Differentially Private Identity and Equivalence Testing of Discrete Distributions

Maryam Aliakbarpour [1] **Ilias Diakonikolas** [2] **Ronitt Rubinfeld** [1 3]

Abstract

We study the fundamental problems of identity and equivalence testing over a discrete population from random samples. Our goal is to develop efficient testers while guaranteeing differential privacy to the individuals of the population. We provide sample-efficient differentially private testers for these problems. Our theoretical results significantly improve over the best known algorithms for identity testing, and are the first results for private equivalence testing. The conceptual message of our work is that there exist private hypothesis testers that are nearly as sample-efficient as their non-private counterparts. We perform an experimental evaluation of our algorithms on synthetic data. Our experiments illustrate that our private testers achieve small type I and type II errors with sample size *sublinear* in the domain size of the underlying distributions.

1. Introduction

We consider the problem of designing sample-efficient algorithms to understand properties of distributions over large discrete domains. Such statistical tests have been traditionally studied in statistics because of their importance in virtually every scientific endeavor that involves data. Recent work in the theoretical computer science community has investigated the setting where the discrete domains are large and no a priori assumptions can be made about the underlying data distribution (for example, when it cannot be assumed that the distribution is normal, Gaussian, or even smooth). In the last few years, optimal methods with sublinear sample complexity have been obtained for testing a range of properties, including whether a distribution is uniform, identical to a known distribution (testing "goodness-of-fit"), equivalence of two distributions (two sample testing), and independence.

While statistical tests are very important for advancing science, when they are performed on sensitive data representing specific individuals, such as data describing medical or other behavioral phenomena, it may be that the outcomes of the tests reveal private information that should not be divulged. Techniques from differential privacy give us hope that one may obtain the scientific benefit of statistical tests without compromising the privacy of the individuals in the study. Concretely, differential privacy requires that similar datasets have statistically close outputs – once this property is achieved, then provable privacy guarantees can be made. Differential privacy is a rich and active area of study, in which techniques have been developed and applied to obtain private algorithms for a range of data analysis tasks.

Our Contributions We study the general problem of hypothesis testing in the setting of differential privacy (Dwork & Roth, 2014). Our emphasis is on the *sublinear* regime, i.e., when the number of samples available is sublinear in the domain size of the underlying distribution(s). We obtain sample-efficient private algorithms for the problems of testing the identity and equivalence of discrete distributions. The main conceptual message of our work is that we can achieve differential privacy with only a small increase in the sample complexity compared to the non-private case. Our theoretical results significantly improve over the best known algorithms for identity testing, and are the first results for private equivalence testing. Our experimental evaluation illustrates that our testers achieve small type I and type II errors with a sublinear number of samples when the domain size is large. The sample complexity of our private identity tester significantly outperforms the sample complexity of recently proposed methods for this problem. For both identity and equivalence testing, our experiments show that differential privacy can be achieved essentially for free, i.e., with a very mild increase in sample complexity.

Technical Overview We now provide a brief overview of our techniques. We start by observing that there is a simple generic method to convert a non-private tester into a private tester with a *multiplicative* overhead in the sample complexity. This method is known in differential privacy, but for the sake of completeness we describe it in Appendix B. It will

[*]Equal contribution [1]CSAIL, MIT, Cambridge, MA 02139, USA [2]Department of Computer Science, USC, Los Angeles, CA 90089, USA [3]TAU, Tel Aviv-Yafo, Israel. Correspondence to: Maryam Aliakbarpour <maryama@mit.edu>, Ilias Diakonikolas <diakonik@usc.edu>, Ronitt Rubinfeld <ronitt@csail.mit.edu>.

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

be useful to contrast the sample complexity of the generic method with the (substantially better) sample complexity of our testers (Sections 3-5). For convenience, throughout our theoretical analysis, we obtain testing algorithms that have failure probability at most $1/3$. As shown in Appendix C, this is without loss of generality: we can achieve error probability δ at the expense of a $\log(1/\delta)$ multiplicative increase in the sample complexity, even in the differentially private setting.

Our algorithm for identity testing is obtained via the following modular approach: First, we adapt a recently discovered reduction of identity testing to uniformity testing (Goldreich, 2016), building on (Diakonikolas & Kane, 2016). We show (Section 3) that this reduction can be adapted to work in the private setting as well. Therefore, we can translate any private uniformity tester to a private identity tester without increasing the sample size by more than a constant factor. It remains to develop sample-efficient private uniformity testers. We develop two such private methods (Section 4): Our first method is a private version of (Paninski, 2008), which relies on the number of domain elements that appear in the sample exactly once. This statistic has low sensitivity, allowing a translation to the private setting via standard techniques. The sample complexity of our aforementioned uniformity tester is

$$\Theta(\sqrt{n}/\epsilon^2 + \sqrt{n}/(\epsilon\sqrt{\xi})) \,, \tag{1}$$

where n is the domain size, ϵ is the accuracy of the tester, and ξ is the privacy parameter. Our experimental results illustrate that this private tester performs exceptionally well in practice, significantly outperforming recently proposed private algorithms for identity testing (see Section 6).

