INSPECTRE: Privately Estimating the Unseen

Jayadev Acharya [*1] **Gautam Kamath** [*2] **Ziteng Sun** [*1] **Huanyu Zhang** [*1]

Abstract

We develop differentially private methods for estimating various distributional properties. Given a sample from a discrete distribution p, some functional f, and accuracy and privacy parameters α and ε, the goal is to estimate $f(p)$ up to accuracy α, while maintaining ε-differential privacy of the sample. We prove almost-tight bounds on the sample size required for this problem for several functionals of interest, including support size, support coverage, and entropy. We show that the cost of privacy is negligible in a variety of settings, both theoretically and experimentally. Our methods are based on a sensitivity analysis of several state-of-the-art methods for estimating these properties with sublinear sample complexities.

1. Introduction

How can we infer a distribution given a sample from it? If data is in abundance, the solution may be simple – the empirical distribution will approximate the true distribution. However, challenges arise when data is scarce in comparison to the size of the domain, and especially when we wish to quantify "rare events." This is frequently the case: for example, it has recently been observed that there are several very rare genetic mutations which occur in humans, and we wish to know how many such mutations exist (Keinan & Clark, 2012; Tennessen et al., 2012; Nelson et al., 2012). Many of these mutations have only been seen once, and we can infer that there are many which have not been seen at all. Over the last decade, a large body of work has focused on developing theoretically sound and effective tools for such settings (Orlitsky et al., 2016) and references therein, including the problem of estimating the frequency distribution of rare genetic variations (Zou et al., 2016).

However, in many settings where one wishes to perform statistical inference, data may contain sensitive information about individuals. For example, in medical studies, where the data may contain individuals' health records and whether they carry some disease which bears a social stigma. Alternatively, one can consider a map application which suggests routes based on aggregate positions of individuals, which contains delicate information including users' residence data. In these settings, it is critical that our methods protect sensitive information contained in the dataset. This does not preclude our overall goals of statistical analysis, as we are trying to infer properties of the population p, and not the samples which are drawn from said population.

That said, without careful experimental design, published statistical findings may be prone to leaking sensitive information about the sample. As a notable example, it was recently shown that one can determine the identity of some individuals who participated in genome-wide association studies (Homer et al., 2008). This realization has motivated a surge of interest in developing data sharing techniques with an explicit focus on maintaining privacy of the data (Johnson & Shmatikov, 2013; Uhler et al., 2013; Yu et al., 2014; Simmons et al., 2016).

Privacy-preserving computation has enjoyed significant study in a number of fields, including statistics and almost every branch of computer science, including cryptography, machine learning, algorithms, and database theory – see, e.g., (Dalenius, 1977; Adam & Worthmann, 1989; Agrawal & Aggarwal, 2001; Dinur & Nissim, 2003; Dwork, 2008; Dwork & Roth, 2014) and references therein. Perhaps the most celebrated notion of privacy, proposed by theoretical computer scientists, is *differential privacy* (Dwork et al., 2006). Informally, an algorithm is differentially private if its outputs on neighboring datasets (differing in a single element) are statistically close (for a more precise definition, see Section 2). Differential privacy has become the standard for theoretically-sound data privacy, leading to its adoption by several large technology companies, including Google and Apple (Erlingsson et al., 2014; Differential Privacy Team, Apple, 2017).

Our focus in this paper is to develop tools for privately performing several distribution property estimation tasks. In

[*]Equal contribution [1]ECE, Cornell University, Ithaca, New York, USA [2]EECS & CSAIL, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence to: Jayadev Acharya <acharya@cornell.edu>, Gautam Kamath <g@csail.mit.edu>, Ziteng Sun <zs335@cornell.edu>, Huanyu Zhang <hz388@cornell.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

particular, we study the tradeoff between statistical accuracy, privacy, and error rate in the sample size. Our model is that we are given sample access to some unknown discrete distribution p, over a domain of size k, which is possibly unknown in some tasks. We wish to estimate the following properties:

- **Support Coverage**: If we take m samples from the distribution, what is the expected number of unique elements we expect to see?
- **Support Size**: How many elements of the support have non-zero probability?
- **Entropy**: What is the Shannon entropy of the distribution?

For more formal statements of these problems, see Section 2.1. We require that our output is α-accurate, satisfies $(\varepsilon, 0)$-differential privacy, and is correct with probability $1-\beta$. The goal is to give an algorithm with minimal sample complexity n, while simultaneously being computationally efficient.

Theoretical Results. Our main results show that privacy can be achieved for all these problems at a very low cost. For example, if one wishes to privately estimate entropy, this incurs an additional additive cost in the sample complexity which is very close to linear in $1/\alpha\varepsilon$. We draw attention to two features of this bound. First, this is independent of k. All the problems we consider have complexity $\Theta(k/\log k)$, so in the primary regime of study where $k \gg 1/\alpha\varepsilon$, this small additive cost is dwarfed by the inherent sample complexity of the non-private problem. Second, the bound is almost linear in $1/\alpha\varepsilon$. We note that performing even the most basic statistical task privately, estimating the bias of a coin, incurs this linear dependence. Surprisingly, we show that much more sophisticated inference tasks can be privatized at almost no cost. In particular, these properties imply that the additive cost of privacy is $o(1)$ in the most studied regime where the support size is large. In general, this is not true – for many other problems, including distribution estimation and hypothesis testing, the additional cost of privacy depends significantly on the support size or dimension (Diakonikolas et al., 2015; Cai et al., 2017; Acharya et al., 2017c; Aliakbarpour et al., 2017). We also provide lower bounds, showing that our upper bounds are almost tight. A more formal statement of our results appears in Section 3.

Experimental Results. We demonstrate the efficacy of our method with experimental evaluations. As a baseline, we compare with the non-private algorithms of (Orlitsky et al., 2016) and (Wu & Yang, 2018). Overall, we find that our algorithms' performance is nearly identical, showing that, in many cases, privacy comes (essentially) for free. We begin with an evaluation on synthetic data. Then, inspired by (Valiant & Valiant, 2013; Orlitsky et al., 2016), we analyze text corpus consisting of words from Hamlet, in order to estimate the number of unique words which occur. Finally, we investigate name frequencies in the US census data. This setting has been previously considered by (Orlitsky et al., 2016), but we emphasize that this is an application where private statistical analysis is critical. This is proven by efforts of the US Census Bureau to incorporate differential privacy into the 2020 US census (Dajani et al., 2017).

