Calibrating Noise to Sensitivity in Private Data Analysis

Cynthia Dwork[1], Frank McSherry[1], Kobbi Nissim[2], and Adam Smith[3]*

[1] Microsoft Research, Silicon Valley. {dwork,mcsherry}@microsoft.com
[2] Ben-Gurion University. kobbi@cs.bgu.ac.il
[3] Weizmann Institute of Science. adam.smith@weizmann.ac.il

Abstract. We continue a line of research initiated in [10, 11] on privacy-preserving statistical databases. Consider a trusted server that holds a database of sensitive information. Given a query function f mapping databases to reals, the so-called *true answer* is the result of applying f to the database. To protect privacy, the true answer is perturbed by the addition of random noise generated according to a carefully chosen distribution, and this response, the true answer plus noise, is returned to the user.

Previous work focused on the case of noisy sums, in which $f = \sum_i g(x_i)$, where x_i denotes the ith row of the database and g maps database rows to $[0, 1]$. We extend the study to general functions f, proving that privacy can be preserved by calibrating the standard deviation of the noise according to the *sensitivity* of the function f. Roughly speaking, this is the amount that any single argument to f can change its output. The new analysis shows that for several particular applications substantially less noise is needed than was previously understood to be the case.

The first step is a very clean characterization of privacy in terms of indistinguishability of transcripts. Additionally, we obtain separation results showing the increased value of interactive sanitization mechanisms over non-interactive.

1 Introduction

We continue a line of research initiated in [10, 11] on privacy in *statistical* databases. A statistic is a quantity computed from a sample. Intuitively, if the database is a representative sample of an underlying population, the goal of a privacy-preserving statistical database is to enable the user to learn properties of the population as a whole while protecting the privacy of the individual contributors.

We assume the database is held by a trusted server. On input a query function f mapping databases to reals, the so-called *true answer* is the result of applying f to the database. To protect privacy, the true answer is perturbed by the addition

* Supported by the Louis L. and Anita M. Perlman Postdoctoral Fellowship.

of random noise generated according to a carefully chosen distribution, and this response, the true answer plus noise, is returned to the user.

Previous work focused on the case of noisy sums, in which $f = \sum_i g(x_i)$, where x_i denotes the ith row of the database and g maps database rows to $[0,1]$. The power of the noisy sums primitive has been amply demonstrated in [6], in which it is shown how to carry out many standard datamining and learning tasks using few noisy sum queries.

In this paper we consider general functions f mapping the database to vectors of reals. We prove that privacy can be preserved by calibrating the standard deviation of the noise according to the *sensitivity* of the function f. This is the maximum amount, over the domain of f, that any single argument to f, that is, any single row in the database, can change the output.

We begin by defining a new notion of privacy leakage, ϵ-indistinguishability. An interaction between a user and a privacy mechanism results in a *transcript*. For now it is sufficient to think of transcripts corresponding to a single query function and response, but the notion is completely general and our results will handle longer transcripts.

Roughly speaking, a privacy mechanism is ϵ-indistinguishable if for all transcripts t and for all databases $\mathbf{x}$ and $\mathbf{x}'$ differing in a single row, the probability of obtaining transcript t when the database is $\mathbf{x}$ is within a $(1+\epsilon)$ multiplicative factor of the probability of obtaining transcript t when the database is $\mathbf{x}'$. More precisely, we require the absolute value of the logarithm of the ratios to be bounded by ϵ. In our work, ϵ is a parameter chosen by *policy*.

We then formally define the sensitivity $S(f)$ of a function f. This is a quantity *inherent* in f; it is not chosen by policy. Note that $S(f)$ is independent of the actual database.

We show a simple method of adding noise that ensures ϵ-indistinguishability of transcripts; the noise depends only on ϵ and $S(f)$, and is independent of the database and hence of its size. Specifically, to obtain ϵ-indistinguishability it suffices to add noise according to the following distribution: $Pr[y] \propto e^{-\epsilon|y|/S(f)}$.

The extension to privacy-preserving approximations to "holistic" functions f that operate on the entire database broadens the scope of private data analysis beyond the orignal motivation of a purely statistical, or "sample population" context. Now we can view the database as an object that is *itself* of intrinsic interest and that we wish to analyze in a privacy-preserving fashion. For example, the database may describe a concrete interconnection network – not a sample subnetwork – and we wish to learn certain properties of the network without releasing information about individual edges, nodes, or subnetworks. The technology developed herein therefore extends the scope of the line of research, beyond privacy-preserving statistical databases to privacy-preserving analysis of data.

1.1 Additional Contributions

Definitions of Privacy Definition of privacy requires care. In addition to our indistinguishability-based definition mentioned above we also consider notions based on semantic security and simulation and prove equivalences among these. A simple hybrid argument shows that utility requires non-negligible information leakage, hence all our definitions differ from their original cryptographic counterparts in that we accommodate non-negligible leakage. In particular, the standard measure of statistical difference is not a sufficiently good metric in our setting, and needs to be replaced with a more delicate one.

In previous work [10, 11, 6], the definitions were based on semantic security but the proofs were based on indistinguishability, so our move to ϵ-indistinguishability is a simplification. Also, semantic security was proved against *informed* adversaries. That is, an adversary with knowledge of the entire database except a single row, say, row i, could not glean any additional information about row i beyond what it knew before interaction with the privacy mechanism. This is fine; it says that without the database, seeing that X smokes does not necessarily increase our gambling odds that X will develop heart disease, but if the database teaches the correlation between smoking and heart disease improving our guessing odds should not be considered a violation of privacy. However, the new formulation immediately gives indistinguishability against an adversary with any amount of prior knowledge, and the above explanation is no longer necessary.

Examples of Sensitivity-Based Analysis To illustrate our approach, we analyze the sensitivity of specific data analysis functions, including histograms, contingency tables, and covariance matrices, all of which have very high-dimensional output, and show that their sensitivities are independent of the dimension. Previous privacy-preserving approximations to these quantities used noise proportional to the dimension; the new analysis permits noise of size $O(1)$. We also give two general classes of functions which have low sensitivity: functions which estimate distance from a set (e.g minimum cut size in a network) and functions which can be approximated from a random sample.

Limits on Non-Interactive Mechanisms. There are two natural models of data sanitization: interactive and non-interactive. In the non-interactive setting, the data collector—a trusted entity—publishes a "sanitized" version of the collected data; the literature uses terms such as "anonymization" and "de-identification". Traditionally, sanitization employed some perturbation and data modification techniques, and may also have included some accompanying synopses and statistics. In the interactive setting, the data collector provides a mechanism with which users may pose queries about the data, and get (possibly noisy) answers.

The first of these seems quite difficult (see [12, 7, 8]), possibly due to the difficulty of supplying utility that has not yet been specified at the time the sanitization is carried out. In contrast, powerful results for the interactive approach have been obtained ([11, 6] and the present paper). We show that for any non-interactive mechanism San satisfying our definition of privacy, there exist low-sensitivity functions $f(\mathbf{x})$ which cannot be approximated at all based on $\mathsf{San}(\mathbf{x})$, unless the database is very large: If each database entry consists of d bits, then the database must have $2^{\Omega(d)}$ entries in order to answer all low-sensitivity queries—even to answer queries from a restricted class called *sum queries*. In other words, a non-interactive mechanism must be tailored to suit certain functions to the exclusion of others. This is not true in the interactive setting, since one can answer the query f with little noise regardless of n[4].

The separation results are significant given that the data-mining literature has focused almost exclusively on non-interactive mechanisms, specifically, randomized response (see Related Work below) and that statisticians have traditionally operated on "tables" and have expressed to us a strong preference for non-interactive "noisy tables" over an interactive mechanism.

1.2 Related Work

The literature in statistics and computer science on disseminating statistical data while preserving privacy is extensive; we discuss only directly relevant work here. See, e.g., [5] for pointers to the broader literature.

Privacy from Perturbation. The venerable idea of achieving privacy by adding noise is both natural and appealing. An excellent and detailed exposition of the many variants of this approach explored in the context of statistical disclosure control until 1989, many of which are still important elements of the toolkit for data privacy today, may be found in the survey of Adam and Wortmann [1]. The "classical" antecedent closest in spirit to our approach is the work of Denning [9].

