hoeffding's inequality to $\langle y, \hat{q}\rangle - \langle z, \hat{q}\rangle$ thus shows that $\mathbb{P}[\langle y, \hat{q}\rangle - \langle z, \hat{q}\rangle \geq O(\sqrt{d \log(1/\delta)})] \leq \delta$. The OUT case of the theorem follows by setting an appropriate choice of $T = O(\sqrt{d \log(1/\delta)})$ and taking expectation over $\hat{q}$.

Now, consider the more difficult case where y is IN the dataset x. The crucial claim to establish is that

$$\mathbb{E}\left[\sum_{i=1}^{n} \langle x_i - z, \hat{q}\rangle\right] \geq \Omega(d). \tag{1}$$

If the inequality in Equation 1 holds, then by a concentration of measure argument we obtain that $\sum_{i=1}^{n} \langle x_i - z, \hat{q}\rangle \geq \Omega(d)$ holds with high probability. Consequently, with high probability, for some $y = x_i$, we have $\langle y, \hat{q}\rangle - \langle z, \hat{q}\rangle \geq \Omega(d/n)$, and we want to ensure that this quantity is larger than the threshold T. Given our choice of $T = O(\sqrt{d \log(1/\delta)})$, the IN case of the theorem follows by taking an appropriately large choice of $d = O(n^2 \log(1/\delta))$.

It remains to justify our claim (Equation 1). By linearity of expectations, it suffices to understand the case of $d = 1$ and show that

$$\mathbb{E}\left[\sum_{i=1}^{n}(x_i - z)\hat{q}\right] \geq \Omega(1).$$

For intuition, consider the case of exact statistics where $\hat{q} = q = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then

$$\mathbb{E}\left[\sum_{i=1}^{n}(x_i - z)\hat{q}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[x_i] \geq \Omega(1),$$

as long as the distribution P has high variance (i.e., the population mean p is bounded away from -1 and 1). It remains to show that introducing error into $\hat{q}$ does not completely break the correlation between $\hat{q}$ and $\sum_{i=1}^{n} x_i$.

Suppose the exact marginal is $q = 1$. In this case, the dataset must be $x_1 = x_2 = \cdots = x_n = 1$. Even if the answers are very noisy, we have that $\hat{q} > +1/2$. Similarly, if the exact marginal is $q = -1$ then $\hat{q} < -1/2$. Now, as a thought experiment, start with the dataset $x_1 = \cdots = x_n = +1$ and change the x_is from $+1$ to -1 one-by-one until the dataset $x_1 = \cdots = x_n = -1$ is reached. The sum of these changes takes $\hat{q}$ from $> +1/2$ to $< -1/2$. Thus, on average over i, changing each x_i changes $\hat{q}$ by $> 1/n$. So on average we have correlation at least $1/n$ between each x_i and $\hat{q}$. To establish our general claim about the correlation between $\sum_{i=1}^{n} x_i$ and $\hat{q}$, we need to show that, when $x_1, \ldots, x_n$ are chosen randomly as in our model, the correlation behaves like the average in this thought experiment, which requires the use of our assumption that $\mathcal{P}$ is a strong distribution.

3.2.4. Additional results. The strength of Theorem 5 is in the weakness of its assumptions. However, as we have described, these weak assumptions lead to somewhat weaker conclusions than what Theorem 4 gives for exact tracing, and certainly weaker than a reconstruction attack. Nonetheless, as the next theorem shows, if we have more accuracy and more reference samples, then we can trace with much lower-dimensional data.

Theorem 6. For every $\alpha \geq \frac{1}{\sqrt{n}}$, there is an attack $A(y, q, z)$ that takes a noisy sample mean $\hat{q}$ of a dataset x of dimension $d = O(\alpha^2 n^2 \log(1/\delta))$, the data $y \in \{\pm 1\}^d$ of a targeted individual, and $m+1 = O(\frac{\log(d)}{\alpha^2})$ reference samples z_0 and $z = \{z_1, \ldots, z_m\} \subseteq \{\pm 1\}^d$ such that if the population mean P is chosen from any strong distribution $\mathcal{P}$, and $\hat{q} = \mathcal{M}(x)$ for any α-accurate $\mathcal{M}$,

1. If y is IN the dataset x, then $\mathbb{P}[A(y, q, z) = \mathsf{IN}] \geq \Omega(1/\alpha^2 n)$.
2. If y is OUT of the dataset x, then $\mathbb{P}[A(y, q, z) = \mathsf{OUT}] \geq 1 - \delta$.