Techniques. Our approach works by choosing statistics for these tasks which possess bounded sensitivity, which is well-known to imply privacy under the Laplace or Gaussian mechanism. We note that bounded sensitivity of statistics is not always something that can be taken for granted. Indeed, for many fundamental tasks, optimal algorithms for the non-private setting may be highly sensitive, thus necessitating crucial modifications to obtain differential privacy (Acharya et al., 2015; Cai et al., 2017). Thus, careful choice and design of statistics must be a priority when performing inference with privacy considerations.

To this end, we leverage recent results of (Acharya et al., 2017a), which studies estimators for non-private versions of the problems we consider. The main technical work in their paper exploits bounded sensitivity to show sharp cutoff-style concentration bounds for certain estimators, which operate using the principle of best-polynomial approximation. They use these results to show that a single algorithm, the Profile Maximum Likelihood (PML), can estimate all these properties simultaneously. On the other hand, we consider the sensitivity of these estimators for purposes of privacy – the same property is utilized by both works for very different purposes, a connection which may be of independent interest.

We note that bounded sensitivity of a statistic may be exploited for purposes other than privacy. For instance, by McDiarmid's inequality, any such statistic also enjoys very sharp concentration of measure, implying that one can boost the success probability of the test at an additive cost which is logarithmic in the inverse of the failure probability. One may naturally conjecture that, if a statistical task is based on a primitive which concentrates in this sense, then it may also be privatized at a low cost. However, this is not true – estimating a discrete distribution in ℓ_1 distance is such a task, but the cost of privatization depends significantly on the support size (Diakonikolas et al., 2015).

One can observe that, algorithmically, our method is quite simple: compute the non-private statistic, and add a relatively small amount of Laplace noise. The non-private statistics have recently been demonstrated to be practical (Orlitsky et al., 2016; Wu & Yang, 2018), and the additional cost of the Laplace mechanism is minimal. This is in contrast to several differentially private algorithms which invoke significant overhead in the quest for privacy. Our algorithms

attain almost-optimal rates (which are optimal up to constant factors for most parameter regimes of interest), while simultaneously operating effectively in practice, as demonstrated in our experimental results.

Related Work. Over the last decade, there have been a flurry of works on the problems we study in this paper by the computer science and information theory communities, including Shannon and Rényi entropy estimation (Paninski, 2003; Valiant & Valiant, 2017; Jiao et al., 2017; Acharya et al., 2017b; Obremski & Skorski, 2017; Wu & Yang, 2018), support coverage and support size estimation (Orlitsky et al., 2016; Wu & Yang, 2018). A recent paper studies the general problem of estimating functionals of discrete distribution from samples in terms of the smoothness of the functional (Fukuchi & Sakuma, 2017). These have culminated in a nearly-complete understanding of the sample complexity of these properties, with optimal sample complexities (up to constant factors) for most parameter regimes.

Recently, there has been significant interest in performing statistical tasks under differential privacy constraints. Perhaps most relevant to this work are (Cai et al., 2017; Acharya et al., 2017c; Aliakbarpour et al., 2017), which study the sample complexity of differentialy privately performing classical distribution testing problems, including identity and closeness testing. Other works investigating private hypothesis testing include (Wang et al., 2015a; Gaboardi et al., 2016; Kifer & Rogers, 2017; Kakizaki et al., 2017; Rogers, 2017; Gaboardi & Rogers, 2017), which focus less on characterizing the finite-sample guarantees of such tests, and more on understanding their asymptotic properties and applications to computing p-values. There has also been study on private distribution learning (Diakonikolas et al., 2015; Duchi et al., 2017; Karwa & Vadhan, 2018; Acharya et al., 2018; Kamath et al., 2018), in which we wish to estimate parameters of the distribution, rather than just a particular property of interest. A number of other problems have been studied with privacy requirements, including clustering (Wang et al., 2015b; Balcan et al., 2017), principal component analysis (Chaudhuri et al., 2013; Kapralov & Talwar, 2013; Hardt & Price, 2014), ordinary least squares (Sheffet, 2017), and much more.

2. Preliminaries

We will start with some definitions.

Let $\Delta \stackrel{\text{def}}{=} \{(p(1), \ldots, p(k)) : p(i) \geq 0, \sum_{i=1}^{k} p(i) = 1, 1 \leq k \leq \infty\}$ be the set of discrete distributions over a countable support. Let Δ_k be the set of distributions in Δ with at most k non-zero probability values. A *property* $f(p)$ is a mapping from $\Delta \to \mathbb{R}$. We now describe the classical distribution property estimation problem, and then state the problem under differential privacy.

Property Estimation. Given α, β, f, and independent samples X_1^n from an unknown distribution p, design an estimator $\hat{f} : X_1^n \to \mathbb{R}$ such that with probability at least $1-\beta$, $\left|\hat{f}(X_1^n) - f(p)\right| < \alpha$. The *sample complexity* of $\hat{f}$, $C_{\hat{f}}(f, \alpha, \beta) \stackrel{\text{def}}{=} \min\{n : \Pr\left(\left|\hat{f}(X_1^n) - f(p)\right| > \alpha\right) < \beta\}$ is the smallest number of samples to estimate f to accuracy α, and error β. We study the problem for $\beta = 1/3$, and by the median trick, we can boost the success probability to $1-\beta$ with an additional multiplicative $\log(1/\beta)$ more samples. Therefore, focusing on $\beta = 1/3$, we define $C_{\hat{f}}(f, \alpha) \stackrel{\text{def}}{=} C_{\hat{f}}(f, \alpha, 1/3)$. The sample complexity of estimating a property $f(p)$ is the minimum sample complexity over all estimators: $C(f, \alpha) = \min_{\hat{f}} C_{\hat{f}}(f, \alpha)$.

An estimator $\hat{f}$ is ε-differentially private (DP) (Dwork et al., 2006) if for any X_1^n and Y_1^n, with $d_{ham}(X_1^n, Y_1^n) \leq 1$, $\frac{\Pr(f(X_1^n) \in S)}{\Pr(f(Y_1^n) \in S)} \leq e^{\varepsilon}$, for all measurable S.