Perturbation techniques are classified into two basic categories: (i) *Input perturbation techniques*, where the underlying data are randomly modified, and answers to questions are computed using the modified data; and (ii) *Output perturbation*, where (correct) answers to queries are computed exactly from the real data, but noisy versions of these are reported. Both techniques suffer from certain inherent limitations (see below); it seems that these limitations caused a decline in interest within the computer science community in designing perturbation techniques for achieving privacy[5].

The work of Agrawal and Srikant [3] rekindled this interest; their principal contribution was an algorithm that, given an input-perturbed database, learns the original input distribution. Subsequent work studied the applicability and limitations of perturbation techniques, and privacy definitions have started to evolve, as we next describe.

Definitional Work. Several privacy definitions have been put forward since [3]. Their definition measured privacy in terms of the noise magnitude added to

[4] It is also not true if one employs weaker definitions of security; the connection between the definitions and the separation between models of interaction is subtle and, in our view, surprising. See Section 4.

[5] The same is not true of the statistics community; see, for example, the work of Roque [14].

a value. This was shown to be problematic, as the definition ignored what an adversary knowing the underlying probability distribution might infer about the data [2]. Evfimievsky et al. [12] noted, however, that such an *average* measure allows for infrequent but noticeable privacy breaches, and suggested measuring privacy in terms of the *worst-case* change in an adversary's a-priori to a-posteriori beliefs. Their definition is a special case of Definition 1 for input perturbation protocols of a limited form. A similar, more general definition was suggested in [10, 11, 6]. This was modeled after semantic security of encryptions.

Our basic definition of privacy, ϵ-indistinguishability, requires that a change in one database entry induce a small change in the distribution on the view of the adversary, under a specific, "worst-case" measure of distance. It is the same as in [12], adapted to general interactive protocols. An equivalent, semantic-security-flavored formulation is a special case of the definition from [10, 11, 6]; those definitions allowed a large loss of privacy to occur with negligible probability.

We note that k-*anonymity* [15] and the similarly motivated notion of protection against *isolation* [7, 8]) have also been in the eye of privacy research. The former is a syntactic characterization of (input-perturbed) databases that does not immediately capture semantic notions of privacy; the latter definition is a geometric interpretation of protection against being brought to the attention of others. The techniques described herein yield protection against isolation.

Sum Queries. A cryptographic perspective on perturbation was initiated by Dinur and Nissim [10]. They studied the amount of noise needed to maintain privacy in databases where a query returns (approximately) the number of 1's in any given subset of the entries. They showed that if queries are not restricted, the amount of noise added to each answer must be very high – linear (in n, the size of the database) for the case of a computationally unbounded adversary, and $\Omega(\sqrt{n})$ for a polynomially (in n) bounded adversary. Otherwise, the adversary can reconstruct the database almost exactly, producing a database that errs on, say, 0.01% of the entries. In contrast, jointly with Dwork, they initiated a sequence of work [10, 11, 6] which showed that limiting the users to a sublinear (in n) number of queries ("SuLQ") allows one to release useful global information while satisfying a strong definition of privacy. For example, it was shown that the computationally powerful noisy sum queries discussed above, that is, $\sum_{i=1}^{n} g(i, x_i)$, where g maps rows to values in $[0, 1]$, can be safely answered by adding $o(\sqrt{n})$ noise (from a gaussian, binomial, or Laplace distribution)— a level well below the sampling error one would expect in the database initially.

2 Definitions

We model the adversary as a probabilistic interactive Turing machine with an advice tape. Given a database access protocol San, an adversary $\mathcal{A}$, and a particular database $\mathbf{x}$, let the random variable $\mathcal{T}_{\mathsf{San},\mathcal{A}}(\mathbf{x})$ denote the transcript. The randomness in $\mathcal{T}_{\mathsf{San},\mathcal{A}}(\mathbf{x})$ comes from the coins of San and of $\mathcal{A}$. Note that for non-interactive schemes, there is no dependence on the adversary $\mathcal{A}$. We will drop either or both of the subscripts San and $\mathcal{A}$ when the context is clear.

We model the database as a vector of n entries from some domain D. We typically consider domains D of the form $\{0,1\}^d$ or $\mathbb{R}^d$. The Hamming distance $\mathsf{d}_H(\cdot,\cdot)$ over D^n is the number of entries in which two databases differ.

Our basic definition of privacy requires that close databases correspond to close distributions on the transcript. Specifically, for every transcript, the probabilities of it being produced with the two possible databases are close. We abuse notation somewhat and use $\Pr[A = a]$ to denote probability density for both continuous and discrete random variables.

Definition 1. *A mechanism is* ϵ-indistinguishable *if for all pairs* $\mathbf{x}, \mathbf{x}' \in D^n$ *which differ in only one entry, for all adversaries* $\mathcal{A}$*, and for all transcripts* t*:*

$$\left|\ln\left(\frac{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}) = t]}{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}') = t]}\right)\right| \leq \epsilon. \tag{1}$$

We sometimes call ϵ the *leakage*. When ϵ is small, $\ln(1+\epsilon) \approx \epsilon$, and so the definition is roughly equivalent to requiring that for all transcripts t, $\frac{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x})=t]}{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}')=t]} \in 1 \pm \epsilon$.

The definition is unusual for cryptography, in that in most cryptographic settings it is sufficient to require that distributions be *statistically close* (i.e. have small total variation distance) or that they be *computationally indistinguishable*. However, the requirement of Definition 1 is much more stringent than statistical closeness: one can have a pair of distributions whose statistical difference is arbitrarily small, yet where the ratio in Eqn. 1 is infinite (by having a point where one distribution assigns probability zero and the other, non-zero). We chose the more stringent notion because (a) it is achievable at very little cost, and (b) more standard distance measures do not yield meaningful guarantees in our context, since, as we will see, the leakage must be non-negligible. As with statistical closeness, Definition 1 also has more "semantic" formulations; these are discussed in Appendix A.

As we will next show, it is possible to release quite a lot of "global" information about the database while satisfying Definition 1. We first define the *Laplace* distribution, $\text{Lap}(\lambda)$. This distribution has density function $h(y) \propto \exp(-|y|/\lambda)$, mean 0, and standard deviation λ.

Example 1 (Noisy Sum). Suppose $\mathbf{x} \in \{0,1\}^n$, and the user wants to learn $f(\mathbf{x}) = \sum_i x_i$, the total number of 1's in the database. Consider adding noise to $f(\mathbf{x})$ according to a Laplace distribution:

$$\mathcal{T}(x_1, \ldots, x_n) = \sum_i x_i + Y, \quad \text{where } Y \sim \text{Lap}(1/\epsilon).$$

This mechanism is ϵ-indistinguishable. To see why, note that for any real numbers y, y' we have $\frac{h(y)}{h(y')} \leq e^{\epsilon|y-y'|}$. For any two databases $\mathbf{x}$ and $\mathbf{x}'$ which differ in a single entry, the sums $f(\mathbf{x})$ and $f(\mathbf{x}')$ differs by one. Thus, for $t \in \mathbb{R}$, the ratio $\frac{\Pr(\mathcal{T}(\mathbf{x})=t)}{\Pr(\mathcal{T}(\mathbf{x}')=t)} = \frac{h(t-f(\mathbf{x}))}{h(t-f(\mathbf{x}'))}$ is at most $e^{\epsilon|f(\mathbf{x})-f(\mathbf{x}')|} \leq e^{\epsilon}$, as desired.

Non-negligible Leakage and the Choice of Distance Measure. In the example above it is clear that even to get a constant-factor approximation to $f(\mathbf{x})$, we must have $\epsilon = \Omega(1/n)$, quite large by cryptographic standards where the usual requirement is for the leakage to drop faster than any polynomial in the lengths of the inputs. However, non-negligible or leakage is *inherent* for statistical utility: If the distance ϵ between the distributions induced by close databases is $o(1/n)$, then the distance between the distributions induced by *any* two databases is $o(1)$ and *no* statistic about the database can be usefully approximated.

