Private Testing of Distributions via Sample Permutations

Maryam Aliakbarpour
CSAIL, MIT
maryama@mit.edu

Ilias Diakonikolas
University of Wisconsin, Madison
ilias.diakonikolas@gmail.com

Daniel Kane
University of California, San Diego
dakane@ucsd.edu

Ronitt Rubinfeld
CSAIL, MIT, TAU
ronitt@csail.mit.edu

October 1, 2019

Abstract

Statistical tests are at the heart of many scientific tasks. To validate their hypotheses, researchers in medical and social sciences use individuals' data. The sensitivity of participants' data requires the design of statistical tests that ensure the privacy of the individuals in the most efficient way. In this paper, we use the framework of property testing to design algorithms to test the properties of the distribution that the data is drawn from with respect to differential privacy. In particular, we investigate testing two fundamental properties of distributions: (1) testing the equivalence of two distributions when we have unequal numbers of samples from the two distributions. (2) Testing independence of two random variables. In both cases, we show that our testers achieve near optimal sample complexity (up to logarithmic factors). Moreover, our dependence on the privacy parameter is an additive term, which indicates that differential privacy can be obtained in most regimes of parameters for free.

1 Introduction

We study questions in *statistical hypothesis testing*, a field of statistics with fundamental importance in scientific discovery. At a high level, given samples from an unknown statistical model, the goal of a hypothesis test is to determine whether the model has a desired property. The first — and arguably the most fundamental — objective in hypothesis testing is to make an accurate determination *with as few samples as possible.* In this work, we focus on understanding the trade off between sample size and an additional important criterion — preserving the *privacy* of the underlying data sets.

Early work in statistics [Pea00, NP33] studied the asymptotic regime, where the sample size goes to infinity. In this paper, we are interested in obtaining finite sample bounds in the *minimax setting* that has been extensively studied in the computer science and information theory communities during the past couple of decades. More specifically, we will work with the formalism of *distribution property testing* [BFR+00, BFR+13]: Given samples from a collection of unknown probability distributions over discrete domains, do the underlying distributions satisfy a desired property $\mathcal{P}$ or are they "far" from satisfying the property? (The definition of "far" is typically quantified via some global error metric, e.g., the total variation distance. See Preliminaries section.) The goal is to develop testers for various properties with information-theoretically optimal sample complexity, which is typically *sublinear* in the domain sizes of the underlying distributions. We note that, in recent years, such sample-optimal methods have been obtained for testing a range of properties, including identity testing ("goodness-of-fit"), closeness testing ("equivalence testing" or "two-sample testing"), and independence testing.

The primary goal in classical statistics theory is to minimize the sample size of inference tasks. In recent years, a wide range of settings involves performing hypothesis testing tasks on sensitive data representing specific individuals, such as data describing medical or other behavioral phenomena. In such cases, the outputs of standard tests may reveal private information that should not be divulged. Differential privacy [Dwo09, DR14a] is a formal framework (see Preliminaries section) that may allow us to obtain the

scientific benefit of statistical tests without compromising the privacy of the underlying individuals. Intuitively, differential privacy postulates that similar data sets should have statistically close outputs; and that once this guarantee is achieved, then provable privacy is preserved. Differentially private data analysis is a very active research area, in which a wealth of techniques have been developed for a range of tasks.

When designing a differentially private hypothesis testing algorithm, there are two criteria to balance. On the one hand, we require that the algorithm satisfies the differential privacy condition on *any* input dataset. On the other hand, we require that the algorithm is a valid statistical hypothesis tester (i.e., it correctly distinguishes inputs satisfying property $\mathcal{P}$ from inputs that are far from satisfying $\mathcal{P}$). These competing criteria suggest that, in general, the sample size required to ensure both of them grows compared to the non-private setting. Recent work [CDK17, ADR18, ASZ18] has shown that for basic tasks, such as identity testing and equivalence testing, the sample size increase of a differentially private tester compared to its non-private analogue is negligible.

In this work, we continue this line of investigation. We give a new general technique that yields sample-efficient differentially private testers and apply it for two fundamental statistical tasks: the problem of equivalence testing *with unequal sized samples* and the problem of *independence testing* (defined in the following paragraph). Notably, prior techniques were inherently unable to provide sample-efficient private testers for either of these problems.

Our Contributions. The main contribution of this work is a general technique for preserving privacy that in particular can be used to obtain sample-efficient and differentially private hypothesis testers based on the algorithmic technique in the work of [DK16]. This technique can be applied to several testing problems and in particular, yields the only known sample optimal testers for the following testing problems:

Equivalence Testing with Unequal Sized Samples: Given a target error parameter $\epsilon > 0$, s_1 independent draws from an unknown distribution p over $[n]$, and s_2 draws from an unknown distribution q over $[n]$, distinguish the case that $p = q$ from the case that $\|p - q\|_1 \geq \epsilon$.

Independence Testing: Given a target error parameter $\epsilon > 0$, and s independent draws from an unknown distribution p over $[n] \times [m]$, distinguish the case that p is a product distribution (i.e., its two coordinates are independent) versus ϵ-far, in ℓ_1-distance, from any product distribution.

These problems have been extensively investigated in distribution testing during the past decade [BFF+01a, LRR11, CDVV14, VV14, AJOS14, ADK15, BV15, DK16] and sample-optimal testers are known for them in the non-private setting [DK16]. In this work, we design the first differentially private testers for these problems with optimal (or near-optimal) sample complexity. In particular, we show that the sample complexity of both these problems in the private setting is nearly the same as in the non-private setting, i.e., privacy comes essentially for free.

In this work, we focus on privatizing the optimal non-private testers for the above problems presented in [DK16]. The algorithmic technique of [DK16] splits samples into two groups — "flattening samples" and "testing samples" — that are used in very different ways. To obtain privacy guarantees, we must design algorithms that have low sensitivity to the samples — that is, changing one sample does not have much effect on the outcome. Previous works in private testers for hypothesis testing are based on algorithms that use the samples analogously to the use of the "testing samples" in [DK16]. One can similarly use those techniques to design algorithms with low sensitivity with respect to the "testing samples" in our setting. However, since using "flattening samples" is a key to achieve testers with the optimal sample complexity, we need to obtain low sensitivity with respect to the "flattening samples". As the output of the testing algorithm can be very sensitive to small changes in the set of "flattening samples", more care is required in designing low sensitivity algorithms for using the "flattening samples", In a nutshell, what we present is a technique for designing private testers which considers the output of the algorithm in [DK16] on *every* permutation of the samples and outputs a result that is based on the aggregate. In order to show that this approach gives the correct answer, we have to show that not only the expectation of the result is the same as that of the non-private tester (which is straightforward), but also the variance is small enough so that the resulting output is correct with high probability (which does not follow from the calculation of the variance of the non-private testers). For the independence testing problem, we need an extra step to reduce sensitivity further. We first show that trying every permutation results in low sensitivity in the typical case (over the random samples), and thus gives a private tester. In the non-typical case, more care must be taken — we give an algorithm for

modifying the samples in such a way that we can reduce to the typical case. We describe the challenges and techniques in more detail in Section 3. Though our methods are mainly tailored for use with the [DK16] testers, there are several other distribution property testing algorithms which use samples in similar sensitive manners, e.g., [CDGR16]— the hope is that these techniques will be fruitful in allowing those algorithms to be made differentially private as well.

Related Work The field of *distribution property testing* [BFR+00] has been extensively investigated in the past couple of decades, see [Rub12, Can15, Gol17]. A large body of the literature has focused on characterizing the sample size needed to test properties of arbitrary discrete distributions. This regime is fairly well understood: for many properties of interest there exist sample-efficient testers [Pan08, CDVV14, VV14, DKN15b, ADK15, CDGR16, DK16, DGPP16, CDS17, Gol17, DGPP17, BC17, DKS18, CDKS17b]. More recently, an emerging body of work has concenterated on leveraging *a priori* structure of the underlying distributions to obtain significantly improved sample complexities [BKR04, DDS+13, DKN15b, DKN15a, CDKS17a, DP17, DKN17, DKP19].

Differential privacy was first introduced in [DMNS06]. Recently, a new line of research studies distribution testing and learning problems with respect to differential privacy [DHS15, CDK17, ADR18, ASZ18]. The focus of these works is on testing identity and closeness of distributions, leaving other properties mostly unexplored. Other models for distribution testing problems with respect to differential privacy have been studied [WLK15, GLRV16, KR17, KFS17]. In most of these latter works, only a type I error guarantee is provided, which is a significantly weaker guarantee compare to ours. In addition, other settings for privacy, e.g, local privacy, is investigated [She18, GR18, ACFT19].

2 Preliminaries

2.1 Definitions

Notation: We use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. We consider discrete distributions over a finite domain, in particular, over $[n]$ without loss of generality. For a distribution p, we write $p(i)$ to denote the probability of element i in $[n]$. One can assume each distribution has an associated probability function $p : [n] \to [0, 1]$ such that $p(i)$'s are non-negative and $\sum_{i\in[n]} p(i) = 1$. For set $S \subseteq [n]$, $p(S)$ denotes the total probability of the elements in S (i.e., $\sum_{i\in S} p(i)$). Note that one can think of each discrete distribution over $[n]$ as a vector in $\mathbb{R}^n$ where the i-th coordinate is the probability of element i. Having said that, we can define the ℓ_k-norm of a distribution in the same manner as of a vector: For a vector $x \in \mathbb{R}^n$ and any $k > 0$, the ℓ_k-norm of x is equal to $\left(\sum_{i\in[n]} |x_i|^k\right)^{\frac{1}{k}}$, and is denoted by $\|x\|_k$. In addition, the ℓ_k-distance between two distributions p and q over $[n]$ is equal to $\|p - q\|_k$. Throughout this paper, we use the ℓ_1-distance to measure the discrepancy between distributions, which is equivalent to the total variation distance up to a factor of two. In particular, we say distribution p is ϵ-far from distribution q if $\|p - q\|_1 \geq \epsilon$.