The probabilities are taken over the same random choices as in Theorem 5.

This result smoothly interpolates between Theorem 5 (minimal accuracy, minimal reference samples, higher dimension) and Theorem 4 (perfect accuracy, many reference samples, lower dimension). For every value of $\alpha > 0$, the dimension required by our attack is essentially optimal, by the positive results we present in Section 4.

The attack in this result is nearly identical to the one presented in **Figure 1**. The only difference is that we are given $m + 1$ reference samples $z_0, z_1, \ldots, z_m$. z_0 acts like the single reference sample in the basic attack, and $\hat{z} = \frac{1}{m}\sum_{i=1}^{m} z_i$ serves as an estimate of the population mean. Specifically, instead of computing $\langle y - z, \hat{q}\rangle$, we compute $\langle y - z, \hat{q} - \hat{z}\rangle$, and apply a suitable threshold.

The first guarantee of this attack can be rephrased as saying that, on average, the attack outputs IN for $1/\alpha^2$ of the n individuals in the sample. In contrast, we can release α-accurate marginals by using a random subsample of size $O(1/\alpha^2)$. This comparison justifies our claim that every algorithm allows almost the same number of individuals to be traced as the subsampling algorithm with comparable accuracy.

In some settings we can make an even stronger claim and trace every individual. Of course, we can only do this if we place restrictions that rule out the subsampling algorithm. One such restriction is to require that the algorithm is symmetric—that is, it treats all users the same, which can be formalized by requiring that the noisy marginals $q \sim \mathcal{M}(x)$ depend only on the vector of exact marginals q. In Markov chain notation, $x \to q \to \hat{q}$. This rules out the subsampling algorithm, and allows us to prove the following theorem.

Theorem 7. There is an attack $A(y, q, z)$ that takes a noisy sample mean $\hat{q}$ of a dataset x of dimension $d = O(n^2 \log(1/\delta))$, the data $y \in \{\pm 1\}^d$ of a targeted individual, and a single reference sample $z \in \{\pm 1\}^d$ such that if the population P's marginals are chosen from any strong distribution $\mathcal{P}$, and $\hat{q} = \mathcal{M}(x)$ for any symmetric 1/2-accurate $\mathcal{M}$,

1. If y is IN the dataset x, then $\mathbb{P}[A(y, q, z) = \mathsf{IN}] \geq 1 - \delta$.
2. If y is OUT of the dataset x, then $\mathbb{P}[A(y, q, z) = \mathsf{OUT}] \geq 1 - \delta$.

In other words, the probabilities of both type I and type II errors are bounded by δ. The probabilities are taken over the same random choices as in Theorem 5.

In this setting the attack is the one described in **Figure 1**, and only the analysis changes.

4. DIFFERENTIAL PRIVACY: FREE AT THE LIMIT

The limits imposed by reconstruction and tracing attacks are absolute: no mechanism protecting against reconstruction and tracing can introduce less noise than is required to stymie the attacks discussed earlier. However, there are other adversarial goals, such as learning the sickle cell status of a specific individual, that do not require reconstruction, re-identification, or tracing, and each of these new goals may have its own set of attack strategies. A privacy solution that rules out reconstruction and tracing may not rule out attacks satisfying these other goals. The cryptographic approach to this dilemma is to first define privacy and then provide techniques that provably satisfy this definition. If the definition is too weak, in that it fails to protect against an important class of adversarial goals, it can be strengthened and new algorithms designed. The advantage to the definitional approach is that, because the definitions are getting stronger, progress is made. Differential privacy was first proposed in 2006 and so far has not required strengthening.

Differential privacy is a very strong definition, and it is not without cost. Nonetheless in one sense it is for free: differential privacy can be achieved by introducing exactly as much noise as is necessary to combat the specific attacks of the previous sections. In other words, the marginal cost of achieving differential privacy and all the protection that entails is zero, if one is protecting against reconstruction and tracing.

4.1. Defining Privacy

Our ultimate privacy goal when releasing information about a sensitive dataset is to ensure that anything that can be learned about an individual from the released information, can be learnt without that individual's data being included. This goal does not ensure that *nothing* about an

individual can be learnt from the released information, which can only be achieved by releasing no information (Dwork 2006, Dwork & Naor 2008). For example, as discussed in the Introduction, releasing the fact that smoking and lung cancer are strongly correlated reveals sensitive information about any individual known to smoke; however, we do not to consider this to be a privacy violation, as learning this correlation has nothing to do with the use of that individual's data. Our goal is only to protect sensitive information that is localized to a single individual's data.

