Lower bounds in Differential Privacy

Anindya De[1]*

University of California at Berkeley
anindya@cs.berkeley.edu

Abstract. This paper is about private data analysis, in which a trusted curator holding a confidential database responds to real vector-valued queries. A common approach to ensuring privacy for the database elements is to add appropriately generated random noise to the answers, releasing only these *noisy* responses. A line of study initiated in [7] examines the amount of distortion needed to prevent privacy violations of various kinds. The results in the literature vary according to several parameters, including the size of the database, the size of the universe from which data elements are drawn, the "amount" of privacy desired, and for the purposes of the current work, the arity of the query. In this paper we sharpen and unify these bounds. Our foremost result combines the techniques of Hardt and Talwar [11] and McGregor *et al.* [13] to obtain linear lower bounds on distortion when providing differential privacy for a (contrived) class of low-sensitivity queries. (A query has low sensitivity if the data of a single individual has small effect on the answer.) Several structural results follow as immediate corollaries:

- We separate so-called *counting* queries from arbitrary *low-sensitivity* queries, proving the latter requires more noise, or distortion, than does the former;
- We separate $(\varepsilon, 0)$-differential privacy from its well-studied relaxation (ε, δ)-differential privacy, even when $\delta \in 2^{-o(n)}$ is negligible in the size n of the database, proving the latter requires less distortion than the former;
- We demonstrate that (ε, δ)-differential privacy is much weaker than $(\varepsilon, 0)$-differential privacy in terms of mutual information of the transcript of the mechanism with the database, even when $\delta \in 2^{-o(n)}$ is negligible in the size n of the database.

We also simplify the lower bounds on noise for counting queries in [11] and also make them unconditional. Further, we use a characterization of (ϵ, δ) differential privacy from [13] to obtain lower bounds on the distortion needed to ensure (ε, δ)-differential privacy for $\epsilon, \delta > 0$. We next revisit the LP decoding argument of [10] and combine it with a recent result of Rudelson [15] to improve on a result of Kasiviswanathan *et al.* [12] on noise lower bounds for privately releasing ℓ-way marginals.

Keywords: Differential privacy, LP decoding

* Supported by NSF grant CCF-1017403, CCF-1118083. Most of the work was done while the author was an intern at Microsoft Research, Silicon Valley.

1 Introduction

This is a paper about private data analysis, in which a trusted curator holding a confidential database responds to real vector-valued queries. Specifically, we focus on the practice of ensuring privacy for the database elements by adding appropriately generated random noise to the answers, releasing only these *noisy* responses. A line of study initiated by Dinur and Nissim examines the amount of distortion needed to prevent privacy violations of various kinds [7]. Dinur and Nissim did not have a definition of privacy; rather, they had a notion that has come to be called *blatant non-privacy*; the modest goal, then, was to add enough distortion to avert blatant non-privacy. Since that time, the community has raised the bar by definining (and achieving) powerful and comprehensive notions of privacy [7, 9, 8], and the goal has been to preserve $(\varepsilon, 0)$-differential privacy and its relaxation, (ε, δ)-differential privacy. A final goal considered herein, *attribute privacy*, has a more complicated description, but may be thought of as preventing blatant non-privacy for a single data attribute [12] in the presence of a certain kind of contingency table query.

The results in the literature vary according to several parameters, including the number n of elements in the database, the size d of the universe from which data elements are drawn, the "amount" and type of privacy desired, and for the purposes of the current work, the arity k of the query. In this paper we strengthen and unify these bounds.

As corollaries of our work, we obtain several "structural" results regarding different types of privacy guarantees:

- We separate so-called *counting* queries from arbitrary *low-sensitivity* queries, proving the latter requires more noise, or distortion, than does the former;
- We separate $(\varepsilon, 0)$-differential privacy from its well-studied relaxation (ε, δ)-differential privacy, even when $\delta \in 2^{-o(n)}$ is negligible in the size n of the database, proving the latter requires less distortion than the former;
- We demonstrate that (ε, δ)-differential privacy is much weaker than $(\varepsilon, 0)$-differential privacy in terms of mutual information of the transcript of the mechanism with the database even when $\delta \in 2^{-o(n)}$ is negligible in the size n of the database.

We also simplify the lower bounds on noise for counting queries in [11] and also make them unconditional removing a technical assumption on the mechanism present in their paper. Next, we use a characterization of (ϵ, δ) differential privacy from [13] to obtain lower bounds on the distortion needed to ensure (ε, δ)-differential privacy for $\epsilon, \delta > 0$. We remark that [12] also obtain quantitatively similar lower bounds on the distortion required to maintain (ϵ, δ) differential privacy for the class of ℓ-way marginals though their proof technique is very different and arguably much more complicated.

After this, we use results of Rudelson [15] and combine it with LP decoding to show that attribute privacy is violated if ℓ-way marginals are released with at least $1 - \eta$ fraction of these marginals are released with $o(\sqrt{n})$ noise for some $\eta > 0$. The results and the technique in [12] required $\eta = 0$ making our results more powerful. Finally, we extend the results of [7] to the case of small universe size achieving stronger lower bounds to prevent blatant non-privacy.

To describe our results even at a high level we must outline the privacy-preserving database model, the notion of *distortion* or *noise* that may be employed in order to preserve privacy, and the meaning of the goals of the adversary: blatant non-privacy, violation of $(\varepsilon, 0)$-differential privacy, violation of (ε, δ)- differential privacy, and attribute non-privacy.

Typically, the curator of a database receives questions to which it responds with potentially noisy answers. There are two possible settings here. One is that the queries are received by the curator one at a time. The other situation is that all the queries are received by the curator at once and it then publishes (noisy) answers to all of them at once. The former is called the interactive setting and the latter is called the non-interactive setting. All our lower bounds are in the non-interactive setting making them applicable to the interactive setting as well.

We now formally describe a database and a query : A database X is an element of $(\mathbb{Z}^+)^d$. Here d is called the universe size and intuitively refers to the number of types of elements present in the database. Also, for a database X, $n = \sum_{i=1}^{d} X_i$ is defined as the size of the database and refers to the number of elements in the database. Note that we are representing databases as histograms. A query (of arity k) is a map $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ such that $\forall i \in [k]$, $\forall x, y \in (\mathbb{Z}^+)^d$, $|F(x+y)_i - F(x)_i| \leq 1$ if $\|y\|_1 = 1$. In other words, every coordinate of the map F is 1-Lipschitz. We say F is a counting query if F is a linear map. The meaning of d, k, n throughout the paper shall be the same as above unless mentioned otherwise.

We now formally introduce the definition of mechanism and privacy.

Definition 1. *Let $\mathcal{F}$ be a family of queries such that $\forall F \in \mathcal{F}$, $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$. Then, a mechanism $M : (\mathbb{Z}^+)^d \times \mathcal{F} \to \mu(\mathbb{R}^k)$ where $\mu(\mathbb{R}^k)$ is simply the set of probability distributions over $\mathbb{R}^k$. On being given a query $F \in \mathcal{F}$ and a database $x \in (\mathbb{Z}^+)^d$, the curator samples z from the probability distribution $M(x, F)$ and returns z.*

We next state the definition of ϵ-differential privacy (introduced by Dwork *et al.* in [9]) and (ϵ, δ)-differential privacy (introduced by Dwork *et al.* in [8]).