Private Property Estimation. Given $\alpha, \varepsilon, \beta$, f, and independent samples X_1^n from an unknown distribution p, design an ε-differentially private estimator $\hat{f} : X_1^n \to \mathbb{R}$ such that with probability at least $1-\beta$, $\left|\hat{f}(X_1^n) - f(p)\right| < \alpha$. Similar to the non-private setting, the *sample complexity* of ε-differentially private estimation problem is $C(f, \alpha, \varepsilon) = \min_{\hat{f}: \hat{f} \text{ is } \varepsilon\text{-DP}} C_{\hat{f}}(f, \alpha, 1/3)$, the smallest number of samples n for which there exists such an ε-DP $\pm\alpha$ estimator with error probability at most 1/3.

In their original paper (Dwork et al., 2006) provides a scheme for differential privacy, known as the Laplace mechanism. This method adds Laplace noise to a non-private scheme in order to make it private. We first define the sensitivity of an estimator, and then state their result in our setting.

Definition 1. *The* sensitivity *of an estimator* $\hat{f} : [k]^n \to \mathbb{R}$ *is* $\Delta_{n,\hat{f}} \stackrel{def}{=} \max_{d_{ham}(X_1^n, Y_1^n) \leq 1} \left|\hat{f}(X_1^n) - \hat{f}(Y_1^n)\right|$. *Let* $D_{\hat{f}}(\alpha, \varepsilon) = \min\{n : \Delta_{n,\hat{f}} \leq \alpha\varepsilon\}$.

Lemma 1.

$$C(f, \alpha, \varepsilon) = O\left(\min_{\hat{f}} \left\{C_{\hat{f}}(f, \alpha/2) + D_{\hat{f}}\left(\frac{\alpha}{4}, \varepsilon\right)\right\}\right).$$

Proof. (Dwork et al., 2006) showed that for a function with sensitivity $\Delta_{n,\hat{f}}$, adding Laplace noise $X \sim Lap(\Delta_{n,\hat{f}}/\varepsilon)$ makes the output ε-differentially private. By the definition of $D_{\hat{f}}(\frac{\alpha}{4}, \varepsilon)$, the Laplace noise we add has parameter at most $\frac{\alpha}{4}$. Recall that the probability density function of $Lap(b)$ is $\frac{1}{2b}e^{-\frac{|x|}{b}}$, hence we have $\Pr(|X| > \alpha/2) < \frac{1}{e^2}$. By the union bound, we get an additive error larger than $\alpha = \frac{\alpha}{2} + \frac{\alpha}{2}$ with probability at most $1/3 + \frac{1}{e^2} < 0.5$. Hence, with the median trick, we can boost the error probability

to $1/3$, at the cost of a constant factor in the number of samples. □

To prove sample complexity lower bounds for differentially private estimators, we observe that the estimator can be used to test between two distributions with distinct property values, hence is a harder problem. For lower bounds on differentially private testing, (Acharya et al., 2017c) gives the following argument based on coupling:

Lemma 2. *Suppose there is a coupling between distributions p and q over $\mathcal{X}^n$, such that $\mathbb{E}\left[d_{ham}(X_1^n, Y_1^n)\right] \leq D$. Then, any ε-differentially private algorithm that distinguishes between p and q with error probability at most $1/3$ must satisfy $D = \Omega\left(\frac{1}{\varepsilon}\right)$.*

2.1. Problems of Interest

Support Size. The support size of a distribution p is $S(p) = |\{x : p(x) > 0\}|$, the number of symbols with non-zero probability values. However, notice that estimating $S(p)$ from samples can be hard due to the presence of symbols with negligible, yet non-zero probabilities. To circumvent this issue, (Raskhodnikova et al., 2009) proposed to study the problem when the smallest probability is bounded. Let $\Delta_{\geq \frac{1}{k}} \stackrel{\text{def}}{=} \{p \in \Delta : p(x) \in \{0\} \cup [1/k, 1]\}$ be the set of all distributions where all non-zero probabilities have value at least $1/k$. For $p \in \Delta_{\geq \frac{1}{k}}$, our goal is to estimate $S(p)$ up to $\pm\alpha k$ with the least number of samples from p.

Support Coverage. For a distribution p, and an integer m, let $S_m(p) = \sum_x (1-(1-p(x))^m)$, be the expected number of symbols that appear when we obtain m independent samples from the distribution p. The objective is to find the least number of samples n in order to estimate $S_m(p)$ to an additive $\pm\alpha m$.

Support coverage arises in many ecological and biological studies (Colwell et al., 2012) to quantify the number of *new* elements (gene mutations, species, words, etc) that can be expected to be seen in the future. Good and Toulmin (Good & Toulmin, 1956) proposed an estimator that for any constant α, requires $m/2$ samples to estimate $S_m(p)$.

Entropy. The Shannon entropy of a distribution p is $H(p) = \sum_x p(x) \log \frac{1}{p(x)}$, $H(p)$ is a central object in information theory (Cover & Thomas, 2006), and also arises in many fields such as machine learning (Nowozin, 2012), neuroscience (Berry et al., 1997; Nemenman et al., 2004), and others. Estimating $H(p)$ is hard with any finite number of samples due to the possibility of infinite support. To circumvent this, a natural approach is to consider distributions in Δ_k. The goal is to estimate the entropy of a distribution in Δ_k to an additive $\pm\alpha$, where Δ_k is all discrete distributions over at most k symbols.

3. Statement of Results

Our theoretical results for estimating support coverage, support size, and entropy are given below. Algorithms for these problems and proofs of these statements are provided in Section 4. Our experimental results are described and discussed in Section 5.