We note that the uniformity tester of (Paninski, 2008) is known to completely fail when the sample size is larger than the domain size (even in the non-private setting). To obtain a uniformity tester that works for the non-sparse regime, we develop our second algorithm: a private version of the collisions-based tester of (Goldreich & Ron, 2000). A collision refers to the event that two random samples drawn from the underlying distribution correspond to the same domain element. The collisions-based tester was recently shown to be sample-optimal in the non-private setting (Diakonikolas et al., 2016). The main difficulty in turning this non-private tester into a private tester is that the underlying statistic (number of collisions) has very high worst-case sensitivity. Hence, the standard approach of adding Laplace noise to the statistic fails in this setting. To overcome this obstacle, we add an appropriate pre-processing step to our tester that rejects when there is a single element that appears many times in the sample. (We note that a similar idea was independently used in (Cai et al., 2017), though the details are somewhat different.) This step allows us to reduce the *effective sensitivity* of the statistic and can be shown to yield a sample-efficient private tester. Specifically, the sample complexity of our collisions-based private tester is

$$\tilde{O}\left(\sqrt{n}/\epsilon^2 + \sqrt{n}/(\epsilon\xi) + 1/(\epsilon^2\xi)\right) \,. \tag{2}$$

For the problem of equivalence testing, we build on the recently developed chi-square tester of (Chan et al., 2014), which is sample-optimal in the non-private testing. We show that this statistic has bounded sensitivity. Hence, developing a sample-efficient private version can be achieved by adding Laplace noise. A careful analysis shows that the noisy statistic is still accurate without substantially increasing the sample complexity. Specifically, the sample complexity of our private equivalence tester is

$$O\left(\sqrt{n}/\epsilon^2 + n^{2/3}/\epsilon^{4/3} + \sqrt{n}/(\sqrt{\xi}\epsilon) + 1/(\xi\epsilon^2)\right) \,. \tag{3}$$

We note that the effect of the privacy constraint on the sample complexity of our testers is in some sense *additive*, as opposed to multiplicative. In particular, for each case, the sample complexity of the private tester equals the sample complexity of the corresponding non-private tester plus a term that depends on the privacy parameter. For reasonable settings of the privacy parameter, this additive term can be negligible compared to the first term, in which case we obtain differential privacy essentially for free. As a concrete example, the second term in (1) is dominated by the first term (which is provably necessary for any identity tester, even in the non-private setting (Paninski, 2008)), as long as $\xi \geq \epsilon^2$. This phenomenon is confirmed in our experimental evaluation.

Related Work During the past two decades, *distribution property testing* (Batu et al., 2000) – whose roots lie in statistical hypothesis testing (Neyman & Pearson, 1933; Lehmann & Romano, 2005) – has received considerable attention by the computer science community, see (Rubinfeld, 2012; Canonne, 2015) for two recent surveys. The majority of the early work in this field has focused on characterizing the sample size needed to test properties of arbitrary distributions of a given support size. After two decades of study, this "worst-case" regime is well-understood: for many properties of interest there exist sample-optimal testers (matched by information-theoretic lower bounds) (Paninski, 2008; Daskalakis et al., 2013; Chan et al., 2014; Valiant & Valiant, 2014; Diakonikolas et al., 2015b;c; Acharya et al., 2015; Diakonikolas & Kane, 2016; Canonne et al., 2016; Diakonikolas et al., 2016; Canonne et al., 2017; Diakonikolas et al., 2017b).

A recent line of work (Wang et al., 2015; Gaboardi et al., 2016; Kifer & Rogers, 2017; Kakizaki et al., 2017; Cai et al., 2017) has studied distribution testing with privacy constraints. The majority of these works (Wang et al., 2015; Gaboardi et al., 2016; Kifer & Rogers, 2017) only obtain

type I error guarantees subject to the privacy constraint, which is a significantly weaker guarantee than ours. The recent work by Cai *et al.* (Cai et al., 2017) provides an identity tester with finite sample guarantees and bounded type I and type II errors. Specifically, (Cai et al., 2017) give a private identity tester with sample complexity

$$\tilde{O}\left(\sqrt{n}/\epsilon^2 + \sqrt{n}/(\epsilon^{3/2}\xi) + n^{1/3}/(\epsilon^{5/3}\xi^{2/3})\right) . \quad (4)$$

The above bound is always asymptotically worse than (1) and worse than (2) for most parameter settings. We remind the reader that (Cai et al., 2017) does not consider the more general problem of equivalence testing, and that ours are the first results in this setting. Finally, (Diakonikolas et al., 2015a) has provided differentially private algorithms for learning various families of discrete distributions. For the case of unstructured discrete distributions, as the ones considered here, such algorithms inherently require sample size at least linear in the domain size, even for constant values of the approximation parameter.

Independent and concurrent work (Acharya et al., 2017) obtained similar upper bounds for private identity and closeness testing. In addition, they obtained nearly matching sample lower bounds in some regimes.

2. Preliminaries

Notation and Basic Definitions. We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. We say that p is a distribution over $[n]$ if $p : [n] \to [0, 1]$ is a function such that $\sum_{i=1}^{n} p(i) = 1$, where $p(i)$ denotes the probability of element i. For a set $S \subseteq [n]$, $p(S)$ denotes the total probability of the elements in S (i.e., $p(S) = \sum_{i \in S} p(i)$). For any integer $k > 0$, the ℓ^k-norm of p is equal to $\left(\sum_{i=1}^{n} |p(i)|^k\right)^{\frac{1}{k}}$, and it is denoted by $\|p\|_k$. The ℓ^k-distance between two distributions p and q over $[n]$ is equal to $\left(\sum_{i=1}^{n} |p(i) - q(i)|^k\right)^{\frac{1}{k}}$. We use $\mathbf{Lap}(b)$ to denote a random variable that is drawn from a Laplace distribution with parameter b and mean zero.

The problem of *identity testing* (or goodness-of-fit) is the following: Given sample access to an unknown distribution p over $[n]$ and an explicit distribution q over $[n]$, we want to distinguish, with probability at least $2/3$, between the cases that $p = q$ (completeness) and $\|p - q\|_1 \geq \epsilon$ (soundness). (If $\|p - q\|_1 \geq \epsilon$, we will say that p and q are ϵ-far from each other.) The special case of this problem when $q = U_n$, the uniform distribution over $[n]$, is called *uniformity testing*. The generalization of identity testing when both p and q are unknown and only accessible via samples is called *equivalence testing.*

Differential Privacy. In our context, a dataset is a multiset of samples drawn from a distribution over $[n]$. We say that X and Y are *neighboring datasets* if they differ in exactly one element.