Average-case distance measures such as statistical difference do not yield meaningful guarantees when $\epsilon = \Omega(1/n)$.

Example 2. Consider the candidate sanitization

$$\mathcal{T}(x_1, ..., x_n) = (i, x_i) \quad \text{where } i \in_R \{1, ..., n\} \, .$$

If $\mathbf{x}$ and $\mathbf{x}'$ differ in a single position, the statistical difference between $\mathcal{T}(\mathbf{x})$ and $\mathcal{T}(\mathbf{x}')$ is $1/n$, and yet it is clear that every transcript reveals private information about some individual.

Indeed, Definition 1 is not satisfied in this example, since if $\mathbf{x}$ and $\mathbf{x}'$ differ, say, in the ith coordinate, then the transcript (i, x_i) has probability zero when the database is $\mathbf{x}'$.

3 Sensitivity and Privacy

We now formally define sensitivity of functions, described informally in the Introduction. We will prove that choosing noise according to $\mathrm{Lap}(S(f)/\epsilon)$ ensures ϵ-indistinguishability when the query function f has sensitivity $S(f)$. We extend the analysis to vector-valued functions f, and even to adaptively chosen series of query functions. Intuitively, if ϵ is a "privacy budget" then this analysis explains how the budget is spent by a sequence of queries.

Definition 2 (L_1 Sensitivity). *The L_1 sensitivity of a function $f : D^n \to \mathbb{R}^d$ is the smallest number $S(f)$ such that for all $\mathbf{x}, \mathbf{x}' \in D^n$ which differ in a single entry,*

$$\|f(\mathbf{x}) - f(\mathbf{x}')\|_1 \leq S(f) \; .$$

Sensitivity is a Lipschitz condition on f: if $\mathsf{d}_H(\cdot, \cdot)$ is the Hamming metric on D^n, then for all pairs of databases $\mathbf{x}, \mathbf{x}' \in D^n$: $\frac{\|f(\mathbf{x}) - f(\mathbf{x}')\|_1}{\mathsf{d}_H(\mathbf{x}, \mathbf{x}')} \leq S(f)$. One can define sensitivity with respect to any metric on the output space; see Section 3.3.

Example 3 (Sums and Histograms). Consider the sum functionality above: if $D = \{0, 1\}$ and $f(\mathbf{x}) = \sum_{i=1}^{n} x_i$ (viewed as an real number), then the sensitivity of f with respect to the usual metric on $\mathbb{R}$ is $S_{L_1}(f) = 1$.

Now consider an arbitrary domain D which has been partitioned into d disjoint bins $B_1, ..., B_d$. The function $f : D^n \to \mathbb{Z}^d$ which computes the number of database points which fall into each bin is called a *histogram* for $B_1, ..., B_m$. Changing one point in the database can change at most two of these counts — one bin loses a point, another bin gains one. The L_1 sensitivity of f is thus 2, independent of d.

3.1 Calibrating Noise According to $S(f)$

Recall that if the noise Y is drawn from the Laplace distribution, then $h(y)/h(y')$ is at most $e^{|y-y'|/\lambda}$. A similar phenomenon holds in higher dimension. If Y is a vector of d independent Laplace variables, the density function at y is proportional to $\exp(-\|y\|_1/\lambda)$. A simple but important consequence is that the random variables $z+Y$ and $z'+Y$ are close in the sense of Definition 1: for all $t \in \mathbb{R}^d$,

$$\frac{\Pr(z+Y=t)}{\Pr(z'+Y=t)} \in \exp(\pm\frac{\|z-z'\|_1}{\lambda}).$$

Thus, to release a (perturbed) value $f(\mathbf{x})$ while satisfying privacy, it suffices to add Laplace noise with standard deviation $S(f)/\epsilon$ in each coordinate.

Proposition 1 (Non-interactive Output Perturbation). *For all $f : D^n \to \mathbb{R}^d$, the following mechanism is ϵ-indistinguishable:*
$\mathsf{San}_f(\mathbf{x}) = f(\mathbf{x}) + (Y_1, ..., Y_d)$ *where the Y_i are drawn i.i.d. from* $\mathrm{Lap}(S(f)/\epsilon)$

The proposition is actually a special case of the privacy of a more general, possibly adaptive, interactive process.

Before continuing with our discussion, we will need to clarify some of the notation to highlight subtleties raised by adaptivity. Specifically, adaptivity complicates the nature of the "query function", which is no longer a predetermined function, but rather a strategy for producing queries based on answers given thus far. For example, an adaptive histogram query might ask to refine those regions with a substantial number of respondents, and we would expect the set of such selected regions to depend on the random noise incorporated into the initial responses.

Recalling our notation, a transcript $t = [Q_1, a_1, Q_2, a_2 \ldots, Q_d, a_d]$ is a sequence of questions and answers. For notational simplicity, we will assume that Q_i is a well defined function of $a_1, \ldots, a_{i-1}$, and that we can therefore truncate our transcripts to be only a vector $t = [a_1, a_2, \ldots, a_d]$[6]. For any transcript t, we will let $f_t : D \to R^d$ be the function whose ith coordinate reflects the query Q_i, which we assume to be determined entirely by the first $i-1$ components of t. As we now see, we can bound the privacy of an adaptive series of questions using the largest diameter among the functions f_t.

Consider a trusted server, holding $\mathbf{x}$, which receives an adaptive sequence of queries $f_1, f_2, f_3, ..., f_d$, where each $f_i : D^n \to \mathbb{R}$. For each query, the server San either (a) refuses to answer, or (b) answers $f_i(\mathbf{x}) + \mathrm{Lap}(\lambda)$. The server can limit

[6] Although as written the choice of query is deterministic, this can be relaxed by adding coins to the transcript.

the queries by refusing to answer when $S(f_t)$ is above a certain threshold. Note that the decision whether or not to respond is based on $S(f_t)$, which can be computed by the user, and hence is not disclosive.

Theorem 1. *For an arbitrary adversary $\mathcal{A}$, let $f_t(\mathbf{x}) : D^n \to \mathbb{R}^d$ be its query function as parameterized by a transcript t. If $\lambda = \max_t S(f_t)/\epsilon$, the mechanism above is ϵ-indistinguishable.*

Proof. Using the law of conditional probability, and writing t_i for the indices of t,

$$\frac{\Pr[\mathsf{San}_f(\mathbf{x}) = t]}{\Pr[\mathsf{San}_f(\mathbf{x}') = t]} = \prod_i \frac{\Pr[\mathsf{San}_f(\mathbf{x})_i = t_i | t_1, \ldots, t_{i-1}]}{\Pr[\mathsf{San}_f(\mathbf{x}')_i = t_i | t_1, \ldots, t_{i-1}]}$$

For each term in the product, fixing the first $i-1$ coordinates of t fixes the values of $f_t(\mathbf{x})_i$ and $f_t(\mathbf{x}')_i$. As such, the conditional distributions are simple laplacians, and we can bound each term and their product as

$$\begin{aligned}\prod_i \frac{\Pr[\mathsf{San}_f(\mathbf{x})_i = t_i | t_1, \ldots, t_{i-1}]}{\Pr[\mathsf{San}_f(\mathbf{x}')_i = t_i | t_1, \ldots, t_{i-1}]} &\leq \prod_i \exp(|f_t(\mathbf{x})_i - f_t(\mathbf{x}')_i|/\lambda) \\ &= \exp(\|f_t(\mathbf{x}) - f_t(\mathbf{x}')\|_1/\lambda)\end{aligned}$$

We complete the proof using the bound $S(f_t) \leq \lambda\epsilon$, for all t.

3.2 Specific Insensitive Functions

We describe specific functionalities which have low sensitivity, and which consequently can be released with little added noise using the protocols of the previous section.