Definition 2. *For a family of queries $\mathcal{F}$, a mechanism $M : (\mathbb{Z}^+)^d \times \mathcal{F} \to \mu(\mathbb{R}^k)$ is said to be ϵ-differentially private if for every $x, y \in (\mathbb{Z}^+)^d$ such that $\|x - y\|_1 \leq 1$, every measurable set $S \subseteq \mathbb{R}^k$ and $\forall F \in \mathcal{F}$, the following holds : Let $M(x, F) = M_{x,F}$ and $M(y, F) = M_{y,F}$ and for a probability distribution Γ, let $\Gamma(S)$ denote the probability of set S under Γ. Then,*

$$2^{-\epsilon} \leq \frac{M_{x,F}(S)}{M_{y,F}(S)} \leq 2^{\epsilon}$$

The mechanism is said to be (ϵ, δ)-differentially private if

$$2^{-\epsilon} \cdot M_{y,F}(S) - \delta \leq M_{x,F}(S) \leq 2^{\epsilon} \cdot M_{y,F}(S) + \delta$$

Typically, δ is set to be negligible in n, k.

We remark that we do not define the notion of noise very precisely here as the notion of noise depends on the context. However, in the context of differential privacy, we use the following definition of noise.

Definition 3. *For a family of queries $\mathcal{F}$, a mechanism $M : (\mathbb{Z}^+)^d \times \mathcal{F} \to \mu(\mathbb{R}^k)$ is said to add noise (at most) η if with high probability (say* 0.99*) over the randomness of M, $\|M(x,F) - F(x)\|_\infty \leq \eta$.*

While differential privacy is a very strong notion of privacy, sometimes one can show that even very modest definitions of privacy get violated. One such notion is that of blatant non-privacy. We say that a mechanism M for answering F over databases of size n and universe size d is blatantly non-private, if there is an attack A such that w.h.p. over the answer y returned by the mechanism M, $A(y)$ differs from the database only at $o(1)$ fraction of the places. Yet another very weak notion of privacy that is interesting to us is that of attribute non-privacy. The formal definition follows :

Definition 4. *For a query $F \in \mathcal{F}$, a mechanism $M : (\{0,1\}^d)^n \times \mathcal{F} \to \mathbb{R}^k$ is said to be attribute non-private if there exists $Y \in (\{0,1\}^{d-1})^n$ and an algorithm A such that for every $x \in \{0,1\}^n$,*

$$\Pr_{z \in M(Y \circ x, F)}[A(z) = x' : \|x - x'\|_1 = o(\|x\|_1)] \geq 1/10$$

where $Y \circ x$ simply denotes the obvious concatenation of Y and x. A need not be computationally efficient and the constant 1/10 *is arbitrary and can be replaced by any positive constant.*

We show the following results :

1. Combining techniques from [11] and [13], we obtain tight lower bounds on the noise for arbitrary (non-counting) low-sensitivity queries for any $(\varepsilon, 0)$-differentially private mechanism. Given positive results of Blum, Ligett, and Roth [3], this separates non-counting queries from counting queries, proving that the former require more distortion than the latter for maintaining differential privacy. Also, given the positive results of [8] for arbitrary low-sensitivity queries, this separates (ε, δ)-differential privacy from $(\varepsilon, 0)$-differential privacy, where $\delta = \delta(n, k)$ denotes a function negligible in its argument. We also use this technique to show that the guarantees in terms of information content is drastically weaker for an (ϵ, δ) differentially private protocol as compared to an ϵ-differentially private protocol. Our technique also simplifies the *volume-based* lower bounds on noise for counting queries in [11]. In addition, we also make the lower bounds unconditional. The lower bound in [11] required the mechanism to be defined on "fractional" databases *i.e.,* on $(\mathbb{R}^+)^d$ as opposed to just $(\mathbb{Z}^+)^d$ while we do not have any such restrictions.
2. We give tight lower bounds on noise for ensuring (ε, δ)-differential privacy for $\delta > 0$. This proof relies on a lemma due to [13] showing that (ε, δ)-differentially private mechanisms yield a certain kind of unpredictable source.

On the other hand, any mechanism that is blatantly non-private cannot yield an unpredictable source. Thus, if the noise is insufficient to prevent blatant non-privacy then it cannot provide (ε, δ)-differential privacy. We subsequently use the lower bounds of [7, 10] for preventing blatant non-privacy to get lower bounds on the distortion for (ϵ, δ) differential privacy.

3. We revisit the LP decoding attack of Dwork, McSherry, and Talwar [10], observing that any linear query matrix yielding a Euclidean section suffices for the attack. The LP decoding attack succeeds even if a certain constant fraction of the responses have wild noise. Armed with the connection to Euclidean sections, and a recent result of Rudelson [15] bounding from below the least singular value of the Hadamard product of certain i.i.d. matrices, we qualitatively strengthen a lower bound of Kasiviswanathan, Rudelson, Smith, and Ullman [12] on the noise needed to avert attribute non-privacy in ℓ-way marginals release by making the attack resilient to a constant fraction of wild responses.

There is an extension of results of [7] when the size of the universe is smaller than the size of the database which can be found in the full version of this paper [5].

2 Lower bound by volume arguments

We now recall the volume based argument of Hardt and Talwar [11] to show lower bounds on the noise required for ϵ differential privacy.

Theorem 1. *Assume $x_1, \ldots, x_{2^s} \in (\mathbb{Z}^+)^d$ such that $\forall i$, $\|x_i\|_1 \leq n$ and for $i \neq j$, $\|x_i - x_j\|_1 \leq \Delta$. Further, let $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ such that for any $i \neq j$, $\|F(x_i) - F(x_j)\|_\infty \geq \eta$. If $\Delta \leq (s-1)/\epsilon$, then any mechanism which is ϵ-differentially private for the query F on databases of size n must add noise $\eta/2$.*

While the line of reasoning in the proof is same as that of [11], we do the proof here as the argument in [11] works only for counting queries *i.e.,* when F is a linear transformation. On the other hand, the statement and proof of our result works for any query F.

Proof. Consider the ℓ_∞ balls of radius $\eta/2$ around each of the $F(x_i)$. By the hypothesis, these balls are disjoint. Now assume, any mechanism M which adds noise $\eta/2$ and consider any x_i. Then, because all the balls are disjoint, we have that there is some $j \neq i$ such that if S is the ℓ_∞ ball of radius $\eta/2$ around $F(x_j)$, then

$$\Pr_{z \in M(x_i, F)}[z \in S] \leq 2^{-s}$$

However, we can also say that because the noise added by the mechanism M is at most η,

$$\Pr_{z \in M(x_j, F)}[z \in S] \geq 1/2$$

Also, because the mechanism M is ϵ-differentially private and $\|x_i - x_j\|_1 \leq \Delta$, then

$$\frac{\Pr_{z \in M(x_i,F)}[z \in S]}{\Pr_{z \in M(x_j,F)}[z \in S]} \geq 2^{-\epsilon \cdot \Delta}$$

This leads to a contradiction if $\Delta \leq (s-1)/\epsilon$ thus proving the assertion.