Definition 2.1. A randomized algorithm $\mathcal{A} : [n]^s \to \mathcal{R}$, is *$\xi$-differentially private if for any $S \subseteq \mathcal{R}$ and any neighboring datasets X, Y, we have that* $\mathbf{Pr}[\mathcal{A}(X) \in S] \leq e^{\xi} \cdot \mathbf{Pr}[\mathcal{A}(Y) \in S]$.

We will say that a tester is (ϵ, ξ)-private, to mean that ϵ is the accuracy parameter, ξ is the privacy parameter, and the tester outputs the right answer with probability at least $2/3$[1]. For conciseness, we use the term ξ-private instead of ξ-differentially private. We provide more details about general techniques in differential privacy in Appendix A.

3. Private Identity Testing: Reduction to Private Uniformity Testing

In this section, we provide a reduction of private identity testing (against a fixed distribution) to its special case of private uniformity testing. Specifically, we prove that a recent reduction (Goldreich, 2016) of (non-private) identity testing to (non-private) uniformity testing can be adapted to work in the private setting as well.

Suppose we want to test identity between an unknown distribution p over $[n]$ and an explicit distribution q. The reduction of (Goldreich, 2016) transforms the distribution p into a new distribution p', over a domain of size $O(n)$, such that if $p = q$ then p' is the uniform distribution, and if p is far from q, p' is also far from uniform. Specifically, the reduction defines a randomized mapping of a sample $i \in [n]$ from p to a sample (j, a) from p' that depends only on the explicit distribution q. This property is crucial as it allows us to show that the transformation preserves differential privacy, as the following theorem states:

Theorem 3.1. *Given an (ϵ, ξ)-private uniformity tester using $s(n, \epsilon, \xi)$ samples, there exists an (ϵ, ξ)-private tester for identity using $s = s(6n, \epsilon/3, \xi)$ samples.*

The detailed proof of the theorem is deferred to Appendix D.

4. Private Uniformity Testing

In this section, we provide two sample-efficient private uniformity testers. Our testers are private versions of two well-studied (non-private) testers, due to Goldreich and Ron (Goldreich & Ron, 2000) and Paninski (Paninski, 2008). Paninski's uniformity tester (Paninski, 2008) relies on the number of unique elements in the sample, while (Goldreich & Ron, 2000) relies on the number of collisions. Both testers are known to be sample-optimal in the non-private setting (Paninski, 2008; Diakonikolas et al., 2016).

[1] We emphasize that the confidence probability 2/3 can be increased to $1-\delta$ at the expense of a $\log(1/\delta)$ multiplicative increase in the sample complexity. See Appendix C.

Algorithm 1 Private Uniformity Testing via Unique Elements: Private-Unique-Elements-Uniformity

Input: Sample access to p, n, ϵ, ξ
$s \leftarrow 5\sqrt{n}/(\epsilon\sqrt{\xi}) + 6\sqrt{n}/\epsilon^2$
$C \leftarrow \frac{s\epsilon^2}{\sqrt{n}}$
$x_1, x_2, \ldots, x_s \leftarrow s$ samples drawn from p
$K \leftarrow$ the number of unique elements in $\{x_1, x_2, \ldots, x_s\}$
$K' \leftarrow K + \mathbf{Lap}(2/\xi)$
$T \leftarrow \mathbf{E}_{\mathcal{U}}[K] - C^2/(2\epsilon^2)$ {where $\mathbf{E}_{\mathcal{U}}[K]$ equals to $s \cdot \left(1 - \frac{1}{n}\right)^{s-1}$}
if $K' < T$ **then**
    Output reject.
end if
Output accept.

We give private versions of both of these algorithms. The sample complexity of our private Paninski uniformity tester is $O(\sqrt{n}/\epsilon^2 + \sqrt{n}/(\epsilon\sqrt{\xi}))$. Therefore, as long as $\xi = \Omega(\epsilon^2)$, the privacy requirement increases the sample complexity by at most a constant factor.

Unfortunately, the aforementioned tester only succeeds when its sample size is smaller than the domain size n. To be able to handle the entire range of parameters, we develop a private version of the collisions-based tester from (Goldreich & Ron, 2000). Our private version of the collisions tester has sample complexity $\tilde{O}\left(\sqrt{n}/\epsilon^2 + \sqrt{n}/(\epsilon\xi) + 1/(\epsilon^2\xi)\right)$. Similarly, the effect of the privacy is mild as long as $\xi = \Omega(\epsilon)$.

4.1. Private Uniformity Tester via Unique Elements

We provide a private tester for uniformity based on the number of unique elements. The number of unique elements is (negatively) related to the number of collisions and the ℓ^2-norm of the distribution. Therefore, the greater the number of unique elements is, the more the distribution appears uniform. To make the algorithm private, we use the Laplace mechanism which adds a small amount of noise to the number of unique elements. Then, we compare the number of unique elements with a threshold to decide if the distribution is uniform or far from uniform. The noise rate is chosen appropriately so that the following conflicting goals are simultaneously achieved: (1) the algorithm is guaranteed to be private, and (2) the accuracy of the tester does not significantly decrease. This is formalized in Theorem 4.1. The algorithm is described in the following pseudocode:

Theorem 4.1. *Given $s = O(\sqrt{n}/(\epsilon\sqrt{\xi}) + \sqrt{n}/\epsilon^2)$ samples from a distribution p over $[n]$, Algorithm 1 is an (ϵ, ξ)-private uniformity tester, if s is sufficiently smaller than the domain size n.*

Let K be the number of unique elements in the sample set. Since changing one sample in the sample set can change the number of unique elements by no more than two, adding Laplace noise with parameter $2/\xi$ to K makes it ξ-private. Using the composition theorem of differential privacy, we conclude that the overall algorithm is ξ-private. To show that the algorithm is an ϵ-tester, we prove that the statistic K' concentrates well around its expectation in both the completeness and soundness cases. To establish this, we exploit the fact that the variance introduced by the added noise is sufficiently small. Since there is a non-trivial gap between the expected values of K' in the two cases, the proof follows by an application of Chebyshev's inequality. See Appendix E for the formal details.

