58th Annual IEEE Symposium on Foundations of Computer Science

Tight Lower Bounds for Differentially Private Selection

Thomas Steinke
IBM Research – Almaden.
topk@thomas-steinke.net.

Jonathan Ullman
Northeastern University, CCIS.
jullman@ccs.neu.edu

***Abstract*—A pervasive task in the differential privacy literature is to select the k items of "highest quality" out of a set of d items, where the quality of each item depends on a sensitive dataset that must be protected. Variants of this task arise naturally in fundamental problems like feature selection and hypothesis testing, and also as subroutines for many sophisticated differentially private algorithms.**

The standard approaches to these tasks—repeated use of the exponential mechanism or the sparse vector technique—approximately solve this problem given a dataset of $n = O(\sqrt{k}\log d)$ samples. We provide a tight lower bound for some very simple variants of the private selection problem. Our lower bound shows that a sample of size $n = \Omega(\sqrt{k}\log d)$ is required even to achieve a very minimal accuracy guarantee.

Our results are based on an extension of the fingerprinting method to sparse selection problems. Previously, the fingerprinting method has been used to provide tight lower bounds for answering an entire set of d queries, but often only some much smaller set of k queries are relevant. Our extension allows us to prove lower bounds that depend on both the number of relevant queries and the total number of queries.

***Keywords*-Differential Privacy, Selection, Fingerprinting, Multiple Hypothesis Testing, Top-k Problem**

I. Introduction

This work studies lower bounds on the sample complexity of differentially private selection problems. Informally, a selection problem consists of a large number of items each with a corresponding value and the task is to select a small subset of those items with large values. In a private selection problem, the values of the items depend on a dataset of sensitive information that must be protected.

Selection problems appear in many natural statistical problems, including private multiple hypothesis testing [1], sparse linear regression [2], [3], finding frequent itemsets [4], and as subroutines in algorithms for answering exponentially many statistical queries [5], [6], [7], [8], [9], approximation algorithms [10], and for establishing the generalization properties of differentially private algorithms [11]. Selection problems appear in many different guises. As we are proving lower bounds, we consider the simplest possible form of selection problems.

More specifically, we consider the following simple selection problem motivated by applications in feature selection and hypothesis testing. There is an unknown probability distribution $\mathcal{P}$ over $\{0,1\}^d$ with mean $p := \mathbb{E}[\mathcal{P}] \in [0,1]^d$, and our goal is to identify a set of coordinates whose mean is large—that is, a set $S \subset [d]$ of size $k \ll d$, such that p^j is large for all $j \in S$. To do this, we obtain n independent samples $X_1, \cdots, X_n \in \{0,1\}^d$ from $\mathcal{P}$. However, each X_i corresponds to the private data of an individual.[1] To protect this data, our procedure for selecting S using the data $X_1, \cdots, X_n$ should satisfy *differential privacy* [12], which is a strong notion of privacy requiring that no individual sample X_i has a significant influence on the set of coordinates S that we select.

For example, suppose $\mathcal{P}$ represents a population of patients suffering from some illness and each coordinate represents the presence of absence of a certain genetic trait. It would be useful for medical researchers to identify genetic traits that are unusually common in this population, but it is also essential not to reveal any individual's genetic information. Thus the researchers would like to obtain genetic data $X_1, \ldots, X_n$ from n random members of this population and run a differentially private selection algorithm on this dataset.

Without privacy, it is necessary and sufficient to draw $n \gtrsim \log d$ samples from $\mathcal{P}$, and compute $\overline{X} = \frac{1}{n}\sum_i X_i$. This ensures that $\|\overline{X} - p\|_\infty$ is small with high probability,[2] so large coordinates of $\overline{X}$ correspond to large coordinates of p. We can ensure differential privacy by adding carefully calibrated noise to the empirical mean $\overline{X}$ to obtain a noisy empirical mean $\tilde{X}$ [13], [14], [15], [12]. Unfortunately, there are strong lower bounds showing that, unless $n \gtrsim \sqrt{d}$, there is no differentially private algorithm whose output $\tilde{X}$ gives a useful approximation to the population mean p [16], [17], [18].

We can avoid this $\sqrt{d}$ lower bound if we only want to identify the k approximately largest coordinates of p, rather than approximating all d values. Specifically, we can use the *exponential mechanism* [19] to identify an approximate largest coordinate of p, and then repeat on the other coordinates. This algorithm provides non-trivial error using just $n \gtrsim \sqrt{k}\log d$ samples. This sample complexity is also achieved by the sparse vector algorithm [20], [21, §3.6] and report noisy max [21, §3.3].

[1] For clarity, we use superscripts to denote the index of a column or item and subscripts to denote the index of a row or individual.

[2] More precisely, if $n \geq \frac{\log(2d/\beta)}{2\alpha^2}$, then $\mathbb{P}\left[\|\overline{X} - p\|_\infty \leq \alpha\right] \geq 1 - \beta$. Since we are proving negative results, we focus on the low-accuracy regime of $\alpha, \beta = \Omega(1)$, where $n = \Theta(\log d)$ samples are both necessary and sufficient.

0272-5428/17 $31.00 © 2017 IEEE
DOI 10.1109/FOCS.2017.57

IEEE computer society

Our first result shows that this sample-complexity is essentially the best possible for the approximate top-k selection problem, even if $\mathcal{P}$ is a product distribution.

Theorem 1 (Informal version of Corollary 13)**.** *Fix $n, d, k \in \mathbb{N}$ with $k \ll d$. Let M be a differentially private algorithm that takes a dataset $X \in (\{0,1\}^d)^n$ of n samples, and outputs an indicator vector $M(X) \in \{0,1\}^d$ such that $\|M(X)\|_1 = k$. Suppose that for every product distribution $\mathcal{P}$ over $\{0,1\}^d$,*

$$\mathop{\mathbb{E}}_{\substack{X \leftarrow \mathcal{P}^n \\ M}} \left[\sum_{j \in [d] \,:\, M(X)^j = 1} p^j \right] \geq \max_{\substack{t \in \{0,1\}^d \\ \|t\|_1 = k}} \sum_{j \in [d] \,:\, t^j = 1} p^j - \frac{k}{10}, \tag{1}$$

where $p = \mathbb{E}[\mathcal{P}]$. Then $n = \Omega(\sqrt{k} \log d)$.

Observe that our lower bound applies whenever the error is at most $k/10$, which is just slightly smaller than the trivial error of k that can be obtained by selecting the first k coordinates.

Scaling with the Privacy and Accuracy Parameters.: For simplicity, we suppress the dependence on the privacy and accuracy parameters in Theorem 1. We assume constant error $\frac{1}{10}$ per selected coordinate, and our lower bound applies to algorithms satisfying $(1, 1/nd)$-differential privacy. Generic reductions can be used to give the appropriate dependence on these parameters in many cases (see e.g. [16], [17]).

Empirical Error vs. Population Error: In Theorem 1, accuracy was defined with respect to the population mean $p = \mathbb{E}[\mathcal{P}]$. This statistical framework is motivated by the fact that we are interested in finding underlying patterns in the population, rather than random empirical deviations.

We could equally well define accuracy with respect to the empirical mean $\overline{X} = \frac{1}{n}\sum_i X_i$. Since $\mathbb{E}[\|\overline{X} - p\|_\infty] \leq \sqrt{\frac{\log(2d)}{2n}}$ and we are interested in settings where $n \gtrsim \log d$, these settings are equivalent.[3] In particular, we can replace the accuracy condition (1) in Theorem 1 with

$$\mathop{\mathbb{E}}_{M} \left[\sum_{j \in [d] \,:\, M(X)^j = 1} \overline{X}^j \right] \geq \max_{\substack{t \in \{0,1\}^d \\ \|t\|_1 = k}} \sum_{j \in [d] \,:\, t^j = 1} \overline{X}^j - \frac{k}{20}. \tag{2}$$

This empirical variant of the problem was first studied in a very recent work by Bafna and Ullman [22]. They proved an optimal lower bound for the empirical variant of the problem in the regime where the error is very small. Specifically they show that, if the empirical error is $\ll k\sqrt{\log(d)/n}$ (i.e. a constant factor smaller than the sampling error), then a dataset of size $n = \Omega(k \log d)$ is necessary.[4] However, their results do not give any lower bound for larger error, or for the statistical problem of approximating the largest entries of the population mean p. Indeed, their lower bounds hold even for uniformly random datasets X. For these datasets we can easily achieve empirical error $k\sqrt{2\log(d)/n}$ and, since $p = (\frac{1}{2}, \ldots, \frac{1}{2})$ is fixed, we can achieve population error 0. Thus, our lower bounds for large error are qualitatively different from the lower bounds of [22] for small error.

Application to Multiple Hypothesis Testing: We can prove an analogous lower bound for a related problem where we do not have a fixed number of coordinates k that we want to select, but instead we want to distinguish coordinates of p that are larger than some threshold τ from those that are smaller than some strictly lower threshold $\tau' < \tau$. This problem is a special case of multiple hypothesis testing in statistics. Without privacy it can be solved using just $n = O_{\tau,\tau'}(\log d)$ samples.