We denote the Poisson distribution with mean λ by $\mathbf{Poi}(\lambda)$. The probability of the non-negative integer x according to the Poisson distribution is $\mathbf{Poi}(x; \lambda) = e^\lambda \lambda^x / x!$. Also, let $\mathbf{Bin}(n, b)$ denote the binomial distribution where n is the number of trials, and b is the bias parameter. Let X be a random variable that indicates the number of "successful" draws after n trials *without replacement*, from a population of size m, in which k elements are considered to be "success." The probability distribution of X is called the hypergeometric distribution, and we denote it by $\mathbf{HG}(m, k, n)$. The probability of $X = x$ according to $\mathbf{HG}(m, k, n)$ is the following: $\mathbf{HG}(x; m, k, n) = \binom{k}{x}\binom{m-k}{n-x}/\binom{m}{n}$. We use $\mathbf{Lap}(\lambda)$ to denote the zero mean Laplace distribution with parameter λ. The probability of $x \in \mathbb{R}$ according to the Laplace distribution is $\mathbf{Lap}(x; \lambda) = e^{-|x|/\lambda}/(2\lambda)$. We use the notation for the distribution and a random variable exchangeably where the difference is clear from the content. For instance, $\mathbf{Poi}(\lambda)$ refers to both the Poisson distribution with mean λ and a Poisson random variable drawn from that distribution.

Distribution Testing: Formally, we define a property $\mathcal{P}$ to be a set of distributions. We say a distribution p has the property $\mathcal{P}$ if $p \in \mathcal{P}$; and, we say p is ϵ-far from having the property $\mathcal{P}$ when p is ϵ-far from all distributions in $\mathcal{P}$. Assume an algorithm has sample access to a distribution p. We say the algorithm is an (ϵ, δ)-*tester for property* $\mathcal{P}$, if the following holds with probability at least $1 - \delta$:

- **Completeness case:** If p has the property $\mathcal{P}$, then the algorithm outputs accept.
- **Soundness case:** If p is ϵ-far from $\mathcal{P}$, then the algorithm outputs reject.

Analogously, one can generalize the above definition to the case that we have sample access to two distributions. In particular, given sample access to two distributions p and q over $[n]$, an (ϵ, δ)-tester for *testing closeness* of p and q distinguishes the following cases with probability at least $1 - \delta$:

- **Completeness case:** If p is equal to q, then the algorithm outputs accept.
- **Soundness case:** If p is ϵ-far from q, then the algorithm outputs reject.

Testing closeness of distributions via flattening technique: In this paper, we build on the non-private closeness tester presented in [DK16, CDVV14]. Here, we give an overview of the closeness tester, and the flattening technique which turn it into a tester with the optimal sample complexity.

Suppose we have sample access to two distributions p and q on the domain $[n]$. Let s be a parameter that determines the expected number of samples. We draw $\mathbf{Poi}(s)$ samples from p and q. For each $i \in [n]$, let X_i and Y_i denote the number of occurrences of element i in the sample sets from p and q respectively. In [CDVV14], the authors proposed the following statistic to test the closeness of p and q: $Z = \sum_{i=1}^{n} (X_i - Y_i)^2 - X_i - Y_i$. The expected value of Z is proportional to the squared ℓ_2-distance of p and q which enables us to use Z for testing closeness of p and q. While this statistic is mainly measuring the ℓ_2-distance, one can use it for testing in ℓ_1-distance as well by using trivial inequalities between the distances. By careful analysis of the variance of Z, it is shown that Z concentrates around its expectation when we draw at least $s = \Theta(n/(\epsilon^2 \max(\|p\|_2, \|q\|_2)))$ samples[1]. More specifically, it is shown that in the case where $p = q$, Z is below a threshold parameter, τ; and when p is ϵ-far from q, Z is at least τ with high probability. Thus, we can test the closeness of p and q by computing Z from a large enough sample set and comparing it with the threshold τ.

The above algorithm would be sample-efficient only when $\max(\|p\|_2, \|q\|_2)$ is not too large. In [DK16], the authors provide a technique, called *flattening*, that decreases the ℓ_2-norm of a distribution. Using this technique, they map p and q to two other distributions at least one of which has smaller ℓ_2-norm. Then, they obtain a sample-optimal ℓ_1-closeness tester for p and q by showing its equivalence to closeness testers for the two distributions obtained after flattening.

We discuss the flattening technique in more detail. To flatten a distribution p, we need a (multi)set of the domain elements denoted by F. This set is usually obtained by drawing samples from the underlying distributions, and the elements in F are called *flattening samples*. Using flattening samples, we transform p to another distribution $p^{(F)}$ over a larger domain in a way that the ℓ_2-norm of $p^{(F)}$ is small. We build the new domain for $p^{(F)}$ as follows: For each element i in the domain of p, we first count the number of occurrences of i in F, namely k_i, and put $b_i := k_i + 1$ elements associated to i in the new domain. We refer to the elements of the new domain as *buckets*. We define $p^{(F)}$ to be the distribution that assigns the probability mass of $p(i)/b_i$ to all the b_i buckets of i for all i in the domain. Note that one can generate a sample from $p^{(F)}$ upon receiving a sample from p: For a fresh sample $X = i$ drawn from p, the flattening procedure maps it to a randomly selected bucket j among the b_i buckets of i. Then, it outputs $X' = (i, j)$ as a sample from $p^{(F)}$. The above procedure has several important properties which help us later in our analysis:

- By the above construction, it is clear that the size of the new domain is $\sum_i b_i = n + |F|$. Thus, as long as $|F|$ is not larger than n (in the regime where we have sublinear number of samples), the size of domain increases only by a constant factor.
- It is shown that if F contains $\mathbf{Poi}(k)$ samples from p, then the expected ℓ_2-norm of $p^{(F)}$ is at most $1/k$.
- If we flatten two distributions using the same ass ignments for the buckets (i.e., the flattening set F), the ℓ_1-distance between the two distributions remains unchanged. Thus, it suffices to test closeness of $p^{(F)}$ and $q^{(F)}$ for testing the closeness of p and q.

[1] In fact, as a byproduct of our analysis, one can show it suffices to have $s \geq \Theta(n/(\epsilon^2 \min(\|p\|_2, \|q\|_2)))$.

Privacy: We say two sample sets, X and X' , from a universe $[n]$ are *neighboring* if and only if their Hamming distance is one (meaning that they differ in exactly *one* sample). A randomized algorithm $\mathcal{A}$ is ξ-private if for any subset S of the possible outputs of the algorithm, $\{\mathsf{accept}, \mathsf{reject}\}$ in the context of this paper, and for any two neighboring X and X', the following holds:

$$\mathbf{Pr}[\mathcal{A}(X) \in S] \leq e^{\xi} \cdot \mathbf{Pr}[\mathcal{A}(X') \in S] .$$

For a function f over sample sets, the *sensitivity of* f is defined as follows:

$$\Delta(f) = \max_{X,X'} |f(X) - f(X')| ,$$

where the maximum is taken over all possible sample sets that differ in only one sample. A standard method for making functions private is the *Laplace mechanism* [DR14b]. In this mechanism, to make a function f ξ-differentially private, one adds Laplace noise to the output of function: $\hat{f}(X) \coloneqq f(X) + \mathbf{Lap}(\Delta(f)/\xi)$. It is easy to show that $\hat{f}$ satisfies the definition of differential privacy with parameter ξ.

3 General approach for making closeness-based testers private

As we mentioned earlier, several properties can be tested via a reduction to the flattening-based closeness tester. These reductions and the resulting optimal testers were presented in [DK16]. In this section, we focus on describing a general approach for making such testers differentially private. We start by explaining the structure of the existing reductions in the non-private setting. Next, we explain our main idea for making the reductions and the testers private, which is to derandomize the non-private tester. Then we give the characteristics of the reductions that can be turned into a private algorithm (see Definition 3.2). In particular, if *any* property can be tested via a desired reduction to the closeness testing problem, it can be also privately tested with a reduction to our general private closeness tester. At the end, we describe our algorithm and prove its correctness in the full version.

3.1 Reduction procedure in non-private setting

In this section, we elaborate on the structure of the reductions to the closeness tester with the use of the flattening technique proposed in [DK16]. Suppose we aim to test whether a distribution[2] has property $\mathcal{P}$ or it is ϵ-far from any distribution in $\mathcal{P}$ via a reduction to the flattening-based closeness tester in the non-private setting. The reduction has the following structure: Upon receiving a sample set from the underlying distribution, the *reduction procedure* splits the samples into two sets *test samples*, denoted by T, and *flattening samples*, denoted by F. The reduction procedure uses these sample sets as follows:

- **Test samples**: The reduction procedure use the test samples to generate samples from two distributions p and q over a domain of size n. The distributions p and q are designed in such a way that if the underlying distribution has the property $\mathcal{P}$, then p and q are the same; and if the underlying distribution is ϵ-far from any distribution that has the property, then p and q are $\Theta(\epsilon)$-far from each other as well. This transformation is essentially the core of the reduction to the testing closeness problem.
- **Flattening samples**: In addition to samples from p and q, the reduction procedure generates n positive integers $b_1, b_2, \ldots, b_n$ that indicate the number of buckets for each domain element. These numbers are used for flattening of p and q. (See the Preliminaries section for more details.)

An example of such reductions is testing independence. Suppose d is a distribution over $[n] \times [m]$, and our goal is to test whether the two coordinates of the samples drawn from d are independent or not. It is not hard to see that this problem is equivalent to testing whether $p \coloneqq d$ is equal to $q \coloneqq d_1 \times d_2$, where d_1 and d_2 are the two marginal distributions of d. For more examples, see [DK16].

After the reduction procedure generates samples from p and q, and the number of buckets for each domain element, we use the flattening-based closeness tester for the testing closeness of p and q. As we explain in

[2]We may have more than one underlying distributions.