4.2. Private Uniformity Tester via Collisions

In this subsection, we describe the private version of our collisions-based uniformity tester. Recall that a collision refers to the event that two random samples drawn from the underlying distribution correspond to the same domain element. The main difficulty in turning this into a private tester is that the underlying statistic (number of collisions) has very high worst-case sensitivity. Specifically, if the sample set contains s copies of a given domain element, by changing just one of the copies to another element, the number of collisions drops by an additive s. So, if we add enough noise to the statistic to cover the sensitivity of s, the tester accuracy substantially degrades.

To overcome this obstacle, we add a pre-processing step to our tester. We notice that the sensitivity of the number of collisions, $f(X)$, for sample set X, depends on the maximum frequency of any element in the sample set. Let $n_i(X)$ denote the number of occurrences of element i in the sample set X, and let $n_{\max}(X)$ denote the maximum $n_i(X)$. We note that for two neighboring sample sets X and Y, the difference of the number of collisions, $|f(X) - f(Y)|$, is at most $n_{\max}(X)$. Therefore, the sensitivity of f is high on X's with large $n_{\max}(X)$. If the underlying distribution is uniform, we do not expect any particular element to show up very frequently. Hence, if $n_{\max}(X)$ is high, the algorithm can output reject regardless of $f(X)$. So, the final output of the algorithm does not change drastically on X and Y, while the number of collisions varies a lot.

This simple idea forms the basis for our modified tester. The algorithm uses two statistics: $n_{\max}$ and f (or more precisely the noisy version of them, $\hat{n}_{\max}$ and $\hat{f}$). If $\hat{n}_{\max}$ is too large, it outputs reject. Otherwise, $\hat{f}(X)$ determines the output. In the second case, since $n_{\max}$ is not too large, f has bounded sensitivity. Therefore, we can make it private by adding a small amount of noise to it.

To prove the privacy guarantee, note that if $f(X)$ has low-sensitivity, then $\hat{f}(X)$ is easily seen to be private. By the

Algorithm 2 Private uniformity tester based on the number of collisions: Private-Collisions-Uniformity

Input: Sample access to p, n, ϵ, ξ
$s \leftarrow \Theta\left(\frac{\sqrt{n}}{\epsilon^2} + \frac{\sqrt{n \log n}}{\epsilon\, \xi^{1/2}} + \frac{\sqrt{n \max(1, \log 1/\xi)}}{\epsilon\, \xi} + \frac{1}{\epsilon^2\, \xi}\right)$.
Let $X = \{x_1, x_2, \ldots, x_s\}$ be a multiset of s samples drawn from p
$n_i(X) \leftarrow |\{j | x_j \in x \text{ and } x_j = i\}|$
$n_{\max}(X) \leftarrow \max_i n_i(X)$
$\hat{n}_{\max}(X) \leftarrow n_{\max}(X) + \mathbf{Lap}(2/\xi)$
$f(X) \leftarrow \text{collisions}(X)$
$\eta_f \leftarrow \max\left(\frac{3s}{2n}, 12\, e^2 \ln 24\, n\right) + (2 \ln 12)/\xi + 2 \max(\ln 3, \ln 3/\xi)/\xi$
$T \leftarrow \max\left(\frac{3s}{2n}, 12\, e^2 \ln 24\, n\right) + (2 \ln 12)/\xi$
$\hat{f}(X) \leftarrow f(X) + \mathbf{Lap}(2\, \eta_f/\xi)$
if $\hat{n}_{\max}(X) < T \;\;\&\;\; \hat{f}(X) < \frac{6+\epsilon^2}{6n}\binom{s}{2}$ **then**
    $O \leftarrow$ accept.
else
    $O \leftarrow$ reject.
end if
With probability 1/6, $O \leftarrow \{\text{accept}, \text{reject}\} \setminus O$. ⟨⟨*flip the answer with probability 1/6.*⟩⟩
Output O.

composition theorem of differential privacy (Lemma A.3), in this case the overall algorithm will be private. The difficulty appears in the complementary case, i.e., when $f(X)$ is highly sensitive. In this case, $n_{\max}(X)$ has to be large. From that we can deduce, that it is very unlikely (over the random noise) that $\hat{n}_{\max}$ is small. Given the above and the fact that our algorithm flips its answer with probability 1/6 in the last step, we can compute a closed form formula for the probability that our tester accepts, which allows us to directly prove the privacy guarantee. With a similar argument, we show the privacy guarantee holds for the case that our tester rejects.

The detailed procedure is explained in Algorithm 2. We have the following (see Appendix F for the proof):

Theorem 4.2. *Algorithm 2 is an (ϵ, ξ)-private tester for uniformity.*

Remark 4.3. Recent work (Diakonikolas et al., 2017a) has obtained an optimal (non-private) uniformity tester based on the ℓ^1-distance of the empirical distribution from the uniform distribution. Since this new tester is Lipschitz, we can make it private by adding Laplace noise to the distribution. The simple proof of this fact will appear in a revised version of this paper. A similar observation was made independently in (Acharya et al., 2017).