As before, we can use the exponential mechanism or the sparse vector technique to obtain a private algorithm for this problem. The algorithm uses $n = O_{\tau,\tau'}(\sqrt{k}\log d)$ samples, where k is an upper bound on the number of coordinates of p that are above the threshold τ'.[5]

Our second result shows that this sample complexity is essentially optimal for the multiple hypothesis testing problem, even if $\mathcal{P}$ is a product distribution.

Theorem 2 (Informal version of Corollary 14)**.** *Fix $n, d, k \in \mathbb{N}$ with $k \ll d$. There exist absolute constants $\tau, \tau', \rho \in (0,1)$, $\tau' < \tau$ such that the following holds. Let M be a differentially private algorithm that takes a dataset $X \in (\{0,1\}^d)^n$ of n samples, and outputs an indicator vector $M(X) \in \{0,1\}^d$. Suppose that for every product distribution $\mathcal{P}$ over $\{0,1\}^d$ such that $|\{j : p^j \geq \tau\}| \leq k$,*

1) $p^j \leq \tau' \Longrightarrow \mathop{\mathbb{P}}_{X \leftarrow \mathcal{P}^n, M}\left[M(X)^j = 1\right] \leq \rho k/d,$
2) $p^j \geq \tau \Longrightarrow \mathop{\mathbb{P}}_{X \leftarrow \mathcal{P}^n, M}\left[M(X)^j = 1\right] \geq 1 - \rho,$

where $p = \mathbb{E}[\mathcal{P}]$. Then $n = \Omega(\sqrt{k}\log d)$.

As before, we remark that the fact that $\tau' = \tau - \Omega(1)$ makes our lower bound stronger. Also, note that we allow the probability of a false positive ($p^j \leq \tau'$ but $M(X)^j = 1$) to be as large as $\rho k/d$, which means that in expectation there can be as many as $\Omega(k)$ of these false positives. In contrast there are only k true positives ($p^j \geq \tau$ and $M(X)^j = 1$), so our lower bound applies even to algorithms for which the number of false positives is a constant fraction of the number of true positives. The accuracy condition in Theorem 2 is closely related to the *false discovery rate*, which is a widely used statistical criterion introduced in the influential work of Benjamini and Hochberg [23]. A recent work by Dwork, Su,

[3] We need $n \gtrsim \log d$ even in the non-private statistical setting to have meaningful statistical accuracy. In the absence of privacy constraints, the empirical accuracy guarantee can be satisfied for every n.

[4] The results of [22] actually use a slightly stronger accuracy requirement, which requires that for every $j \in [d]$, if $M(X)^j = 1$ then $\overline{X}^{(k)} - \overline{X}^j \ll \sqrt{\log(d)/n}$ where $\overline{X}^{(k)}$ is the k-th largest entry of $\overline{X}$. This technical distinction is not crucial for this high-level discussion.

[5] We assume that the upper bound k is specified as part of the problem input. If k is not specified, the problem and the accuracy guarantee can be formulated differently, but this is not relevant for the current high-level discussion.

and Zhang [1] introduced the problem of privately controlling the false discovery rate.

A. Techniques

Our techniques build on the recent line of work proving lower bounds in differential privacy and related problems via either fingerprinting codes or techniques inspired by fingerprinting codes [16], [24], [25], [17], [18], [26]. Our results follow from the following very general lower bound that refines and generalizes several of the results from those works.

Theorem 3 (Main Lower Bound)**.** *Let $\beta, \gamma, \Delta, k > 0$ and $n, d \in \mathbb{N}$ be a fixed set of parameters. Let $P^1, \cdots, P^d$ be independent draws from $\mathsf{Beta}(\beta, \beta)$ and let $X \in (\{0,1\}^d)^n$ be a random dataset such that every X_i^j is independent (conditioned on P) and $\mathbb{E}[X_i^j] = P^j$ for every $i \in [n]$ and $j \in [d]$.*

Let $M : (\{0,1\}^d)^n \to \mathbb{R}^d$ be a $(1, \beta\gamma k/n\Delta)$-differentially private algorithm and assume that M satisfies the conditions $\underset{P,X,M}{\mathbb{E}}\left[\|M(X)\|_2^2\right] = k$ and $\forall x\ \mathbb{P}\left[\|M(x)\|_1 \leq \Delta\right] = 1$ and the accuracy condition

$$\underset{P,X,M}{\mathbb{E}}\left[\sum_{j \in [d]} M(X)^j \cdot \left(P^j - \frac{1}{2}\right)\right] \geq \gamma k.$$

Then $n \geq \gamma\beta\sqrt{k}$.

In our applications, $\gamma = \Omega(1)$ is a small constant, whereas $\beta = \omega(1)$ is large, namely $\beta = \Theta(\log(d/k))$ for the top-k lower bound. For the purposes of this introduction, it suffices to know that $\mathsf{Beta}(\alpha, \beta)$ is a family of probability distributions over $[0,1]$ with mean $\frac{\alpha}{\alpha+\beta}$. For simplicity, we restrict our attention to the symmetric case where $\alpha = \beta$. The distribution $\mathsf{Beta}(1,1)$ is the uniform distribution on $[0,1]$ and the distribution becomes more concentrated around $1/2$ as $\beta \to \infty$, specifically the variance of $\mathsf{Beta}(\beta, \beta)$ is $\Theta(\frac{1}{\beta})$. The necessary technical details about the beta distribution are in Section II-C.

Observe that in our lower bound the population mean P is itself random. If the population mean were fixed, then we could obtain a private algorithm with perfect accuracy by ignoring the sample and outputting a fixed function of P. Thus to obtain lower bounds then we must assume that the distribution $\mathcal{P}$ is chosen randomly and that M is accurate for these distributions $\mathcal{P}$.

We now describe informally how Theorem 3 implies Theorem 1. First, observe that any algorithm for approximate top-k selection by definition satisfies $\mathbb{E}\left[\|M(X)\|_2^2\right] = k$, since it outputs an indicator vector with exactly k non-zero coordinates. By Theorem 3, to prove an $n = \Omega(\sqrt{k}\log d)$ lower bound, it suffices to show that for some $\beta = \Omega(\log d)$, if M solves the approximate top-k problem, then $\mathbb{E}\left[\sum_j M(X)^j \cdot (P^j - \frac{1}{2})\right] = \Omega(k)$. By the accuracy assumption (1), it suffices to show

$$\underset{P}{\mathbb{E}}\left[\max_{\substack{t \in \{0,1\}^d \\ \|t\|_1 = k}} \sum_{j : t^j = 1} \left(P^j - \frac{1}{2}\right) - \frac{k}{10}\right] \geq \Omega(k). \tag{3}$$

This is simply a property of the beta distribution and our choice of β. We give a simple anti-concentration result for beta distributions showing that the required bound (3) holds for some choice of $\beta = \Omega(\log d)$.

We remark that previous fingerprinting-based lower bounds in differential privacy [16], [27], [28], [17], [18], [3] essentially correspond to setting $\beta = O(1)$ in Theorem 3. Thus the key novelty of our result is that we obtain stronger lower bounds by setting $\beta = \omega(1)$.

Overview of the Analysis: We will sketch the argument for our lower bound in the case of approximate top-k selection. Inspired by prior lower bounds [16], [17], [18], [22] we consider the quantity

$$Z := \sum_{i \in [n]} \langle M(X), (X_i - P)\rangle = n \cdot \left(\sum_{j : M(X)^j = 1} \overline{X}^j - P^j\right).$$

We then use the privacy and accuracy assumptions to establish conflicting upper and lower bounds on the quantity $\mathbb{E}\left[Z\right]$. Combining the two bounds yields the result.

Firstly, we use the differential privacy of M to get an upper bound on $\mathbb{E}\left[Z\right]$. Specifically, for any $i \in [n]$, $M(X)$ should have approximately the same distribution as $M(X_{\sim i})$, where $X_{\sim i}$ is the dataset we obtain by replacing X_i with an independent sample from $\mathcal{P}$. However, X_i and $M(X_{\sim i})$ are independent (conditioned on P) and, therefore, $\mathbb{E}\left[\langle M(X_{\sim i}), X_i - P\rangle\right] = 0$. By differential privacy, $\mathbb{E}\left[\langle M(X), X_i - P\rangle\right] \approx \mathbb{E}\left[\langle M(X_{\sim i}), X_i - P\rangle\right] = 0$. More precisely, we obtain $\mathbb{E}\left[\langle M(X), X_i - P\rangle\right] \leq O(\sqrt{k})$ and, thus, $\mathbb{E}[Z] \leq O(n\sqrt{k})$ (Lemma 8).