2.1 Linear lower bound for arbitrary queries

In this subsection, we prove the following theorem.

Theorem 2. *For any $k, d, n \in \mathbb{N}$ and $1/40 \geq \epsilon > 0$, where $n \geq \min\{k/\epsilon, d/\epsilon\}$, there is a query $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ such that any mechanism M which is ϵ-differentially private adds noise $\Omega(\min\{d/\epsilon, k/\epsilon\})$.*

If $\epsilon > 1$, then there is a query $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ such that any mechanism M which is ϵ-differentially private adds noise $\Omega(\min\{d/(\epsilon \cdot 2^{5\epsilon}), k/\epsilon\})$ as long as $n \geq \min\{k/\epsilon, d/(\epsilon \cdot 2^{5\epsilon})\}$

Before starting the proof, we make a couple of observations. First of all, note that the statement of the theorem does not give any lower bound for $1 \geq \epsilon > 1/40$. However, any mechanism which is ϵ-differentially private for ϵ in the aforementioned range is also ϵ'-differentially private for $\epsilon' = 10/9$. Hence, the noise lower bounds for ϵ'-differential privacy for $\epsilon' = 10/9$ are also applicable for the range of $1 \geq \epsilon > 1/40$. It is easy to see that up to constant factors, the lower bounds with $\epsilon' = 10/9$ are optimal for ϵ in the aforementioned range.

Secondly, Laplacian mechanism maintains ϵ-differential privacy while adding only $O(k/\epsilon)$ noise. Also, because the databases are of size n, it is enough to add noise $O(n)$ to maintain ϵ-differential privacy for any $\epsilon \geq 0$. Thus, as long as $k = O(d)$, our lower bounds are tight up to constant factors. Next, we do the proof of Theorem 2.

Also, in the subsequent proofs, the databases shall be constructed in clever ways. The full details of these constructions can be found in [5]. We will be referring to the appropriate claims whenever necessary.

Proof. Our proof strategy is to construct a set of databases and a query which meets the conditions stated in the hypothesis of Theorem 1 and then get the desired lower bound on the noise. We first deal with the case when $0 < \epsilon < 1/40$. Let $\ell = \min\{d, k\}$. We can now use Claim A.2 in [5] to construct 2^s databases $x_1, \ldots, x_{2^s}$ (for $s = \ell/400$) such that $x_i \in (\mathbb{Z}^+)^d$ with the property that $\forall i \neq j$, $\|x_i - x_j\|_1 \geq n'/10$ and $\|x_i\|_1 \leq n'$ where $n' = \ell/(1280\epsilon)$ (Application of Claim A.2 uses $d' = \ell/320$). Note that our databases are of size bounded by $n' \leq n$. We now describe a mapping $\mathcal{L} : (\mathbb{Z}^+)^d \to \mathbb{R}^{2^s}$ which is related to a construction in [13]. The mapping is as follows :

- For every x_i, there is a coordinate i in the mapping.
- The i^{th} coordinate of $\mathcal{L}(z)$ is $\max\{n'/30 - \|x_i - z\|_1, 0\}$.

Claim. The map $\mathcal{L}$ is 1-Lipschitz *i.e.,* if $\|z_1 - z_2\|_1 = 1$, then $\|\mathcal{L}(z_1) - \mathcal{L}(z_2)\|_1 \leq 1$.

Proof. We observe that for any z_1, z_2 such that $\|z_1 - z_2\| \leq 1$, if A denotes the set of coordinates where at least one of $\mathcal{L}(z_1)$ or $\mathcal{L}(z_2)$ are non-zero, then A is either empty or is a singleton set. Given this, the statement in the claim is obvious, since the mapping corresponding to any particular coordinate is clearly 1-Lipschitz.

We now describe the queries. Corresponding to any $r \in \{-1,1\}^{2^s}$, we define $f_r : (\mathbb{Z}^+)^d \to \mathbb{R}$, as

$$f_r(x) = \sum_{i=1}^{d} \mathcal{L}(x)_i \cdot r_i$$

Now, we define a random map $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ as follows. Pick $r_1, \ldots, r_k \in \{-1,1\}^{2^s}$ independently and uniformly at random and define F as follows :

$$F(x) = (f_{r_1}(x), \ldots, f_{r_k}(x))$$

Now consider any $x_h, x_j \in S$ such that $h \neq j$. Because of the way $\mathcal{L}$ is defined, it is clear that for any r_i,

$$\Pr_{r_i}[|f_{r_i}(x_h) - f_{r_i}(x_j)| \geq n'/15] \geq 1/2$$

A basic application of the Chernoff bound implies that

$$\Pr_{r_1,\ldots,r_k}[\text{For at least } 1/10 \text{ of the } r_i\text{'s}, \quad |f_{r_i}(x_h) - f_{r_i}(x_j)| \geq n'/15] \geq 1 - 2^{-k/30}$$

Now, note that the total number of pairs (x_i, x_j) of databases such that $x_i, x_j \in S$ is at most $2^{2s} \leq 2^{\ell/200} \leq 2^{k/200}$. This implies (via a union bound)

$$\Pr_{r_1,\ldots,r_k}[\forall h \neq j, \geq 1/10 \text{ of the } r_i\text{'s}, \quad |f_{r_i}(x_h) - f_{r_i}(x_j)| \geq n'/15] \geq 1 - 2^{-k/40}$$

This implies that we can fix $r_1, \ldots, r_k$ such that the following is true.

$$\forall h \neq j, \quad \text{For at least } 1/10 \text{ of the } r_i\text{'s}, \quad |f_{r_i}(x_h) - f_{r_i}(x_j)| \geq n'/15$$

This implies that for any $x_h \neq x_j \in S$, $\|F(x_h) - F(x_j)\|_\infty \geq n'/15$. In fact, $\|F(x_h) - F(x_j)\|_2 \geq n'\sqrt{k}/150$ which is a much stronger assumption than what we require and is quantitatively similar to the results in [11] where they consider ℓ_2 noise as opposed to ℓ_∞ noise.

We can now apply Theorem 1 by putting $\Delta = 2n'$ and $s = \ell/400 > 3\epsilon n'$ and $\eta = n'/15$ and observe that $\Delta \leq (s-1)/\epsilon$ thus proving the result.

We next deal with the case when $\epsilon > 1$. This part of the proof differs from the case when $\epsilon < 1$ only in the construction of $x_1, \ldots, x_{2^s}$. We also emphasize that had we not insisted on integral databases, our proof would have been identical to the first part. We construct the databases $x_1, \ldots, x_{2^s}$ using combinatorial designs. More precisely, for some sufficiently large constant C, let $\ell = \min\{d/(C \cdot 2^{5\epsilon}), k\}$. We can now use Claim A.3 from [5] to construct 2^s databases $x_1, \ldots, x_{2^s}$ (for $s = \ell/400$) such that $x_i \in (\mathbb{Z}^+)^d$ with the property that $\forall i \neq j$, $\|x_i - x_j\|_1 \geq$

$n'/10$ and $\|x_i\|_1 \le n'$ where $n' = \ell/(1280\epsilon)$ (using $d' = \ell/320$ in Claim A.3). Again, we note here that the databases constructed are of size n'.