Differential privacy (Dwork et al. 2006b) is such a quantitative privacy goal. Differential privacy is a property of a procedure or mechanism $\mathcal{M}$ that takes a sensitive dataset x and releases the output $\mathcal{M}(x)$. We compare the output $\mathcal{M}(x)$ with a hypothetical output $\mathcal{M}(y)$ in which the input x is changed to y by removing, adding, or modifying the data of a single individual. The requirement of differential privacy is that $\mathcal{M}(x)$ should be indistinguishable from $\mathcal{M}(y)$ for any inputs x and y differing only on the data of a single individual:

Definition 7 (Differential Privacy, Dwork et al. 2006b). A mechanism $\mathcal{M}$ satisfies ε-differential privacy if, for any datasets x and y differing only on the data of a single individual and any potential outcome $\hat{q}$,

$$\mathbb{P}[\mathcal{M}(x) = \hat{q}] \leq e^{\varepsilon} \cdot \mathbb{P}[\mathcal{M}(y) = \hat{q}]. \tag{2}$$

Setting $\varepsilon = 0$ corresponds to revealing no information [$\mathcal{M}(x)$ and $\mathcal{M}(y)$ are identically distributed], whereas setting $\varepsilon > 0$ permits revealing some information about individuals. The parameter ε (sometimes called the bound on privacy loss) should be thought of as a small constant no larger than 1. The definition of differential privacy (Equation 2) is inherently probabilistic; as in cryptography, randomness is used to hide or obscure the individual information we wish not to reveal. Thus any nontrivial differentially private release of information requires randomization.

There are several variants of this definition of differential privacy, which are similar in spirit to what we discuss here, but have important quantitative differences. A common generalization of differential privacy (Dwork et al. 2006a) introduces a second parameter δ and replaces Equation 2 with $\mathbb{P}[T(\mathcal{M}(x)) = 1] \leq e^{\varepsilon} \cdot \mathbb{P}[T(\mathcal{M}(y)) = 1] + \delta$, which is required to hold for all functions T. For clarity, we only discuss the simplest definition.

Differential privacy is a very robust definition—as we would expect of a meaningful privacy guarantee. In particular, it satisfies the following important properties.

- *Postprocessing:* Additional analysis of the released information or the inclusion of information from other sources will not change the differential privacy guarantee. In particular, if an attack (such as those discussed earlier) were applied to a differentially private release, then the guarantee of differential privacy would apply to the output of the attack, which precludes successful reconstruction or tracing.
- *Composition:* If the same individual's data is used in multiple releases, then, as long as each release satisfies differential privacy on its own, the combination of these releases also satisfies differential privacy. However, the quantitative privacy guarantee degrades—namely, if each release satisfies ε-differential privacy, then the combination of k such releases satisfies $k\varepsilon$-differential privacy.
- *Group privacy:* If information is shared by several individuals (such as a family), differential privacy continues to protect this information. If we view the dataset as a random sample, this corresponds to having some correlated samples, rather than i.i.d. samples. Again, the privacy guarantee degrades with the number of individuals we wish to protect simultaneously. That

is, if x differs from y by the addition, removal, or modification of the data of at most k individuals and $\mathcal{M}$ satisfies ε-differential privacy, then $\mathbb{P}[\mathcal{M}(x) = \hat{q}] \leq e^{k\varepsilon} \cdot \mathbb{P}[\mathcal{M}(y) = \hat{q}]$ for all possible outcomes $\hat{q}$.

Composition is arguably the signature property of differential privacy, as it permits differentially private analyses to be viewed as part of a larger system. Privacy-preserving data analysis does not occur in a vacuum—a single individual's data may be used multiple times over her lifetime. Furthermore, simple mechanisms can be composed to perform complex analytical tasks. The richness of the literature on differential privacy largely stems from the fact that composition permits an algorithmic approach to differential privacy, whereby simple building blocks can be combined in sophisticated ways to carry out a wide variety of analytical tasks.

Perhaps the most surprising property of differential privacy is that, despite its protective strength, it is compatible with meaningful data analysis. An extensive literature has been developed showing that a wide range of useful analyses can be carried out subject to differential privacy and its variants (for an introduction to and overview of differential privacy, see the textbook on the subject by Dwork & Roth 2014). Indeed, the parameter regime where the attacks of Sections 2 and 3 break down is very close to the setting where it becomes possible to release approximate aggregate statistics while satisfying differential privacy.