Secondly, if $M(X)$ solves the approximate top-k selection problem, then $\mathbb{E}[Z]$ must be large (Lemma 11). This is the technical heart of our result and requires extending the analysis of fingerprinting codes. We give some imprecise intuition for why we should expect $\mathbb{E}\left[Z\right] \geq \Omega(k\beta)$.

The beta distribution has the following "conjugate prior" property. Suppose we sample $P \leftarrow \mathsf{Beta}(\beta, \beta)$, independently sample $Y_1, \ldots, Y_n \in \{0,1\}$ with mean P, and let $\overline{Y} = \frac{1}{n}\sum_i Y_i$. Then the conditional distribution of P given $\overline{Y}$ is

$$(P \mid \overline{Y} = \overline{y}) \sim \mathsf{Beta}\left(\beta + n\overline{y}, \beta + n(1 - \overline{y})\right),$$

so that

$$\mathbb{E}\left[P \mid \overline{Y} = \overline{y}\right] = \frac{\beta + n\overline{y}}{2\beta + n}.$$

Thus, if $\overline{y} \geq \frac{1}{2} + \Omega(1)$, then

$$\mathbb{E}\left[\overline{Y} - P \mid \overline{Y} = \overline{y}\right] = \frac{(2\overline{y} - 1)\beta}{2\beta + n} = \Omega(\beta/n).$$

We connect this back to Z by observing that, if M accurately solves the approximate top-k selection problem, then it will identify a set of k coordinates such that $\overline{X}^j = \frac{1}{2} + \Omega(1)$ on average over the selected indices j. Applying this analysis and summing over the k selected coordinates yields

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left[Z\right] =& \mathbb{E}\left[\sum_{j\in[d]\ :\ M(X)^j=1} \overline{X}^j - P^j\right] \\ \approx& k\cdot\mathbb{E}\left[\overline{X}^j - P^j \;\middle|\; \overline{X}^j \geq \frac{1}{2} + \Omega(1)\right] \\ \geq& \Omega(\beta k/n),\end{aligned}$$

as desired. Unfortunately, our actual proof is somewhat more technical and deviates significantly from this intuition, but also gives a more versatile result.

Finally, combining the bounds $\Omega(k\beta) \leq \mathbb{E}\left[Z\right] \leq O(n\sqrt{k})$ yields $n \geq \Omega(\sqrt{k}\beta)$ (Theorem 3).

B. Relationship to Previous Lower Bounds and Attacks

Our argument is closely related to the work on *tracing attacks* [29], [30], [16], [17], [18], [26], [31]. In a tracing attack, the adversary is given (i) the output $M(X)$ (where $X \leftarrow \mathcal{P}^n$ consists of n independent samples of individuals' data), (ii) an approximate population mean $p \approx \mathbb{E}\left[\mathcal{P}\right]$, and (iii) the data Y of a "target" individual. The target individual is either a random member of the dataset X or an independent random sample from the population $\mathcal{P}$, and the attacker's goal is to determine which of these two is the case. Although we don't state our attack in this model, our attack has essentially this format. Specifically, we consider the quantity $Z = Z_{Y,M(X),p} = \langle M(X), Y - p\rangle$. If $Y \leftarrow \mathcal{P}$ is a fresh sample from the population, Z is zero in expectation and small with high probability. Whereas, when $Y = X_i$ for a random $i \in [n]$, Z is large in expectation, thus we have some ability to distinguish between these two cases.

Another line of work proves lower bounds in differential privacy via *reconstruction attacks* [13], [32], [33], [34], [35], [36], [37]. At a high-level, in a reconstruction attack, each sample X_i contains some public information and an independent, random sensitive bit. The attacker is given $M(X)$ and the public information, and must determine the sensitive bit for 99% of the samples. These attacks do not give any asymptotic separations between the sample complexities of private and non-private problems, because it is easy to prevent reconstruction attacks without providing meaningful privacy by simply throwing out half of the samples and then running a non-private algorithm on the remaining samples. This subsampling prevents reconstruction and only increases the sample complexity by a factor of two compared to the non-private setting.

The work of Bun, Ullman, and Vadhan [16] combines tracing attacks with reconstruction attacks to prove tight lower bounds for large, structured sets of queries (e.g. all k-wise conjunctions). In particular, their work demonstrates that the private multiplicative weights algorithm [7] is nearly optimal. Since selection is a subroutine of private multiplicative weights, this implies a lower bound for private selection. However, this implicit lower bound for private selection only holds in a complex *adaptive* setting [26], where the algorithm must select k items one at a time and the values of the available items change after each selection is made. In contrast, our lower bound is stronger, as it holds for a simple set of items with fixed values.

For the special case of pure differential privacy (i.e. (ε, δ)-differential privacy with $\delta = 0$) lower bounds can be proved using the "packing" technique [38], [39], [40]. The sample complexity of the top-k selection problem becomes $n = \Theta(k \log d)$ under pure differential privacy. (The upper bound is still attained by repeated use of the exponential mechanism, but the stricter privacy requirement changes the analysis and increases the sample complexity.) Packing arguments do not provide any non-trivial lower bounds for general differentially private algorithms (i.e. (ε, δ)-differential privacy with $\delta > 0$).

II. Preliminaries

A. Notational Conventions

We will use the following notational conventions extensively throughout our analysis. We use $[n] = \{1, 2, \cdots, n\}$ to denote the first n natural numbers. We use $X \leftarrow \mathcal{D}$ to denote that X is sampled from the probability distribution $\mathcal{D}$. We also use the shorthand $X_{1\cdots n} \leftarrow \mathcal{D}$ to denote that $X_1, \cdots, X_n$ are drawn independently from $\mathcal{D}$. Given a probability $p \in [0, 1]$, we use the shorthand $X \leftarrow p$ to denote that $X \leftarrow \mathsf{Bernoulli}(p)$ is a sample from a Bernoulli distribution. Likewise, $X_{1\cdots n} \leftarrow p$ denotes that $X_1, \ldots, X_n$ are independent samples from $\mathsf{Bernoulli}(p)$. We follow the convention that upper case (non-caligraphic) letters represent random variables and lower case letters represent their realizations. We will treat $X \in (\{0,1\}^d)^n$ and $X \in \{0,1\}^{n\times d}$ equivalently. For $i \in [n], j \in [d]$, we will subscript X_i to denote the i^{th} row, superscript X^j to denote the j^{th} column, and X_i^j to denote the entry in the i^{th} row and j^{th} column, for $i \in [n]$ and $j \in [d]$. We use $\log$ to denote the natural logarithm, i.e. $\log(z) := \log_e(z)$.

B. Differential Privacy

A dataset $x = (x_1, \ldots, x_n) \in (\{0,1\}^d)^n$ is an $n \times d$ matrix. We say that two datasets x, x' are *neighbors* if they differ on at most one row.

Definition 4 (Differential Privacy [12])**.** *Fix $n, d \in \mathbb{N}$, $\varepsilon, \delta > 0$. A (randomized) algorithm $M : (\{0,1\}^d)^n \to \mathcal{R}$ is (ε, δ)-differentially private if, for every pair of neighboring datasets x, x', and every $R \subseteq \mathcal{R}$,*

$$\mathbb{P}\left[M(x) \in R\right] \leq e^{\varepsilon}\mathbb{P}\left[M(x') \in R\right] + \delta.$$

This definition provides meaningful privacy roughly when $\varepsilon \leq 1$ and $\delta \ll \frac{1}{n}$ [41]. Since our lower bounds allow for

$\varepsilon = 1$ and δ almost as large as $\frac{1}{n}$, they apply to nearly the entire range of parameters for which differential privacy is meaningful.

C. Beta Distributions

Our results make heavy use of the properties of beta distributions. A *beta distribution*, denoted $\mathsf{Beta}(\alpha, \beta)$, is a continuous distribution on $[0, 1]$ with two parameters $\alpha > 0$ and $\beta > 0$ and probability density at p proportional to $p^{\alpha-1}(1-p)^{\beta-1}$. More precisely, the cumulative distribution function is described by $\forall \alpha > 0\ \forall \beta > 0\ \forall p_* \in [0, 1]$

$$\underset{P \leftarrow \mathsf{Beta}(\alpha,\beta)}{\mathbb{P}} [P \leq p_*] = \int_0^{p_*} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathsf{B}(\alpha, \beta)} \mathrm{d}p,$$

where $\mathsf{B}(\alpha, \beta) := \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}\mathrm{d}p$ is the beta function. For all $\alpha, \beta > 0$,

$$\underset{P \leftarrow \mathsf{Beta}(\alpha,\beta)}{\mathbb{E}} [P] = \frac{\alpha}{\alpha + \beta}$$

and

$$\underset{P \leftarrow \mathsf{Beta}(\alpha,\beta)}{\mathsf{Var}} [P] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Note that $\mathsf{Beta}(1, 1)$ is simply the uniform distribution on $[0, 1]$, and $\mathsf{Beta}(\beta, \beta)$ becomes more concentrated around its mean of $1/2$ as β gets larger.