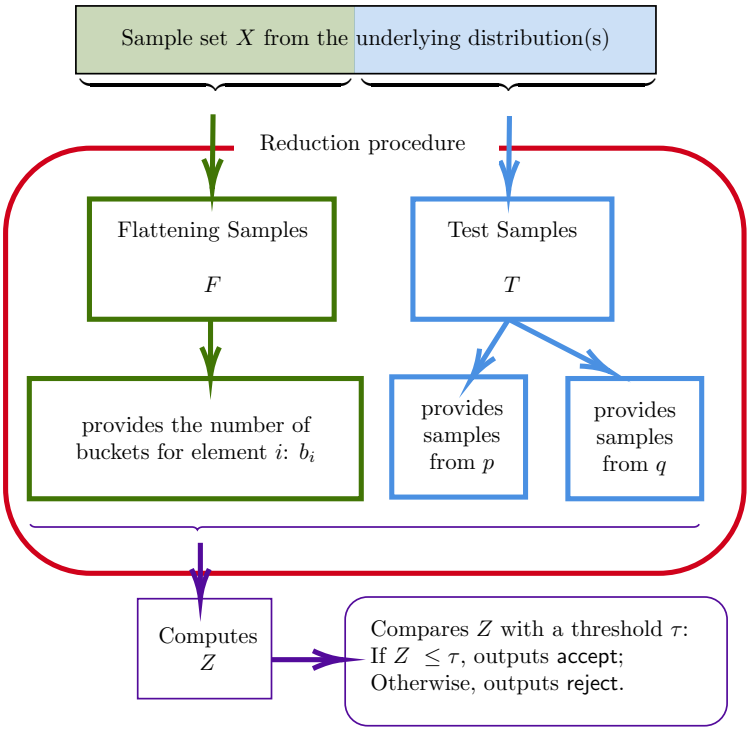

Figure 1: Standard reduction procedure to testing closeness of two distributions.

the Preliminaries section, first, we transform p and q to $p^{(F)}$ and $q^{(F)}$ on the new domain:

$$D = \{(i,j) | i \in [n] \text{ and } j \in [b_i]\}\,.$$

The hope is that after flattening the ℓ_2-norm of at least one of the resulting distributions is small. Recall that the probability of each element (i,j) in D, i.e., a bucket, according to $p^{(F)}$ is $p(i)/b_i$.

The closeness tester transforms samples from p to samples from $p^{(F)}$ (and similarly for q). For each sample i from p, the closeness tester assigns it to the bucket (i,j), where j is picked uniformly at random from $[b_i]$. We denote the number of occurrences of bucket (i,j) in the sample set from $p^{(F)}$ (and $q^{(F)}$) by $v_{i,j,1}$ (and $v_{i,j,2}$). The flattening-based closeness tester computes the following statistic Z:

$$Z := \sum_{i=1}^{n} \sum_{j=1}^{b_i} (v_{i,j,1} - v_{i,j,2})^2 - v_{i,j,1} - v_{i,j,2}\ , \tag{1}$$

and compares it with a threshold to establish whether $p^{(F)}$ and $q^{(F)}$ are equal or $\Theta(\epsilon)$-far from each other. Since the transformation to $p^{(F)}$ and $q^{(F)}$ does not change the ℓ_1-distance between p and q, the output of the tester determines whether d has the property $\mathcal{P}$ or ϵ-far from it. See Figure 1 for an overview of this process.

3.2 Derandomizing the non-private tester

To develop a differentially private algorithm for closeness testing with flattening, we "derandomize" the standard non-private closeness tester provided in [DK16, CDVV14]. The derandomization of the tester results in a stable statistic, which means if we change one sample in the sample set, the value of the statistics does not change drastically. The stability implies that the statistic has low sensitivity, and it can be privatized using fewer number of extra samples.

In the previous section, we explained how the reduction and the tester work. Note that there are two steps that the algorithm can make random choices: (i) The algorithm splits the samples into two sets

F and T (flattening and test samples): Upon receiving a set of $s+k$ samples, $X = \{x_1, x_2, \ldots, x_{s+k}\}$, where s and k is the number of test samples and the number of flattening samples respectively, usually the algorithm assigns the first s samples to the test set, and the rest to the flattening set. Equivalently, we can view this step as follows: the algorithm permutes the sample according to a *random* permutation π. Then, it splits them into F and T. In other words, T is equal to $\{x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(s)}\}$, and F is equal to $\{x_{\pi(s+1)}, x_{\pi(s+2)}, \ldots, x_{\pi(s+k)}\}$. (ii) The algorithm randomly selects a bucket for each sample from p and q. Let r denote the string of random bits that the algorithm uses to choose the buckets. We eliminate the randomness of these two steps by setting our new statistic, $\overline{Z}$, to be the *expected value* of the statistic Z over the random choices of the algorithm. More precisely, for a given input sample set X, we define:

$$\overline{Z}(X) \coloneqq \mathbf{E}_{\pi,r}[\,Z|X\,]\,.$$

We can simplify the above statistic a step further by computing the closed-form of the expected value of Z over all the random choices of r, i.e., $\mathbf{E}_r[\,Z|X,\pi\,]$. We provide the exact value in the following lemma, and it is proved in the full version.

Lemma 3.1. *Let $s_{i,1}$ (similarly $s_{i,2}$) be the number of occurrences of element i in the sample sets of p (q). Assume b_i is the number of buckets assigned to element i. Let $v_{i,j,1}$ (similarly $v_{i,j,2}$) be the number of occurrences of bucket (i,j) in the sample set from $p^{(F)}$ ($q^{(F)}$). Then, we have:*

$$\mathbf{E}_r\left[\sum_{j=1}^{b_i}(v_{i,j,1}-v_{i,j,2})^2 - v_{i,j,1} - v_{i,j,2}\,\middle|\, b_i, s_{i,1}, s_{i,2}\right] = \frac{(s_{i,1}-s_{i,2})^2 - s_{i,1} - s_{i,2}}{b_i}\,,$$

where the expectation is taken over all random assignments of the samples to the buckets.

The lemma above immediately implies the following equation:

$$\overline{Z}(X) = \mathbf{E}_\pi[\,\mathbf{E}_r[\,Z|X,\pi\,]|X\,] = \mathbf{E}_\pi\left[\sum_{i=1}^{n}\frac{(s_{i,1}-s_{i,2})^2 - s_{i,1} - s_{i,2}}{b_i}\,\middle|\, X\right]. \tag{2}$$

For the rest of this paper, we work with this latter form of $\overline{Z}$.

3.3 Designing a general private tester

Note that the algorithm we design has to satisfy two guarantees: First, it should be an accurate tester, i.e., it should output the correct answer with high probability. Second, it should be a differentially private algorithm.

For the accuracy guarantee, we first, need to show that the proposed statistic, $\overline{Z}$, is sufficient for testing closeness of p and q, and ultimately testing property $\mathcal{P}$. At first glance, the claim seems trivially true: We know that for *any* sample set X and *any* random choice of π, and r, the statistic Z will be a sufficient statistic for testing property $\mathcal{P}$ just because of the properties of the non-private tester. Since $\overline{Z}$ is essentially the expected value of a group of sufficient statistics, it must be immediate that $\overline{Z}$ is a sufficient statistic as well. However, there is a subtle difference between the guarantees we require for the statistics Z and $\overline{Z}$. To analyze the standard tester in [DK16], the authors showed that with high probability the set of flattening samples decreases the ℓ_2-norm of one of the two distributions p and q. The low ℓ_2-norm property is sufficient to show that Z has low variance, and one can use it for closeness testing. In fact, the authors show that the statistic Z works for some flattening test which occurs with high constant probability. However, in our setting, we wish to show the variance of $\overline{Z}$ is low over the random choices of both the flattening samples and the test samples. Hence, while we are taking out the randomness of r, or π, we are introducing a new source of randomness, which is the set F. Thus, it is unclear whether $\overline{Z}$ can be used as a statistic for the problem or not.

The goal for the rest of this section is to prove that if the reduction procedure has the desired characteristics, then $\overline{Z}$ is a sufficient statistic for testing property $\mathcal{P}$. We say a reduction procedure, $\mathcal{A}$, is a *proper procedure* if it has these desired guarantees, which we formally define in the following definition:

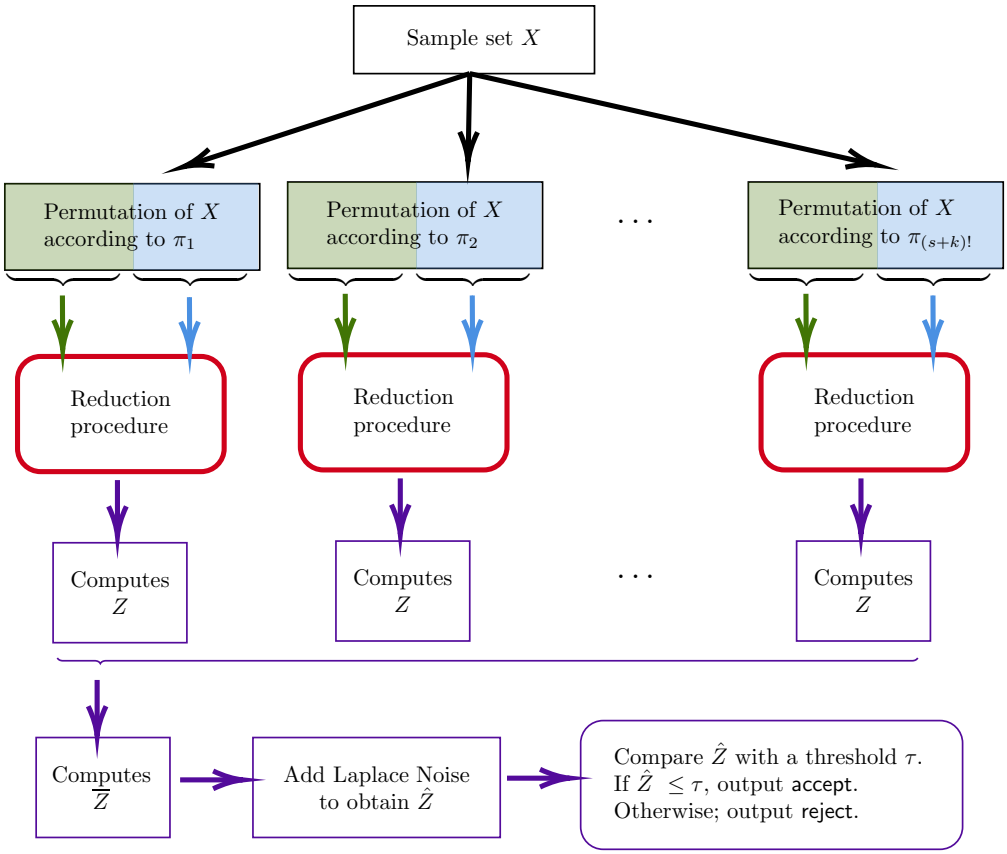

Figure 2: Our approach which uses $\overline{Z}$ instead of Z to reduce sensitivity.