Theorem 1. *The sample complexity of support coverage estimation $C(S_m, \alpha, \varepsilon)$ is*

$$\begin{cases} O\left(\frac{m\log(1/\alpha)}{\log m} + \frac{m\log(1/\alpha)}{\log(2+\varepsilon m)}\right), & \text{when } m \geq \frac{1}{\alpha\varepsilon} \\ O\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right), & \text{when } \frac{1}{\alpha} \leq m \leq \frac{1}{\alpha\varepsilon} \\ O\left(m^2 + \frac{m}{\varepsilon}\right). & \text{when } m \leq \frac{1}{\alpha} \end{cases}$$

Furthermore,

$$C(S_m, \alpha, \varepsilon) = \Omega\left(\frac{m\log(1/\alpha)}{\log m} + \frac{1}{\alpha\varepsilon}\right).$$

Theorem 2. *The sample complexity of support size estimation $C(S, \alpha, \varepsilon)$ is*

$$\begin{cases} O\left(\frac{k\log^2(1/\alpha)}{\log k} + \frac{k\log^2(1/\alpha)}{\log(2+\varepsilon k)}\right), & \text{when } k \geq \frac{1}{\alpha\varepsilon} \\ O\left(k\log(1/\alpha) + \frac{1}{\alpha\varepsilon}\right), & \text{when } \frac{1}{\alpha} \leq k \leq \frac{1}{\alpha\varepsilon} \\ O\left(k\log k + \frac{k}{\varepsilon}\right). & \text{when } k \leq \frac{1}{\alpha} \end{cases}$$

Furthermore,

$$C(S, \alpha, \varepsilon) = \begin{cases} \Omega\left(\frac{k\log^2(1/\alpha)}{\log k} + \frac{1}{\alpha\varepsilon}\right), & \text{when } k \geq \frac{1}{\alpha} \\ \Omega\left(k\log k + \frac{k}{\varepsilon}\right). & \text{when } k \leq \frac{1}{\alpha} \end{cases}$$

Theorem 3. *Let $\lambda > 0$ be* any *small fixed constant. For instance, λ can be chosen to be any constant between 0.01 and 1. We have the following upper bounds on the sample complexity of entropy estimation $C(H, \alpha, \varepsilon)$:*

$$O\left(\frac{k}{\alpha} + \frac{\log^2(\min\{k, n\})}{\alpha^2} + \frac{1}{\alpha\varepsilon}\log\left(\frac{1}{\alpha\varepsilon}\right)\right)$$

and

$$O\left(\frac{k}{\lambda^2\alpha\log k} + \frac{\log^2(\min\{k, n\})}{\alpha^2} + \left(\frac{1}{\alpha\varepsilon}\right)^{1+\lambda}\right).$$

Furthermore,

$$C(H, \alpha, \varepsilon) = \Omega\left(\frac{k}{\alpha\log k} + \frac{\log^2(\min\{k, n\})}{\alpha^2} + \frac{\log k}{\alpha\varepsilon}\right).$$

We provide some discussion of our results. At a high level, we wish to emphasize the following two points:

1. Our upper bounds show that the cost of privacy in these settings is often negligible compared to the sample complexity of the non-private statistical task, especially when we are dealing with distributions over a large support. Furthermore, our upper bounds are almost tight in all parameters.
2. The algorithmic complexity introduced by the requirement of privacy is minimal, consisting only of a single step which noises the output of an estimator. In other words, our methods are realizable in practice, and we demonstrate the effectiveness on several synthetic and real-data examples.

Before we continue, we emphasize that, in Theorems 1 and 2, we consider the "sublinear" regime to be of primary interest (when $m \geq \frac{1}{\alpha\varepsilon}$ or $k \geq \frac{1}{\alpha\varepsilon}$, respectively), both technically, and in terms of parameter regimes which may be of greatest interest in practice. We include results for other regimes mostly for completeness.

First, we examine our results on support coverage and support size estimation in the sublinear regime, when $m \geq \frac{1}{\alpha\varepsilon}$ (focusing on support coverage for simplicity, but support size is similar). In this regime, if $\varepsilon = \Omega(m^\gamma/m)$ for any constant $\gamma > 0$, then up to constant factors, our upper bound is within a constant factor of the optimal sample complexity without privacy constratints. In other words, for most meaningful values of ε, privacy comes for free. In the non-sublinear regime for these problems, we provide upper and lower bounds which match in a number of cases. We note that in this regime, the cost of privacy may not be a lower order term – however, this regime only occurs when one requires very high accuracy, or unreasonably large privacy, which we consider to be of somewhat lesser interest.

Next, we turn our attention to entropy estimation. We note that the second upper bound in Theorem 3 has a parameter λ that indicates a tradeoff between the sample complexity incurred in the first and third term. This parameter determines the degree of a polynomial to be used for entropy estimation. As the degree becomes smaller (corresponding to a large λ), accuracy of the polynomial estimator decreases, however, at the same time, low-degree polynomials have a small sensitivity, allowing us to privatize the outcome.

In terms of our theoretical results, one can think of $\lambda = 0.01$. With this parameter setting, it can be observed that our upper bounds are almost tight. For example, one can see that the upper and lower bounds match to either logarithmic factors (when looking at the first upper bound), or a very small polynomial factor in $1/\alpha\varepsilon$ (when looking at the second upper bound). For our experimental results, we empirically determined an effective value for the parameter λ on a single synthetic instance. We then show that this choice of parameter generalizes, giving highly-accurate private estimation in other instances, on both synthetic and real-world data.

4. Algorithms and Analysis

We now prove our results for support coverage estimation, Theorem 1, while support size and entropy estimation appear in the supplementary material. We first describe and analyze our algorithms, and then go on to describe and analyze a lower bound construction, showing that our upper bounds are almost tight.

All our algorithms fall into the following simple framework:

1. Compute a non-private estimate of the property;
2. Privatize this estimate by adding Laplace noise, where the parameter is determined through analysis of the estimator and potentially computation of the estimator's sensitivity.

4.1. Support Coverage Estimation

4.1.1. Upper Bound for Support Coverage Estimation

We split the analysis into two regimes. First, we focus on the case where $m \leq \frac{1}{\alpha\varepsilon}$, and we prove the upper bound $O\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)$. Note that the problem is identical for any $\alpha < \frac{1}{m}$, since this corresponds to estimating the support coverage exactly, and the above bound simplifies to $O\left(m^2 + \frac{m}{\varepsilon}\right)$. The algorithm in this case is simple: since $n = \Omega(m)$, we group the dataset into n/m batches of size m. Let Y_j be the number of unique symbols observed in batch j. Our estimator is $\hat{S}_m(X_1^n) = \frac{m}{n}\sum_{j=1}^{n/m} Y_j$. Observe that $\mathbb{E}[Y_j] = S_m(p)$, and that $\mathrm{Var}[Y_j] \leq m$. The latter can be seen by observing that Y_j is the sum of m negatively correlated indicator random variables, each one being the indicator of whether that sample in the batch is the first time the symbol is observed. This gives that $\hat{S}_m(X_1^n)$ is an unbiased estimator of $S_m(p)$, with variance $O(m^2/n)$. By Chebyshev's inequality, since we want an estimate which is accurate up to $\pm\alpha m$, this gives us that $C_{\hat{S}_m}(S_m(p), \alpha/2) = O\left(\frac{1}{\alpha^2}\right)$. Furthermore, we can see that the sensitivity of $\hat{S}_m(X_1^n)$ is at most $2m/n$. By Lemma 1, there is a private algorithm for support coverage estimation as long as $\Delta\left(\frac{\hat{S}_m(X_1^n)}{m}\right) \leq \alpha\varepsilon$. With the above bound on sensitivity, this is true with $n = O(1/\alpha\varepsilon)$, giving the desired upper bound.