The key result we need is a form of anti-concentration for beta distributions, which says that if we draw d independent samples from a certain beta distribution with mean $1/2$, then in expectation the k largest samples are at least $3/4$.

Proposition 5. *Fix $\beta > 0$ and $d, k \in \mathbb{N}$. Let $P^1, \cdots, P^d$ be independent samples from $\mathsf{Beta}(\beta, \beta)$. If $k \geq 1$ and $1 \leq \beta \leq 1 + \frac{1}{2}\log\left(\frac{d}{8\max\{2k, 28\}}\right)$, then*

$$\underset{P^{1 \cdots d}}{\mathbb{E}} \left[\max_{s \subset [d]\ :\ |s| = k} \sum_{j \in s} P^j \right] \geq \frac{3}{4}k.$$

The above proposition follows from an anti-concentration lemma for the beta distribution.

Lemma 6. *For all $\beta \geq 1$ and all $p_* \in [0, 1/2]$,*

$$\begin{aligned} \underset{P \leftarrow \mathsf{Beta}(\beta,\beta)}{\mathbb{P}} [P > 1 - p_*] =& \underset{P \leftarrow \mathsf{Beta}(\beta,\beta)}{\mathbb{P}} [P < p_*] \\ \geq& \left(4p_*(1-p_*)\right)^{\beta-1} \frac{p_*}{\beta} \\ \geq& p_* \cdot e^{(\log(4p_*(1-p_*))-1)(\beta-1)}. \end{aligned}$$

Proof of Lemma 6: The equality in the statement follows from the fact that $\mathsf{Beta}(\beta, \beta)$ is symmetric around $1/2$. Now we prove the inequalities. We use two bounds: Firstly, $p(1-p) \leq 1/4$ for all $p \in [0, 1]$. Secondly, $p(1-p) \geq p(1-p_*)$ for all $p \in [0, p_*]$. Thus

$$\begin{aligned} \underset{P \leftarrow \mathsf{Beta}(\beta,\beta)}{\mathbb{P}} [P < p_*] =& \frac{\int_0^{p_*} (p(1-p))^{\beta-1}\mathrm{d}p}{\int_0^1 (p(1-p))^{\beta-1}\mathrm{d}p} \\ \geq& \frac{\int_0^{p_*} (p(1-p_*))^{\beta-1}\mathrm{d}p}{\int_0^1 (1/4)^{\beta-1}\mathrm{d}p} \\ =& \frac{(1-p_*)^{\beta-1}\int_0^{p_*} p^{\beta-1}\mathrm{d}p}{(1/4)^{\beta-1}} \\ =& \left(4(1-p_*)\right)^{\beta-1} \frac{p_*^\beta}{\beta} \\ =& \left(4p_*(1-p_*)\right)^{\beta-1} \frac{p_*}{\beta} =: f_{p_*}(\beta). \end{aligned}$$

This proves the first inequality. Now we prove the second inequality by applying calculus to the function f_{p_*} we have just defined.

We have $f_{p_*}(1) = p_*$ and

$$\begin{aligned} f'_{p_*}(\beta) =& \left(4p_*(1-p_*)\right)^{\beta-1} \frac{p_*}{\beta} \left(\log(4p_*(1-p_*)) - \frac{1}{\beta}\right) \\ \geq& f_{p_*}(\beta) \left(\log(4p_*(1-p_*)) - 1\right). \end{aligned}$$

This differential inequation implies

$$f_{p_*}(\beta) \geq p_* e^{(\log(4p_*(-p_*))-1)(\beta-1)},$$

as required. ■

III. Proof of the Main Lower Bound (Theorem 3)

The goal of this section is to prove the following theorem from the introduction.

Theorem 7 (Theorem 3 restated)**.** *Let $\beta, \gamma, \Delta, k > 0$ and $n, d \in \mathbb{N}$ be a fixed set of parameters. Let $P^1, \cdots, P^d$ be independent draws from $\mathsf{Beta}(\beta, \beta)$ and let $X \in (\{0,1\}^d)^n$ be a random dataset such that every X_i^j is independent (conditioned on P) and $\mathbb{E}[X_i^j] = P^j$ for every $i \in [n]$ and $j \in [d]$.*

Let $M : (\{0,1\}^d)^n \to \mathbb{R}^d$ be a $(1, \beta\gamma k/n\Delta)$-differentially private algorithm and assume that M satisfies the conditions $\underset{P,X,M}{\mathbb{E}} \left[\|M(X)\|_2^2\right] = k$ and $\forall x\ \mathbb{P}\left[\|M(x)\|_1 \leq \Delta\right] = 1$ and the accuracy condition

$$\underset{P,X,M}{\mathbb{E}} \left[\sum_{j \in [d]} M(X)^j \cdot \left(P^j - \frac{1}{2}\right) \right] \geq \gamma k.$$

Then $n \geq \gamma\beta\sqrt{k}$.

For the remainder of this section, we will fix the following parameters and variables. Fix $\beta, \gamma, \varepsilon, \delta, \Delta > 0$ and $n, d \in \mathbb{N}$. Let $M : (\{0,1\}^d)^n \to \mathbb{R}^d$ satisfy (ε, δ)-differential privacy. Let $P^1, \cdots, P^d \leftarrow \mathsf{Beta}(\beta, \beta)$ be independent samples. Define a random variable $X \in (\{0,1\}^d)^n$ to have independent entries (conditioned on P) where $\mathbb{E}[X_i^j] = P^j$ for all $i \in [n]$ and $j \in [d]$.

The crux of the proof is to analyze the expected value of $\sum_{i\in[n],j\in[d]} M(X)^j(X_i^j - P^j)$. To this end, for every $i \in [n]$ and $j \in [d]$, we define the random variables

$$\begin{aligned} Z_i^j =& M(X)^j \cdot (X_i^j - P^j), & Z_i =& \sum_{j\in[d]} Z_i^j, \\ Z^j =& \sum_{i\in[n]} Z_i^j, & Z =& \sum_{\substack{i\in[n]\\ j\in[d]}} Z_i^j. \end{aligned}$$

At a high level, we will show that when the size of the dataset n is too small, we obtain contradictory upper and lower bounds on $\mathop{\mathbb{E}}_{P,X,M}[Z]$.

A. Upper Bound via Privacy

First we prove that (ε,δ)-differential privacy of M implies an upper bound on $\mathop{\mathbb{E}}_{P,X,M}[Z]$.

Lemma 8. *Suppose that $\|M(X)\|_1 \le \Delta$ with probability 1. Then*

$$\mathop{\mathbb{E}}_{P,X,M}[Z] \le \frac{n}{2}\cdot\left(e^{\varepsilon}\sqrt{\mathop{\mathbb{E}}_{P,X,M}[\|M(X)\|_2^2]} + \Delta\delta\right).$$

Proof: Fix $i \in [n]$ and also fix the vector $P \in [0,1]^d$. Since $\sum_{j\in[d]} |M(X)^j| \le \Delta$, we have

$$Z_i = \sum_{j\in[d]} M(X)^j\left(P^j - \frac{1}{2}\right) \le \frac{1}{2}\Delta.$$

Let $X_{\sim i} \in (\{0,1\}^d)^n$ denote X with X_i replaced with an independent draw from P. In particular, the marginal distribution of $X_{\sim i}$ is the same as X. However, conditioned on P, X_i is independent from $X_{\sim i}$. By the differential privacy assumption $M(X)$ and $M(X_{\sim i})$ are indistinguishable. We can use this fact to bound the expectation of Z_i in the following calculation.

$$\begin{aligned} \mathop{\mathbb{E}}_{X,M}[Z_i] \le& \mathop{\mathbb{E}}_{X,M}[\max\{0, Z_i\}] \\ =& \int_0^{\Delta/2} \mathop{\mathbb{P}}_{X,M}[Z_i \ge z]\mathrm{d}z \\ =& \int_0^{\Delta/2} \mathop{\mathbb{P}}_{X,M}\left[\sum_{j\in[d]} M(X)^j\left(X_i^j - P^j\right) \ge z\right]\mathrm{d}z \end{aligned} \quad (4)$$

Now, defining the event

$$T(z) := \left\{ y \in \mathbb{R}^d : \sum_{j\in[d]} y^j\left(X_i^j - P^j\right) \ge z\right\},$$

we can apply (ε,δ)-differential privacy to (4) to obtain

$$\begin{aligned} (4) =& \int_0^{\Delta/2} \mathop{\mathbb{P}}_{X,M}[M(X) \in T(z)]\mathrm{d}z \\ \le& \int_0^{\Delta/2} \min\left\{1, e^{\varepsilon}\mathop{\mathbb{P}}_{X,X_{\sim i},M}[M(X_{\sim i}) \in T(z)] + \delta\right\}\mathrm{d}z \\ =& \int_0^{\Delta/2} \min\left\{1, e^{\varepsilon}\mathbb{P}\left[\sum_j M(X_{\sim i})^j\left(X_i^j - P^j\right) \ge z\right] + \delta\right\}\mathrm{d}z \end{aligned} \quad (5)$$