Definition 3.2 (Proper procedure)**.** *Let $\mathcal{A}$ be a procedure that reduces testing property $\mathcal{P}$ to testing closeness of two distributions p and q over $[n]$ given an input sample set X. We say $\mathcal{A}$ is a* proper procedure *if $\mathcal{A}$ flattens p and q in such a way that the following holds for two non-negative constants $c_0 < 1$, and $c_1 \geq 1$:*

$$\mathbf{Pr}_X\Big[\mathbf{E}_\pi\Big[\, \|p^{(F)} - q^{(F)}\|_2^2 \,\Big|\, X, \pi\Big] \geq 4\, c_0 \cdot \mathbf{E}_F\Big[\, \|p^{(F)} - q^{(F)}\|_2^2 \,\Big]\Big] \geq 0.9\,, \tag{3}$$

$$\mathbf{E}_F\Big[\, \|p^{(F)} - q^{(F)}\|_4^4 \,\Big] \leq c_1 \cdot \Big(\mathbf{E}_F\Big[\, \|p^{(F)} - q^{(F)}\|_2^2 \,\Big]^2\Big)\,. \tag{4}$$

For the privacy guarantee, using the statistic $\overline{Z}$, we design a ξ-private algorithm. To make the statistic ξ-differentially private, we use a standard technique in differential privacy, the Laplace mechanism: we add $\mathbf{Lap}(\Delta(\overline{Z})/\xi)$ to the statistic, to make it ξ-differentially private. Note that $\Delta(\overline{Z})$ is the sensitivity of $\overline{Z}$. The importance of the idea of taking $\overline{Z}$ instead of Z as the statistic is that it results in a stable statistic with low sensitivity. Thus, the magnitude of the noise we add is small, and we can achieve nearly optimal sample complexity. Note that the exact value of $\Delta(\overline{Z})$ depends on the reduction procedure. We bound this quantity for each of the properties we consider separately. However, in this section, we state our result in terms of $\Delta(\overline{Z})$.

Finally, we propose Algorithm 1 for differentially private testing of property $\mathcal{P}$. We analyze the correctness of the algorithm in Theorem 3.3. We provide an illustration of our approach in Figure 2 in comparison with the standard approach in Figure 1.

Theorem 3.3. *Let $\mathcal{A}$ be a proper procedure for testing property $\mathcal{P}$ as defined in Definition 3.2. Suppose the*

Algorithm 1 A private procedure for property testing

procedure PRIVATE TESTER(n, ϵ, s, k)
  $X \leftarrow$ Draw $s+k$ samples, $x_1, x_2, \ldots, x_{s+k}$ from the underlying distribution(s).
  $\overline{Z} \leftarrow 0$
  for each permutation π **do**
    $T \leftarrow x_{\pi(1)}, \ldots, x_{\pi(s)}$
    $F \leftarrow x_{\pi(s+1)}, \ldots, x_{\pi(s+k)}$
    $S_p \leftarrow \mathcal{A}$ determines a set of samples from p using the test samples T
    $S_q \leftarrow \mathcal{A}$ determines a set of samples from q using the test samples T
    for $i = 1, 2, \ldots, n$ **do**
      $b_i \leftarrow \mathcal{A}$ determines the number of buckets for element i using flattening samples F.
      $s_{i,1} \leftarrow$ Number of occurrences of element i in S_p
      $s_{i,2} \leftarrow$ Number of occurrences of element i in S_q
      $\overline{Z} \leftarrow \overline{Z} + \frac{(s_{i,1}-s_{i,2})^2 - s_{i,1} - s_{i,2}}{b_i} \cdot \mathbf{Pr}[\text{probability of picking } \pi]$
  $n \leftarrow \mathcal{A}$ determines an upper bound for the new domain size.
  $\hat{Z} \leftarrow \overline{Z} + \mathbf{Lap}(\Delta(\overline{Z})/\xi)$
  if $\hat{Z} \leq \tau$ **then**
    Output accept.
  else
    Output reject.

expected number of test samples, s, is bounded from below:

$$s \geq \Theta\left(\frac{n \cdot \sqrt{\mathbf{E}_F\left[\min\left(\|p^{(F)}\|_2^2, \|q^{(F)}\|_2^2\right)\right]}}{\epsilon^2} + \frac{\sqrt{n\Delta(\overline{Z})}}{\epsilon\sqrt{\xi}}\right).$$

Then Algorithm 1 is a ξ-differentially private $(\epsilon, 3/4)$-tester for testing property $\mathcal{P}$.

Remark 3.4 (Sample complexity of the algorithm)**.** Our algorithm receives two parameter s and k for the number of test and flattening samples. Our analysis is based on the Poissonization method, so the algorithm is required to generate $\hat{s} := \mathbf{Poi}(s)$ samples from p and q, and $\hat{k} := \mathbf{Poi}(k)$ samples for flattening. In this section, for simplicity, we use s and k as the number of samples. Note that we assume that one can generate a sample from p and q using $\Theta(1)$ samples from the underlying distribution, and with probability 0.99, $\hat{s}$ and $\hat{k}$ are within a constant factor of their expectation. Thus, the sample complexity remains $\Theta(s+k)$.

Remark 3.5. Although the running time of Algorithm 1 is exponential as stated, one can run it in Poly(s) time as follows: For each domain element i and three numbers a and b, and c, one can calculate the probability of $S_{i,1} = a, s_{i,2} = b, b_i = c$, so once can compute $\mathbf{E}_\pi\left[(s_{i,1} - s_{i,2})^2 - s_{i,1} - s_{i,2})/b_i\right]$, and $\overline{Z}$ without trying all permutations π.

3.4 Applications of Our Framework

We use our general private tester to achieve differentially private algorithms for the two distribution testing problems we mentioned earlier: (i) Testing closeness of two distributions with unequal sized sample sets, (ii) Testing independence. As stated earlier, our approach is to use the non-private testers for these problems and show they satisfy the proper procedure definition (Definition 3.2). Then, using our general methodology, we achieve a near sample-optimal private testers for these problems. In particular, we have the following theorems. For more information, see the full version.

Theorem 3.6. *Suppose p and q are two distributions over $[n]$. There exists a ξ-differentially private $(\epsilon, 2/3)$-tester for closeness of p and q that uses $k_1 = \Omega(\max(n^{2/3}/\epsilon^{4/3}, \sqrt{n}/\epsilon\sqrt{\xi}))$ samples from p, $\Theta(\max(n/(\epsilon^2\sqrt{\min(n,k_1)}), \sqrt{n}/\epsilon^2, \sqrt{n}/\epsilon\sqrt{\xi}, 1/\epsilon^2\xi))$ from both p and q.*

Theorem 3.7. *Let p be a distribution over $[n] \times [m]$. There exists a ξ-differentially private $(\epsilon, 2/3)$-tester for the testing independence of p that uses $\Theta(s)$ samples where s is:*

$$s = \Theta\left(\frac{n^{2/3}m^{1/3}}{\epsilon^{4/3}} + \frac{(m\,n)^{1/2}}{\epsilon^2} + \frac{(m\,n\log n)^{1/2}}{\epsilon\sqrt{\xi}} + \frac{\log n}{\epsilon^2\xi}\right).$$

4 Proof of Theorem 3.3

Theorem 3.3. *Let $\mathcal{A}$ be a proper procedure for testing property $\mathcal{P}$ as defined in Definition 3.2. Suppose the expected number of test samples, s, is bounded from below:*

$$s \geq \Theta\left(\frac{n \cdot \sqrt{\mathbf{E}_F\left[\min\left(\|p^{(F)}\|_2^2, \|q^{(F)}\|_2^2\right)\right]}}{\epsilon^2} + \frac{\sqrt{n\Delta(\overline{Z})}}{\epsilon\sqrt{\xi}}\right).$$

Then Algorithm 1 is a ξ-differentially private $(\epsilon, 3/4)$-tester for testing property $\mathcal{P}$.

Proof: Note that Algorithm 1 computes $\overline{Z}$ based on the reduction procedure $\mathcal{A}$. Then, it obtains $\hat{Z}$ by adding Laplace noise to $\overline{Z}$. Finally, it compares $\hat{Z}$ with a threshold parameter τ (which we specify later) to makes its decision to output accept or reject. We prove that (1) the output of the algorithm is ξ-differentially private, and (2) the algorithm outputs the correct answer with high constant probability, i.e., the algorithm is an $(\epsilon, 3/4)$ tester.

Privacy Guarantee: By the standard properties of the Laplace mechanism, since the noise is drawn from a Laplace distribution with parameter $\Delta(\overline{Z})/\xi$, $\hat{Z}$ and consequently, the output of the algorithm are ξ-differentially private quantities. Thus, our algorithm satisfies the required privacy guarantee.

Correctness Guarantee: At a high level, to prove the correctness of the algorithm, we need to show that with high probability $\hat{Z} \leq \tau$ when $p = q$, and $\hat{Z} > \tau$ when p and q are ϵ-far from each other. Set the threshold τ to be $2c_0 s^2\epsilon^2/n$ where c_0 is the constant we have in Equation (3) for the proper procedure $\mathcal{A}$.

Recall that $\hat{Z}$ is defined $\overline{Z}$ plus noise. We start off by showing that the magnitude of the noise is at most $\tau/2$ with probability 0.99. Using the definition of the cumulative density function of the Laplace distribution, we have:

$$\mathbf{Pr}\left[|\hat{Z} - \overline{Z}| \geq \frac{\tau}{2}\right] \leq \exp\left(\frac{c_0\, s^2\epsilon^2\xi}{n\Delta(\overline{Z})}\right) \leq 0.01$$

where the last inequality is true if s is bounded from below as follows for a sufficiently large constant c_2:

$$s \geq c_2 \cdot \frac{\sqrt{n\Delta(\overline{Z})}}{\epsilon\sqrt{\xi}}.$$

Next, by analyzing the variance of $\overline{Z}$, and using Chebyshev's inequality, we show that $\overline{Z}$ is close to its expectation. More specifically, we prove the following claims:

- Completeness case: If p is equal to q, then $\overline{Z}$ at most than $\tau/2$ with high probability.
- Soundness case: If p is ϵ-far from q, then $\overline{Z}$ is at least $3\,\tau/2$ with high probability.