From this point onwards, we define the map $\mathcal{L}$ and the query F as we did in the proof of Theorem 2 and the proof proceeds identically. In particular, we get a query $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ such that for any $i \neq j$, $\|F(x_i) - F(x_j)\|_2 \ge n'\sqrt{k}/150$. As before, we can now apply Theorem 1 by putting $\varDelta = 2n'$ and $s = \ell/100 > 3\epsilon n'$ and $\eta = n'/15$ and observe that $\varDelta \le (s-1)/\epsilon$ thus proving the result

For the subsequent part of this paper, we only consider lower bounds on ϵ-differential privacy for $0 < \epsilon < 1$ as opposed to $\epsilon > 1$. This is because the privacy guarantees one gets becomes unmeaningful when ϵ is large. However, we do remark that the results can be carried in a straightforward way to the regime of $\epsilon > 1$ using combinatorial designs (like we did for Theorem 2).

Consequences of the linear lower bound We briefly describe the two consequences of the linear lower bound on the noise proven in Theorem 2. The first is separation of counting queries from non-counting queries. While our separation gives quantitatively the same results as long as $d = k^{O(1)}$ and $n = \varTheta(k/\epsilon)$, for simplicity, we consider the setting when $k = d$ and $n = k/\epsilon$. In this case, Theorem 2 shows existence of a (non-counting) query such that maintaining ϵ-differential privacy requires noise $\varOmega(n)$. On the other hand, [3] had proven that for any counting query with the same setting of parameters, there is a mechanism which adds noise $\tilde{O}(n^{2/3})$ and maintains ϵ-differential privacy. This shows that maintaining ϵ-differential privacy inherently requires more distortion in case of non-counting queries than counting queries.

The next consequence is a separation of (ϵ, δ) differential privacy from $(\epsilon, 0)$ differential privacy for $\delta = 2^{-o(n)}$. We note that Hardt and Talwar [11] had shown such a separation but that was only when $k = O(\log n)$ and $\delta = n^{-O(1)}$. Again, we use the setting of parameters when $k = d$ and $n = k/\epsilon$. The gaussian mechanism of [8] shows that to maintain (ϵ, δ) differential privacy for any k queries, it sufficies to add noise $O(\sqrt{k \log(1/\delta)}/\epsilon) = o(n)$. However, Theorem 2 shows that there is a query which requires adding noise $\varOmega(n)$ to maintain $(\epsilon, 0)$ differential privacy.

The last consequence of our result is more indirect and is explained next.

2.2 Information loss in differentially private protocols

In [13], a connection was established between differentially private protocols and the notion of mutual information from information theory. In fact, as [13] was dealing with 2-party protocols, the connection was actually between differentially private protocols and that of information content [1, 2] which is a symmetric variant of mutual information useful in 2-party protocols. In that paper, it was shown that the information content (which simplifies to mutual information in our setting) between transcript of a ϵ-differentially private mechanism and the database vector is bounded by $O(\epsilon n)$. Using the construction used in the previous subsection, we show that in case of (ϵ, δ) differentially private protocols (for any $\delta = 2^{-o(n)}$), there is no non-trivial bound on the mutual information between the transcript of the mechanism and the database vector. Thus as far as information theoretic guarantees go, the situation is drastically different for pure differentially private protocols vis-a-vis approximately differentially private protocols. The contents of this subsection are a result of personal communication between the author and Salil Vadhan [6].

We first define the notion of mutual information (can be found in standard information theory textbooks).

Definition 5. *Given two random variables X and Y, their mutual information $I(X;Y)$ is defined as*

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y)$$

where $H(X)$ denotes the Shannon entropy of X.

The next claim establishes an upper bound on the mutual information between transcript of a differentially private protocol and the database vector.

Claim. Let $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ be a query and $M : (\mathbb{Z}^+)^d \to \mu(\mathbb{R}^k)$ be an ϵ-differentially private protocol for answering F for databases of size n. If X is a distribution over the inputs in $(\mathbb{Z}^+)^d$, then $I(M(X);X) \leq 3\epsilon n$.

Proof. We first note that since the databases are of size bounded by n, hence instead of assuming that μ is a distribution over the inputs $X \in (\mathbb{Z}^+)^d$, we can assume that μ is a distribution over the inputs $X \in [n]^d$ where $[n] = \{0, 1, \ldots, n\}$. Now, we can apply Proposition 7 from [13]. We note that the aforesaid proposition is in terms of information content for 2-party protocols but we observe that we can simply make the second party's input as a constant and get that $I(M(X);X) \leq 3\epsilon n$.

Next, we state the following claim which says that for (ϵ, δ) differentially private protocols, even for an exponentially small δ, the mutual information between the transcript and the input can be as large as $n(1-\eta)$ for any value of $0 < \epsilon, \eta < 1$. In other words, an (ϵ, δ) differentially private protocol does not imply any effective bound on the mutual information between the input and the transcript even as $\epsilon \to 0$ and δ is exponentially small.

Lemma 1. *For $n \in \mathbb{N}$ and $0 < \epsilon, \eta < 1$, there is a constant $C = C(\epsilon, \eta) > 0$ and a distribution X over $(\mathbb{Z}^+)^n$ with a support over databases of size n and a query $F : (\mathbb{Z}^+)^n \to \mathbb{R}^k$ and an (ϵ, δ)-differentially private protocol M for answering F such that $I(X; M(X)) \geq n(1 - 2\eta)$ if $\delta \geq 2^{-C(\epsilon,\eta)n}$.*

Proof. We first construct 2^s vectors in $\{0, 1\}^n$ (for $s = n(1 - \eta)$) with the property that for any x_i, x_j $(i \neq j)$, $\|x_i - x_j\|_1 \geq \eta^2 n/8$. It is easy to guarantee the existence of such a set of vectors by a simple application of the probabilistic method. The distribution X is simply the uniform distribution over the set $\{x_1, \ldots, x_{2^s}\}$. By construction, all the databases in X are of size bounded by n.

Next, we define the query $F : (\mathbb{Z}^+)^n \to \mathbb{R}^k$ be defined in the same way as the query F in the proof of Theorem 2. Following, exactly the same calculations, we can show that if we set $k = 80n$, we get a query $F : (\mathbb{Z}^+)^n \to \mathbb{R}^k$ such that for any $i \neq j$, $\|F(x_i) - F(x_j)\|_2 \geq \eta^2 n\sqrt{k}/50$. We now recall the Gaussian mechanism of [8] which maintains (ϵ, δ) differential privacy.