Observe that in (5), $M(X_{\sim i})$ is independent of X_i, which allows us to bound (5) as follows

$$\begin{aligned} (5) \le& \int_0^{\Delta/2} e^{\varepsilon}\mathbb{P}\left[\sum_{j\in[d]} M(X_{\sim i})^j\left(X_i^j - P^j\right) \ge z\right] + \delta\ \mathrm{d}z \\ =& e^{\varepsilon}\mathbb{E}\left[\max\left\{0, \sum_{j\in[d]} M(X_{\sim i})^j\left(X_i^j - P^j\right)\right\}\right] + \frac{\Delta}{2}\delta \\ \le& e^{\varepsilon}\sqrt{\mathbb{E}\left[\left(\sum_{j\in[d]} M(X_{\sim i})^j\left(X_i^j - P^j\right)\right)^2\right]} + \frac{\Delta}{2}\delta \\ =& e^{\varepsilon}\sqrt{\mathop{\mathbb{E}}_{X_{\sim i},M}\left[\sum_j (M(X_{\sim i})^j)^2 \mathop{\mathbb{E}}_{X_i}\left[\left(X_i^j - P^j\right)^2\right]\right]} + \frac{\Delta}{2}\delta \\ \le& e^{\varepsilon}\sqrt{\mathop{\mathbb{E}}_{X_{\sim i},M}\left[\sum_{j\in[d]} (M(X_{\sim i})^j)^2 \frac{1}{4}\right]} + \frac{1}{2}\Delta\delta \\ =& e^{\varepsilon}\frac{1}{2}\sqrt{\mathop{\mathbb{E}}_{X,M}[\|M(X)\|_2^2]} + \frac{1}{2}\Delta\delta. \end{aligned}$$

Finally, we sum over i and average over P to obtain

$$\begin{aligned} \mathop{\mathbb{E}}_{P,X,M}[Z] \le& n \cdot \mathop{\mathbb{E}}_{P}\left[e^{\varepsilon}\frac{1}{2}\sqrt{\mathop{\mathbb{E}}_{X,M}[\|M(X)\|_2^2]} + \frac{1}{2}\Delta\delta\right] \\ \le& n \cdot \left(e^{\varepsilon}\frac{1}{2}\sqrt{\mathop{\mathbb{E}}_{P,X,M}[\|M(X)\|_2^2]} + \frac{1}{2}\Delta\delta\right), \end{aligned}$$

where the final inequality follows from Jensen's inequality and the concavity of the function $x \mapsto \sqrt{x}$. ■

B. Lower Bound via Accuracy

The more involved part of the proof is to use the accuracy assumption

$$\mathop{\mathbb{E}}_{P,X,M}\left[\sum_{j\in[d]} M(X)^j \cdot \left(P^j - \frac{1}{2}\right)\right] \ge \gamma \cdot \mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)\|_2^2\right].$$

to prove a lower bound on $\mathop{\mathbb{E}}_{P,X,M}[Z]$. In order to do so we need to develop a technical tool that we call a "fingerprinting lemma," which is a refinement of similar lemmas from prior work [17], [18], [26] that more carefully exploits the properties of the distribution P.

Fingerprinting Lemma: To keep our notation compact, throughout this section we will use the shorthand $X_{1\cdots n} \leftarrow p$ to denote that $X_1, X_2, \cdots, X_n \in \{0,1\}$ are independent random variables each with mean p.

Lemma 9 (Rescaling of [25], [18])**.** *Let $f : \{0,1\}^n \to \mathbb{R}$. Define $g : [0,1] \to \mathbb{R}$ by*

$$g(p) = \mathop{\mathbb{E}}_{X_{1\cdots n} \leftarrow p}[f(X)].$$

Then

$$\mathop{\mathbb{E}}_{X_{1\cdots n} \leftarrow p}\left[f(X)\sum_{i\in[n]}(X_i - p)\right] = p(1-p)g'(p)$$

for all $p \in [0,1]$.

Proof of Lemma 9: Firstly, $\mathop{\mathbb{P}}_{X\leftarrow p}[X=1] = p$ and $\mathop{\mathbb{P}}_{X\leftarrow p}[X=0] = 1-p$. Thus

$$p(1-p)\frac{\mathrm{d}}{\mathrm{d}p}\mathop{\mathbb{P}}_{X\leftarrow p}[X=1] = p(1-p) = (1-p)\mathop{\mathbb{P}}_{X\leftarrow p}[X=1] \tag{6}$$

and

$$p(1-p)\frac{\mathrm{d}}{\mathrm{d}p}\mathop{\mathbb{P}}_{X\leftarrow p}[X=0] = -p(1-p) = (0-p)\mathop{\mathbb{P}}_{X\leftarrow p}[X=0]. \tag{7}$$

Hence

$$\begin{aligned}
&p(1-p)g'(p)\\
&= p(1-p)\frac{\mathrm{d}}{\mathrm{d}p}\sum_{x\in\{0,1\}^n}\mathop{\mathbb{P}}_{X_{1\cdots n}\leftarrow p}[X=x]f(x)\\
&= \sum_{x\in\{0,1\}^n} f(x)p(1-p)\frac{\mathrm{d}}{\mathrm{d}p}\prod_{i\in[n]}\mathop{\mathbb{P}}_{X\leftarrow p}[X=x_i]\\
&= \sum_{x\in\{0,1\}^n} f(x)\sum_{i\in[n]}\left(\prod_{j\in[n]\setminus\{i\}}\mathop{\mathbb{P}}_{X\leftarrow p}[X=x_j]\right)\\
&\qquad\qquad\cdot\left(p(1-p)\frac{\mathrm{d}}{\mathrm{d}p}\mathop{\mathbb{P}}_{X\leftarrow p}[X=x_i]\right)\\
&= \sum_{x\in\{0,1\}^n} f(x)\sum_{i\in[n]}\left(\prod_{j\in[n]\setminus\{i\}}\mathop{\mathbb{P}}_{X\leftarrow p}[X=x_j]\right)\\
&\text{(by (6) and (7))}\qquad\cdot\left((x_i-p)\mathop{\mathbb{P}}_{X\leftarrow p}[X=x_i]\right)\\
&= \sum_{x\in\{0,1\}^n} f(x)\sum_{i\in[n]}(x_i-p)\left(\prod_{j\in[n]}\mathop{\mathbb{P}}_{X\leftarrow p}[X=x_j]\right)\\
&= \mathop{\mathbb{E}}_{X_{1\cdots n}\leftarrow p}\left[f(X)\sum_{i\in[n]}(X_i-p)\right].
\end{aligned}$$

■

Lemma 10. *Let $f : \{0,1\}^n \to \mathbb{R}$ and let $\alpha, \beta > 0$. Define $g : [0,1] \to \mathbb{R}$ by*

$$g(p) = \mathop{\mathbb{E}}_{X_{1\cdots n} \leftarrow p}[f(X)].$$

Then

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{P\leftarrow\mathsf{Beta}(\alpha,\beta)\\ X_{1\cdots n}\leftarrow P}}\left[f(X)\sum_{i\in[n]}(X_i-P)\right]\\
&\qquad = (\alpha+\beta)\mathop{\mathbb{E}}_{P\leftarrow\mathsf{Beta}(\alpha,\beta)}\left[g(P)\left(P-\frac{\alpha}{\alpha+\beta}\right)\right].
\end{aligned}$$

This is the form of the lemma we use. Note that $\mathbb{E}[P] = \alpha/(\alpha+\beta)$.

Proof of Lemma 10: The proof is a calculation using integration by parts. Using Lemma 9 and the fundamental theorem of calculus, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\substack{P\leftarrow\mathsf{Beta}(\alpha,\beta)\\ X_{1\cdots n}\leftarrow P}}\left[f(X)\sum_{i\in[n]}(X_i-P)\right]\\
&= \mathop{\mathbb{E}}_{P\leftarrow\mathsf{Beta}(\alpha,\beta)}[P(1-P)g'(P)]\\
&= \int_0^1 p(1-p)g'(p)\cdot\frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathsf{B}(\alpha,\beta)}\mathrm{d}p\\
&= \frac{1}{\mathsf{B}(\alpha,\beta)}\int_0^1 g'(p)\cdot p^{\alpha}(1-p)^{\beta}\mathrm{d}p\\
&= \frac{1}{\mathsf{B}(\alpha,\beta)}\int_0^1\left(\begin{array}{c}\frac{\mathrm{d}}{\mathrm{d}p}\left(g(p)\cdot p^{\alpha}(1-p)^{\beta}\right)\\ -g(p)\cdot\frac{\mathrm{d}}{\mathrm{d}p}\left(p^{\alpha}(1-p)^{\beta}\right)\end{array}\right)\mathrm{d}p\\
&= \frac{1}{\mathsf{B}(\alpha,\beta)}\left(g(1)\cdot 1^{\alpha}(1-1)^{\beta} - g(0)\cdot 0^{\alpha}(1-0)^{\beta}\right)\\
&\quad - \frac{1}{\mathsf{B}(\alpha,\beta)}\int_0^1 g(p)\cdot\frac{\mathrm{d}}{\mathrm{d}p}\left(p^{\alpha}(1-p)^{\beta}\right)\mathrm{d}p\\
&= \frac{-1}{\mathsf{B}(\alpha,\beta)}\int_0^1 g(p)\cdot(\alpha-(\alpha+\beta)p)p^{\alpha-1}(1-p)^{\beta-1})\mathrm{d}p\\
&= \mathop{\mathbb{E}}_{P\leftarrow\mathsf{Beta}(\alpha,\beta)}[g(P)((\alpha+\beta)P-\alpha)].
\end{aligned}$$