It is not difficult to prove the correctness of the algorithm if the above claims hold. Thus, for the rest of the proof, we show the two bullet points above about $\overline{Z}$. We introduce an auxiliary random variable W corresponding to each random variable Z which helps us in analyzing $\overline{Z}$:

$$W \coloneqq Z - s^2 \cdot \|p^{(F)} - q^{(F)}\|_2^2,$$

where F is the set of flattening samples. We define $\overline{W}$ to be the expected of W over the random choice of π and r: $\overline{W}(X) = \mathbf{E}_{\pi,r}[W|Z]$. We analyze the expected value, and the variance of $\overline{W}$ over the random

choice of X. Then, we use the concentration of $\overline{W}$ around its expectation to prove that $\overline{Z}$ must be around its expected value as well, and achieve the desired bound for $\overline{Z}$ with high probability via Chebyshev's inequality.

We define the following notations, $d_{\max}^{(F)}$ and $d_{\min}^{(F)}$ to indicate the maximum and the minimum of the two quantities $\|p^{(F)}\|_2^2$ and $\|q^{(F)}\|_2^2$ respectively:

$$d_{\max}^{(F)} = \max\left(\|p^{(F)}\|_2^2, \|q^{(F)}\|_2^2\right), \quad \text{and} \quad d_{\min}^{(F)} = \min\left(\|p^{(F)}\|_2^2, \|q^{(F)}\|_2^2\right).$$

The expected value and the variance of Z, as defined in Equation (1), is given in the proof of Proposition 3.1 in [CDVV14], if we fix the set of flattening samples F:

$$\mathbf{E}_{T,r}[Z|F] = s^2\|p^{(F)} - q^{(F)}\|_2^2, \text{ and} \qquad \mathbf{Var}_{T,r}[Z|F] \le 8s^3\sqrt{d_{\max}^{(F)}}\|p^{(F)} - q^{(F)}\|_4^2 + 8s^2 d. \tag{5}$$

Note that since the samples in X are independent, the order of the samples cannot change the expected value or the variance. Thus, by symmetrization, we can fix an order of the samples, namely π_0, and the expected value of W for any other permutation must be the same as for π_0. Since T and F are completely separated and independent, by Equation (5), we have:

$$\begin{aligned}\mathbf{E}_X\left[\overline{W}\right] &= \mathbf{E}_X[\mathbf{E}_{\pi,r}[W|X]] = \mathbf{E}_X[\mathbf{E}_r[W|X,\pi_0]] \\ &= \mathbf{E}_F\left[\mathbf{E}_{T,r}\left[\, Z - s^2 \cdot \|p^{(F)} - q^{(F)}\|_2^2 \,\middle|\, F\right]\right] = 0\end{aligned} \tag{6}$$

Moreover, given the variance bound in the Equation (5), we obtain the following bound for the variance of $\overline{W}$:

$$\begin{aligned}\mathbf{Var}_X\left[\overline{W}\right] &= \mathbf{Var}_X[\mathbf{E}_{\pi,r}[W|X]] = \mathbf{Var}_X\left[\sum_{\pi}\mathbf{E}_r[W|X,\pi]\cdot\mathbf{Pr}[\pi]\right] \\ &= \sum_{\pi_1}\sum_{\pi_2}\mathbf{Pr}[\pi_1]\cdot\mathbf{Pr}[\pi_2]\cdot\mathbf{Cov}_X(\mathbf{E}_r[W|X,\pi_1],\, \mathbf{E}_r[W|X,\pi_2]) \\ &\le \frac{1}{2}\sum_{\pi_1}\sum_{\pi_2}\mathbf{Pr}[\pi_1]\cdot\mathbf{Pr}[\pi_2]\cdot\left(\mathbf{Var}_X[\mathbf{E}_r[W|X,\pi_1]] + \mathbf{Var}_X[\mathbf{E}_r[W|X,\pi_2]]\right) \\ &= \mathbf{Var}_X[\mathbf{E}_r[W|X,\pi_0]] = \mathbf{Var}_{F,T}[\mathbf{E}_r[W|F,T]] \\ &= \mathbf{E}_F\left[\mathbf{Var}_T[\mathbf{E}_r[W|F]]\right] + \mathbf{Var}_F[\mathbf{E}_{T,r}[W|F]] \\ &\le \mathbf{E}_F\left[8\,s^3\cdot\sqrt{d_{\max}^{(F)}}\cdot\|p^{(F)} - q^{(F)}\|_4^2 + 8\,s^2\cdot d_{\max}^{(F)}\right] + 0\,.\end{aligned} \tag{7}$$

Now, we are ready to analyze $\overline{Z}$ using the expected value and the variance of $\overline{W}$, which we have bounded above. We analyze the completeness case and the soundness case separately as below.

Completeness case: p is equal to q. If p is equal to q, for any flattening set F, $p^{(F)}$ and $q^{(F)}$ are equal. Thus, $\|p^{(F)} - q^{(F)}\|_2^2$ is zero, and $\overline{W}$ is always equal to $\overline{Z}$ by definition. In fact, we have $\overline{Z} = \overline{W} = \overline{W} - \mathbf{E}_X\left[\overline{W}\right]$ using Equation (6). Also, the ℓ_2-norms of $p^{(F)}$ and $q^{(F)}$ are the same. Thus, the minimum and the maximum of $\|p^{(F)}\|_2^2$ and $\|q^{(F)}\|_2^2$ are equal which implies: $d_{\min}^{(F)} = d_{\max}^{(F)}$. Now, by applying Chebyshev's inequality for $\overline{W}$, the probability of $\overline{Z}$ be above $\tau/2$ is bounded as follows:

$$\mathbf{Pr}_X\left[\overline{Z} \ge \frac{\tau}{2}\right] \le \mathbf{Pr}_X\left[\left|\overline{W} - \mathbf{E}_X\left[\overline{W}\right]\right| \ge \frac{c_0\, s^2\epsilon^2}{n}\right] \le \frac{n^2\mathbf{Var}_X\left[\overline{W}\right]}{c_0^2\, s^4\, \epsilon^4} \le \frac{8\, n^2\cdot\mathbf{E}_F\left[d_{\max}^{(F)}\right]}{c_0^2\, s^2\, \epsilon^4}$$

Clearly, the above quantity is at most 0.01 if for a large constant c_3, s is at least:

$$s \ge c_3\cdot\frac{n\cdot\sqrt{\mathbf{E}_F\left[d_{\min}^{(F)}\right]}}{\epsilon^2}\,. \tag{8}$$

Soundness case: p is ϵ-far from q. Before showing that $\overline{W}$ is close to zero with high probability, we establish two inequalities below. First, we show that the expected ℓ_2-distance between $p^{(F)}$ and $q^{(F)}$. Observe that flattening does not change the ℓ_1-distance between two distributions due to the following:

$$\|p-q\|_1 = \sum_{i=i}^{n} |p(i)-q(i)| = \sum_{i=1}^{n}\sum_{j=1}^{b_i} \frac{|p(i)-q(i)|}{b_i} = \|p^{(F)}-q^{(F)}\|\,.$$

Thus, by the Cauchy-Schwarz inequality, we have the following lower bound for the expected ℓ_2-distance between $p^{(F)}$ and $q^{(F)}$ for any F:

$$\mathbf{E}_F\Big[\,\|p^{(F)}-q^{(F)}\|_2^2\Big] \geq \mathbf{E}_F\left[\frac{\|p^{(F)}-q^{(F)}\|_1^2}{n}\right] \geq \frac{\|p-q\|_1^2}{n} \geq \frac{\epsilon^2}{n}\,. \tag{9}$$

Second, we provide the following lemma to show a bound for $\mathbf{E}_F\Big[d_{\max}^{(F)}\Big]$. This term appears in the bound for the variance of W (Equation (7)), and the following bound will help us to prove concentration of $\overline{W}$ and $\overline{Z}$. This lemma is a crucial tool to show that the sample complexity only depends on $d_{\min}^{(F)}$, the minimum ℓ_2-norm of $p^{(F)}$ and $q^{(F)}$. Therefore, we are only required to make sure one of the two distributions has a small ℓ_2-norm.

Lemma 4.1. *Assume F is a random set of samples to be used for flattening. Then, we have:*

$$\mathbf{E}_F\Big[d_{\max}^{(F)}\Big] \leq 6\left(\mathbf{E}_F\Big[d_{\min}^{(F)}\Big] + \mathbf{E}_F\Big[\,\|p^{(F)}-q^{(F)}\|_2^2\Big]\right)$$

Proof: Consider an arbitrary F, and two distributions $p^{(F)}$ and $q^{(F)}$. Without loss of generality suppose $\left\|p^{(F)}\right\|_2^2$ is at least $\left\|q^{(F)}\right\|_2^2$. To prove our bound, it is sufficient to prove the following inequality:

$$d_{\max}^{(F)} = \left\|p^{(F)}\right\|_2^2 \leq 3\left\|q^{(F)}\right\|_2^2 + 6\left\|p^{(F)}-q^{(F)}\right\|_2^2 = 3\,d_{\min}^{(F)} + 6\left\|p^{(F)}-q^{(F)}\right\|_2^2$$