Lemma 2. *[8] Let $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ be a query. Let $Y = (Y_1, \ldots, Y_k)$ be a distribution over $\mathbb{R}^k$ such that each Y_i is an i.i.d. $\mathcal{N}(0, \sigma)$ random variable. Here $\sigma^2 = \frac{k \log(1/\delta)}{\epsilon^2}$. Then the mechanism M which for a database x and query F, which samples Y_0 from Y and responds by $F(x) + Y_0$ is an (ϵ, δ) differentially private mechanism.*

Note that for the above mechanism M, and database x, if Z is sampled from $M(x)$, then the distribution of $M(x) - F(x)$ is same as $(Y_1, \ldots, Y_k)$ where each Y_i is an i.i.d. $\mathcal{N}(0, \sigma)$ random variable. Thus,

$$\|M(x) - F(x)\|_2^2 \sim Y_1^2 + \ldots + Y_k^2$$

As the following fact shows, the distribution on the right hand side is concentrated around its mean. The fact is possibly well-known but we could not find a reference and hence we prove it in Appendix C in [5].

Fact 3 *If $Y_1, \ldots, Y_k$ are i.i.d. $\mathcal{N}(0, \sigma)$ random variables, then,*

$$\Pr_{Y_1, \ldots, Y_k}[Y_1^2 + \ldots + Y_k^2 > 2(1+\xi) \cdot k \cdot \sigma^2] \leq 2^{-\frac{k\xi}{2}}$$

Using the above fact, we get

$$\Pr\left[\|M(x) - F(x)\|_2^2 > \frac{2(1+\xi)k^2 \log(1/\delta)}{\epsilon^2}\right] \leq 2^{\frac{-\xi k}{2}}$$

Here the probability is over the randomness of the mechanism. Putting $\xi = 1$ and $\delta = 2^{-C(\epsilon,\eta)n}$ for an appropriate constant $C(\epsilon, \eta)$, we get that

$$\Pr\left[\|M(x) - F(x)\|_2 > \frac{\eta^2 n\sqrt{k}}{200}\right] \leq 2^{-40n}$$

As we know, for any $i \neq j$, $\|F(x_i) - F(x_j)\|_2 \geq \eta^2 n\sqrt{k}/50$. Hence, with probability at least $1 - 2^{-n}$ over the randomness of the mechanism, for any database $x_i \in supp(X)$, if y is sampled from $M(x_i)$,

$$\forall j \neq i \quad \|F(x_j) - y\|_2 > \|F(x_i) - y\|_2$$

Thus, for any x_i, given $M(x_i)$, we can recover x_i with high probability and hence, we can say

$$\Pr_{y \sim M(X)}[H(X|M(X) = y) = 0] > 1 - 2^{-n}$$

This means that

$$H(X|M(X)) \leq 2^{-n} n < 1$$

Recall that $I(X; M(X)) = H(X) - H(X|M(X)) \geq H(X) - 1 = (1-\eta)n - 1 \geq (1-2\eta)n$. This completes the proof of the Lemma 1.

3 Lower bound on noise for counting queries

In the last section, we proved that to preserve ϵ differential privacy for k queries, one may need to add $\Omega(k/\epsilon)$ noise provided $d, n \gg k$. However, these queries were not counting queries. It is interesting to derive lower bounds on noise required to preserve privacy for counting queries as these are the queries mostly used in practice. While one might initially hope to prove a similar lower bound for counting queries, [3] states that there is a ϵ-differentially private mechanism which adds $\tilde{O}(n^{2/3}/\epsilon)$ noise per query and can answer $O(n)$ counting queries (when $d = n^{O(1)}$).

Still, Hardt and Talwar [11] showed that to answer k counting queries, any mechanism which is ϵ-differentially private must add $\min\{k/\epsilon, \sqrt{k \log(d/k)}/\epsilon\}$ noise (in fact, this is true for k random queries). However, [11] make a technical assumption that the mechanism has a smooth extension which works for "fractional" databases as well. In other words, they require the domain of the mechanism to be $(\mathbb{R}^+)^d$ as opposed to $(\mathbb{Z}^+)^d$. However, it is not clear if this is always true *i.e.,* if given a mechanism which is defined only over true (integral) databases, one can get a mechanism which is defined over "fractional" databases with similar privacy guarantees.

Next, we prove the same result without making any such technical assumptions. Again, our constructions are dependent on combinatorial designs [14]. First, we prove the following simple but useful claim.

Claim. Let $a \in \mathbb{Z}$ and assume $x_1, x_2, \ldots, x_{2^s} \in (\mathbb{Z}^+)^d$ such that $\forall i$, every entry of x_i is either 0 or a. Also, for every $i \neq \ell$, $\|x_i - x_\ell\|_1 \geq \Delta$. Then, for $k \geq 20s$, there is a linear query $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ such that for every $i, \ell \in [2^s]$ and $i \neq \ell$, the following holds :

$$\Pr_{j \in [k]}[|F(x_i)_j - F(x_\ell)_j| \geq \Delta'/10] \geq 1/40$$

where $\Delta' = \sqrt{\Delta \cdot a}$.

Proof. Consider any x_i, x_ℓ such that $i \neq \ell$. Note that, z defined as $z = x_i - x_\ell$ is such that all its entries are $0, \pm a$ and also that z has at least Δ/a or more non-zero entries. If we choose $r \in \{-1,1\}^d$ u.a.r., then note that

$$Y = \sum_{i=1}^{d} z_i \cdot r_i = \sum_{z_i = \pm a} z_i \cdot r_i$$

Note that the total number of summands is $\ell' \geq \Delta/a$ and hence the distribution of the random variable Y is same as choosing $r' \in \{-1,1\}^d$ and considering the random variable

$$Y' = a \cdot \left(\sum_{i=1}^{\ell'} r'_i \right)$$

However using Corollary B.2 from [5], we get

$$\Pr\left[|Y'| \geq \frac{\sqrt{\Delta \cdot a}}{10}\right] = \Pr\left[|\sum_{i=1}^{\ell'} r'_i| \geq \frac{\sqrt{\Delta/a}}{10}\right] \geq \frac{9}{10} \tag{1}$$

Now, let us choose $r'_1, \ldots, r'_k$ uniformly and independently at random from $\{-1,1\}^d$ and consider the linear query $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ defined as

$$F(x) = \left(\sum_{j=1}^{d} x_j \cdot r'_{1j}, \ldots, \sum_{j=1}^{d} x_j \cdot r'_{kj}\right)$$

Set $\Delta' = \sqrt{\Delta \cdot a}$. Now, (1) and an application of Chernoff bound implies that for any x_i, x_ℓ $(i \neq \ell)$

$$\Pr_{r'_1,\ldots,r'_k}\left[\Pr_{j \in [k]}[|F(x_i)_j - F(x_\ell)_j| \geq \Delta'/10] \geq 1/40\right] > 1 - 2^{-k/10}$$

We now observe that the total number of pairs (x_i, x_ℓ) $(i \neq \ell)$ is at most $2^{2s} \leq 2^{k/10}$. Applying a union bound, we get that there is some choice of $r'_1, \ldots, r'_k$ (and hence a fixed F) such that

$$\Pr_{j \in [k]}[|F(x_i)_j - F(x_\ell)_j| \geq \Delta'/10] \geq 1/40$$

We now prove a lower bound on the noise required to maintain privacy for random counting queries. As we have said before, Hardt and Talwar [11] proved the same result under an additional assumption that the mechanism defined over integral databases can be smoothly extended to fractional databases as well.