Now, we turn our attention to the case where $m \geq \frac{1}{\alpha\varepsilon}$, and we prove the upper bound $O\left(\frac{m\log(1/\alpha)}{\log m} + \frac{m\log(1/\alpha)}{\log(2+\varepsilon m)}\right)$. Let φ_i be the number of symbols that appear i times in X_1^n. We will use the following non-private support coverage estimator from (Orlitsky et al., 2016):

$$\hat{S}_m(X_1^n) = \sum_{i=1}^{n} \varphi_i\big(1 - (-t)^i \cdot \Pr(Z \geq i)\big),$$

where Z is a Poisson random variable with mean r (which

is a parameter to be instantiated later), and $t = (m-n)/n$.

Our private estimator of support coverage is derived by adding Laplace noise to this non-private estimator with the appropriate noise parameter, and thus the performance of our private estimator, is analyzed by bounding the sensitivity and the bias of this non-private estimator according to Lemma 1.

The sensitivity and bias of this estimator is bounded in the following lemmas.

Lemma 3. *Suppose $m > 2n$, then the maximum coefficient of φ_i in $\hat{S}_m(p)$ is at most $1 + e^{r(t-1)}$.*

Proof. By the definition of Z, we know $\Pr(Z \geq i) = \sum_{k=i}^{\infty} e^{-r} \frac{r^k}{k!}$, hence we have: $|1 + (-t)^i \cdot \Pr(Z \geq i)| \leq 1 + t^i \sum_{k=i}^{\infty} e^{-r} \frac{r^k}{k!} \leq 1 + e^{-r} \sum_{k=i}^{\infty} \frac{(rt)^k}{k!} \leq 1 + e^{-r} \sum_{k=0}^{\infty} \frac{(rt)^k}{k!} = 1 + e^{r(t-1)}$. □

The bias of the estimator is bounded in Lemma 4 of (Acharya et al., 2017a):

Lemma 4. *If $m > 2n$, then $\left|\mathbb{E}\left[\hat{S}_m(X_1^n)\right] - S_m(p)\right| \leq 2 + 2e^{r(t-1)} + \min(m, S(p)) \cdot e^{-r}$.*

Using these results, letting $r = \log(1/\alpha)$, (Orlitsky et al., 2016) showed that there is a constant C, such that with $n = C\frac{m}{\log m}\log(1/\alpha)$ samples, with probability at least 0.9, $\left|\frac{\hat{S}_m(X_1^n)}{m} - \frac{S_m(p)}{m}\right| \leq \alpha$.

Our upper bound in Theorem 1 is derived by the following analysis of the sensitivity of $\frac{\hat{S}_m(X_1^n)}{m}$. If we change one sample in X_1^n, at most two of the φ_j's change. Hence by Lemma 3, the sensitivity of the estimator satisfies $\Delta\left(\frac{\hat{S}_m(X_1^n)}{m}\right) \leq \frac{2}{m} \cdot \left(1 + e^{r(t-1)}\right)$. By Lemma 1, there is a private algorithm for support coverage estimation as long as $\Delta\left(\frac{\hat{S}_m(X_1^n)}{m}\right) \leq \alpha\varepsilon$, which, by the inequality above, holds if $2(1 + \exp(r(t-1))) \leq \alpha\varepsilon m$. Let $r = \log(3/\alpha)$, note that $t - 1 = \frac{m}{n} - 2$. Suppose $\alpha\varepsilon m > 2$, then, the condition above reduces to $\log\left(\frac{3}{\alpha}\right) \cdot \left(\frac{m}{n} - 2\right) \leq \log\left(\frac{1}{2}\alpha\varepsilon m - 1\right)$. This is equivalent to $n \geq \frac{m\log(3/\alpha)}{\log(\frac{1}{2}\alpha\varepsilon m - 1) + 2\log(3/\alpha)} = \frac{m\log(3/\alpha)}{\log(\frac{3}{2}\varepsilon m - 3/\alpha) + \log(3/\alpha)}$.

Suppose $\alpha\varepsilon m > 2$, then the condition above reduces to the requirement that $n = \Omega\left(\frac{m\log(1/\alpha)}{\log(2+\varepsilon m)}\right)$.

4.1.2. Lower Bound for Support Coverage Estimation

We now prove the lower bound described in Theorem 1. Note that the first term in the lower bound is the sample complexity of non-private support coverage estimation, shown in (Orlitsky et al., 2016). Therefore, we turn our attention to prove the last term in the sample complexity.

Consider the following two distributions. u_1 is uniform over $[m(1+\alpha)]$. u_2 is distributed over $m+1$ elements $[m] \cup \{\triangle\}$ where $u_2[i] = \frac{1}{m(1+\alpha)} \forall i \in [m]$ and $u_2[\triangle] = \frac{\alpha}{1+\alpha}$. Moreover, $\triangle \notin [m(1+\alpha)]$. Then, $S_m(u_1) = m(1+\alpha) \cdot \left(1 - \left(1 - \frac{1}{m(1+\alpha)}\right)^m\right)$, and $S_m(u_2) = m \cdot \left(1 - \left(1 - \frac{1}{m(1+\alpha)}\right)^m\right) + \left(1 - \left(1 - \frac{\alpha}{1+\alpha}\right)^m\right)$. Therefore, $S_m(u_2) - S_m(u_1) = m\alpha \cdot \left(1 - \left(1 - \frac{1}{m(1+\alpha)}\right)^m\right) - \left(1 - \left(1 - \frac{\alpha}{1+\alpha}\right)^m\right) = \Omega(\alpha m)$.