Histograms and Disjoint Analyses. There are many types of analyses that first partition the input space into disjoint regions, before proceeding to analyze each region separately. One very simple example of such an analysis is a histogram, which simply counts the number of elements that fall into each region. Imagining that D is subdivided into d disjoint regions, and that $f : D^n \to \mathbb{Z}^d$ is the function that counts the number of elements in each region, we saw in Example 3 that $S(f) = 2$. Notice that the output dimension, d, does not play a role in the sensitivity, and hence in the noise needed in an ϵ-indistinguishable implementation of a histogram. Comparing this with what one gets by applying the framework of [6] we note a significant improvement in the noise. Regarding each bin value as a query, the noise added in the original framework to each coordinate is $O(\sqrt{d}/\epsilon)$, and hence the total L_1 error is an $O(\sqrt{d})$ factor larger than in our scheme. This factor is especially significant in applications where bins outnumber the data points (which is often the case with contingency tables).

Clearly any analysis that can be run on a full data set can be run on a subset, and we can generalize the above observation in the following manner. Letting

D be partitioned into d disjoint regions, let $f : D^n \to \mathbb{R}^d$ be a function whose output coordinates $f(x)_i$ depend only on those elements in the ith region. We can bound $S(f) \leq 2 \max_i S(f_i)$. Again, and importantly, the value of d does not appear in this bound.

Linear Algebraic Functions. One very common analysis is measuring the mean and covariance of attributes of the data. If $v : D \to \mathbb{R}^d$ is some function mapping rows in the database to column vectors in $\mathbb{R}^d$, the mean vector μ and covariance matrix C are defined as

$$\begin{aligned} \mu \;&=\; \mu(x_1, \ldots, x_n) \;=\; \operatorname*{avg}_i v(x_i) \\ \text{and}\;\; C \;&=\; C(x_1, \ldots, x_n) \;=\; \operatorname*{avg}_i v(x_i)v(x_i)^T - \mu\mu^T \;. \end{aligned}$$

These two objects have dimension d and d^2, respectively, and a crude bound on their sensitivity would incorporate these terms. However, if we are given an upper bound $\gamma = \max_x \|v(x)\|_1$, then we can incorporate it into our sensitivity bounds.

Specifically, the mean is simply a sum, and an arbitrary change to a single term results in a vector $\mu + \delta$ where $\|\delta\|_1 \leq 2\gamma/n$. The covariance matrix is more complicated, but is also a sum at heart. Using the L_1 norm on matrices as one might apply the L_1 to a d^2 dimensional vector, we see that an arbitrary change to a single x_i can change the $\mu\mu^T$ term by at most

$$\mu\mu^T - (\mu + \delta)(\mu + \delta)^T = \mu\delta^T + \delta\mu^T + \delta\delta^T \tag{2}$$

$$= \mu\delta^T + \delta(\mu + \delta)^T \tag{3}$$

The first and second terms each have L_1 norm at most $2\gamma^2/n$. An arbitrary change to x_i can alter a $v(x_i)v(x_i)^T$ term by at most $4\gamma^2$. Hence a total $L1$ change of $8\gamma^2/n$.

Again, we witness an improvement in the noise magnitude when compared to applying the framework of [6]. As computing C amounts to performing d^2 queries, we get $L1$ noise that is $O(d)$ factor larger than with the current analysis.

Distance from a Property. The functionalities discussed until now had a simple representation as sums of vectors, and the sensitivity was then (at most) twice the maximum L_1 norm of one of these vectors. However, one can bound the sensitivity of much more complex functions.

Given a set $S \subseteq D^n$, the distance $f_S(\mathbf{x})$ between a particular database $\mathbf{x}$ and S is the Hamming distance (in D^n) between $\mathbf{x}$ and the nearest point $\mathbf{x}'$ in S. For any set S, $f_S(\mathbf{x})$ has sensitivity 1. We can safely release $f_S(\mathbf{x}) + Y$ where $Y \sim \mathrm{Lap}(1/\epsilon)$.

As an example, we could imagine social network described as a database of links between pairs of individuals. We might like to measure how "robust" the network is: how many social links would have to change (either added or removed) for the graph to become disconnected, non-expansive, or poorly clustered? Each of these counts, which change by at most one when a single edge is altered, can be released with only a small amount of noise added.

Suppose that $n = \binom{m}{2}$, $D = [0, 1]$, and we interpret the entries of the database as giving the weights of edges in a graph with m vertices (that is, the "individuals" here are the edges). Then the weight of the minimum edge-cut in the graph, which is the distance from the nearest disconnected graph, is a 1-sensitive function. It is easily computable, and so one can safely release this information about a network (approximately) without violating the privacy of the component edges.

Other interesting graph functionalities also have low sensitivity. For example, if $D = [0, 1]$, the weight of the minimum spanning tree is 1-sensitive.

Functions with Low Sample Complexity. Any function f which can be accurately approximated by an algorithm which looks only at a small fraction of the database has low sensitivity, and so the value can be released safely with relatively little noise. In particular, functions which can be approximated based on a random sample of the data points fit this criterion.

Lemma 1. *Let $f : D^n \to \mathbb{R}^d$. Suppose there is a randomized algorithm A such that for all inputs $\mathbf{x}$, (1) for all i, the probability that A reads x_i is at most α and (2) $\|A(\mathbf{x}) - f(\mathbf{x})\|_1 \le \sigma$ with probability at least $\beta = \frac{1+\alpha}{2}$. Then $S(f) \le 2\sigma$.*

The lemma translates a property of f related to ease of computation into a combinatorial property related to privacy. It captures many of the low-sensitivity functions described in the preceding sections, although the bounds on sensitivity given by the lemma are often quite loose.

Proof. For any particular entry $i \in \{1, ..., n\}$, denote by $A(\mathbf{x})\big|_{-i}$ the distribution on the outputs of A conditioned on the event that A does not read position i. By the definition of conditional probability, we get that for all $\mathbf{x}$ the probability that $A(\mathbf{x})\big|_{-i}$ is within distance σ of $f(\mathbf{x})$ is strictly greater than $(\beta - \alpha)/(1 - \alpha) \ge \frac{1}{2}$. Pick any $\mathbf{x}, \mathbf{x}'$ which only differ in the ith position. By the union bound, there exists some point p in the support of $A(\mathbf{x})\big|_{-i}$ which is within distance σ of *both* $f(\mathbf{x})$ and $f(\mathbf{x}')$, and hence $\|f(\mathbf{x}) - f(\mathbf{x}')\|_1 \le \|f(\mathbf{x}) - p\|_1 + \|p - f(\mathbf{x}')\|_1 \le 2\sigma$.

One might hope for a converse to Lemma 1, but it does not hold. Not all functions with low sensitivity can be approximated by an algorithm with low sample complexity. For example, let $D = GF(2^{\lceil \log n \rceil})$ and let $f(\mathbf{x})$ denote the Hamming distance between $\mathbf{x}$ and the nearest codeword in a Reed-Solomon code of dimension $k = n(1 - o(1))$. One cannot learn anything about $f(\mathbf{x})$ using fewer than k queries, and yet f has sensitivity 1 [4].

3.3 Sensitivity in General Metric Spaces

The intuition that *in*sensitive functions of a database can be released privately is not specific to the L_1 distance. Indeed, it seems that if changing one entry in $\mathbf{x}$ induces a small change in $f(\mathbf{x})$ — under any measure of distance on $f(\mathbf{x})$

— then we should be able to release $f(\mathbf{x})$ privately with relatively little noise. We formalize this intuition for (almost) any metric $\mathsf{d}_{\mathcal{M}}$ on the output $f(\mathbf{x})$. We will use symmetry, i.e. $\mathsf{d}_{\mathcal{M}}(x,y) = \mathsf{d}_{\mathcal{M}}(y,x)$, and the triangle inequality: $\mathsf{d}_{\mathcal{M}}(x,y) \leq \mathsf{d}_{\mathcal{M}}(x,z) + \mathsf{d}_{\mathcal{M}}(z,y)$.