Differential privacy is not only useful for privacy: A major concern in empirical science is the danger of overfitting data and reaching conclusions that are specific to the dataset, rather than generalizing to the larger population from which that dataset was drawn. This problem is exacerbated by adaptivity—that is, when an analysis of a dataset is informed by prior exploration of the same dataset, standard hypothesis testing techniques may misrepresent the significance of a hypothesis owing to the dependence between the hypothesis and the dataset that has been introduced by prior use (e.g., through model selection). However, differential privacy also offers protection from such overfitting (Dwork et al. 2015a,b,c; Bassily et al. 2016). Namely, if a dataset is only used in a differentially private manner, then any conclusion drawn from that information cannot overfit the dataset. This is especially useful in the adaptive setting, as the composition property of differential privacy holds even for adaptive data analysis. Indeed, differentially private algorithms provide nearly optimal results for adaptive data analysis (Hardt & Ullman 2014, Steinke & Ullman 2015). Thus we see that differential privacy can, in fact, be an aid to analysis, even when privacy is not a concern.

4.2. Example: The Gaussian Mechanism

Having defined a formal privacy goal, we now discuss an example technique for releasing aggregate statistics about a dataset while protecting privacy. We restrict our attention to aggregate statistics of the form "what fraction of people in the dataset have property q?" For example, q may be the property "smoke and have cancer." This is a simple, yet powerful, class of aggregate statistics, often called counting queries.

We assume that k properties $q = (q_1, \ldots, q_k)$ are specified, and we will release approximate answers for all of them on a given dataset x containing the data of n individuals. Let $q_j(x)$ denote the fraction of individuals in the dataset x having property q_j and $q(x) = (q_1(x), \ldots, q_k(x))$.

Definition 8 (Gaussian mechanism). Given properties $q = (q_1, \ldots, q_k)$, the Gaussian mechanism $\mathcal{M}_{q,\sigma^2}$ takes x as input and releases $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_k)$ where each $\hat{q}_j$ is an independent sample from $\mathcal{N}(q_j(x), \sigma^2)$, for an appropriate variance σ^2.

We first ask whether $\mathcal{M}_{q,\sigma^2}(x)$ releases useful information about x: We would like to know $q_j(x)$, but the Gaussian mechanism only gives us an approximation $\hat{q}_j$. Whether this error with standard deviation σ is acceptable depends on the context. However, in many situations, the dataset x is itself a random sample of size n from a larger distribution, in which case $q_j(x)$ is also only an approximation to the quantity of interest. In particular, the sampling error of $q_j(x)$ has variance $p_j(1-p_j)/n$, where p_j is the probability that a random sample from the larger distribution has property q_j. Taking the unavoidable sampling error as a comparison point, we argue that the error introduced by the Gaussian mechanism is tolerable as long as $\sigma \approx 1/\sqrt{n}$. Furthermore, the output of the Gaussian mechanism remains useful as long as $\sigma \ll 1$. For example, if $\sigma \leq 1/100$, then each y_j is an estimate of $q_j(x)$ whose standard deviation is 1% of the size of the dataset.

We give a self-contained analysis of the privacy guarantee afforded by the Gaussian mechanism using the Neyman-Pearson lemma. Rather than satisfying differential privacy as defined earlier, the Gaussian mechanism satisfies a variant called concentrated differential privacy (Bun & Steinke 2016, Dwork & Rothblum 2016). The properties of concentrated differential privacy are similar to those of differential privacy listed earlier. However, concentrated differential privacy gives a tighter and more elegant analysis of composition.

We demonstrate that, given the released output of $\mathcal{M}_{q,\sigma^2}(x)$, we cannot infer much about an individual in the dataset x, where "much" will be parameterized by $\rho > 0$. We formulate this in the language of hypothesis testing by showing that any hypothesis about a single individual cannot be tested accurately.[5] However, the Gaussian mechanism is still useful in the sense that hypotheses about the population as a whole can be tested accurately given the output of the Gaussian mechanism. Thus we must distinguish hypotheses about individuals.

Consider simple null and alternate hypotheses H_0 and H_1 about an individual in the dataset, where H_0 is the hypothesis that the dataset is x and H_1 is the hypothesis that the dataset is y. Here x and y differ only in the addition, removal, or modification of the data of one individual, who is the subject or target of this test. We show that it is impossible to accurately test H_0 versus H_1.