Hence we know there support coverage differs by $\Omega(\alpha m)$. Moreover, their total variation distance is $\frac{\alpha}{1+\alpha}$. The following lemma is folklore, based on the coupling interpretation of total variation distance, and the fact that total variation distance is subadditive for product measures.

Lemma 5. *For any two distributions p, and q, there is a coupling between n i.i.d. samples from the two distributions with an expected Hamming distance of $d_{\mathrm{TV}}(p,q) \cdot n$.*

Using Lemma 5 and $d_{\mathrm{TV}}(u_1, u_2) = \frac{\alpha}{1+\alpha}$, we have

Lemma 6. *Suppose u_1 and u_2 are as defined before, there is a coupling between u_1^n and u_2^n with expected Hamming distance equal to $\frac{\alpha}{1+\alpha}n$.*

Moreover, given n samples, we must be able to privately distinguish between u_1 and u_2 given an α accurate estimator of support coverage with privacy considerations. Thus, according to Lemma 2 and 6, we have $\frac{\alpha}{1+\alpha}n \geq \frac{1}{\varepsilon} \Rightarrow n = \Omega\left(\frac{1}{\varepsilon\alpha}\right)$.

5. Experiments

We evaluated our methods for entropy estimation and support coverage on both synthetic and real data. Overall, we found that privacy is quite cheap: private estimators achieve accuracy which is comparable or near-indistinguishable to non-private estimators in many settings. Our results on entropy estimation and support coverage appear in Sections 5.1 and 5.2, respectively. Code of our implementation is available at `https://github.com/HuanyuZhang/INSPECTRE`.

5.1. Entropy

We compare the performance of our entropy estimator with a number of alternatives, both private and non-private. Non-private algorithms considered include the plug-in estimator (`plug-in`), the Miller-Madow Estimator (`MM`) (Miller, 1955), the sample optimal polynomial approximation estimator (`poly`) of (Wu & Yang, 2016). We analyze the privatized versions of plug-in, and `poly` in the supplementary material. The implementation of the latter is based on

code from the authors of (Wu & Yang, 2016)[1]. We compare performance on different distributions including uniform, a distribution with two steps, Zipf(1/2), a distribution with Dirichlet-1 prior, and a distribution with Dirichlet-$1/2$ prior, and over varying support sizes.

While `plug-in`, and `MM` are parameter free, `poly` (and its private counterpart) have to choose the degree L of the polynomial to use, which manifests in the parameter λ in the statement of Theorem 3. (Wu & Yang, 2016) suggests the value of $L = 1.6 \log k$ in their experiments. However, since we add further noise, we choose a single L as follows: (i) Run privatized `poly` for different L values and distributions for $k = 2000$, $\varepsilon = 1$, (b) Choose the value of L that performs well across different distributions (See Figure 1). We choose $L = 1.2 \cdot \log k$ from this, and use it for all other experiments. To evaluate the sensitivity of `poly`, we computed the estimator's value at all possible input values, computed the sensitivity, (namely, $\Delta = \max_{d_{ham}(X_1^n, Y_1^n) \leq 1} |\texttt{poly}(X_1^n) - \texttt{poly}(Y_1^n)|$), and added noise distributed as Lap $\left(0, \frac{\Delta}{\varepsilon}\right)$.

The RMSE of various estimators for $k = 1000$, and $\varepsilon = 1$ for various distributions are illustrated in Figure 2. The RMSE is averaged over 100 iterations in the plots.

We observe that the performance of our private-`poly` is near-indistinguishable from the non-private `poly`, particularly as the number of samples increases. It also performs significantly better than all other alternatives, including the non-private Miller-Madow and the plug-in estimator. The cost of privacy is minimal for several other settings of k and ε, additional experiments appear in the supplementary material.

5.2. Support Coverage

We investigate the cost of privacy for the problem of support coverage. We provide a comparison between the Smoothed Good-Toulmin estimator (SGT) of (Orlitsky et al., 2016) and our algorithm, which is a privatized version of their statistic (see Section 4.1.1). Our implementation is based on code provided by the authors of (Orlitsky et al., 2016). As shown in our theoretical results, the sensitivity of SGT is at most $2(1 + e^r(t-1))$, necessitating the addition of Laplace noise with parameter $2(1 + e^{r(t-1)})/\varepsilon$. Note that while the theory suggests we select the parameter $r = \log(1/\alpha)$, α is unknown. We instead set $r = \frac{1}{2t} \log_e \frac{n(t+1)^2}{t-1}$, as previously done in (Orlitsky et al., 2016).

[1]See `https://github.com/Albuso0/entropy` for their code for entropy estimation.

5.2.1. Evaluation on Synthetic Data

In our synthetic experiments, we consider different distributions over different support sizes k. We generate $n = k/2$ samples, and then estimate the support coverage at $m = n \cdot t$. For large t, estimation is harder. Some results of our evaluation on synthetic are displayed in Figure 3. We compare the performance of SGT, and privatized versions of SGT with parameters $\varepsilon = 1, 2,$ and 10. For this instance, we fixed the domain size $k = 20000$. We ran the methods described above with $n = k/2$ samples, and estimated the support coverage at $m = nt$, for t ranging from 1 to 10. The performance of the estimators is measured in terms of RMSE over 1000 iterations.

We observe that, in this setting, the cost of privacy is relatively small for reasonable values of ε. This is as predicted by our theoretical results, where unless ε is extremely small (less than $1/k$) the non-private sample complexity dominates the privacy requirement. However, we found that for smaller support sizes (as shown in the supplementary material), the cost of privacy can be significant. We provide an intuitive explanation for why no private estimator can perform well on such instances. To minimize the number of parameters, we instead argue about the related problem of support-size estimation. Suppose we are trying to distinguish between distributions which are uniform over supports of size 100 and 200. We note that, if we draw $n = 50$ samples, the "profile" of the samples (i.e., the histogram of the histogram) will be very similar for the two distributions. In particular, if one modifies only a few samples (say, five or six), one could convert one profile into the other. In other words, these two profiles are almost-neighboring datasets, but simultaneously correspond to very different support sizes. This pits the two goals of privacy and accuracy at odds with each other, thus resulting in a degradation in accuracy.