This completes the proof. ■

Using the Fingerprinting Lemma: Now we can use Lemma 10 to prove a lower bound

Lemma 11.

$$\mathop{\mathbb{E}}_{P,X,M}[Z] \geq 2\beta\mathop{\mathbb{E}}_{P,X,M}\left[\sum_{j\in[d]}M(X)^j\left(P^j-\frac{1}{2}\right)\right]$$

Proof of Lemma 11: Fix a column $j \in [d]$. Define $f : \{0,1\}^n \to [0,1]$ and $g : [0,1] \to [0,1]$ to be

$$f(x^j) := \mathop{\mathbb{E}}_{P^{-j},X^{-j}}\left[M(x^{-j}\|X^j)^j\right]$$

and define g to be

$$g(p^j) := \mathop{\mathbb{E}}_{P^{-j},X^j_{1\ldots n}\sim P^j}\left[f(X^j)\right].$$

That is $f(x)$ is the expectation of $M(X)^j$ conditioned on $X^j = x$, where the expectation is over the randomness of M and the randomness of $P^{j'}$ and $X^{j'}$ for $j' \neq j$. Similarly $g(p)$ is the expectation of $M(X)^j$ conditioned on $P^j = p$, where the expectation is over M and $P^{j'}$ and $X^{j'}$ for $j' \neq j$ and also over X^j. Now we can calculate

$$\begin{aligned}
\mathop{\mathbb{E}}_{P,X,M}\left[Z^j\right] &= \mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j \sum_{i\in[n]} (X_i^j - P^j)\right] \\
&= \mathop{\mathbb{E}}_{\substack{P^j \leftarrow \mathsf{Beta}(\beta,\beta)\\ X^j_{1\cdots n} \leftarrow P^j}}\left[f(X^j) \sum_{i\in[n]} (X_i^j - P^j)\right] \\
&= 2\beta \mathop{\mathbb{E}}_{P^j \leftarrow \mathsf{Beta}(\beta,\beta)}\left[g(P^j)\left(P^j - \frac{1}{2}\right)\right] \\
&\qquad\qquad\qquad\qquad \text{(Lemma 10)} \\
&= 2\beta \mathop{\mathbb{E}}_{\substack{P^j \leftarrow \mathsf{Beta}(\beta,\beta)\\ X^j_{1\cdots n} \leftarrow P^j}}\left[f(X^j)\left(P^j - \frac{1}{2}\right)\right] \\
&= 2\beta \mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j\left(P^j - \frac{1}{2}\right)\right],
\end{aligned}$$

The result now follows by summation over $j \in [d]$. ■

C. Putting it Together

We can now combine the upper bound (Lemma 8) and the lower bound (Lemma 11) that we've proven on the expectation of Z to complete the proof of Theorem 3.

Proof of Theorem 3: By our accuracy assumption and Lemmas 11 and 8,

$$\begin{aligned}
&2\beta\gamma \cdot \mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)\|_2^2\right] \\
\le\ & 2\beta \mathop{\mathbb{E}}_{P,X,M}\left[\sum_{j\in[d]} M(X)^j\left(P^j - \frac{1}{2}\right)\right] \\
\le\ & \mathop{\mathbb{E}}_{P,X,M}[Z] \\
\le\ & \frac{n}{2}\cdot e^{\varepsilon}\sqrt{\mathop{\mathbb{E}}_{P,X,M}[\|M(X)\|_2^2]} + \frac{n}{2}\Delta\delta.
\end{aligned}$$

This implies

$$\begin{aligned}
n \ge& \frac{4\beta\gamma \mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)\|_2^2\right] - n\Delta\delta}{e^{\varepsilon}\sqrt{\mathop{\mathbb{E}}_{P,X,M}[\|M(X)\|_2^2]}} \\
\ge& \frac{3\beta\gamma}{e}\sqrt{\mathop{\mathbb{E}}_{P,X,M}[\|M(X)\|_2^2]} \\
\ge& \beta\gamma\sqrt{k},
\end{aligned}$$

where the final inequality follows from $\mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)\|_2^2\right] = k$, $\varepsilon = 1$, and $\delta = \beta\gamma k/n\Delta$. ■

IV. Using our Lower Bound

In this section we show how to apply the lower bound of Theorem 3 to natural problems, and thereby prove Theorems 1 and 2 from the introduction. We can also use it to derive known lower bounds for releasing the dataset mean [16], [17], [18], which we detail in Section IV-C.

A. Application to Approximate Top-k Selection

We first state an upper bound for the top-k selection problem:

Theorem 12. *Fix $d, k \in \mathbb{N}$ and $\alpha, \varepsilon, \delta > 0$. For every $n \ge \frac{1}{\alpha\varepsilon}\sqrt{8k\log(\frac{e^\varepsilon}{\delta})}\log(d)$, there is an (ε,δ)-differentially private algorithm $M : (\{0,1\}^d)^n \to \{0,1\}^d$ such that for every $x \in (\{0,1\}^d)^n$, such that*

$$\mathop{\mathbb{E}}_{M}\left[\sum_{j\in[d]} M(x)^j \overline{x}^j\right] \ge \max_{s\subset[d]:|s|=k} \sum_{j\in s} \overline{x}^j - \alpha k.$$

This theorem follows immediately by using the exponential mechanism [19] (see [21, Theorem 3.10] and [11, Lemma 7.1] for the analysis) to repeatedly "peel off" the column of x with the approximately largest mean $\overline{x}^j$, and applying the composition theorem for differential privacy [12], [42], [43].

Alternatively, Theorem 12 can provide ρ-concentrated differential privacy [43] (instead of (ε,δ)-differential privacy) for $n \ge \frac{\log d}{\alpha}\sqrt{\frac{2k}{\rho}}$ (with the same accuracy guarantee).

Using Theorem 3 we can obtain a nearly matching lower bound that is tight up to a factor of $O(\sqrt{\log(1/\delta)})$ in most parameter regimes.[6] The lower bound actually holds even for algorithms M that provide just average case accuracy guarantees.

Corollary 13. *Fix $n, d, k \in \mathbb{N}$ with $d \ge \max\{16k, 224\}$. Set $\beta = 1 + \frac{1}{2}\log\left(\frac{d}{8\max\{2k,28\}}\right)$. Let $P^1, \cdots, P^j$ be independent draws from $\mathsf{Beta}(\beta,\beta)$ and let $X \in (\{0,1\}^d)^n$ be such that each X_i^j is independent (conditioned on P) and $\mathbb{E}\left[X_i^j\right] = P^j$ for all $i \in [n]$ and $j \in [d]$. Let $M : (\{0,1\}^d)^n \to \{0,1\}^d$ be $(1, 3\beta/20n)$-differentially private. Suppose $\|M(x)\|_1 = \|M(x)\|_2^2 = k$ for all x with probability 1. Suppose*

$$\mathop{\mathbb{E}}_{P,X,M}\left[\sum_{j\in[d]} M(X)^j P^j\right] \ge \mathop{\mathbb{E}}_{P}\left[\max_{s\subset[d]:|s|=k} \sum_{j\in s} P^j\right] - \frac{k}{10}.$$

Then $n \ge \frac{3}{40}\sqrt{k}\log(\frac{d}{16k+208})$.

Although Corollary 13 is stated for accuracy guarantees that hold with respect to the population mean P, since

[6] The lower bound can be made to include a $\log(1/\delta)$ factor using a group privacy reduction [17]. For the sake of clarity, we do not delve into this issue.

$\mathop{\mathbb{E}}_{X}\left[\|\overline{X} - P\|_\infty\right] \le \sqrt{\frac{\log(2d)}{2n}}$, we can replace the accuracy condition with

$$\begin{aligned}
&\mathop{\mathbb{E}}_{P,X,M}\left[\sum_{j\in[d]} M(X)^j \overline{X}^j\right] \\
&\quad \ge \mathop{\mathbb{E}}_{P,X}\left[\max_{s\subset[d]:|s|=k}\sum_{j\in s}\overline{X}^j\right] - k\left(\frac{1}{10} - \sqrt{\frac{2\log(2d)}{n}}\right)
\end{aligned}$$

to get a theorem that is more directly comparable to Theorem 12.