Definition 3. *Let $\mathcal{M}$ be a metric space with a distance function $\mathsf{d}_{\mathcal{M}}(\cdot,\cdot)$. The sensitivity $S_{\mathcal{M}}(f)$ of a function $f : D^n \to \mathcal{M}$ is the amount that the function value varies when a single entry of the input is changed.*

$$S_{\mathcal{M}}(f) \stackrel{def}{=} \sup_{\mathbf{x},\mathbf{x}' :\ \mathsf{d}_H(\mathbf{x},\mathbf{x}')=1} \mathsf{d}_{\mathcal{M}}(f(\mathbf{x}), f(\mathbf{x}'))$$

Given a point $z \in \mathcal{M}$, (and a measure on $\mathcal{M}$) we can attempt to define a probability density function

$$h_{z,\epsilon}(y) \propto \exp\left(\frac{\epsilon \cdot \mathsf{d}_{\mathcal{M}}(y,z)}{2 \cdot S_{\mathcal{M}}(f)}\right).$$

There may not always exist such a density function, since the right-hand expression may not integrate to a finite quantity. However, if it is finite then the distribution given by $h_{z,\epsilon}()$ is well-defined.

To reveal an approximate version of $f(\mathbf{x})$ with sensitivity S, one can sample a value according to $h_{f(\mathbf{x}),\epsilon/S}()$.

$$\Pr[\mathcal{T}(\mathbf{x}) = y] = \frac{\exp\left(\frac{\epsilon}{2S_{\mathcal{M}}(f)} \cdot \mathsf{d}_{\mathcal{M}}(y, f(\mathbf{x}))\right)}{\int_{y \in \mathcal{M}} \exp\left(\frac{\epsilon}{2S_{\mathcal{M}}(f)} \cdot \mathsf{d}_{\mathcal{M}}(y, f(\mathbf{x}))\right) dy}. \tag{4}$$

Theorem 2. *In a metric space where $h_{f(\mathbf{x}),\epsilon}()$ is well-defined, adding noise to $f(\mathbf{x})$ as in Eqn. 4 yields an ϵ-indistinguishable scheme.*

Proof. Let $\mathbf{x}$ and $\mathbf{x}'$ be two databases differing in one entry. The distance $\mathsf{d}_{\mathcal{M}}(f(\mathbf{x}), f(\mathbf{x}'))$ is at most $S(f)$. For any y, the ratio $\frac{\exp(\mathsf{d}_{\mathcal{M}}(y,f(\mathbf{x})))}{\exp(\mathsf{d}_{\mathcal{M}}(y,f(\mathbf{x}')))}$ is thus at most $e^{S(f)}$, by the triangle inequality. Similarly, the ratio $\frac{\exp\left(\frac{\epsilon}{2S(f)} \cdot \mathsf{d}_{\mathcal{M}}(y,f(\mathbf{x}))\right)}{\exp\left(\frac{\epsilon}{2S(f)} \cdot \mathsf{d}_{\mathcal{M}}(y,f(\mathbf{x}'))\right)}$ is at most $e^{\epsilon/2}$. Finally, the normalization constant $\int_{y \in \mathcal{M}} \exp\left(\frac{\epsilon \cdot \mathsf{d}_{\mathcal{M}}(y,f(\mathbf{x}))}{2S(f)}\right) dy$ also differs by a factor of at most $e^{\epsilon/2}$ between $\mathbf{x}$ and $\mathbf{x}'$, since at all points in the space the integrand differs by at most $e^{\epsilon/2}$. The total ratio $h_{f(\mathbf{x}),\epsilon}(y) \Big/ h_{f(\mathbf{x}'),\epsilon}(y)$ differs by at most $e^{\epsilon/2} \cdot e^{\epsilon/2} = e^{\epsilon}$, as desired.

Remark 1. One can get rid of the factor of 2 in the definition of $h_{z,\epsilon}()$ in cases where the normalization factor does not depend on z. This introduces slightly less noise.

As a simple example, consider a function whose output lies in the Hamming cube $\{0,1\}^d$. By Theorem 2, one can release $f(\mathbf{x})$ safely by flipping each bit of the output $f(\mathbf{x})$ independently with probability roughly $\frac{1}{2} - \frac{\epsilon}{2S(f)}$.

4 Separating Interactive Mechanisms from Non-interactive Ones

In this section, we show a strong separation between interactive and non-interactive database access mechanisms. Consider the interactive setting of [10, 11, 6], that answers queries of the form $f_g(\mathbf{x}) = \sum_{i=1}^{n} g(i, x_i)$ where $g : [n] \times D \to [0, 1]$. As the sensitivity of any f_g is 1, an interactive access mechanism can answer any such query with accuracy about $1/\epsilon$. This gives a good approximation to $f(\mathbf{x})$ as long as ϵ is larger than $1/n$.

Suppose the domain D is $\{0, 1\}^d$. We show below that for any non-interactive, ϵ-indistinguishable mechanism San, there are many functions f_g which cannot be answered by $\mathcal{T}_{\mathsf{San}}$ unless the database consists of at least $2^{\Omega(d)}$ points. For these queries, it is not possible to distinguish the sanitization of a database in which *all* of the n entries satisfy $g(i, x_i) = 0$ from a database in which all of the entries satisfy $g(i, x_i) = 1$. We will consider Boolean functions $g_{\mathbf{r}}$ of a specific form. Given n non-zero binary strings $\mathbf{r} = (r_1, r_2, ..., r_n)$, $r_i \in \{0, 1\}^d$, we define $g_{\mathbf{r}}(i, x)$ to be the inner product, modulo 2, of r_i and x, that is $g_{\mathbf{r}}(i, x) = \bigoplus_j x^{(j)} r_i^{(j)}$, denoted $r_i \odot x$. In the following we will usually drop the subscript $\mathbf{r}$ and write g for $g_{\mathbf{r}}$.

Theorem 3 (Non-interactive Schemes Require Large Databases). *Suppose that* San *is an ϵ-indistinguishable*non-interactive *mechanism with domain $D = \{0, 1\}^d$. For at least $2/3$ of the functions of the form $f_g(\mathbf{x}) = \sum_i g(i, x_i)$, the following two distributions have statistical difference $O(n^{4/3} \epsilon^{2/3} 2^{-d/3})$:*

$$\text{Distribution 0: } \mathcal{T}_{\mathsf{San}}(\mathbf{x}) \text{ where } \mathbf{x} \in_R \{\mathbf{x} \in D^n : \ f_g(\mathbf{x}) = 0\}$$
$$\text{Distribution 1: } \mathcal{T}_{\mathsf{San}}(\mathbf{x}) \text{ where } \mathbf{x} \in_R \{\mathbf{x} \in D^n : \ f_g(\mathbf{x}) = n\}$$

In particular, if $n = o(\frac{2^{d/4}}{\sqrt{\epsilon}})$, for most functions $g(i, x) = r_i \odot x$, it is impossible to learn the relative frequency of database items satisfying the predicates $g(i, x_i)$. We prove Theorem 3 below. First, a few remarks:

1. The order of the quantifiers is important: for any particular $f_g()$, it is easy to design a non-interactive scheme which answers that query accurately. However, no single non-interactive scheme can answer most queries of this form, unless $n \in \exp(d)$.
2. The strong notion of ϵ-indistinguishability in Definition 1 is essential to Theorem 3. For example, consider the candidate sanitization which outputs m pairs (i, x_i) chosen at random from the database. When $m = \theta(1)$ this is essentially Example 2; it fails to satisfy Definition 1 but yields $O(1/n)$-close distributions $O(1/n)$ on neighboring databases. However, it does permit estimating f_g with accuracy about $n/\sqrt{m}$ (the order of quantifiers is again important: for any particular query, the sample will be good with high probability). Thus, even for constant m, this is better than what is possible for any ϵ-indistinguishable scheme with $n = 2^{o(d)}$.

4.1 A Stronger Separation for Randomized Response Schemes

"Randomized response" refers to a special class of non-interactive schemes, in which each user's data is perturbed individually, and then the perturbed values are published. That is, there exists a randomization operator $Z : D \to \{0,1\}^*$ such that

$$\mathcal{T}_{\mathsf{San}}(x_1, ..., x_n) = Z(x_1), ..., Z(x_n).$$

This approach means that no central server need ever see the users' private data: each user i computes $Z(x_i)$ and releases only that.