Theorem 4. *For every $k, d \in \mathbb{N}$ and $1 > \epsilon > 0$, there is a counting query $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ such that any mechanism which maintains ϵ-differential privacy adds noise $\Omega(\min\{k/\epsilon, \sqrt{k \log(d/k)}/\epsilon\})$. The size of the database i.e., $n = O(k/\epsilon)$.*

Proof. The proof strategy is to come up with databases meeting the hypothesis of Claim 3 and use Claim 3 to get a counting query F. We then use Theorem 1 to get a lower bound on the distortion required by any private mechanism to answer F. We consider two cases : $k \leq \log d$ and $k > \log d$.

The first case is trivial : Namely, consider databases $x_1, \ldots, x_{2^{k/20}}$ such that each $x_i = \lfloor (k/80\epsilon) \rfloor \cdot e_i$ where e_i is the standard unit vector in the i^{th} direction. This is possible as there are $d \geq 2^k$ different unit vectors. Note that for any $i \neq \ell$, $\|x_i - x_\ell\|_1 = 2 \cdot \lfloor k/(80\epsilon) \rfloor$. We can now apply Claim 3 and get that there is a linear query $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ (using $\Delta = 2 \cdot \lfloor k/(80\epsilon) \rfloor$ and $a = \lfloor k/(80\epsilon) \rfloor$) such that

$$\Pr_{j \in [k]}\left[|F(x_i)_j - F(x_\ell)_j| \geq \frac{\sqrt{2}}{10}\lfloor k/(80\epsilon) \rfloor \geq \frac{k}{800\epsilon}\right] \geq 1/40$$

We see that there are $2^{k/20} = 2^s$ databases which differ by exactly $2 \cdot \lfloor k/(80\epsilon) \rfloor = \Delta$. Note that $\Delta \leq (s-1)/\epsilon$. Hence we can apply Theorem 1 to note that to maintain ϵ-differential privacy, any mechanism needs to add $k/(800\epsilon)$ noise. In fact, we note that the ℓ_2 error of the answer returned by the mechanism needs to be $\Omega(k^{3/2}/\epsilon)$ which is quantitatively the same as the result in [11].

The second case is slightly more complicated. We use Claim A.1 from [5] to construct $x_1, \ldots, x_{2^{k/20}} \in (\mathbb{Z}^+)^d$ with the following properties :

- Every entry of any of the x_i's is either 0 or $a \in \mathbb{Z}$ such that $a \geq \log(d/k)/160\epsilon$.
- $\forall i$, $\|x_i\|_1 \leq k/80\epsilon$ and $\forall i \neq j$, $\|x_i - x_j\|_1 \geq k/160\epsilon$

Again, we can apply Claim 3 and get that there is a linear query $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ (using $\Delta \geq k/(160\epsilon)$ and $a \geq (\log(d/k)/160\epsilon)$) such that $\forall i \neq \ell$

$$\Pr_{j \in [k]} \left[|F(x_i)_j - F(x_\ell)_j| \geq \frac{1}{10} \cdot \frac{\sqrt{k \log(d/k)}}{160\epsilon} \right] \geq 1/40$$

Again, we have $2^{k/20}$ databases which differ by at most $k/(40\epsilon)$ and hence we can apply Theorem 1 to get that to maintain ϵ-differential privacy, any mechanism needs to add $\Omega\left(\frac{\sqrt{k \log(d/k)}}{\epsilon}\right)$ noise.

4 Lower bounds for approximate differential privacy

In this section, we prove lower bounds on the noise required to maintain (ϵ, δ) differential privacy for $\epsilon, \delta > 0$. Our lower bounds are valid for any positive $\delta > 0$ and are in fact tight for a constant ϵ and δ. We note that a quantitatively similar lower bound was proven for the class of ℓ-way marginals by [12] though our proof (for random queries) is arguably much simpler.

In this section, we consider databases which are elements of $\{0,1\}^n$ or in other words we consider the case when the universe size $d = n$ and the databases are allowed to have exactly one element of each type. We note that restricting databases to bit vectors is a well-considered model in literature including [7, 10, 13] among others.

We prove the following theorem.

Theorem 5. *For any $n \in \mathbb{N}$, $\epsilon > 0$ and $1/20 > \delta > 0$, there exist positive constants α, γ and η such that there is a counting query $F : \{0,1\}^n \to \mathbb{R}^k$ with $k = \alpha n$ such that any mechanism M that satisfies*

$$\Pr_M[\Pr_{i \in [k]}[|M(x,F)_i - F(x)_i| \leq \eta\sqrt{n}] \geq 1/2 + \gamma] \geq 3\sqrt{\delta}$$

is not (ϵ, δ) differentially private. In other words, any mechanism M which with significant probability i.e., $3\sqrt{\delta}$ answers at least $1/2 + \gamma$ fraction of the k queries with at most $\eta\sqrt{n}$ noise, is not (ϵ, δ) differentially private.

An immediate corollary is that there exists a positive constant α and a counting query $F : \{0,1\}^n \to \mathbb{R}^k$ where $k = \alpha n$ such that any mechanism which adds $o(\sqrt{n})$ noise is not (ϵ, δ) differentially private for $\epsilon > 0$ and $\delta < 1/20$.

To do the proof of Theorem 5, we first need to introduce some definitions previously discussed in [13]. We do note that the paper [13] deals with the two-party setting but the relevant definitions and the lemma we use here easily extend to the standard (curator-client) setting of privacy.

Definition 6. *A random variable $Y = (y_1, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots, y_n) \in \{0,1\}^n$ is said to be δ-approximate strongly α-unpredictable bit source (for $\alpha \geq 1$) if with probability $1 - \delta$ over $i \in [n]$*

$$\frac{1}{\alpha} \leq \frac{\Pr[Y_i = 1 | Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}, Y_{i+1} = y_{i+1}, \ldots, Y_n = y_n]}{\Pr[Y_i = 0 | Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}, Y_{i+1} = y_{i+1}, \ldots, Y_n = y_n]} \leq \alpha$$

The next lemma (proven in [13] for the two-party setting) roughly says that for any (ϵ, δ) private mechanism, conditioned on the transcript of the mechanism, the distribution of the database is a δ-approximate strong 2^{ϵ}-unpredictable source. More precisely, we have the following lemma.

Lemma 3. *Let $F : \{0,1\}^n \to \mathbb{R}^k$ be a query and M be a (ϵ, δ)-differentially private mechanism for answering F. Let X be the uniform distribution over $\{0,1\}^n$ and Γ be the probability distribution over the transcripts of $M(x)$ when x is drawn from X. Then for any $\mu > 0$ and $t \leftarrow \Gamma$, the distribution $X|_{\Gamma = t}$ is δ_t approximate strongly $2^{\epsilon+\mu}$-unpredictable sources such that*

$$\mathop{\mathbb{E}}_{t \in \Gamma} [\delta_t] \leq 2\delta \cdot \frac{1 + e^{-\epsilon-\mu}}{1 - e^{-\mu}}$$

.