Proof of Corollary 13: By Proposition 5 and our choice of β,

$$\mathop{\mathbb{E}}_{P^{1\cdots d}}\left[\max_{s\subset[d]\;:\;|s|=k}\sum_{j\in s}P^j\right] \ge \frac{3}{4}k.$$

Thus

$$\begin{aligned}
&\mathop{\mathbb{E}}_{P,X,M}\left[\sum_{j\in[d]} M(X)^j\left(P^j - \frac{1}{2}\right)\right] \\
&\quad \ge \frac{1}{4}k - \frac{k}{10} \;=\; \frac{3}{20}\mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)\|_2^2\right].
\end{aligned}$$

Thus, by Theorem 3,

$$n \ge \beta\frac{3}{20}\sqrt{k} = \frac{3}{20}\sqrt{k}\left(1 + \frac{1}{2}\log\left(\frac{d}{8\max\{2k, 28\}}\right)\right).$$

This completes the proof. ■

B. Application to Multiple Hypothesis Testing

Corollary 14. *Fix $n, d, k \in \mathbb{N}$ with $d \ge 16k \ge 32$. Set $\beta = 1 + \frac{1}{2}\log\left(\frac{d}{16k}\right)$. Let $P^1, \cdots, P^j$ be independent draws from $\mathsf{Beta}(\beta, \beta)$ and let $X \in (\{0,1\}^d)^n$ be such that each X_i^j is independent (conditioned on P) and $\mathbb{E}\left[X_i^j\right] = P^j$ for all $i \in [n]$ and $j \in [d]$. Let $M : (\{0,1\}^d)^n \to \{0,1\}^d$ be $(1, 1/8nd)$-differentially private. Suppose that M is such that for every j,*

(1) $P^j \le \frac{7}{8} - \frac{3}{16} \Longrightarrow \mathop{\mathbb{P}}_{X,M}\left[M(X)^j = 1\right] \le \frac{k}{16d}$, *and*

(2) $P^j \ge \frac{7}{8} \Longrightarrow \mathop{\mathbb{P}}_{X,M}\left[M(X)^j = 1\right] \ge 1 - \frac{1}{16}$,

then $n \ge \frac{1}{16}\sqrt{k}\log(\frac{d}{16k})$.

The assumptions of the theorem may seem a bit confusing, so we will clarify a bit. Note that for every $j \in [d]$ we have $\mathbb{P}\left[P^j > \frac{7}{8}\right] \ge \frac{2k}{d}$ (Lemma 6), so in expectation there are at least $2k$ such values P^j. Thus, the second assumption implies that on average $M(X)^j$ must have at least $\frac{30}{16}k$ non-zero entries. Thus, the parameter k plays roughly the same role in this problem as it does for the top-k selection problem. Furthermore, since $P^1, \cdots, P^d$ are independent, $\left|\left\{j \in [d] \;:\; P^j > \frac{7}{8}\right\}\right|$ concentrates around its expectation.

Proof of Corollary 14: First, we can lower bound the expected norm of $M(X)$ by

$$\begin{aligned}
&\mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)\|_2^2\right] \\
&\quad = \mathop{\mathbb{E}}_{P,X,M}\left[\sum_{j\in[d]} M(X)^j\right] \\
&\quad = \sum_{j\in[d]}\mathop{\mathbb{P}}_{P,X,M}\left[M(X)^j = 1\right] \qquad (M(X)^j \in \{0,1\}) \\
&\quad \ge \sum_{j\in[d]}\mathop{\mathbb{P}}_{P}\left[P^j \ge \frac{7}{8}\right]\cdot\mathop{\mathbb{P}}_{P,X,M}\left[M(X)^j = 1 \;\middle|\; P^j \ge \frac{7}{8}\right] \\
&\quad \ge \sum_{j\in[d]}\frac{2k}{d}\cdot\frac{15}{16} \qquad \text{(Lemma 6 and Assumption 2)} \\
&\quad \ge k
\end{aligned}$$

We need to relate this quantity to $\mathbb{E}\left[\sum_j M(X)^j(P^j - \frac{1}{2})\right]$. As a shorthand, let $\tau := \frac{7}{8} - \frac{3}{16} = \frac{11}{16}$ be the constant from assumption 2. We start by writing

$$\begin{aligned}
&\mathop{\mathbb{E}}_{P,X,M}\left[\sum_{j\in[d]} M(X)^j\left(P^j - \frac{1}{2}\right)\right] \\
&\quad = \sum_{j\in[d]}\mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j\left(P^j - \frac{1}{2}\right)\right] \\
&\quad = \sum_{j\in[d]}\mathop{\mathbb{P}}_{P}\left[P^j \le \tau\right]\cdot(A^j) + \mathop{\mathbb{P}}_{P}\left[P^j > \tau\right]\cdot(B^j)
\end{aligned}$$

where we define

$$\begin{aligned}
(A^j) &:= \mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j\left(P^j - \frac{1}{2}\right) \;\middle|\; P^j \le \tau\right] \\
(B^j) &:= \mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j\left(P^j - \frac{1}{2}\right) \;\middle|\; P^j > \tau\right]
\end{aligned}$$

We will manipulate each of the three terms separately. First, for (A^j), using our first assumption on M we can calculate

$$\begin{aligned}
(A^j) &= \mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j\left(P^j - \frac{1}{2}\right) \;\middle|\; P^j \le \tau\right] \\
&\ge \mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j \mid P^j \le \tau\right]\cdot\left(0 - \frac{1}{2}\right) \\
&\qquad\qquad (M(X)^j \ge 0) \\
&= \mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j \mid P^j \le \tau\right]\cdot\left(\tau - \frac{1}{2}\right) \\
&\quad - \tau\cdot\mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j \mid P^j \le \tau\right] \\
&\ge \mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j \mid P^j \le \tau\right]\cdot\left(\tau - \frac{1}{2}\right) - \tau\cdot\frac{k}{16d} \\
&\qquad\qquad \text{(Assumption 1)} \\
&\ge \mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j \mid P^j \le \tau\right]\cdot\left(\tau - \frac{1}{2}\right) - \frac{k}{16d} \\
&\qquad\qquad (\tau \le 1)
\end{aligned}$$

And, for (B^j), we can calculate

$$\begin{aligned}(B^j) &= \mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j\left(P^j-\frac{1}{2}\right)\;\middle|\; P^j>\tau\right]\\ &\geq \mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j \mid P^j>\tau\right]\cdot\left(\tau-\frac{1}{2}\right)\end{aligned}$$

Combining our inequalities for (A^j) and (B^j) we have

$$\begin{aligned}&\sum_{j\in[d]}\mathop{\mathbb{P}}_{P}\left[P^j\leq\tau\right]\cdot(A^j)+\mathop{\mathbb{P}}_{P}\left[P^j>\tau\right]\cdot(B^j)\\ &\geq \sum_{j\in[d]}\mathop{\mathbb{E}}_{P,X,M}\left[M(X)^j\right]\left(\tau-\frac{1}{2}\right)-\frac{k}{16d}\\ &=\left(\tau-\frac{1}{2}\right)\cdot\mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)\|_2^2\right]-\frac{k}{16}\\ &\qquad\qquad\qquad\qquad\qquad\qquad (M(X)^j\in\{0,1\})\\ &\geq\left(\tau-\frac{1}{2}-\frac{1}{16}\right)\cdot\mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)\|_2^2\right]\\ &\qquad\qquad\qquad\qquad\qquad\qquad (\mathbb{E}\left[\|M(X)\|_2^2\right]\geq k)\\ &=\frac{1}{8}\cdot\mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)\|_2^2\right]\qquad\qquad (\tau=\tfrac{11}{16})\end{aligned}$$

Applying Theorem 3 completes the proof. ■

C. Releasing the Dataset Mean

To illustrate the versatility of Theorem 3, we show how it implies known lower bounds for releasing the mean of the dataset [16], [17], [18].