We can strengthen Theorem 3 for randomized response schemes. We can consider functions f_g where the *same* predicate $g : D \to \{0,1\}$ is applied to all the entries in $\mathbf{x}$. I.e. $f(\mathbf{x}) = \sum_i g(x_i)$ (e.g. "how many people in the database have blue eyes?"). For most vectors r, the parity check $g_r(x) = r \odot x$ will be difficult to learn from $Z(x)$, and so $f(\mathbf{x})$ will be difficult to learn from $\mathcal{T}_{\mathsf{San}}(\mathbf{x})$ unless n is very large.

Proposition 2 (Randomized Response). *Suppose that* San *is a ϵ-indistinguishable randomized response mechanism. For at least $2/3$ of the values $r \in \{0,1\}^d \setminus \{0^d\}$, the following two distributions have statistical difference $O(n\epsilon^{2/3}2^{-d/3})$:*

Distribution 0: $\mathcal{T}_{\mathsf{San}}(\mathbf{x})$ where each $x_i \in_R \left\{x \in \{0,1\}^d : \ r \odot x = 0\right\}$

Distribution 1: $\mathcal{T}_{\mathsf{San}}(\mathbf{x})$ where each $x_i \in_R \left\{x \in \{0,1\}^d : \ r \odot x = 1\right\}$

In particular, if $n = o(2^{d/3}/\epsilon^{2/3})$, no user can learn the relative frequency of database items satisfying the predicate $g_r(x) = r \odot x$, for most values r.

4.2 Proving the Separation Results

The two proofs have the same structure: a hybrid argument with a chain of length $2n$, in which the bound on statistical distance at each step in the chain is given by Lemma 2 below. Adjacent elements in the chain will differ according to the domain from which one of the entries in the database is chosen, and the elements in the chain are the probability distributions of the sanitizations when the database is chosen according to the given n-tuple of distributions.

For any r, partition the domain D into two sets: $D_r = \left\{x \in \{0,1\}^d : r \odot x = 0\right\}$, and $\bar{D}_r = D \setminus D_r = \left\{x \in \{0,1\}^d : r \odot x = 1\right\}$. We abuse notation and let D_r also stand for a random vector chosen uniformly from that set (similarly for D and $\bar{D}_r$).

The intuition for the key step is as follows. Given a randomized map $Z : D \to \{0,1\}^*$, the quantity $\Pr[Z(D_r) = z]$ is with high probability an estimate for $\Pr[Z(D) = z]$. That is because when r is chosen at random, D_r consists of 0^d, along with $2^{d-1} - 1$ points chosen pairwise independently in $\{0,1\}^d$. This allows us to show that the variance of the estimator $\Pr[Z(D_r) = z]$ is very small, as long as Z satisfies a strong indistinguishability condition implied by ϵ-indistinguishability. As a result, the distribution $Z(D_r)$ will be very close to $Z(D)$.

Lemma 2. *Let $Z : D \to \{0,1\}^*$ be a randomized map such that for all pairs $x, x' \in D$, and all outputs z, $\frac{\Pr[Z(x)=z]}{\Pr[Z(x')=z]} \in \exp(\pm\epsilon)$. For all $\alpha > 0$: with probability at least $1-\alpha$ over $r \in \{0,1\}^d \setminus \{0^d\}$,*

$$\mathbf{SD}\left(Z(D_r)\ ,\ Z(D)\right) \le O\Big(\frac{\epsilon^2}{\alpha \cdot 2^d}\Big)^{1/3}.$$

The same statement holds for $\bar{D}_r$.

The lemma is proved below, in Section 4.3. We first use it to prove the two separation results.

Proof (Proof of Theorem 3). "Distribution 0" in the statement is $\mathcal{T}_{\mathsf{San}}(D_{r_1}, ..., D_{r_n})$. We show that with high probability over the choice of the r_i's, this is close the transcript distribution induced by a uniform input, i.e. $\mathcal{T}(D, ..., D)$. We proceed by a hybrid argument, adding one constraint at a time. For each i, we want to show

$$\begin{aligned}&\mathcal{T}_{\mathsf{San}}(D_{r_1}, ..., D_{r_i}, \quad D \quad, D, ..., D) \quad \text{is close to}\\ &\mathcal{T}_{\mathsf{San}}(D_{r_1}, ..., D_{r_i}, \quad D_{r_{i+1}}\ , D, ..., D).\end{aligned}$$

Suppose that we have chosen $r_1, ..., r_i$ already. For any $x \in \{0,1\}^d$, consider the randomized map where the $(i+1)$-th coordinate is fixed to x:

$$Z(x) = \mathcal{T}_{\mathsf{San}}(D_{r_1}, ..., D_{r_i}, \quad x \quad, D, ..., D) \tag{5}$$

Note that $Z(D)$ is equal to the i-th step in the hybrid, and $Z(D_{r_{i+1}})$ is equal to the $(i+1)$-st step.

The ϵ-indistinguishabilityof San implies that $Z()$ satisfies $\frac{\Pr[Z(x)=z]}{\Pr[Z(x')=z]} \in \exp(\pm\epsilon)$. Applying Lemma 2 shows that with probability at least $1-\frac{1}{6n}$ over r_{i+1}, $Z(D_{r_i})$ is within statistical difference σ of $Z(D)$, where $\sigma = O(\sqrt[3]{n\epsilon^2 2^{-d}})$. That is, adding the i-th constraint on the inputs changes the output distribution by at most σ. By a union bound, all the steps in the hybrid have size at most σ with probability at least $\frac{5}{6}$. In that case, the total distance is $n\sigma$.

We can apply exactly the same reasoning to a hybrid starting with Distribution 1, and ending with $\mathcal{T}(D, ..., D)$. Again, with probability at least $\frac{5}{6}$, the total distance is $n\sigma$. With probability at least $2/3$, both chains of hybrids accumulate statistical difference bounded by $n\sigma$, and the distance between Distributions 0 and 1 is at most $2n\sigma = O(n^{4/3}\epsilon^{2/3}2^{-d/3})$.

Proof (Proof of Proposition 2). If $\mathcal{T}_{\mathsf{San}}$ is a randomized response scheme, then there is a randomized map $Z()$ from D to $\{0,1\}^*$, such that $\mathcal{T}_{\mathsf{San}}(x_1, ..., x_n) = Z(x_1), ..., Z(x_n)$. If $\mathcal{T}_{\mathsf{San}}$ is ϵ-indistinguishable, then for all pairs $x, x' \in D$, and for all outputs z, $\frac{\Pr[Z(x)=z]}{\Pr[Z(x')=z]} \in \exp(\pm\epsilon)$.

It is sufficient to show that with probability at least $2/3$ over a random choice r, $r \neq 0^d$, the distributions $Z(D_r)$ and $Z(\bar{D}_r)$ are within statistical difference $O(\epsilon^{2/3}2^{-d/3})$. This follows by applying Lemma 2 with $\alpha = 1/3$. By a hybrid argument, the difference between Distributions 0 and 1 above is then $O(n\epsilon^{2/3}2^{-d/3})$.

4.3 Proving that Random Subsets Approximate the Output Distribution

Proof (Proof of Lemma 2). Let $p(z|x)$ denote the probability that $Z(x) = z$. If x is chosen uniformly in $\{0,1\}^d$, then the probability of outcome z is $p(z) = \frac{1}{2^d}\sum_x p(z|x)$.

For symmetry, we will pick not only the string r but an offset bit b, and look at the set $D_{r,b} = \left\{x \in \{0,1\}^d : r \odot x = b\right\}$. This simplifies the calculations somewhat.

One can think of $\Pr[Z(D_{r,b}) = z]$ as estimating $p(z)$ by pairwise-independently sampling $2^d/2$ values from the set D and only averaging over that subset. Since, by the assumption on Z, the values $p(z|x)$ all lie in an interval of width about $\epsilon \cdot p(z)$ around $p(z)$, this estimator will have small standard deviation. We will use this to bound the statistical difference.