The above lemma trivially follows from Lemma 20 of [13] (full version) and hence we do not prove it here. Before, proving Theorem 5, we need to recall the following theorem from [10] (Theorem 24 in the paper).

Theorem 6. *For any $\gamma > 0$ and any $\nu = \nu(n)$, there is a constant $\alpha = \alpha(\gamma) > 0$ such that for $k = \alpha n$, there is a counting query $F : \{0,1\}^n \to \mathbb{R}^k$ and an algorithm A such that given $\tilde{y}$ which satisfies*

$$\Pr_{i \in [k]} [|\tilde{y}_i - F(x)_i| \leq \nu] \geq \frac{1}{2} + \gamma$$

the output of A on $\tilde{y}$ i.e., $A(\tilde{y}) = x'$ such that $x' \in \{0,1\}^n$ and $\|x - x'\|_1 \leq \frac{4\nu^2}{\gamma^2}$

The following corollary follows immediately from Theorem 6.

Corollary 1. *For any $\delta' > 0$, there are positive constants $\gamma = \gamma(\delta'), \eta = \eta(\delta'), \alpha = \alpha(\delta')$ such that for $k = \alpha n$, there is a counting query $F : \{0,1\}^n \to \mathbb{R}^k$ and an algorithm A such that given $\tilde{y}$ which satisfies*

$$\Pr_{i \in [k]} [|\tilde{y}_i - F(x)_i| \leq \eta\sqrt{n}] \geq \frac{1}{2} + \gamma$$

the output of A on $\tilde{y}$ i.e., $A(\tilde{y}) = x'$ such that $x' \in \{0,1\}^n$ and $\|x - x'\|_1 \leq \delta' n$.

We now prove Theorem 5.

Proof (of Theorem 5).

Let X denote the uniform distribution over $\{0,1\}^n$. First, using Lemma 3, we get that over the randomness of the mechanism M and the choice of $x \in X$, if we sample a transcript t from $M(x,F)$, then for any positive μ, the distribution $X|_{M(x,F)=t}$ is a δ_t-approximate strongly $2^{\epsilon+\mu}$-unpredictable sources where δ_t satisfies

$$\mathop{\mathbb{E}}_{t \in M(x,F)} [\delta_t] \leq 2\delta \cdot \frac{1+e^{-\epsilon-\mu}}{1-e^{-\mu}}$$

. Clearly, we can put $\mu = 10$ and get that the distribution $X|_{M(x,F)=t}$ is a δ_t-approximate strongly $2^{\epsilon+10}$-unpredictable sources where $\mathbb{E}_{t \in M(x,F)} [\delta_t] \leq 3\delta$. By an application of Markov's inequality, we get that with probability $1-2\sqrt{\delta}$ over the choice of x and the randomness of the mechanism M, the distribution $X|_{M(x,F)=t}$ is $2\sqrt{\delta}$-approximate strongly $2^{\epsilon+10}$-unpredictable source.

We now apply corollary 1. In particular, we put $\delta' = \sqrt{\delta}$ and get that for some positive γ, η, α (which are functions of δ' and hence δ), there is a counting query $F : \{0,1\}^n \to \mathbb{R}^{\alpha n}$ and an algorithm A such that given $\tilde{y}$ which satisfies

$$\Pr_{i \in [k]}[|\tilde{y}_i - F(x)_i| \leq \eta\sqrt{n}] \geq \frac{1}{2} + \gamma$$

the output of A on $\tilde{y}$ *i.e.,* $A(\tilde{y}) = x'$ such that $x' \in \{0,1\}^n$ and $\|x-x'\|_1 \leq \sqrt{\delta} \cdot n$. Now, consider a mechanism M which satisfies

$$\Pr_M[\Pr_{i \in [k]}[|M(x,F)_i - F(x)_i| \leq \eta\sqrt{n}] \geq 1/2 + \gamma] \geq \beta$$

for $\beta = 3\sqrt{\delta}$. Clearly such a mechanism M is not (ϵ, δ) differentially private because with probability at least $\beta = 3\sqrt{\delta}$, the algorithm A will be able to predict at least $1-\sqrt{\delta}$ fraction of the positions which contradicts that with probability $1-2\sqrt{\delta}$, the distribution $X|_{M(x,F)=t}$ is a $2\sqrt{\delta}$ -approximate strongly $2^{\epsilon+10}$-unpredictable source.

5 LP decoding, Euclidean sections and hardness of releasing ℓ-way marginals

In this section, we consider attacks on privacy using linear programming. In particular, we use the technique of LP decoding (previously used in [10] in context of privacy) to give attacks which violate even minimal notions of privacy when $1-\epsilon_0$ (for some $\epsilon_0 > 0$) fraction of the queries are released with insufficient noise. We do this by establishing a connection between Euclidean sections and use of LP decoding in context of privacy which does not seem to have explicitly appeared in the literature before. We remark that the relation between LP decoding and Euclidean spaces is very well known in context of compressed sensing [4]. However, in case of privacy, the adversary is allowed to add small error to say 99% of the entries and arbitrary error to the remaining 1% of the entries. In context of compressed sensing however, the adversary is allowed to add error to only 1% of the entries.

We first describe how to use linear programming in context of privacy. Assume $x \in {\mathbb{Z}^+}^d$ is a database and $A : \mathbb{R}^d \to \mathbb{R}^k$ is a linear map which represents a counting query with arity k made on the database x. Further, the right set of answers is given by $y = A \cdot x$. (To make sure that the queries are 1-Lipschitz, all the entries of A come from $[-1, 1]$.) Suppose, $\tilde{y} \in \mathbb{R}^k$ is the answer returned by the mechanism. Then, consider the following optimization problem (which can be written as a linear program) :

$$\text{Minimize } \|y - \tilde{y}\|_1 \text{ subject to } y = A \cdot \tilde{x} \tag{2}$$

The following theorem states the necessary conditions such that the solution to the above linear program, call it $\tilde{x}$, is such that $\|x - \tilde{x}\|_1$ is small. To state the theorem, we will need the definition of a Euclidean section.

Definition 7. *$V \subseteq \mathbb{R}^k$ is said to be a (δ, d, k) euclidean section if V is a linear subspace of dimension d and for every $x \in V$, the following holds:*

$$\sqrt{k}\|x\|_2 \geq \|x\|_1 \geq \delta\sqrt{k}\|x\|_2$$

Theorem 7. *Let $A : \mathbb{R}^d \to \mathbb{R}^k$ be a full rank linear map ($k > d$) and all the singular values of A are at least σ. Further, the range of A (denoted by $\mathcal{L}(A)$) is a (δ, d, k) Euclidean section. Let $F : (\mathbb{Z}^+)^d \to \mathbb{R}^k$ the query corresponding to A. Then, there exists $\gamma = \gamma(\delta)$ such that if*

$$\Pr_{i \in [k]}[|F(x)_i - \tilde{y}_i| \leq \alpha] \geq 1 - \gamma$$

then, any solution $\tilde{x}$ to the linear program (2) satisfies $\|\tilde{x} - x\|_1 \leq O(\alpha\sqrt{kd}/\sigma)$ where the constant inside the $O(\cdot)$ notation depends on δ.