5. Private Equivalence Testing

In this section, we give a private algorithm for testing equivalence of two unknown discrete distributions. Our tester relies on the chi-squared type sample-optimal (non-private) equivalence tester of (Chan et al., 2014). The equivalence tester relies on the following statistic:

$$Z := \sum_i \frac{(X_i - Y_i)^2 - X_i - Y_i}{X_i + Y_i}\,,$$

where X_i is the number of occurrences of element i in the sample set from p, and Y_i is the number of occurrences of element i in the sample set from q. The statistic Z is chosen in a way so that its expected values in the completeness and soundness cases differ substantially. The challenging part of the analysis involves a tight upper bound on the variance, which allows to show that Z is well-concentrated after an appropriate number of samples. More precisely, the following statements were shown in (Chan et al., 2014):

$$\begin{aligned}\mathbf{E}[Z] &= \sum_i \frac{(p(i) - q(i))^2}{p(i) + q(i)} m \left(1 - \frac{1 - e^{m(p(i)+q(i))}}{m(p(i) + q(i))}\right) \\ &\geq \frac{m^2}{4n + 2m} \|p - q\|_1^2\,. \end{aligned} \tag{5}$$

and

$$\mathbf{Var}[Z] \leq 2 \min\{m, n\} + \sum_i 5m \frac{(p(i) - q(i))^2}{p(i) + q(i)}\,. \tag{6}$$

The private version of the above statistic is simple: We add noise to the random variable Z and work with the noisy statistic, denoted by Z'. We need to show that we still can infer the correct answer from Z', and the noise does not incapacitate our tester. The main reason that this is indeed possible is because the statistic Z has bounded sensitivity.

Algorithm 3 is our private equivalence tester and we prove its correctness in Theorem 5.1.

Theorem 5.1. *Given sample access to two distributions p and q, Algorithm 3 is an (ϵ, ξ)-private tester for equivalence of p and q.*

Since the sensitivity of Z is small, we can add a small amount of noise to it to make it private, using the Laplace mechanism. Then, we show that adding the noise to Z does not increase its variance drastically. Finally, we prove by the Chebyshev inequality that, with high probability, Z concentrates well around its expected value given the size of the sample set. The proof of the theorem is in Appendix G.

Algorithm 3 Private Equivalence Tester: Private-Equivalence-Test

Input: Sample access to p and q, n, ϵ, ξ
$m \leftarrow C \cdot \max\left(\frac{\sqrt{n}}{\epsilon^2}, \frac{n^{2/3}}{\epsilon^{4/3}}, \frac{\sqrt{n}}{\sqrt{\xi}\epsilon}, \frac{1}{\xi\epsilon^2}\right)$
Draw m samples from distributions p and q.
$X_i \leftarrow$ the number of occurrences of the i-th element in the samples from p
$Y_i \leftarrow$ the number of occurrences of the i-th element in the samples from q
$Z \leftarrow \sum_i \frac{(X_i - Y_i)^2 - X_i - Y_i}{X_i + Y_i}$ $\langle\langle$*for* $X_i + Y_i \neq 0$.$\rangle\rangle$
$\eta \leftarrow \mathbf{Lap}(8/\xi)$
$Z' = Z + \eta$
$T \leftarrow \frac{m^2\epsilon^2}{8n + 4m}$
if $Z' \leq T$ **then**
    Output accept.
else
    Output reject.
end if

6. Experiments

We provide an empirical evaluation of the proposed algorithms on synthetic data. All experiments were performed on a computer with a 1.6 GHz Intel(R) Core(TM) i5-4200U CPU and 3 GB of RAM.

Before we describe our methodology and experimental results in detail, we make two crucial remarks. First, as we explain in more detail below, we note that our synthetic datasets include the provably *hardest instances* of the corresponding testing problems in the non-private setting. That is, we provide as input to our algorithms sets of samples from pairs of discrete distributions that are the hardest to distinguish information-theoretically. The related work (Cai et al., 2017) evaluated the empirical performance of their identity tester on essentially identical synthetic inputs. Second, since theoretical sample upper bounds in distribution testing typically use big-O notation, the practical performance of the various algorithms depends on the hidden absolute constants in these bounds (which are notoriously hard to pin-down theoretically). As a result, our experimental evaluation reveals phenomena which are not directly implied by our theoretical upper bounds.

We now briefly describe our methodology. To measure the accuracy of our algorithms, we empirically estimate the error probability, i.e., the probability that our algorithms output the wrong answer. We run our algorithms on input of s samples from a distribution q (or a pair of distributions) that either satisfies the property (completeness) or is ϵ-far in ℓ^1-distance from satisfying the property. We denote the distribution q in these two cases by q^+ and q^- respectively. We repeatedly run our algorithm r times and compute the ratio of the incorrect answers among these r trials for both q^+ and q^-. This gives as estimates for the type I and II errors of our algorithm. We want to understand how fast the error probability converges to 0 as the sample size increases.

For the case of uniformity testing, we observe that Private-Unique-Elements-Uniformity (Algorithm 1) performs significantly better on our datasets than algorithm Private-Collisions-Uniformity (Algorithm 2), especially when the domain size n is very large (Figure 1). For the case of private identity testing, we show (Figures 3 and 4) that our identity tester obtained by combining our reduction with Private-Unique-Elements-Uniformity significantly outperforms all previous algorithms for this problem. Finally, for the case of equivalence testing, our experiments illustrate (Figure 5) that we can obtain differential privacy essentially for free.

Private Uniformity Testing. We implemented Private-Unique-Elements-Uniformity (Algorithm 1) and Private-Collisions-Uniformity (Algorithm 2) to test the uniformity of a distribution in ℓ^1-distance.

Let q^+ be the uniform distribution on $[n]$ and q^- be a distribution that has probability $(1+\epsilon)/n$ on half of the domain and probability $(1-\epsilon)/n$ on the other half. Note that q^- is ϵ-far from uniform in ℓ^1-distance. It is known that q^- is the hardest distribution to distinguish from uniform among all distributions that are ϵ-far (Paninski, 2008), without losing any constant factor (Diakonikolas et al., 2017a).