Let $\hat{p}(z) = \Pr[Z(D_{r,b}) = z]$, where the probability is taken over the coin flips of Z and the choice of $x \in D_{r,b}$. For a fixed z, $\hat{p}(z)$ is a random variable depending on the choice of r, b, and $\mathbb{E}_{r,b}\left[\hat{p}(z)\right] = p(z)$.

Claim 1. $\mathsf{Var}_{r,b}\left[\hat{p}(z)\right] \le \dfrac{2\cdot\tilde{\epsilon}^2\cdot p(z)^2}{2^d}$, *where* $\tilde{\epsilon} = e^\epsilon - 1$.

The proof of Claim 1 appears below. We now complete the proof of Lemma 2. We say that a value z is *δ-good* for a pair (r,b) if $\hat{p}(z) - p(z) \le \delta\cdot p(z)$. By the Chebyshev bound, for all z,

$$\Pr_{r,b}\left[z \text{ is not } \delta\text{-good for } (r,b)\right] \le \frac{\mathsf{Var}\left[\hat{p}(z)\right]}{\delta^2 p(z)^2} \le \frac{2\tilde{\epsilon}^2}{\delta^2 2^d}.$$

If we take the distribution on z given by $p(z)$, then with probability at least $1-\alpha$ over pairs (r,b), the fraction of z's (under $p(\cdot)$) which are good is at least $1 - \frac{2\tilde{\epsilon}^2}{\alpha\delta^2 2^d}$.

Finally, if a $1-\gamma$ fraction of the z's are δ-good for a particular pair (r,b), then the statistical difference between the distribution $\hat{p}(z)$ and $p(z)$ is at most $2(\gamma+\delta)$. Setting $\delta = \sqrt[3]{\frac{2\alpha\tilde{\epsilon}^2}{2^d}}$, we get a total statistical difference of at most 4δ. Since $\tilde{\epsilon} < 2\epsilon$ for $\epsilon \le 1$, the total distance between $\hat{p}(\cdot)$ and $p(\cdot)$ is at most $4\sqrt[3]{12\epsilon^2 2^{-d}}$, for at least a $1-\alpha$ fraction of the pairs (r,b). The bit b is unimportant here since it only switches D_r and its complement $\bar{D}_r$. The distance between $Z(D_r)$ and $Z(D)$ is exactly the same as the distance between $Z(\bar{D}_r)$ and $Z(D)$, since $Z(D)$ is the mid-point between the two. Thus, the statement holds even over pairs of the form $(r,0)$. This proves Lemma 2.

Proof (Proof of Claim 1). Let p^* be the minimum over x of $p(z|x)$. Let $q_x = p(z|x) - p^*$ and $\bar{q} = p(z) - p^*$. The variance of $\hat{p}(z)$ is the same as the variance of $\hat{p}(z) - p^*$. We can write $\hat{p}(z) - p^*$ as $\frac{2}{2^d}\sum_x q_x\chi_0(x)$, where $\chi_0(x)$ is 1 if $x \in D_{r,b}$. The expectation of $\hat{p}(z) - p^*$ is $\bar{q}$, which we can write $\frac{1}{2^d}\sum_x q_x$.

$$\mathop{\mathsf{Var}}_{r,b}\left[\hat{p}(z)\right] = \mathop{\mathbb{E}}_{r,b}\left[\left(\frac{2}{2^d}\sum_x q_x\chi_0(x) - \frac{1}{2^d}\sum_x q_x\right)^2\right] = \mathop{\mathbb{E}}_{r,b}\left[\left(\frac{1}{2^d}\sum_x q_x\left(2\chi_0(x) - 1\right)\right)^2\right] \tag{6}$$

Now $\big(2\chi_0(x)-1\big) = (-1)^{r\odot x \oplus b}$. This has expectation 0. Moreover, for $x \neq y$, the expectation of $\big(2\chi_0(x)-1\big)\big(2\chi_0(y)-1\big)$ is exactly $1/2^d$ (if we chose r with no restriction it would be 0, but we have the restriction that $r \neq 0^d$). Expanding the square in Eqn. 6,

$$\underset{r,b}{\mathsf{Var}}\left[\hat{p}(z)\right] = \tfrac{1}{2^{2d}} \sum_x q_x^2 + \tfrac{1}{2^{3d}} \sum_{x \neq y} q_x q_y$$

$$= \tfrac{1-\frac{1}{2^d}}{2^{2d}} \sum_x q_x^2 + \tfrac{1}{2^d} \left(\tfrac{1}{2^d} \sum_x q_x \right)^2 \leq \tfrac{1}{2^d} \left(\max_x q_x^2 + \bar{q}^2 \right).$$

By the indistinguishability condition, both $(\max_x q_x)$ and $\bar{q}$ are at most $(e^\epsilon - 1)p^* \leq \tilde{\epsilon} \cdot p(z)$. Plugging this into the last equation proves Claim 1.

References

[1] N. R. Adam and J. C. Wortmann. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys*, 25(4), December 1989.

[2] Dakshi Agrawal and Charu C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. ACM, 2001.

[3] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In Weidong Chen, Jeffrey F. Naughton, and Philip A. Bernstein, editors, *SIGMOD Conference*, pages 439–450. ACM, 2000.

[4] Eli Ben-Sasson, Prahladh Harsha, and Sofya Raskhodnikova. Some 3cnf properties are hard to test. In *STOC*, pages 345–354. ACM, 2003.

[5] Web page for the Bertinoro CS-Statistics workshop on privacy and confidentiality. Available from `http://www.stat.cmu.edu/~hwainer`, July 2005.

[6] Avrim Blum, Cynthia Dwork, Frank McSherry, and Kobbi Nissim. Practical privacy: The sulq framework. In *PODS*, 2005.

[7] Shuchi Chawla, Cynthia Dwork, Frank McSherry, Adam Smith, and Hoeteck Wee. Toward privacy in public databases. In *Theory of Cryptography Conference (TCC)*, pages 363–385, 2005.

[8] Shuchi Chawla, Cynthia Dwork, Frank McSherry, and Kunal Talwar. On the utility of privacy-preserving histograms. In *21st Conference on Uncertainty in Artificial Intelligence (UAI)*, 2005.

[9] Dorothy E. Denning. Secure statistical databases with random sample queries. *ACM Transactions on Database Systems*, 5(3):291–315, September 1980.

[10] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 202–210, 2003.

[11] Cynthia Dwork and Kobbi Nissim. Privacy-preserving datamining on vertically partitioned databases. In Matthew K. Franklin, editor, *CRYPTO*, volume 3152 of *Lecture Notes in Computer Science*, pages 528–544. Springer, 2004.

[12] Alexandre V. Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 211–222, 2003.

[13] Shafi Goldwasser and Silvio Micali. Probabilistic encryption. *Journal of Computer and System Sciences*, 28(2):270–299, April 1984.
[14] Gina Roque. *Masking microdata with mixtures of normal distributions*. University of California, Riverside, 2000. Doctoral Dissertation.
[15] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, 2002.

Appendix

A "Semantically" Flavored Implications of Definition 1

Definition 1 equates privacy with the inability to distinguish two close databases. Indistinguishability is a convenient notion to work with (as is indistinguishability of encryptions [13]); however, it does not directly say what an adversary may do and learn. In this section we present some "semantically" flavored definitions of privacy, and their equivalence to Definition 1.

Because of the need to have some utility conveyed by the database, it is not possible to get as strong a notion of security as we can, say, with encryption. We discuss two definitions which we consider meaningful, suggestively named *simulatability* and *semantic security*. The natural intuition is that if the adversary learns very little about x_i for all i, then privacy is satisfied. Recall the discussion of smoking and heart disease, from the Introduction. What is actually shown is that the adversary cannot learn much more about any x_i than she could learn from knowing almost all the data points except x_i.

Extending terminology from Blum et al. [6], we say an adversary is *informed* if she knows some set of $n-k$ database entries before interacting with the mechanism, and tries to learn about the remaining ones. The parameter k measures her remaining uncertainty.