The proof of this theorem can be found in [5]. The specific problem we are interested in is the application of LP decoding to violate attribute privacy when ℓ-way marginals of a contingency table are released. Informally, attribute privacy refers to the situation in a contingency table when all but one of the attributes are public and attacks on privacy amount to revealing the last attribute given the responses to the queries and knowledge of all the other attributes. Releasing the ℓ-way marginals is simply the following : For every subset of size ℓ of the attributes and every configuration of these ℓ-attributes, a count of how many entries in the database have that specific configuration on those ℓ-attributes is released. Due to the lack of space, we refer the reader to [12, 5] for the precise definitions of attribute privacy and ℓ-way marginals. We will also need the definition of row products of matrices which can be found in [5]. The next theorem (proven in [5]) shows how if the range of row product of matrices is Euclidean and all the singular values of the row product are large, one can violate attribute privacy when noisy ℓ-way marginals are released.

Lemma 4. *Let $A_1, \ldots, A_{\ell-1} \in \{0,1\}^{d' \times n}$. Let $A = A_1 \circ A_2 \ldots \circ A_{\ell-1}$ (with $d'^{\ell-1} > n$) be their row product. Also, all the singular values of A are at least σ and the range of A i.e., $\mathcal{L}(A)$ is a $(\delta, n, d'^{\ell-1})$ Euclidean section. Then, there exists a constant $\gamma = \gamma(\delta) > 0$ such that any mechanism which answers at least $1-\gamma$ fraction of the ℓ-way marginals with noise bounded by α is attribute non-private provided $\frac{\alpha\sqrt{d^{\ell'-1} \cdot n}}{\sigma} = o(n)$ or in other words, $\alpha = o(\sqrt{n}\sigma/\sqrt{d^{\ell'-1}})$*

The main technical tool for us is the following theorem of Rudelson [15].

Theorem 8. *[15] Let $q, \ell \in \mathbb{N}$ be constants. Also, let $D \sim \mathbb{R}^{d' \times n}$ be a distribution over matrices such that every entry of the matrix is an independent and unbiased $\{0,1\}$ random variable. Let $A_1, \ldots, A_{\ell-1}$ be i.i.d. copies of random matrices drawn from the distribution D and A be the Hadamard product of $A_1, \ldots, A_{\ell-1}$. Then, provided that $d'^{\ell-1} \gg n \log_{(q)} n$, with probability $1 - o(1)$, the smallest singular value of A denoted by $\sigma_n(A)$ satisfies $\sigma_n(A) = \Omega(\sqrt{d'^{\ell-1}})$ Also, the range of A is a $(n, d'^{\ell-1}, \gamma(q,\ell))$ Euclidean section for some $\gamma(q,\ell) > 0$.*

The above theorem uses the notion of iterated logarithm which is defined as :For $r \in \mathbb{N}$, we define $\log_{(r)} n$ as follows : $\log_{(1)} n = \max\{\log_2 n, 1\}$ and for $r > 1$, $\log_{(r)} n = \log_{(1)} (\log_{(r-1)} n)$. Combining Theorem 8 and Lemma 4, we have the main theorem of this section.

Theorem 9. *Let $q, \ell \in \mathbb{N}$ be constant integers. Then, there exists a constant $\gamma = \gamma(q,\ell) > 0$ such that any mechanism which releases the ℓ-way marginals of a table of size n over d' attributes and $n \leq d'^{\ell-1} \log_{(q)} n$ by adding at most η noise to $1-\gamma$ fraction of the queries where*

$$\eta = o(\sqrt{n})$$

is attribute non-private. Further, the algorithm which violates attribute privacy is efficient and uses LP decoding.

This improves upon the following result of Kasiviswanathan *et al.* [12] who could violate attribute privacy only when all the queries were allowed $o(\sqrt{n})$ noise.

Theorem 10. *[12] Let $\ell \in \mathbb{N}$ be a constant and $n, d \in \mathbb{N}$ such that $d'^{\ell-1} \gg n \cdot \log^{2\ell-4} n$. Then, for every mechanism M which releases ℓ-way marginals of a database of size n (and universe $\{0,1\}^{d'}$) such that the noise for every single query is bounded by η where $\eta \ll \frac{\sqrt{n}}{\log^{\ell^2-\ell+1} n}$ is attribute non-private. The attack is an efficient algorithm based on ℓ_2 norm minimization.*

The details of the results in this section can be found in [5].

Acknowledgements

I would like to thank Cynthia Dwork for her contributions to this paper. Even though she declined to co-author, without her contributions, the paper would not have existed. I would also like to thank Salil Vadhan for his kind permission to include the results of subsection 2.2 in this paper. Moritz Hardt and Mark Rudelson answered countlessly many questions. I also had useful conversations about this work with Ilya Mironov, Elchanan Mossel, Omer Reingold, Adam Smith, Alexandre Stauffer, Kunal Talwar, and Salil Vadhan. I would also like to thank the SODA 2012 and TCC 2012 reviewers for many useful comments including pointing out an error in the earlier proof of Lemma 1.

References

1. Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, 68(4):702–732, 2004.
2. Boaz Barak, Mark Braverman, Xi Chen, and Anup Rao. How to compress interactive communication. In *Proceedings of the 42nd ACM Symposium on Theory of Computing*, pages 67–76, 2010.
3. Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th ACM Symposium on Theory of Computing*, pages 609–618, 2008.
4. Emmanuel J. Candès, Mark Rudelson, Terence Tao, and Roman Vershynin. Error Correction via Linear Programming. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science*, pages 295–308, 2005.
5. Anindya De. Lower bounds in Differential Privacy. 2011. arXiv:1107.2183v1.
6. Anindya De and Salil Vadhan. Personal Communication, 2010.
7. Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Principles of Database Systems*, pages 202–210, 2003.
8. Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In *Proceedings of EUROCRYPT*, pages 486–503, 2006.
9. Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography Conference*, pages 265–284, 2006.
10. Cynthia Dwork, Frank McSherry, and Kunal Talwar. The price of privacy and the limits of LP decoding. In *Proceedings of the 39th ACM Symposium on Theory of Computing*, pages 85–94, 2007.
11. Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the 42nd ACM Symposium on Theory of Computing*, pages 705–714, 2010.
12. Shiva Prasad Kasiviswanathan, Mark Rudelson, Adam Smith, and Jonathan Ullman. The Price of privately releasing Contingency tables and the spectra of random matrices with correlated rows. In *Proceedings of the 42nd ACM Symposium on Theory of Computing*, pages 775–784, 2010.
13. Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil P. Vadhan. The Limits of Two-Party Differential Privacy. In *Proceedings of the 51st IEEE Symposium on Foundations of Computer Science*, pages 81–90, 2010.
14. Paul Erdős, Peter Frankl and Zoltán Füredi. Families of finite sets in which no set is covered by the union of r others. *Israel Journal of Mathematics*, 51(1,2):79–89, 1985.
15. Mark Rudelson. Row products of random matrices. arXiv:1102.1947, 2011.