We show if $\left\|p^{(F)}\right\|_2^2$ is at least $3\left\|q^{(F)}\right\|_2^2$, then $\left\|p^{(F)}\right\|_2^2$ is at most $6\left\|p^{(F)}-q^{(F)}\right\|_2^2$. By the Cauchy-Schwarz inequality, we have:

$$\begin{aligned}
\frac{4}{9}\cdot\left\|p^{(F)}\right\|_2^4 &= \left(\left\|p^{(F)}\right\|_2^2 - \frac{1}{3}\left\|p^{(F)}\right\|_2^2\right)^2 \leq \left(\left\|p^{(F)}\right\|_2^2 - \left\|q^{(F)}\right\|_2^2\right)^2 \leq \left(\sum_i \left(p^{(F)}(i)\right)^2 - \left(q^{(F)}(i)\right)^2\right)^2 \\
&= \left(\sum_i (p^{(F)}(i)-q^{(F)}(i))(p^{(F)}(i)+q^{(F)}(i))\right)^2 \leq \left(\sum_i (p^{(F)}(i)-q^{(F)}(i))^2\right)\cdot\left(\sum_i (p^{(F)}(i)+q^{(F)}(i))^2\right) \\
&= \left\|p^{(F)}-q^{(F)}\right\|_2^2\cdot\left(\sum_i (p^{(F)}(i)+q^{(F)}(i))^2\right) \leq \left\|p^{(F)}-q^{(F)}\right\|_2^2\cdot\left(\sum_i 2\left(\left(p^{(F)}(i)\right)^2+\left(q^{(F)}(i)\right)^2\right)\right) \\
&= \left\|p^{(F)}-q^{(F)}\right\|_2^2\cdot\left(2\left\|p^{(F)}\right\|_2^2 + 2\left\|q^{(F)}\right\|_2^2\right) \leq \left\|p^{(F)}-q^{(F)}\right\|_2^2\cdot\left(2+\frac{2}{3}\right)\cdot\left\|p^{(F)}\right\|_2^2\,,
\end{aligned}$$

where the first inequality is due to our assumption that $3\left\|q^{(F)}\right\|_2^2 \leq \left\|p^{(F)}\right\|_2^2$. The above equation implies that

$$\left\|p^{(F)}\right\|_2^2 \leq 6\left\|p^{(F)}-q^{(F)}\right\|_2^2$$

Therefore, one can conclude either $\|p\|_2^2$ has to be less than $3\,\|q\|_2^2$, or it will be at least $6\,\|p^{(F)}-q^{(F)}\|_2^2$. Since it is true for any F, the statement of the lemma is concluded. □

Now, we used the two previous inequalities to show that the probability of $\overline{W}$ being far from its expec-

tation, i.e., zero, is bounded. By Chebyshev's inequality, we have:

$$\begin{aligned}
\mathbf{Pr}_X\Big[\Big|\overline{W} - \mathbf{E}_X[\overline{W}]\Big| \geq c_0 \cdot \mathbf{E}_F\Big[s^2\|p^{(F)} - q^{(F)}\|_2^2\Big]\Big] &\leq \frac{\mathbf{Var}_X[\overline{W}]}{c_0^2 \cdot \mathbf{E}_F\Big[s^2\|p^{(F)} - q^{(F)}\|_2^2\Big]^2} \\
&\leq \frac{\mathbf{E}_F\Big[8s^3\sqrt{d_{\max}^{(F)}}\|p^{(F)} - q^{(F)}\|_4^2 + 8s^2\, d_{\max}^{(F)}\Big]}{c_0^2 \cdot \mathbf{E}_F\Big[s^2\|p^{(F)} - q^{(F)}\|_2^2\Big]^2} \\
&\leq \Theta\left(\frac{\sqrt{\mathbf{E}_F\Big[d_{\max}^{(F)}\Big]} \cdot \sqrt{\mathbf{E}_F\Big[\|p^{(F)} - q^{(F)}\|_4^4\Big]}}{s \cdot \mathbf{E}_F\Big[\|p^{(F)} - q^{(F)}\|_2^2\Big]^2} + \frac{\mathbf{E}_F\Big[d_{\max}^{(F)}\Big]}{s^2 \cdot \mathbf{E}_F\Big[\|p^{(F)} - q^{(F)}\|_2^2\Big]^2}\right),
\end{aligned}$$

where the last inequality is due to the Cauchy-Schwarz inequality. Note that we assume $\mathcal{A}$ was a proper procedure, so Equation (4) holds: there exists a constant c_1 such that

$$\mathbf{E}_F\Big[\|p^{(F)} - q^{(F)}\|_4^4\Big] \leq c_1 \cdot \Big(\mathbf{E}_F\Big[\|p^{(F)} - q^{(F)}\|_2^2\Big]^2\Big).$$

Hence, we obtain:

$$\begin{aligned}
&\mathbf{Pr}_X\Big[\Big|\overline{W} - \mathbf{E}_X[\overline{W}]\Big| \geq c_0 \cdot \mathbf{E}_F\Big[s^2\|p^{(F)} - q^{(F)}\|_2^2\Big]\Big] \\
&\qquad \leq \Theta\left(\frac{\sqrt{\mathbf{E}_F\Big[d_{\max}^{(F)}\Big]}}{s \cdot \mathbf{E}_F\Big[\|p^{(F)} - q^{(F)}\|_2^2\Big]} + \frac{\mathbf{E}_F\Big[d_{\max}^{(F)}\Big]}{s^2 \cdot \mathbf{E}_F\Big[\|p^{(F)} - q^{(F)}\|_2^2\Big]^2}\right).
\end{aligned}$$

Note that in the last line above, the second term is the squared of the first term, so it is sufficient to make sure that the first term is smaller than 0.01 by a constant factor. Using this bound and Lemma 4.1, we get:

$$\begin{aligned}
\mathbf{Pr}_X\Big[\Big|\overline{W} - \mathbf{E}_X[\overline{W}]\Big| \geq c_0 \cdot \mathbf{E}_F\Big[s^2\|p^{(F)} - q^{(F)}\|_2^2\Big]\Big] &\leq \Theta\left(\frac{\sqrt{\mathbf{E}_F\Big[d_{\max}^{(F)}\Big]}}{s \cdot \mathbf{E}_F\Big[\|p^{(F)} - q^{(F)}\|_2^2\Big]}\right) \\
&\leq \Theta\left(\frac{\sqrt{\mathbf{E}_F\Big[d_{\min}^{(F)}\Big]} + \sqrt{\mathbf{E}_F\Big[\|p^{(F)} - q^{(F)}\|_2^2\Big]}}{s \cdot \mathbf{E}_F\Big[\|p^{(F)} - q^{(F)}\|_2^2\Big]}\right) \\
&\leq \Theta\left(\frac{\sqrt{\mathbf{E}_F\Big[d_{\min}^{(F)}\Big]}}{s \cdot \mathbf{E}_F\Big[\|p^{(F)} - q^{(F)}\|_2^2\Big]} + \frac{1}{s\sqrt{\mathbf{E}_F\Big[\|p^{(F)} - q^{(F)}\|_2^2\Big]}}\right).
\end{aligned}$$

Recall that we show earlier in Equation (9) that the expected value of the ℓ_2-distance squared is at least ϵ^2/n. This bound results in the following upper bound for the probability of $\overline{W}$ being far from its expectation:

$$\begin{aligned}
&\mathbf{Pr}_X\Big[\Big|\overline{W} - \mathbf{E}_X[\overline{W}]\Big| \geq c_0 \cdot \mathbf{E}_F\Big[s^2\|p^{(F)} - q^{(F)}\|_2^2\Big]\Big] \\
&\qquad \leq \Theta\left(\frac{n \cdot \sqrt{\mathbf{E}_F\Big[d_{\min}^{(F)}\Big]}}{s\,\epsilon^2} + \frac{\sqrt{n}}{s\,\epsilon}\right) = \Theta\left(\frac{n \cdot \sqrt{\mathbf{E}_F\Big[d_{\min}^{(F)}\Big]}}{s\,\epsilon^2}\right)
\end{aligned}$$

where the last equality is true, because $d_{\min}^{(F)}$ is never smaller than $1/n$. Also, the above probability is bounded

by 0.01 if s is larger than the bound given below for a sufficiently large constant, c_4.

$$s \geq c_4 \cdot \left(\frac{n \cdot \sqrt{\mathbf{E}_F\left[d_{\min}^{(F)}\right]}}{\epsilon^2} \right) .$$

Now, we analyze $\overline{Z}$ and show $\overline{Z}$ has to be at least $3\,\tau/2$. Recall that since $\mathcal{A}$ is a proper procedure, we know that Equation (3) holds with high probability:

$$\mathbf{E}_\pi\left[\|p^{(F)} - q^{(F)}\|_2^2 \middle| X, \pi \right] \geq 4\,c_0 \cdot \mathbf{E}_F\left[\|p^{(F)} - q^{(F)}\|_2^2 \right] .$$

We show earlier that with high probability $\left|\overline{W}\right| = \left|\overline{W} - \mathbf{E}_X\left[\overline{W}\right]\right| \leq c_0 \cdot \mathbf{E}_F\left[s^2\,\|p^{(F)} - q^{(F)}\|\right]$. Now, by definition of $\overline{W}$ and its concentration around its expectation, we achieve the desired lower bound for $\overline{Z}$:

$$\begin{aligned}\overline{Z} &= \overline{W} + \mathbf{E}_\pi\left[s^2\|p^{(F)} - q^{(F)}\|_2^2\right] \geq -c_0 \cdot \mathbf{E}_F\left[s^2\|p^{(F)} - q^{(F)}\|_2^2\right] + 4\,c_0 \cdot \mathbf{E}_F\left[s^2\|p^{(F)} - q^{(F)}\|_2^2\right] \\ &\geq 3\,c_0 \cdot \mathbf{E}_F\left[s^2\|p^{(F)} - q^{(F)}\|_2^2\right] \geq \frac{3\,c_0\,s^2\epsilon^2}{n} \geq \frac{3\,\tau}{2} .\end{aligned}$$

By taking the union bound, the probability of having too large Laplace noise or a too large $\left|\overline{W}\right|$ is at most 0.02. Moreover, Equation (4) and Equation (3) do not hold with probability at most 0.1 each. Thus, with probability at least 3/4, $\hat{Z}$ is on the correct side of the threshold τ, and the algorithm output the correct answer. □

5 Testing Closeness of Distributions with Unequal Sample Sizes

In this section, we prove that there exists a non-private algorithm for testing closeness using unequal sample sizes which is a proper procedure. We use the non-private algorithm in [DK16] with some small modifications. Then, we turn it into a private tester using our general private closeness tester provided in Section 3. We also analyze the sensitivity of the statistic $\overline{Z}$, and the exact sample complexity of the tester.

We introduce a proper procedure $\mathcal{A}$ for testing closeness of p and q using unequal sample sizes. First, we explain how $\mathcal{A}$ works. To generate a sample from p (or q) the algorithm simply draw an i.i.d. sample from p (or q). Assume k_1, k_2, s_1, and s_2 are four parameters that we determine later. $\mathcal{A}$ draws $s_1 + k_1$ from p and $s_2 + k_2$ samples from q. For the number of buckets, $\mathcal{A}$ uses the following process. Let F be the number of a set of k_1 samples from p and k_2 samples from q. The number of buckets for element i is determined by the number of instances of i in F plus one.