Lemma 1 (Neyman & Pearson 1933). Fix simple hypotheses H_0 and H_1. Define the log-likelihood ratio test statistic by

$$\mathrm{LLR}(\hat{q}) = \log\left(\frac{\mathbb{P}[\hat{q}|H_0]}{\mathbb{P}[\hat{q}|H_1]}\right).$$

Then any test T is dominated by the likelihood ratio test. That is,

$$\mathbb{P}\left[T(\hat{q}) = \text{reject}|H_0\right] \geq \mathbb{P}[\mathrm{LLR}(\hat{q}) < \eta|H_0]$$

and

$$\mathbb{P}\left[T(\hat{q}) = \text{reject}|H_1\right] \leq \mathbb{P}[\mathrm{LLR}(\hat{q}) \leq \eta|H_1]$$

for some threshold η depending on T.

Lemma 1 tells us that, rather than considering all possible tests T for H_0 versus H_1, we need only consider the likelihood ratio test. Under H_0, $\hat{q} = \mathcal{M}_{q,\sigma^2}(x)$ is distributed according to

[5] For more about formulating differential privacy using hypothesis testing, see theorem 2.4 in Wasserman & Zhou (2010).

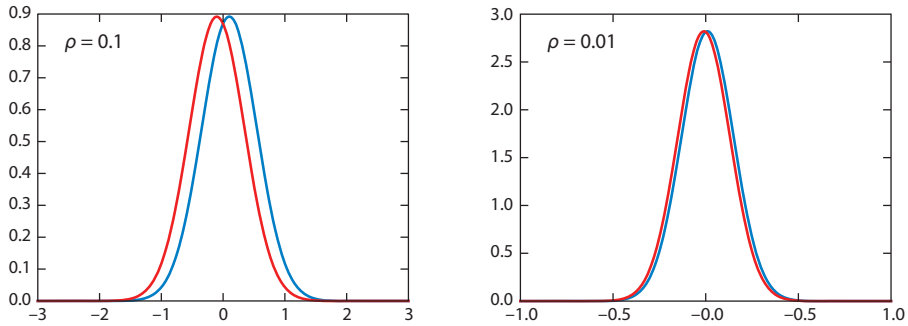

Figure 2

Probability density, under H_0, of $\mathrm{LLR}(\hat{q}) \sim \mathcal{N}(\rho, 2\rho)$ (*blue*) and probability density, under H_1, of $\mathrm{LLR}(\hat{q}) \sim \mathcal{N}(-\rho, 2\rho)$ (*red*).

$\mathcal{N}(q(x), \sigma^2 I)$ and, under H_1, $\hat{q} = \mathcal{M}_{q,\sigma^2}(y)$ is distributed according to $\mathcal{N}(q(y), \sigma^2 I)$. We calculate

$$\begin{aligned}
\mathrm{LLR}(\hat{q}) &= \log\left(\frac{(2\pi)^{-k/2}\cdot\exp\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{k}(\hat{q}_j - q_j(x))^2\right)}{(2\pi)^{-k/2}\cdot\exp\left(\frac{-1}{2\sigma^2}\sum_{j=1}^{k}(\hat{q}_j - q_j(y))^2\right)}\right)\\
&= \frac{-1}{2\sigma^2}\sum_{j=1}^{k}\left((\hat{q}_j - q_j(x))^2 - (\hat{q}_j - q_j(y))^2\right)\\
&= \frac{1}{2\sigma^2}\sum_{j=1}^{k}\left(q_j(x) - q_j(y)\right)\left(2\hat{q}_j - q_j(x) - q_j(y)\right).
\end{aligned}$$

Thus, under H_0, $\mathrm{LLR}(\hat{q})$ is distributed according to $\mathcal{N}(\rho, 2\rho)$, where

$$\rho = \frac{1}{2\sigma^2}\left\|q(x) - q(y)\right\|_2^2 = \frac{1}{2\sigma^2}\sum_{j=1}^{k}(q_j(x) - q_j(y))^2.$$

Under H_1, $\mathrm{LLR}(\hat{q})$ is distributed according to $\mathcal{N}(-\rho, 2\rho)$. Because x and y differ only on the data of one individual, the fractions $q_j(x)$ and $q_j(y)$ differ by at most $1/n$. (For simplicity, assume x and y are datasets of the same size with the data of one individual modified, rather than removed or added.) Hence,

$$\rho \le \frac{k}{2n^2\sigma^2}.$$

If ρ is small (say, 0.1 or 0.01), then the distribution $\mathcal{N}(\rho, 2\rho)$ is very close to $\mathcal{N}(-\rho, 2\rho)$.[6] By the Neyman-Pearson lemma, any test distinguishing H_0 from H_1 fares no better than distinguishing these distributions by means of a threshold. For example, if $\rho = 0.01$ and a test T has significance $\mathbb{P}\left[T(\hat{q}) = \text{reject}|H_0\right] \le 0.05$, then we conclude that the power of T is bounded by $\mathbb{P}\left[T(\hat{q}) = \text{reject}|H_1\right] \le 0.067$. **Figure 2** shows how close these distributions are.