We run our two algorithms using samples from q^+ and q^- with the following parameters: $n = 800,000$, $\epsilon = 0.3$, $r = 300$, and $\xi = 0.2$. We estimate how the empirical error probability of the tester changes by increasing the number of samples. As shown in Figure 1, for such a large domain, the algorithm Private-Unique-Elements-Uniformity reaches empirical error of almost zero with sample size *sub-linear* in the size of the domain. We emphasize that none of the previous algorithms in this setting was able to obtain meaningful guarantees in this sparse sample regime.

As predicted by our theoretical bounds, the tester Private-Unique-Elements-Uniformity completely fails if it uses more samples than the domain size. This fact is important when the domain size n and accuracy ϵ are such that the quantity $\sqrt{n}/\epsilon^2$ is comparable to the domain size n. For example, if $n = 1000$ and $\epsilon = 0.1$, Private-Unique-Elements-Uniformity is unable to provide any meaningful guarantees, hence we need to resort to Private-Collisions-Uniformity. This is illustrated in Figure 2.

A plausible interpretation for the apparent superiority of Private-Unique-Elements-Uniformity in the sparse regime

is that: (1) The non-private version of this tester is known to achieve *optimal constants* in the big-O of the sample complexity (Huang & Meyn, 2013). (2) The non-private tester has low sensitivity, hence we do not require a pre-processing phase (as in Private-Collisions-Uniformity) to obtain a private algorithm.

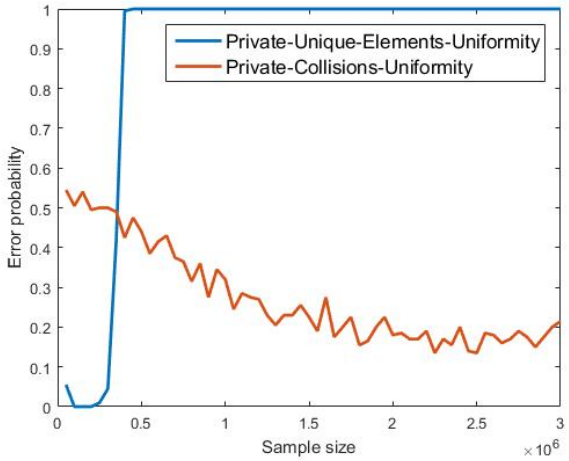

Figure 1. The estimated error probability of Private-Unique-Elements-Uniformity and Private-Collisions-Uniformity when $\sqrt{n}/\epsilon^2 = o(n)$.

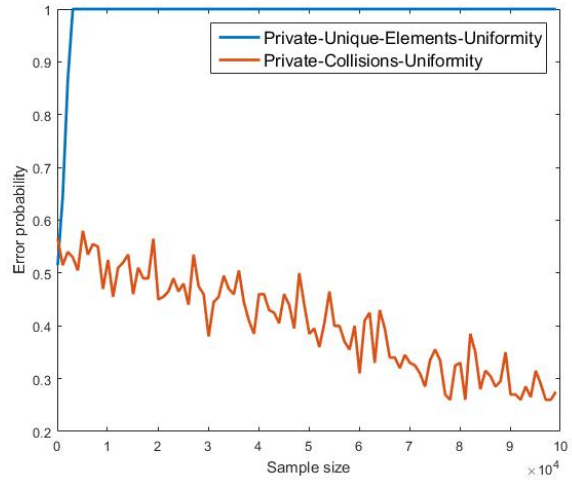

Figure 2. The estimated error probability of Private-Unique-Elements-Uniformity and Private-Collisions-Uniformity when $\sqrt{n}/\epsilon^2 = \Omega(n)$.

Private Identity Testing. We now describe our two private identity testers and experimentally compare them to previous private identity testers developed in the recent literature. As explained in Section 3, we proceed to reduce private identity testing to private uniformity testing. More specifically, our identity testers work by first mapping the sample set S to a new set S' on a somewhat larger domain, and then testing uniformity on the new domain using samples in S'. Since we have two uniformity testers, Private-Unique-Elements-Uniformity and Private-Collisions-Uniformity, we thus obtain two identity testers based on which uniformity tester we use. We term these two private identity testers Private-Unique-Elements-Identity and Private-Collisions-Identity respectively.

We compare our algorithms with the two recent algorithms: Priv'IT proposed in (Cai et al., 2017) and MCGOF proposed in (Gaboardi et al., 2016). It should be noted that our algorithms (and those of (Cai et al., 2017)) provides significantly stronger guarantees compared to (Gaboardi et al., 2016). More specifically, (Gaboardi et al., 2016) only provides type I error guarantees: the algorithm outputs reject with small probability when the distribution is identical to the given distribution. In contrast, our identity testers provably provide small type I and type II error probabilities.

We evaluate the various identity testers on two different pairs of distributions: (1) q_1^+ is the uniform distribution on $[n]$, while q_1^- assigns probability $(1+\epsilon)/n$ on half of the domain and probability $(1-\epsilon)/n$ on the other half. (2) q_2^+ is a 4-histogram distribution, i.e., the probability mass function is piecewise constant with 4 pieces, and q_2^- is obtained from q_2^+ by perturbing the probability of each element by $\pm\epsilon/n$. Testing uniformity is a special case of identity testing, and it is known to be essentially the hardest instance of this more general problem.

For (1), we explicitly give the uniform distribution, q_1^+, to our identity testing algorithm, and draw samples from q_1^+ or q_1^-. We use the parameters $n = 800,000$, $\epsilon = 0.3$, and $\xi = 0.2$. We vary the sample size staring from $50,000$ and up to 3×10^6, increasing it by $50,000$ at each step, and repeat the algorithm for $r = 200$ times to estimate the maximum of type I and type II errors. We repeat the same process for all the testers we compared against. The results are shown in Figure 3.