Definition 4. *A mechanism* San *is* (k,ϵ)-simulatable *if for every adversary* $\mathcal{A}$*, and for every set* $I \subseteq [n]$ *of size* $n-k$*, there exists an* informed *adversary* $\mathcal{A}'$ *such that for any* $\mathbf{x} \in D^n$*:*

$$\left|\ln\left(\frac{\Pr[\ \mathcal{T}_{\mathsf{San},\mathcal{A}}(\mathbf{x}) = t\ ]}{\Pr[\ \mathcal{A}'(\mathbf{x}\big|_I) = t\ ]}\right)\right| \le \epsilon$$

where $\mathbf{x}\big|_I$ *denotes the restriction of* $\mathbf{x}$ *to the index set* I*.*

For convenience in stating implications among definitions, we extend the definition of indistinguishability (Definition 1) to pairs of databases at Hamming distance k:

Definition 5. *A mechanism is* (k,ϵ)*-indistinguishable if for all pairs* $\mathbf{x}, \mathbf{x}'$ *which differ in at most* k *entries, for all adversaries* $\mathcal{A}$ *and for all transcripts* t*,* $\left|\ln\left(\frac{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x})=t]}{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}')=t]}\right)\right| \le \epsilon$*.*

Any $(1, \frac{\epsilon}{k})$-indistinguishable mechanism is also (k, ϵ)-indistinguishable. To see why, consider a chain of at most k databases connecting $\mathbf{x}$ and $\mathbf{x}'$, where only one entry changes at each step. The probabilities change by a factor of $\exp(\pm\epsilon/k)$ at each step, so $\frac{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x})=t]}{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}')=t]} \in \exp(\pm\epsilon/k)^k = \exp(\pm\epsilon)$.

Claim 2.

1. *A (k, ϵ)-indistinguishable mechanism is (k, ϵ)-simulatable.*
2. *A (k, ϵ)-simulatable mechanism is $(k, 2\epsilon)$-indistinguishable.*

Proof. (1) A mechanism that is (k, ϵ)-indistinguishable is (k, ϵ)-simulatable. The simulator fills in the missing entries of $\mathbf{x}$ with default values to obtain $\mathbf{x}'$ which differs from $\mathbf{x}$ in at most k entries, then simulates an interaction between $\mathsf{San}(\mathbf{x}')$ and $\mathcal{A}$.
(2) A mechanism that is (k, ϵ)-simulatable is $(k, 2\epsilon)$-indistinguishable. Suppose that $\mathbf{x}', \mathbf{x}''$ agree in a set I of $n-k$ positions. Definition 4 says that for all $\mathcal{A}$ and all subsets I of $n-k$ indices, there exists an $\mathcal{A}'$ that, seeing only the rows indexed by I, can relatively accurately simulate the distribution of transcripts induced when $\mathcal{A}$ interacts with the full database. Since $\mathbf{x}'\big|_I = \mathbf{x}''\big|_I$ the behavior of $\mathcal{A}'$ is close to both that of the privacy mechanism interacting with $\mathcal{A}$ on $\mathbf{x}'$ and $\mathcal{A}$ on $\mathbf{x}''$:

$$\left|\ln\left(\frac{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}') = t]}{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}'') = t]}\right)\right| \le \left|\ln\left(\frac{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}') = t]}{\Pr[\mathcal{A}'(\mathbf{x}'\big|_I) = t]}\right)\right| + \left|\ln\left(\frac{\Pr[\mathcal{A}'(\mathbf{x}''\big|_I) = t]}{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}'') = t]}\right)\right| \le 2\epsilon. \tag{7}$$

Simulatability states that for any i, little more is learned about individual i by an adversary interacting with the access mechanism than what she might learn from studying the rest of the world.

Simulatability still leaves implicit what, exactly, the adversary can compute about the database. *Semantic security* captures a more computationally-flavored meaning of privacy. Given an informed adversary, who knows $\mathbf{x}\big|_I$, we say a $\mathbf{x}' \in D^n$ is *consistent* if it agrees with the adversary's knowledge; i.e. $\mathbf{x}'\big|_I = \mathbf{x}\big|_I$. A *consistent probability distribution* $\mathcal{D}$ is a probability distribution over consistent databases.

Definition 6. *A mechanism is (k, ϵ)-*semantically secure *if every interaction with an informed adversary results in a bounded change in the a-posteriori probability distribution. That is, for all informed adversaries $\mathcal{A}$, for all consistent distributions $\mathcal{D}$, for all transcripts t, and for all predicates $f : D^n \to \{0, 1\}$:*

$$\left|\ln\left(\frac{\Pr[f(\mathbf{x}') = 1]}{\Pr[f(\mathbf{x}') = 1 | \mathcal{T}_{\mathcal{A}}(\mathbf{x}') = t]}\right)\right| \le \epsilon. \tag{8}$$

The probabilities are taken over the coins of $\mathcal{A}, \mathsf{San}$ and choices of consistent $\mathbf{x}'$ according to $\mathcal{D}$.

Claim 3. *A mechanism is (k,ϵ)-indistinguishable iff it is (k,ϵ)-semantically-secure.*

Proof. (1) Let San be a (k,ϵ)-indistinguishable mechanism, and assume San is not (k,ϵ)-semantically-secure. Using Bayes' rule, we get that for some f and t:

$$\ln\left(\frac{\Pr[f(\mathbf{x})=1]}{\Pr[f(\mathbf{x})=1|\mathcal{T}_{\mathcal{A}}(\mathbf{x})=t]}\right)=\ln\left(\frac{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x})=t]}{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x})=t|f(\mathbf{x})=1]}\right)>\epsilon. \tag{9}$$

Pick a consistent $\mathbf{x}_0$ that maximizes $\Pr[\mathcal{T}(\mathbf{x}_0)=t]$ subject to $f(\mathbf{x}_0)=0$. Clearly, $\Pr[\mathcal{T}(\mathbf{x}_0)=t]\geq\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x})=t]$. Similarly, pick a consistent $\mathbf{x}_1\in D$ that minimizes $\Pr[\mathcal{T}(\mathbf{x}_1)=t]$ subject to $f(\mathbf{x}_1)=1$. We get that

$$\ln\left(\frac{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}_0)=t]}{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}_1)=t]}\right)>\epsilon. \tag{10}$$

Noting that $\mathsf{d}_H(\mathbf{x}_1,\mathbf{x}_2)\leq k$ we get a contradiction to the mechanism being (k,ϵ)-indistinguishable.
(2) Let San be a (k,ϵ)-semantically-secure mechanism, and assume San is not (k,ϵ')-indistinguishable. That is, there exist $\mathbf{x}_0,\mathbf{x}_1$ such that $\mathsf{d}_H(\mathbf{x}_0,\mathbf{x}_1)\leq k$ and a possible transcript t such that

$$\left|\ln\left(\frac{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}_1)=t]}{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}_0)=t]}\right)\right|>\epsilon. \tag{11}$$

Wlog, assume $\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}_0)=t]>\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}_1)=t]$, and that $\mathbf{x}_0,\mathbf{x}_1$ agree on their first $K=n-k$ coordinates. Let $\mathcal{A}$ be an informed adversary that knows these entries, and $\mathcal{D}$ be a consistent distribution that assigns probability α to $\mathbf{x}_0$ and $1-\alpha$ to $\mathbf{x}_1$. Finally, take f to be any predicate such that $f(\mathbf{x}_b)=b$. We get that

$$\Pr[f(\mathbf{x}')=1|\mathcal{T}_{\mathcal{A}}(x')=t]=\frac{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}_1)=t]\cdot\Pr[f(\mathbf{x}')=1]}{\alpha\cdot\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}_0)=t]+(1-\alpha)\cdot\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}_1)=t]}, \tag{12}$$

and hence

$$\ln\left(\frac{\Pr[f(\mathbf{x}')=1]}{\Pr[f(\mathbf{x}')=1|\mathcal{T}_{\mathcal{A}}(\mathbf{x}')=t]}\right)=\ln\left(1-\alpha+\alpha\frac{Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}_0)=t]}{\Pr[\mathcal{T}_{\mathcal{A}}(\mathbf{x}_1)=t]}\right)>\ln(1-\alpha+\alpha e^{\epsilon}). \tag{13}$$

Taking $\alpha\to 1$ yields the claim.