Similarly to differential privacy, we can show that, for any test T and any $\varepsilon \ge \rho$,

$$\mathbb{P}\left[T(\hat{q}) = \text{reject}|H_1\right] \le e^{\varepsilon}\cdot\mathbb{P}\left[T(\hat{q}) = \text{reject}|H_0\right] + e^{-(\varepsilon-\rho)^2/4\rho}.$$

[6] By dividing $\mathrm{LLR}(\hat{q})$ by $\sqrt{2\rho}$ we can rescale these to $\mathcal{N}(\sqrt{\rho/2}, 1)$ versus $\mathcal{N}(-\sqrt{\rho/2}, 1)$.

For $\rho = 0.01$, the statistical distance (also called total variation distance) between $(\hat{q})$ under H_0 versus H_1 is 0.06. Hence, if $\rho \leq 0.01$, then no test can correctly guess whether H_0 or H_1 holds with probability greater than 53% in both cases.

A privacy attack can also be thought of as a test T. A tracing attack yields a test T to determine whether a target individual is included in the dataset, whereas a reconstruction attack entails multiple tests to determine the sensitive attribute of each user. Because the Gaussian mechanism ensures that these tests cannot be accurate, these attacks must fail. Thus, if $k \ll n^2\sigma^2$, then $\rho \ll 1$ and we can be assured that the output of $\mathcal{M}_{q,\sigma^2}(x)$ does not reveal sensitive information about the dataset x.

Theorem 8. The Gaussian mechanism can provide answers to k counting queries given a dataset of size n with error standard deviation σ whilst protecting privacy, as long as $\rho = \frac{k}{2n^2\sigma^2} \ll 1$.[7] In particular, we can answer $k \approx n^2$ queries with constant relative error (e.g., $\sigma = 0.01$) or we can answer $k \approx n$ queries with error comparable to the sampling error (i.e., $\sigma \approx 1/\sqrt{n}$).

The bounds of Theorem 8 almost match the attacks in Sections 2 and 3: Theorem 2 shows that answering $k = 2n$ queries with error $\sigma \ll 1/\sqrt{n}$ makes a reconstruction attack possible. In contrast, the Gaussian mechanism can answer $k = 2n$ queries with error $\sigma = 1/\sqrt{\rho n}$ and privacy $\rho \ll 1$. Similarly, Theorem 5 shows that answering $k \gg n^2$ queries with error $\sigma = 0.01$ opens the possibility of a tracing attack. Again, Theorem 8 shows that answering $k \approx n^2$ queries with error $\sigma = 0.01$ is possible whilst protecting privacy. Therefore, the results we have surveyed essentially pin down (in this simple setting) the boundary between what information can be released subject to a strong privacy guarantee like differential privacy versus when the released information permits a privacy attack.

We see that there is a tradeoff here—smaller values of σ yield less noisy answers, whereas larger values of σ provide greater privacy protection. This is a fundamental and inescapable dilemma; differential privacy provides the language in which to quantify and formally study the tension between privacy and utility.

4.3. Beyond Noise Addition

The Gaussian mechanism is but one of many differentially private mechanisms. It is extremely simple and versatile, but more sophisticated techniques can be used to obtain a better privacy-utility tradeoff in certain circumstances. If the properties $q = (q_1, \ldots, q_k)$ are structured in some way (such as being m-way marginals), then carefully correlated noise (instead of independent Gaussian noise) sometimes yields better results. In particular, if the data is inherently low-dimensional (e.g., the data of each individual is described by d bits), there are differentially private mechanisms that can answer many more queries (e.g., $k \approx 2^{\sigma^2 n/\sqrt{d}}$ queries, each with error σ) (Blum et al. 2008, Hardt & Rothblum 2010).