Theorem 5.1. *There exists a ξ-differentially private algorithm that uses $k_1 = \Omega(\max(n^{2/3}/\epsilon^{4/3}, \sqrt{n}/\epsilon\sqrt{\xi}))$ samples from p, $\Theta(\max(n/(\epsilon^2\sqrt{\min(n,k_1)}), \sqrt{n}/\epsilon^2, \sqrt{n}/\epsilon\sqrt{\xi}, 1/\epsilon^2\xi))$ from both p and q and distinguishes the following cases with probability at least 0.8:*

- *Completeness case: $p = q$*
- *Soundness case: $\|p - q\|_1 > \epsilon$.*

Proof: The goal is to transform the problem to the generate tester we provided in Section 3. First, in Lemma 5.2 we show that the non-private algorithm in [DK16] is a "proper procedure". Using Theorem 3.3, the existence of the tester with the sample complexity s for the test part is immediate where s is at least the bound bellow

$$s \geq \Theta\left(\frac{n' \cdot \sqrt{\mathbf{E}_F\left[\min\left(\|p^{(F)}\|_2^2, \|q^{(F)}\|_2^2\right)\right]}}{\epsilon^2} + \frac{\sqrt{n'\Delta(\overline{Z})}}{\epsilon\sqrt{\xi}} \right) . \tag{10}$$

We first show the relationship between s above and the rest of the parameters we have. Then we set the parameters k_1, k_2, and s and analyze the sample complexity. Without loss of generality assume $k_1 \geq k_2$.

Note that after flattening the size of the domain increases to $n' = \Theta(n + k_1 + k_2)$ with high probability. Then, in Lemma 5.5, we show that the proposed statistic, $\overline{Z}$, has a bounded sensitivity:

$$\Delta(\overline{Z}) \leq \Theta\left(\frac{k_1}{k_1 + s} \cdot \left(\frac{s + k_2}{k_2}\right)^2\right).$$

In addition, it is shown in [DK16] that the probability of the expected ℓ_2-norm of p after flattening is at most $1/k_1$. Moreover, adding more flattening samples does not increase this quantity. Hence, we have:

$$\mathbf{E}_F\Big[\min\left(\|p^{(F)}\|_2^2, \|q^{(F)}\|_2^2\right)\Big] \leq \frac{1}{k_1}$$

We consider the following cases:

- **case 1: $\epsilon = \Omega(\mathbf{n^{-1/4}})$ and $\epsilon^2\xi = \Omega(\mathbf{n^{-1}})$:** In this case, we have the following properties:

$$\Theta\left(\frac{\sqrt{n}}{\epsilon^2}\right) \leq \Theta\left(\frac{n^{2/3}}{\epsilon^{4/3}}\right) \leq \Theta(n), \quad \text{and} \quad \Theta\left(\frac{1}{\epsilon^2\xi}\right) \leq \Theta\left(\frac{\sqrt{n}}{\epsilon\sqrt{\xi}}\right) \leq \Theta(n).$$

  Let k_1 be a number in the range below:

$$\Theta\left(\max\left(\frac{n^{2/3}}{\epsilon^{4/3}} + \frac{\sqrt{n}}{\epsilon\sqrt{\xi}}\right)\right) \leq k_1 \leq \Theta(n).$$

  Hence, n' is $\Theta(n)$. Then, we set s and k_2 as follows:

$$k_2 := s := \Theta\left(\max\left(\frac{n}{\epsilon^2\sqrt{k_1}}, \frac{\sqrt{n}}{\epsilon\sqrt{\xi}}\right)\right)$$

  $\Delta(\overline{Z})$ is $O(1)$ in this case. Therefore, s is $\Omega(n/\epsilon^2\sqrt{k_1} + \sqrt{n\Delta(\overline{Z})}/(\epsilon\sqrt{\xi}))$ and the condition in Equation (10) holds.

- **case 2: $\epsilon = \mathbf{o}(\mathbf{n^{-1/4}})$:** In this case, $\sqrt{n}/\epsilon^2$ is $\Omega(n)$. Thus, we cannot avoid sample complexity of $\Omega(n)$. We set k_1 and k_2 to be equal to n, and we set s to be the following:

$$s := \Theta\left(\max\left(\frac{\sqrt{n}}{\epsilon^2}, \frac{1}{\epsilon^2\xi}\right)\right).$$

  Clearly n' is still $\Theta(n)$, and s is $\Omega(n/\sqrt{k_1}\epsilon^2)$. In this case $\Delta(\overline{Z})$ is $\Theta(s/n)$. Hence, in order to have s at least $\Omega\sqrt{n'\Delta(\overline{Z})}/\epsilon\sqrt{\xi}$, it suffices to have $s = \Omega(1/\epsilon\sqrt{\xi})$.

□

5.1 Non-Private Closeness Tester Is a Proper Procedure

Lemma 5.2. *Procedure $\mathcal{A}$ explained above is a proper procedure according to Definition 3.2.*

Proof: First, we show the number of samples we generate is not too far from their expectation. Hence X can be a set with a bounded number of samples. In the following lemma we show if the means are larger than a fixed constant, then with probability 0.01 we can assume the number of samples from each of distributions is at most three times larger than their means.

Lemma 5.3. *Assume random variable x is drawn from* $\mathbf{Poi}(\lambda)$. *If λ is at least $1.5 \cdot \ln(1/c)$, then the probability of x being larger than $3\,\lambda$ is at most $1 - c$.*

Now, we only need to show that inequalities in Equation (3) and Equation (4) are correct. Before proving the equations, we provide an insightful information about the distribution over the b_i's. It is clear that for

a fixed i, $b_i - 1$ is an independent Poisson random variable with mean $k_1 p(i) + k_2 q(i)$. More precisely, we can think of $b_i - 1$ as the sum of two random variables $b_{i,1} \sim \mathbf{Poi}(k_1 p(i))$ and $b_{i,2} \sim \mathbf{Poi}(k_2 q(i))$ plus one. However, assume a random set of samples, X, is given to us with $t_{i,1}$ instances of i from p, and $t_{i,2}$ instances of i from q. Then, considering a random permutation of samples, then $b_{i,j}$ is a binomial random variable from $\mathbf{Bin}(t_{i,j}, k_j/(k_j + s))$ for $j = 1, 2$.

Now, we focus on proving Equation (3). Fix a set of sample X and a domain element i. Using Jensen's inequality, we have

$$\mathbf{E}_\pi\left[\frac{1}{b_i(X,\pi)}\right] = \mathbf{E}_{b_{i,1},b_{i,2}}\left[\frac{1}{b_i}\right] \geq \frac{1}{\mathbf{E}_{b_{i,1},b_{i,2}}[b_i]} = \frac{1}{t_{i,1}\, k_1/(k_1+s) + t_{i,2}\, k_2/(k_2+s) + 1}\,.$$

Note that by Markov's inequality the probability of any of $t_{i,1}$ or $t_{i,2}$ being 50 times[3] larger than their expectations is at most 0.04. Therefore, with probability 0.96 assume they are at most 50 times their expectation. Since $t_{i,1}$ and $t_{i,2}$ are two Poisson random variables with means $p(i)(k_1 + s)$ and $q(i)(k_2 + s)$, we can bound the above quantity as follows:

$$\mathbf{E}_\pi\left[\frac{1}{b_i(\pi)}\right] \geq \frac{1}{50\, p(i)k_1 + 50\, q(i)k_2 + 1} = \frac{1}{50\lambda + 1}$$

where we use λ to denote $p(i)k_1 + q(i)k_2$. On the other hand, the expected value of $1/b_i$ when X has not be observed is the following:

$$\mathbf{E}_F\left[\frac{1}{b_i}\right] = \mathbf{E}_{x\sim\mathbf{Poi}(\lambda)}\left[\frac{1}{x+1}\right] = \frac{1-e^{-\lambda}}{\lambda} \leq \frac{65}{50\lambda+1}\,.$$

Putting all of the above facts together, we conclude:

$$\mathbf{Pr}_{t_{i,1},t_{i,2}}\left[\mathbf{E}_\pi\left[\frac{(p(i)-q(i))^2}{b_i(\pi)}\right] \geq \frac{1}{65}\cdot\mathbf{E}_F\left[\frac{(p(i)-q(i))^2}{b_i}\right]\right] \geq 0.96\,. \tag{11}$$

We define a random variable x_i over the randomness of $t_{i,1}$ and $t_{i,2}$ to be the following:

$$x_i \coloneqq \mathbf{E}_\pi\left[\frac{(p(i)-q(i))^2}{b_i(\pi)}\right]$$

Note that by the *Poissonization method*, all the number of instances of a particular element are independent form the rest. Hence, x_i's are independent given the independence of $t_{i,j}$'s. In addition, we prove the following lemma, to bound the sum of x_i's from below:

Lemma 5.4. *Assume we have n independent random variables $x_1, x_2, \ldots, x_n$ in the range $[0, +\infty)$. Suppose each x_i is at least A_i with probability $p \geq 0.95$ where A_i is a fixed number. Then, with probability at least 0.9, $\sum_{i=1}^n x_i$ is at least $0.1\sum_{i=1}^n A_i$.*

For the proof of the lemma, see Section 7.

Using Equation (11), and Lemma 5.4, with probability 0.9 we have:

$$\begin{aligned}\mathbf{E}_\pi\left[\|p^{(F)} - q^{(F)}\|_2^2\right] = \mathbf{E}_\pi\left[\frac{(p(i)-q(i))^2}{b_i(\pi)}\right] &\geq \frac{1}{650}\cdot\sum_{i=1}^n \mathbf{E}_F\left[\frac{(p(i)-q(i))^2}{b_i}\right] \\ &= 4c_0\cdot\mathbf{E}_F\left[\|p^{(F)} - q^{(F)}\|_2^2\right]^2.\end{aligned}$$

where $c_0 = 1/26000$. Hence, the proof of Equation (3) is complete.