For (2), we use the same methodology on input the 4-histogram distribution q_2^+ with interval pieces I_1, I_2, I_3, I_4 each of size $n/4$ such that $q_2^+(I_1) = 4q_2^+(I_4)$, $q_2^+(I_2) = 3q_2^+(I_4)$, and $q_2^+(I_3) = 2q_2^+(I_4)$. The results are shown in Figure 4.

In both cases, we observe that our identity tester Private-Unique-Elements-Identity converges much faster than all other algorithms.

We remind the reader that (Gaboardi et al., 2016) did not provide any type II error guarantees for their tester MCGOF. We included two different curves in our plots illustrating the empirical type I and type II errors of the (Gaboardi et al., 2016) tester.

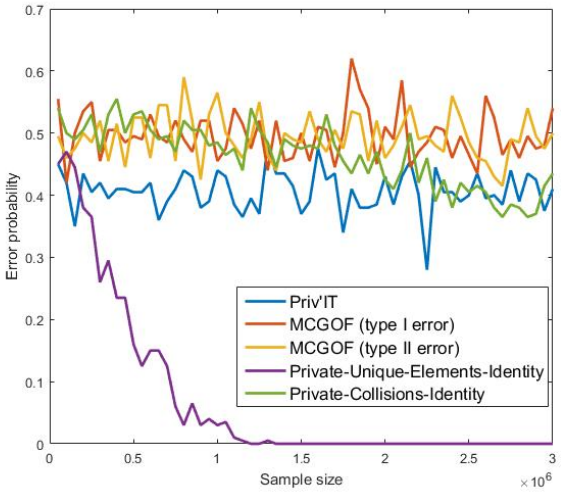

Figure 3. The estimated error probability of various private identity testers given samples form q_1^+ and q_1^-.

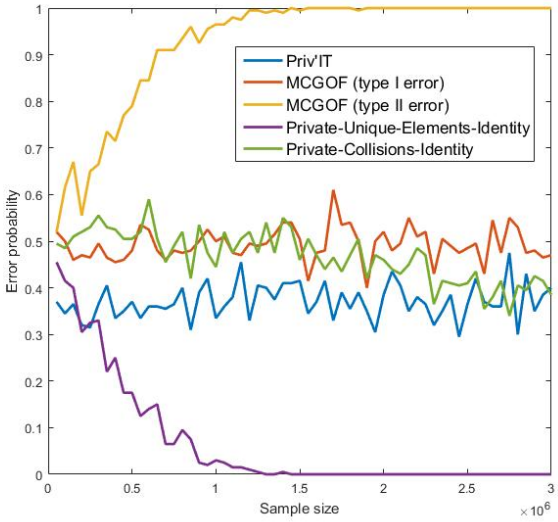

Figure 4. The estimated error probability of various identity testers given samples form q_2^+ and q_2^-.

Private Equivalence Testing. The focus of the experimental evaluation for equivalence testing is as follows: For a range of increasing domain sizes, n, we want to find the smallest sample size such that the error probability (maximum of type I and type II errors) drops below $1/3$.

We implemented Algorithm 3 to test equivalence of two unknown distributions. We show that for sufficiently large domain size n, our private algorithm succeeds with a *sublinear* number of samples.

For given domain size n, to find the (approximately) minimum number of samples such that the error probability of the algorithm drops below $1/3$, we proceed as follows: We start with an initial number of samples s. Then, we estimate the empirical error of the algorithm for these sample sets. If it is more than $1/3$, we increase s appropriately and repeat the process until we find s that results in an error of at most $1/3$.

We choose our input distributions to be the information-theoretically hardest distributions to distinguish in the non-private setting (Batu et al., 2013; Chan et al., 2014). In particular, p is defined to be the distribution such that $n^{2/3}$ of the domain elements have probability $(1-\epsilon/2)/n^{2/3}$ (the "heavy elements") and $n/4$ "light" elements have probability $2\epsilon/n$. (The rest of the domain elements have mass 0.) Similarly, q is defined to be be a distribution that has probability $(1-\epsilon/2)/n^{2/3}$ on the same set of heavy elements as p, and for a disjoint set of $n/4$ light elements assigns probability $2\epsilon/n$. Since the light elements are disjoint, it is clear that p is ϵ-far from q.

To evaluate the sample complexity of our algorithm, we use the tester to distinguish the following pairs: (q, q) and (p, q). We set $\epsilon = 0.3$, $r = 200$, and $\xi = 0.2$. We calculate the required number of samples of this tester in order to achieve accuracy at least $2/3$, for n raging from 10^4 up to 2×10^6, increasing n by 10^4 at each step.

As a point of comparison, we also implemented the non-private equivalence tester of (Chan et al., 2014). As shown in Figure 5, the sample complexities of private and non-private equivalence testing are very close to each other. This result was expected given the theoretical sample complexity of our equivalence tester, since the dependence on the privacy parameter ξ appears in an additive term, and is dominated by the other term, when ξ is a constant.

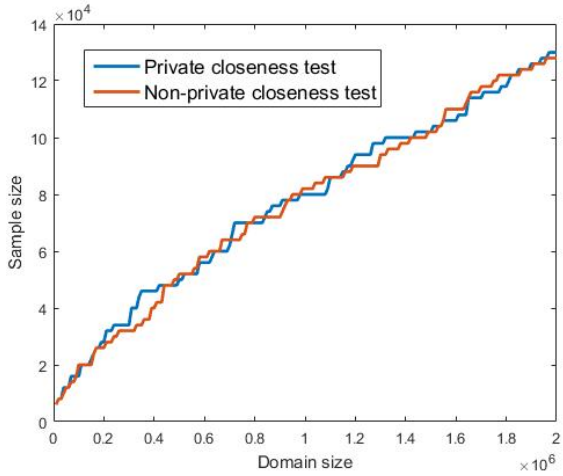

Figure 5. Sample complexity of private and non-private equivalence testers.