There is now a rich algorithmic and statistical literature on the design of differentially private mechanisms, introducing a wide array of techniques. The reader is directed to Hardt et al. (2012), Ligett (2013), Dwork & Roth (2014), and Vadhan (2016) for recent tutorials.

[7] Formally, the Gaussian mechanism satisfies concentrated differential privacy (Bun & Steinke 2016, Dwork & Rothblum 2016) with parameter ρ, which implies that it satisfies (ε, δ)-differential privacy (Dwork et al. 2006a) with $\varepsilon = \rho + 2\sqrt{\rho \log(1/\delta)}$ for every $\delta > 0$.

DISCLOSURE STATEMENT

C.D. is a Microsoft employee, and Microsoft holds patents on differential privacy. The other authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We are grateful to Steve Fienberg for his critical questioning since the inception of differential privacy. His intellectual curiosity and generosity of spirit have helped to lessen the distance between our two communities, and we thank him for soliciting this survey. We also wish to acknowledge Salil Vadhan for many helpful conversations. A.S. is supported in part by a grant from the Sloan foundation, a Google Faculty Research Award, and NSF grant IIS-1447700. T.S. was supported by NSF grants CCF-1420938 and CNS-1237235.

LITERATURE CITED

Bassily R, Nissim K, Smith A, Stemmer U, Ullman J. 2016. Algorithmic stability for adaptive data analysis. *Proc. 48th Annu. ACM SIGACT Symp. Theory Comput.*, pp. 1046–59. New York: ACM

Blum A, Ligett K, Roth A. 2008. A learning theory approach to non-interactive database privacy. In *Proc. 40th Annu. ACM Symp. Theory Comput.*, pp. 609–18. New York: ACM

Bun M, Steinke T. 2016. Concentrated differential privacy: Simplifications, extensions, and lower bounds. arXiv:1605.02065 [cs.CR]

Calandrino J, Kilzer A, Narayanan A, Felten E, Shmatikov V. 2011. "You might also like:" privacy risks of collaborative filtering. *Proc. 2011 IEEE Symp. Secur. Priv.*, pp. 231–46. Washington, DC: IEEE

Chor B, Fiat A, Naor M. 1994. Tracing traitors. *Proc. 14th Annu. Int. Cryptol. Conf. Adv. Cryptol.*, pp. 257–70. London: Springer-Verlag

De A. 2012. Lower bounds in differential privacy. *Theory of Cryptography: 9th Theory of Cryptography Conference, TCC 2012, Taormina, Sicily, Italy, March 19–21, 2012. Proceedings*, ed. R Cramer, pp. 321–38. Berlin: Springer-Verlag

Dinur I, Nissim K. 2003. Revealing information while preserving privacy. *Proc. 22nd ACM SIGMOD-SIGACT-SIGART Symp. Princ. Database Syst.*, pp. 202–10. New York: ACM

Dwork C. 2006. Differential privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10–14, 2006, Proceedings, Part II*, eds. M Bugliesi, B Preneel, V Sassone, I Wegener, pp. 1–13. Berlin: Springer-Verlag

Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A. 2015a. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems 28*, ed. C Cortes, ND Lawrence, DD Lee, M Sugiyama, R Garnett, pp. 2350–58. Red Hook, NY: Curran

Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth AL. 2015b. Preserving statistical validity in adaptive data analysis. *Proc. 47th Annu. ACM Symp. Theory Comput.*, pp. 117–26. New York: ACM

Dwork C, Feldman V, Hardt M, Pitassi T, Reingold O, Roth A. 2015c. The reusable holdout: preserving validity in adaptive data analysis. *Science* 349:636–38

Dwork C, Kenthapadi K, McSherry F, Mironov I, Naor M. 2006a. Our data, ourselves: privacy via distributed noise generation. *Advances in Cryptology - EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006. Proceedings*, ed. S Vaudenay, pp. 486–503. Berlin: Springer

Dwork C, McSherry F, Nissim K, Smith A. 2006b. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4–7, 2006. Proceedings*, ed. S Halevi, T Rabin, pp. 265–84. Berlin: Springer

Dwork C, McSherry F, Talwar K. 2007. The price of privacy and the limits of LP decoding. *Proc. 39th Annu. ACM Symp. Theory Comput.*, pp. 85–94. New York: ACM

Dwork C, Naor M. 2008. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *J. Priv. Confid.* 2(1):8

Dwork C, Naor M, Reingold O, Rothblum GN, Vadhan SP. 2009. On the complexity of differentially private data release: efficient algorithms and hardness results. *Proc. 41st Annu. ACM Symp. Theory Comput.*, pp. 381–90. New York: ACM