5.2.2. Evaluation on Census Data and Hamlet

We conclude with experiments for support coverage on two real-world datasets, the 2000 US Census data and the text of Shakespeare's play Hamlet, inspired by investigations in (Orlitsky et al., 2016) and (Valiant & Valiant, 2017). Our investigation on US Census data is also inspired by the fact that this is a setting where privacy is of practical importance, evidenced by the proposed adoption of differential privacy in the 2020 US Census (Dajani et al., 2017).

The Census dataset contains a list of last names that appear at least 100 times. Since the dataset is so oversampled, even a small fraction of the data is likely to contain almost all the names. As such, we make the task non-trivial by subsampling $m_{total} = 86080$ individuals from the data, obtaining 20412 distinct last names. We then sample n of the m_{total} individuals without replacement and attempt to

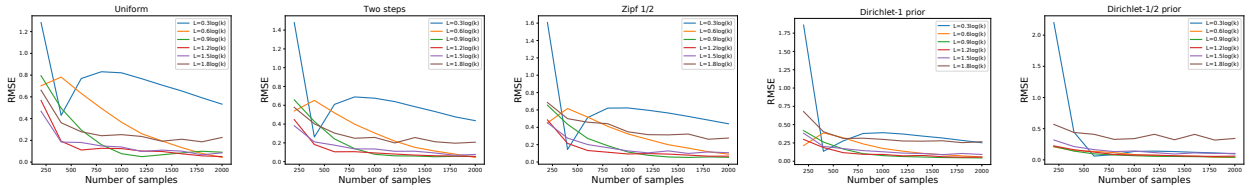

Figure 1. RMSE comparison between private Polynomial Approximation Estimators for entropy with various values for degree L, $k = 2000$, $\varepsilon = 1$. The degree L represents a bias-variance tradeoff: a larger degree decreases the bias but increases the sensitivity, necessitating the addition of Laplace noise with a larger variance.

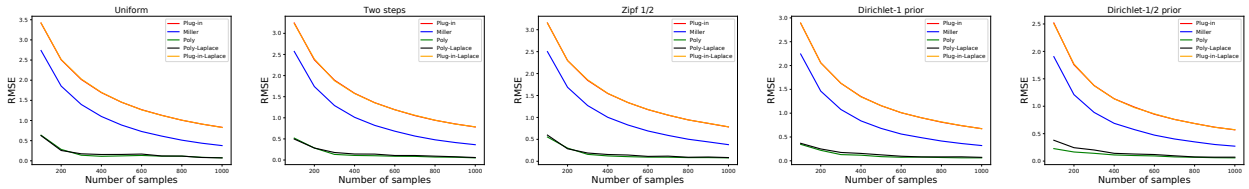

Figure 2. Comparison of various estimators for entropy estimation, $k = 1000$, $\varepsilon = 1$.

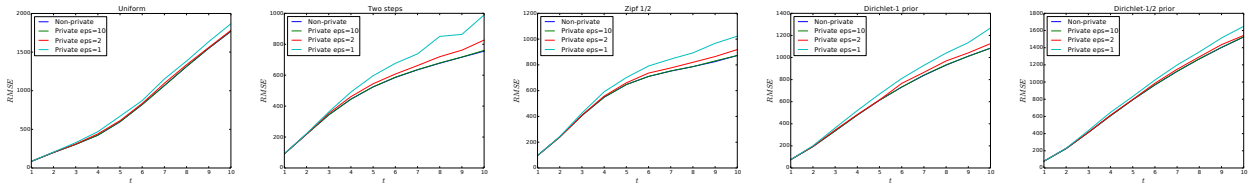

Figure 3. Comparison between the private support coverage estimator with the non-private SGT when $k = 20000$

estimate the total number of last names. Figure 4 displays the RMSE over 100 iterations of this process. We observe that even with an exceptionally stringent privacy budget of $\varepsilon = 0.5$, the performance is almost indistinguishable from the non-private SGT estimator.

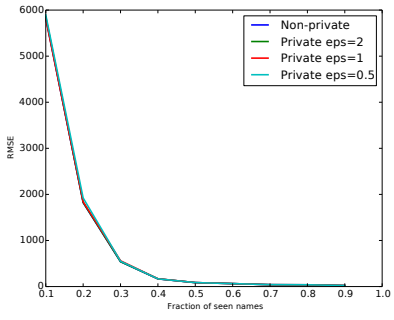

Figure 4. Comparison between our private support coverage estimator with the SGT on Census Data.

The Hamlet dataset has $m_{total} = 31,999$ words, of which 4804 are distinct. Since the distribution is not as oversampled as the Census data, we do not need to subsample the data. Besides this difference, the experimental setup is identical to that of the Census dataset. Once again, as we can see in Figure 5, we get near-indistinguishable performance between the non-private and private estimators, even for very small values of ε. Our experimental results demonstrate that privacy is realizable in practice, with particularly accurate performance on real-world datasets.

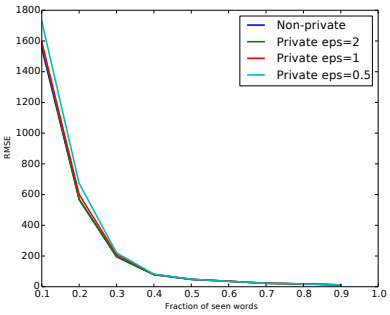

Figure 5. Comparison between our private support coverage estimator with the SGT on Hamlet.

Acknowledgements

JA, ZS, and HZ are supported by NSF CCF-1657471 and a Cornell University startup grant. GK is supported by ONR N00014-12-1-0999, NSF CCF-1617730, CCF-1650733, and CCF-1741137. Work partially done while author was an intern at Microsoft Research, New England.

References

Acharya, J., Daskalakis, C., and Kamath, G. Optimal testing for properties of distributions. In *NIPS '15*, pp. 3577–3598, 2015.

Acharya, J., Das, H., Orlitsky, A., and Suresh, A. T. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *ICML '17*, 2017a.

Acharya, J., Orlitsky, A., Suresh, A. T., and Tyagi, H. Estimating rényi entropy of discrete distributions. *IEEE Transactions on Information Theory*, 63(1):38–56, 2017b.

Acharya, J., Sun, Z., and Zhang, H. Differentially private testing of identity and closeness of discrete distributions. *arXiv preprint arXiv:1707.05128*, 2017c.