Acknowledgements

M.A. and R.R. were supported by the National Science Foundation Award under Grant No. CCF-1733808, CCF-1650733, and IIS-1741137. In addition, R.R. was supported by National Science Foundation Award under Grant No. CCF-1740751, and ISF grant 1536/14. I.D. was supported by National Science Foundation Award No. CCF-1652862 (CAREER) and a Sloan Research Fellowship.

References

Acharya, J., Daskalakis, C., and Kamath, G. Optimal testing for properties of distributions. In *Conference on Neural Information Processing Systems, NIPS*, pp. 3591–3599, 2015.

Acharya, J., Sun, Z., and Zhang, H. Differentially private testing of identity and closeness of discrete distributions. *CoRR*, abs/1707.05128, 2017. URL http://arxiv.org/abs/1707.05128.

Batu, T., Fortnow, L., Rubinfeld, R., Smith, W. D., and White, P. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science, FOCS*, pp. 259–269, 2000. URL citeseer.ist.psu.edu/batu00testing.html.

Batu, T., Fortnow, L., Rubinfeld, R., Smith, W. D., and White, P. Testing closeness of discrete distributions. *Journal of the ACM*, 60(1):4, 2013.

Cai, B., Daskalakis, C., and Kamath, G. Priv'it: Private and sample efficient identity testing. In *International Conference on Machine Learning, ICML*, pp. 635–644, 2017. URL http://proceedings.mlr.press/v70/cai17a.html.

Canonne, C., Diakonikolas, I., Gouleakis, T., and Rubinfeld, R. Testing shape restrictions of discrete distributions. In *Theoretical Aspects of Computer Science, STACS*, pp. 25:1–25:14, 2016.

Canonne, C., Diakonikolas, I., Kane, D. M., and Stewart, A. Testing bayesian networks. In *Conference on Learning Theory, COLT*, pp. 370–448, 2017. URL http://proceedings.mlr.press/v65/canonne17a.html.

Canonne, C. L. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015.

Chan, S., Diakonikolas, I., Valiant, P., and Valiant, G. Optimal algorithms for testing closeness of discrete distributions. In *ACM-SIAM Symposium on Discrete Algorithms, SODA*, pp. 1193–1203, 2014.

Daskalakis, C., Diakonikolas, I., Servedio, R., Valiant, G., and Valiant, P. Testing k-modal distributions: Optimal algorithms via reductions. In *ACM-SIAM Symposium on Discrete Algorithms, SODA*, pp. 1833–1852, 2013.

Diakonikolas, I. and Kane, D. M. A new approach for testing properties of discrete distributions. In *IEEE Symposium on Foundations of Computer Science, FOCS*, pp. 685–694, 2016. Full version available at abs/1601.05557.

Diakonikolas, I., Hardt, M., and Schmidt, L. Differentially private learning of structured discrete distributions. In *Conference on Neural Information Processing Systems, NIPS*, pp. 2566–2574, 2015a.

Diakonikolas, I., Kane, D. M., and Nikishkin, V. Testing identity of structured distributions. In *ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pp. 1841–1854, 2015b. doi: 10.1137/1.9781611973730.123. URL https://doi.org/10.1137/1.9781611973730.123.

Diakonikolas, I., Kane, D. M., and Nikishkin, V. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *IEEE Symposium on Foundations of Computer Science, FOCS*, pp. 1183–1202, 2015c.

Diakonikolas, I., Gouleakis, T., Peebles, J., and Price, E. Collision-based testers are optimal for uniformity and closeness. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:178, 2016.

Diakonikolas, I., Gouleakis, T., Peebles, J., and Price, E. Sample-optimal identity testing with high probability. *Electronic Colloquium on Computational Complexity (ECCC)*, 24:133, 2017a. URL https://eccc.weizmann.ac.il/report/2017/133. To appear in ICALP 2018.

Diakonikolas, I., Kane, D. M., and Nikishkin, V. Near-optimal closeness testing of discrete histogram distributions. In *International Colloquium on Automata, Languages, and Programming, ICALP*, pp. 8:1–8:15, 2017b. doi: 10.4230/LIPIcs.ICALP.2017.8. URL https://doi.org/10.4230/LIPIcs.ICALP.2017.8.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Gaboardi, M., Lim, H. W., Rogers, R. M., and Vadhan, S. P. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International Conference on Machine Learning, ICML*, pp. 2111–2120, 2016.

Goldreich, O. The uniform distribution is complete with respect to testing identity to a fixed distribution. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:15, 2016. URL http://eccc.hpi-web.de/report/2016/015.

Goldreich, O. and Ron, D. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium on Computational Complexity (ECCC), 2000.

Huang, D. and Meyn, S. Generalized error exponents for small sample universal hypothesis testing. *IEEE Trans. Inf. Theor.*, 59(12):8157–8181, December 2013.

Kakizaki, K., Fukuchi, K., and Sakuma, J. Differentially private chi-squared test by unit circle mechanism. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1761–1770, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/kakizaki17a.html.

Kifer, D. and Rogers, R. A new class of private chi-square hypothesis tests. In *International Conference on Artificial Intelligence and Statistics, AISTATS*, pp. 991–1000, 2017. URL http://proceedings.mlr.press/v54/rogers17a.html.

Lehmann, E. L. and Romano, J. P. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, 2005.

Neyman, J. and Pearson, E. S. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706): 289–337, 1933. doi: 10.1098/rsta.1933.0009. URL http://rsta.royalsocietypublishing.org/content/231/694-706/289.short.

Paninski, L. A coincidence-based test for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54:4750–4755, 2008.

Rubinfeld, R. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

Valiant, G. and Valiant, P. An automatic inequality prover and instance optimal identity testing. In *IEEE Symposium on Foundations of Computer Science, FOCS*, 2014.

Wang, Y., Lee, J., and Kifer, D. Differentially private hypothesis testing, revisited. *CoRR*, abs/1511.03376, 2015.