Dwork C, Roth A. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9:211–407

Dwork C, Rothblum G. 2016. Concentrated differential privacy. arXiv:1603.01887 [cs.DS]

Dwork C, Smith A, Steinke T, Ullman J, Vadhan S. 2015d. Robust traceability from trace amounts. In *IEEE 56th Ann. Symp. on Foundations of Computer Science (FOCS), Berkeley, CA, Oct. 18–20*. **http://ieeexplore.ieee.org/document/7354420/**

Dwork C, Yekhanin S. 2008. New efficient attacks on statistical disclosure control mechanisms. *Proc. 28th Annu. Conf. Cryptology: Adv. Cryptol.*, pp. 469–480. Berlin: Springer-Verlag

Hardt M, Miklau G, Pierce B, Roth A. 2012. *Slides and video for tutorial presentations*. DIMACS Worksh. Recent Work on Differ. Priv. across Comput. Sci., Oct. 24–26. **http://dimacs.rutgers.edu/Workshops/DifferentialPrivacy/Slides/slides.html**

Hardt M, Rothblum G. 2010. A multiplicative weights mechanism for privacy-preserving data analysis. *Proc. 2010 IEEE 51st Ann. Symp. Found. Comput. Sci.*, pp. 61–70. Washington, DC: IEEE

Hardt M, Ullman J. 2014. Preventing false discovery in interactive data analysis is hard. *2014 IEEE 55th Annu. Symp. Found. Comput. Sci.*, pp. 454–63. Washington, DC: IEEE

Homer N, Szelinger S, Redman M, Duggan D, Tembe W, et al. 2008. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLOS Genet.* 4:e1000167

Kasiviswanathan SP, Rudelson M, Smith A. 2013. The power of linear reconstruction attacks. *Proc. 24th Annu. ACM-SIAM Symp. Discret. Algorithms*, pp. 1415–33. Philadelphia: SIAM

Kasiviswanathan SP, Rudelson M, Smith A, Ullman J. 2010. The price of privately releasing contingency tables and the spectra of random matrices with correlated rows. *Proc. 42nd ACM Symp. Theory Comput.*, pp. 775–84. New York: ACM

Ligett K. 2013. *Slides and archived video tutorials.* Simons Inst. Worksh. on Big Data and Differ. Priv., Berkeley, CA, Dec. Dec. 11–14. **https://simons.berkeley.edu/workshops/schedule/78**

Muthukrishnan S, Nikolov A. 2012. Optimal private halfspace counting via discrepancy. *Proc. 44th Annu. ACM Symp. Theory Comput.*, pp. 1285–92. New York: ACM

Narayanan A, Shmatikov V. 2008. Robust de-anonymization of large sparse datasets. *Proc. 2008 IEEE Symp. Secur. Priv.*, pp. 111–125. Washington, DC: IEEE

Neyman J, Pearson ES. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc. A* 231:289–337

Nikolov A, Talwar K, Zhang L. 2013. The geometry of differential privacy: the sparse and approximate cases. *Proc. 45th Annu. ACM Symp. Theory Comput.*, pp. 351–60. New York: ACM

Pres. Counc. Advis. Sci. Technol. 2014. *Report to the President: Big Data and Privacy: A Technological Perspective*. Washington, DC: Executive Off. Pres.

Quadrianto N, Smola AJ, Caetano TS, Le QV. 2009. Estimating labels from label proportions. *J. Mach. Learn. Res.* 10:2349–74

Sankararaman S, Obozinski G, Jordan MI, Halperin E. 2009. Genomic privacy and limits of individual detection in a pool. *Nat. Genet.* 41:965–67

Steinke T, Ullman J. 2015. Interactive fingerprinting codes and the hardness of preventing false discovery. *JMLR Worksh. Conf. Proc.* 40:1–41

Sweeney L. 1997. Weaving technology and policy together to maintain confidentiality. *J. Law Med. Ethics* 25:98–110

Vadhan S. 2016. *The complexity of differential privacy*. Work. Pap., Cent. Res. Comput. Soc., Harvard Univ. **http://privacytools.seas.harvard.edu/publications/complexity-differential-privacy**

Wasserman L, Zhou S. 2010. A statistical framework for differential privacy. *J. Am. Stat. Assoc.* 105:375–89

Yu F. 2015. Scalable privacy-preserving data sharing methodology for genome-wide association studies. PhD thesis, Carnegie Mellon Univ.