Acharya, J., Sun, Z., and Zhang, H. Communication efficient, sample optimal, linear time locally private discrete distribution estimation. *arXiv preprint arXiv:1802.04705*, 2018.

Adam, N. R. and Worthmann, J. C. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys (CSUR)*, 21(4):515–556, 1989.

Agrawal, D. and Aggarwal, C. C. On the design and quantification of privacy preserving data mining algorithms. In *PODS '01*, 2001.

Aliakbarpour, M., Diakonikolas, I., and Rubinfeld, R. Differentially private identity and closeness testing of discrete distributions. *arXiv preprint arXiv:1707.05497*, 2017.

Balcan, M.-F., Dick, T., Liang, Y., Mou, W., and Zhang, H. Differentially private clustering in high-dimensional euclidean spaces. In *ICML '17*, 2017.

Berry, M. J., Warland, D. K., and Meister, M. The structure and precision of retinal spike trains. *Proceedings of the National Academy of Sciences*, 94(10):5411–5416, 1997.

Cai, B., Daskalakis, C., and Kamath, G. Priv'it: Private and sample efficient identity testing. In *ICML '17*, pp. 635–644, 2017.

Chaudhuri, K., Sarwate, A. D., and Sinha, K. A near-optimal algorithm for differentially-private principal components. *Journal of Machine Learning Research*, 14(Sep):2905–2943, 2013.

Colwell, R. K. et al. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5(1):3–21, 2012.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. Wiley, 2006.

Dajani, A. N. et al. The modernization of statistical disclosure limitation at the U.S. census bureau, 2017. Presented at the September 2017 meeting of the Census Scientific Advisory Committee.

Dalenius, T. Towards a methodology for statistical disclosure control. *Statistisk Tidskrift*, 15:429–444, 1977.

Diakonikolas, I., Hardt, M., and Schmidt, L. Differentially private learning of structured discrete distributions. In *NIPS '15*, 2015.

Differential Privacy Team, Apple. Learning with privacy at scale, December 2017.

Dinur, I. and Nissim, K. Revealing information while preserving privacy. In *PODS '03*, 2003.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Minimax optimal procedures for locally private estimation. *JASA*, 2017.

Dwork, C. Differential privacy: A survey of results. In *TAMC '08*, 2008.

Dwork, C. and Roth, A. *The Algorithmic Foundations of Differential Privacy*. Now Publishing, Inc., 2014.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *TCC '06*, 2006.

Erlingsson, Ú., Pihur, V., and Korolova, A. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *CCS '14*, 2014.

Fukuchi, K. and Sakuma, J. Minimax optimal estimators for additive scalar functionals of discrete distributions. In *ISIT '17*, 2017.

Gaboardi, M. and Rogers, R. Local private hypothesis testing: Chi-square tests. *arXiv preprint arXiv:1709.07155*, 2017.

Gaboardi, M., Lim, H., Rogers, R. M., and Vadhan, S. P. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *ICML '16*, 2016.

Good, I. and Toulmin, G. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

Hardt, M. and Price, E. The noisy power method: A meta algorithm with applications. In *NIPS '14*, 2014.

Homer, N. et al. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genetics*, 4(8):1–9, 2008.

Jiao, J., Venkat, K., Han, Y., and Weissman, T. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2017.

Johnson, A. and Shmatikov, V. Privacy-preserving data exploration in genome-wide association studies. In *KDD '13*, 2013.

Kakizaki, K., Sakuma, J., and Fukuchi, K. Differentially private chi-squared test by unit circle mechanism. In *ICML '17*, 2017.

Kamath, G., Li, J., Singhal, V., and Ullman, J. Privately learning high-dimensional distributions. *arXiv preprint arXiv:1805.00216*, 2018.

Kapralov, M. and Talwar, K. On differentially private low rank approximation. In *SODA '13*, 2013.

Karwa, V. and Vadhan, S. Finite sample differentially private confidence intervals. In *ITCS '18*, 2018.

Keinan, A. and Clark, A. G. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, 336(6082):740–743, 2012.

Kifer, D. and Rogers, R. M. A new class of private chi-square tests. In *AISTATS '17*, 2017.

Miller, G. A. Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods*, 2:95–100, 1955.

Nelson, M. R. et al. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*, 337 (6090):100–104, 2012.

Nemenman, I., Bialek, W., and de Ruyter van Steveninck, R. Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69(5):056111:1–056111:6, 2004.

Nowozin, S. Improved information gain estimates for decision tree induction. In *ICML '12*, 2012.

Obremski, M. and Skorski, M. Renyi entropy estimation revisited. In *APPROX '17*, 2017.

Orlitsky, A., Suresh, A. T., and Wu, Y. Optimal prediction of the number of unseen species. *PNAS*, 113(47):13283–13288, 2016.

Paninski, L. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

Raskhodnikova, S., Ron, D., Shpilka, A., and Smith, A. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39 (3):813–842, 2009.

Rogers, R. M. *Leveraging Privacy in Data Analysis*. PhD thesis, University of Pennsylvania, May 2017.

Sheffet, O. Differentially private ordinary least squares. In *ICML '17*, 2017.

Simmons, S., Sahinalp, C., and Berger, B. Enabling privacy-preserving GWASs in heterogeneous human populations. *Cell Systems*, 3(1):54–61, 2016.

Tennessen, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, 337(6090):64–69, 2012.

Uhler, C., Slavković, A., and Fienberg, S. E. Privacy-preserving data sharing for genome-wide association studies. *The Journal of Privacy and Confidentiality*, 5(1):137–166, 2013.

Valiant, G. and Valiant, P. Estimating the unseen: Improved estimators for entropy and other properties. In *NIPS '13*, 2013.

Valiant, G. and Valiant, P. Estimating the unseen: Improved estimators for entropy and other properties. *Journal of the ACM*, 64(6):37:1–37:41, 2017.

Wang, Y., Lee, J., and Kifer, D. Revisiting differentially private hypothesis tests for categorical data. *arXiv preprint arXiv:1511.03376*, 2015a.

Wang, Y., Wang, Y.-X., and Singh, A. Differentially private subspace clustering. In *NIPS '15*, 2015b.

Wu, Y. and Yang, P. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

Wu, Y. and Yang, P. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 2018.

Yu, F. et al. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50:133–141, 2014.

Zou, J. et al. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nature Communications*, 7, 2016.