Corollary 15. *Let $M : (\{0,1\}^d)^n \to [0,1]^d$ be $(1, 1/10n)$-differentially private. Let$P^1, \cdots, P^j$ be independent draws from the uniform distribution on $[0,1]$ and let $X \in (\{0,1\}^d)^n$ be such that each X_i^j is independent (conditioned on P) and $\mathbb{E}\left[X_i^j\right] = P^j$ for all $i \in [n]$ and $j \in [d]$. Assume $\mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)-P\|_2^2\right] \leq \alpha^2 d$. If $\alpha \leq 1/18$, then $n \geq \sqrt{d}/5$.*

Note that we can use empirical values $\overline{X} = \frac{1}{n}\sum_{i\in[n]} X_i$ instead of population values P, as we have $\mathop{\mathbb{E}}_{P,X}\left[\|\overline{X}-P\|_2^2\right] = \frac{1}{n}\sum_{j\in[d]}\mathop{\mathbb{E}}_{P}\left[P^j(1-P^j)\right] \leq \frac{d}{4n}$. In this case the accuracy assumption would be replaced with $\mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)-\overline{X}\|_2^2\right] \leq (\alpha^2-\frac{1}{4n})d$

Proof: Let $k = \mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)\|_2^2\right]$. We have

$$\begin{aligned}&|k-d/3|\\ &=\left|\mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)\|_2^2-\|P\|_2^2\right]\right|\\ &=\left|\mathop{\mathbb{E}}_{P,X,M}\left[\left(\|M(X)\|_2-\|P\|_2\right)\left(\|M(X)\|_2+\|P\|_2\right)\right]\right|\\ &\leq\left|\mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)-P\|_2\cdot 2\sqrt{d}\right]\right|\\ &\leq 2\sqrt{d\mathop{\mathbb{E}}_{P,X,M}\left[\|M(X)-P\|_2^2\right]}\\ &\leq 2\alpha d.\end{aligned}$$

So $d(1/3-2\alpha)\leq k\leq d(1/3+2\alpha)$. Furthermore,

$$\begin{aligned}&\mathop{\mathbb{E}}_{P,X,M}\left[\sum_{j\in[d]}M(X)^j\cdot\left(P^j-\frac{1}{2}\right)\right]\\ &=\mathop{\mathbb{E}}_{P,X,M}\left[\sum_{j\in[d]}P^j\cdot\left(P^j-\frac{1}{2}\right)\right]\\ &\quad-\mathop{\mathbb{E}}_{P,X,M}\left[\sum_{j\in[d]}(P^j-M(X)^j)\cdot\left(P^j-\frac{1}{2}\right)\right]\\ &\geq\frac{d}{4}-\sqrt{\mathop{\mathbb{E}}_{P,X,M}\left[\sum_{j\in[d]}(P^j-M(X)^j)^2\right]}\\ &\qquad\cdot\sqrt{\mathop{\mathbb{E}}_{P,X,M}\left[\sum_{j\in[d]}\left(P^j-\frac{1}{2}\right)^2\right]}\quad\text{(Cauchy-Schwartz)}\\ &\geq\frac{d}{4}-\sqrt{\alpha^2 d}\cdot\sqrt{\frac{d}{4}}\\ &=d\left(\frac{1}{4}-\frac{\alpha}{2}\right)\\ &\geq\frac{1}{2}d\left(\frac{1}{3}+2\alpha\right)\\ &\geq\frac{1}{2}k,\end{aligned}$$

as long as $\alpha\leq 1/18$. Hence, if $\alpha\leq 1/18$, by Theorem 3 (with $\beta=1$ and $\gamma=1/2$), we have

$$n\geq\frac{1}{2}\sqrt{k}\geq\frac{\sqrt{d/3-2\alpha d}}{2}\geq\frac{\sqrt{\frac{2}{9}d}}{2}\geq\frac{\sqrt{d}}{5}.$$

This completes the proof. ■

Acknowledgements

We thank Adam Smith for his instrumental role in the early stages of this research.

References

[1] C. Dwork, W. Su, and L. Zhang, "Private false discovery rate control," *arXiv preprint arXiv:1511.03803*, 2015.

[2] A. Smith and A. Thakurta, "Differentially private model selection via stability arguments and the robustness of the lasso," *J Mach Learn Res Proc Track*, vol. 30, pp. 819–850, 2013.

[3] K. Talwar, A. Thakurta, and L. Zhang, "Nearly optimal private lasso," in *Advances in Neural Information Processing Systems*, 2015, pp. 3025–3033.

[4] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, "Discovering frequent patterns in sensitive data," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 503–512.

[5] A. Blum, K. Ligett, and A. Roth, "A learning theory approach to noninteractive database privacy," *J. ACM*, vol. 60, no. 2, p. 12, 2013.

[6] A. Roth and T. Roughgarden, "Interactive privacy via the median mechanism," in *STOC*. ACM, June 5–8 2010, pp. 765–774.

[7] M. Hardt and G. N. Rothblum, "A multiplicative weights mechanism for privacy-preserving data analysis," in *FOCS*, 2010.

[8] A. Gupta, A. Roth, and J. Ullman, "Iterative constructions and private data release," in *TCC*, 2012.

[9] J. Ullman, "Private multiplicative weights beyond linear queries," in *PODS*, 2015.

[10] A. Gupta, K. Ligett, F. McSherry, A. Roth, and K. Talwar, "Differentially private combinatorial optimization," in *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2010, pp. 1106–1125.

[11] R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman, "Algorithmic stability for adaptive data analysis," in *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2016, pp. 1046–1059.

[12] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *TCC*, 2006.

[13] I. Dinur and K. Nissim, "Revealing information while preserving privacy," in *PODS*, 2003.

[14] C. Dwork and K. Nissim, "Privacy-preserving datamining on vertically partitioned databases," in *CRYPTO*, 2004.

[15] A. Blum, C. Dwork, F. McSherry, and K. Nissim, "Practical privacy: the SuLQ framework," in *PODS*, 2005.

[16] M. Bun, J. Ullman, and S. P. Vadhan, "Fingerprinting codes and the price of approximate differential privacy," in *STOC*, 2014.

[17] T. Steinke and J. Ullman, "Between pure and approximate differential privacy," *Journal of Privacy and Confidentiality*, vol. 7, no. 2, 2017.

[18] C. Dwork, A. D. Smith, T. Steinke, J. Ullman, and S. P. Vadhan, "Robust traceability from trace amounts," in *FOCS*, 2015.

[19] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *FOCS*, 2007.

[20] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. P. Vadhan, "On the complexity of differentially private data release: efficient algorithms and hardness results," in *STOC*, 2009.

[21] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[22] M. Bafna and J. Ullman, "The price of selection in differential privacy," *arXiv preprint arXiv:1702.02970*, 2017.

[23] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[24] M. Hardt and J. Ullman, "Preventing false discovery in interactive data analysis is hard," in *FOCS*. IEEE, 2014.

[25] T. Steinke and J. Ullman, "Interactive fingerprinting codes and the hardness of preventing false discovery," in *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, 2015, pp. 1588–1628.

[26] M. Bun, T. Steinke, and J. Ullman, "Make up your mind: The price of online queries in differential privacy," in *SODA*. Society for Industrial and Applied Mathematics, 2017.

[27] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*. IEEE, 2014, pp. 464–473.

[28] C. Dwork, K. Talwar, A. Thakurta, and L. Zhang, "Analyze gauss: optimal bounds for privacy-preserving principal component analysis," in *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. ACM, 2014, pp. 11–20.

[29] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays," *PLoS genetics*, vol. 4, no. 8, p. e1000167, 2008.

[30] S. Sankararaman, G. Obozinski, M. I. Jordan, and E. Halperin, "Genomic privacy and limits of individual detection in a pool," *Nature genetics*, vol. 41, no. 9, pp. 965–967, 2009.

[31] C. Dwork, A. Smith, T. Steinke, and J. Ullman, "Exposed a survey of attacks on private data," *Annual Review of Statistics and Its Application*, 2017.

[32] C. Dwork, F. McSherry, and K. Talwar, "The price of privacy and the limits of lp decoding," in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*. ACM, 2007, pp. 85–94.

[33] C. Dwork and S. Yekhanin, "New efficient attacks on statistical disclosure control mechanisms," in *Annual International Cryptology Conference*. Springer, 2008, pp. 469–480.

[34] S. P. Kasiviswanathan, M. Rudelson, A. Smith, and J. Ullman, "The price of privately releasing contingency tables and the spectra of random matrices with correlated rows," in *STOC*, 2010.

[35] S. P. Kasiviswanathan, M. Rudelson, and A. Smith, "The power of linear reconstruction attacks," in *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2013, pp. 1415–1433.

[36] S. Muthukrishnan and A. Nikolov, "Optimal private halfspace counting via discrepancy," in *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. ACM, 2012, pp. 1285–1292.

[37] A. Nikolov, K. Talwar, and L. Zhang, "The geometry of differential privacy: the sparse and approximate cases," in *STOC*, 2013.

[38] D. Feldman, A. Fiat, H. Kaplan, and K. Nissim, "Private coresets," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*. ACM, 2009, pp. 361–370.

[39] A. Beimel, S. P. Kasiviswanathan, and K. Nissim, "Bounds on the sample complexity for private learning and private data release," in *Theory of Cryptography Conference*. Springer, 2010, pp. 437–454.

[40] M. Hardt and K. Talwar, "On the geometry of differential privacy," in *Proceedings of the forty-second ACM symposium on Theory of computing*. ACM, 2010, pp. 705–714.

[41] S. Kasiviswanathan and A. Smith, "On the "semantics" of differential privacy: A bayesian formulation," *Journal of Privacy and Confidentiality*, vol. 6, no. 1, 2014.

[42] C. Dwork, G. N. Rothblum, and S. P. Vadhan, "Boosting and differential privacy," in *FOCS*. IEEE, 2010.

[43] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory of Cryptography Conference*. Springer, 2016, pp. 635–658.