Now, we focus on proving Equation (4). To prove the inequality, it suffices to show that $\mathbf{E}_F\left[1/b_i^3\right]$ is $O\left(\mathbf{E}_F[1/b_i]^2\right)$. Note one can think of b_i to be equal to $x + 1$ where x is a Poisson random variable with

[3] Needless to say, we are not optimizing constants here.

mean $\lambda' = p(i)k_1 + q(i)k_2$. Thus, we have:

$$\begin{aligned}\mathbf{E}_F\left[\frac{1}{b_i^3}\right] &= \mathbf{E}_{x\sim\mathbf{Poi}(\lambda')}\left[\frac{1}{(x+1)^3}\right] \leq \mathbf{E}\left[\frac{6}{(x+1)(x+2)(x+3)}\right] \leq 6\cdot\sum_{x=0}^{\infty}\frac{e^{-\lambda'}\lambda'^x}{(x+3)!} \\ &= \frac{6}{\lambda'^3}\sum_{y=3}^{\infty}\frac{e^{-\lambda'}\lambda'^y}{y!} = \frac{6\left(1-e^{-\lambda'}-e^{-\lambda'}\lambda'-e^{-\lambda'}\lambda'^2/2\right)}{\lambda'^3} \leq 6\cdot\left(\frac{1-e^{-\lambda'}}{\lambda'}\right)^2.\end{aligned} \tag{12}$$

On the other hand, we can compute the expected value of $1/b_i$ as follows:

$$\left(\mathbf{E}_F\left[\frac{1}{b_i}\right]\right)^2 = \left(\sum_{x=0}^{\infty}\frac{e^{-\lambda'}\lambda'^x}{(x+1)!}\right)^2 = \left(\frac{1}{\lambda'}\sum_{y=1}^{\infty}\frac{e^{-\lambda'}\lambda'^y}{y!}\right)^2 = \left(\frac{1-e^{-\lambda'}}{\lambda'}\right)^2.$$

Putting these two equations together, one can conclude the Equation (4):

$$\begin{aligned}\mathbf{E}_F\left[\|p^{(F)}-q^{(F)}\|_4^4\right] &= \sum_{i=1}^{n}\mathbf{E}_F\left[\frac{(p(i)-q(i))^4}{b_i^3}\right] \leq 6\cdot\sum_{i=1}^{n}\left(\mathbf{E}_F\left[\frac{(p(i)-q(i))^2}{b_i}\right]\right)^2 \\ &\leq 6\cdot\left(\sum_{i=1}^{n}\mathbf{E}_F\left[\frac{(p(i)-q(i))^2}{b_i}\right]\right)^2 = 6\cdot\mathbf{E}_F\left[\|p^{(F)}-q^{(F)}\|_2^2\right]^2.\end{aligned}$$

Therefore, the statement of the lemma is concluded. □

5.2 Bounding the Sensitivity

In this section, we provide an upper bound for the sensitivity of $\overline{Z}$, which is the amount that the statistic changes if we change one sample in the input. We use the following notation in this section: We use a $j \in 1, 2$ subscript to indicate corresponding quantities to the distribution p and q. For example, X_1 denotes the sample set from p, and X_2 denotes the sample set form q. Let m_j denote the number of samples in X_j. We indicate the number of instances of i in X_j by $t_{i,j}$. We use k_j and s_j for $j \in \{1, 2\}$ to indicate the number of flattening samples, and the number of the test samples from p and q. In addition, for a fixed permutation π and the corresponding flattening and test sets F_j and T_j, we use the following notation for an element i: $k_{i,j}$ is the number of instances of element i in the set F_j. And, $s_{i,j}$ is the number of instances of element i in the set T_j.

We consider another sample set, which differs with $X_1 \cup X_2$ in one location. We use the prime symbol to designate that the values are corresponding to the second sample set. Without loss of generality, we assume that if an instance of i in X_1 has been replaced by i', we will obtain X_1'. Since we assumed that the sample sets are different only in one location, X_2 has to be equal to X_2'.

Recall the definition of $\overline{Z}$ as in Equation (2):

$$\overline{Z} = \mathbf{E}_\pi\left[\sum_{i=1}^{n}\frac{(s_{i,1}-s_{i,2})^2 - s_{i,1} - s_{i,2}}{b_i}\middle| X\right].$$

Note that $\overline{Z}$ and $\overline{Z}'$ only differ in the i-th and the i'-th terms in the sum. We bound the difference of the i-th term in $\overline{Z}$ and $\overline{Z}'$), and we can bound the i'-th term from above similarly.

Now, we formally state the upper bound for the sensitivity of $\overline{Z}$ in the following lemma:

Lemma 5.5. *Let $\overline{Z}(X)$ be a function of a sample set X as defined in Equation 2. The sensitivity of $\overline{Z}(X)$ is bounded by:*

$$\Delta(Z) \leq \Theta\left(\frac{k_1}{k_1+s}\cdot\left(\frac{s+k_2}{k_2}\right)^2 + \frac{k_2}{k_2+s}\cdot\left(\frac{s+k_1}{k_1}\right)^2\right)$$

Proof: Since we assumed that X_1 has one more instance of i, we have $t_{i,1} = t'_{i,t} - 1$. For a fixed permutation

π, the extra instances of i can end up in the flattening set or the test set. Thus, we consider the following cases, and compute the conditional expected value in each case:

Case 1: The extra instance of i is in the flattening set F_1, which implies $s_{i,1} = s'_{i,1}$ and $k_{i,1} = k'_{i,1} + 1$. Therefore, we have:

$$\begin{aligned}
&\left|\frac{(s_{i,1}-s_{i,2})^2 - s_{i,1} - s_{i,2}}{k_{i,1}+k_{i,2}+1} - \frac{(s'_{i,1}-s'_{i,2})^2 - s'_{i,1} - s'_{i,2}}{k'_{i,1}+k'_{i,2}+1}\right| \\
&\quad = \left|\frac{(s_{i,1}-s_{i,2})^2 - s_{i,1} - s_{i,2}}{(k'_{i,1}+1)+k_{i,2}+1} - \frac{(s_{i,1}-s_{i,2})^2 - s_{i,1} - s_{i,2}}{k'_{i,1}+k_{i,2}+1}\right| \\
&\quad \le 2\,t_{i,\ell}^2 \cdot \left|\frac{1}{k'_{i,1}+k_{i,2}+2} - \frac{1}{k'_{i,1}+k_{i,2}+1}\right| \le \frac{2\,t_{i,\ell}^2}{\left(k'_{i,1}+k_{i,2}+2\right)\cdot\left(k'_{i,1}+k_{i,2}+1\right)} \\
&\quad \le \min\left(\frac{2\,t_{i,\ell}^2}{\left(k'_{i,1}+2\right)\cdot\left(k'_{i,1}+1\right)}, \frac{2\,t_{i,\ell}^2}{(k_{i,2}+2)\cdot(k_{i,2}+1)}\right)
\end{aligned}$$

where Let ℓ be the index in $\{1,2\}$ for which $t_{i,\ell}$ denotes the maximum of $t_{i,1}$ and $t_{i,2}$.

Now, we focus on computing the conditional expected value. We use the above inequality. Depending on ℓ being one or two, we pick the first or the second term in the last line of the equation above. Note that $t_{i,\ell}$ is the same, no matter which permutation we pick. Also, it is not too hard to see $k'_{i,1}$ can be viewed as a random variable from hypergeometric distribution: $\mathbf{HG}(m_1, t'_{i,1}, k_2)$. Similarly, $k_{i,2}$ is a random variable drawn from $\mathbf{HG}(m_2, t_{i,2}, k_2)$. Using Lemma 7.5, we can bound the conditional expectation of the differences of these two terms:

$$\mathbf{E}_\pi\left[\left|\frac{(s_{i,1}-s_{i,2})^2 - s_{i,1} - s_{i,2}}{k_{i,1}+k_{i,2}+1} - \frac{(s'_{i,1}-s'_{i,2})^2 - s'_{i,1} - s'_{i,2}}{k'_{i,1}+k'_{i,2}+1}\right| \,\middle|\, \text{extra copy of } i \text{ is in } F_1\right] \le \Theta\left(\frac{m_\ell^2}{k_\ell^2}\right)$$

Case 2: The extra instance of i is in the test set, which implies $s_{i,1} = s'_{i,1} + 1$ and $k_{i,1} = k'_{i,1}$.

Therefore, we have:

$$\begin{aligned}
&\left|\frac{(s_{i,1}-s_{i,2})^2 - s_{i,1} - s_{i,2}}{k_{i,1}+k_{i,2}+1} - \frac{(s'_{i,1}-s'_{i,2})^2 - s'_{i,1} - s'_{i,2}}{k'_{i,1}+k'_{i,2}+1}\right| \\
&\quad = \left|\frac{(s_{i,1}-s_{i,2})^2 - s_{i,1} - s_{i,2}}{k_{i,1}+k_{i,2}+1} - \frac{(s_{i,1}-1-s_{i,2})^2 - (s_{i,1}-1) - s_{i,2}}{k_{i,1}+k_{i,2}+1}\right| \\
&\quad \le \Theta\left(\frac{t_{\ell,i}}{k_{i,\ell}}\right).
\end{aligned}$$

Now, we focus on computing the conditional expected value in this case. We use the above inequality. Note that $t_{i,\ell}$ is the same, no matter which permutation we pick. Also, it is not too hard to see $k_{i,\ell}$ can be viewed as a random variable from hypergeometric distribution: $\mathbf{HG}(m_\ell, t_{i,\ell}, k_\ell)$. Using Lemma 7.4, we can bound the conditional expectation of the differences of these two terms:

$$\mathbf{E}_\pi\left[\left|\frac{(s_{i,1}-s_{i,2})^2 - s_{i,1} - s_{i,2}}{k_{i,1}+k_{i,2}+1} - \frac{(s'_{i,1}-s'_{i,2})^2 - s'_{i,1} - s'_{i,2}}{k'_{i,1}+k'_{i,2}+1}\right| \,\middle|\, \text{extra copy of } i \text{ is in not } F_1\right] \le \frac{m_\ell}{k_\ell}$$

By Bayes' Rule, we can bound the expected difference over the random choice of π.

$$\begin{aligned}
&\mathbf{E}_\pi\left[\left|\frac{(s_{i,1}-s_{i,2})^2-s_{i,1}-s_{i,2}}{k_{i,1}+k_{i,2}+1}-\frac{(s'_{i,1}-s'_{i,2})^2-s'_{i,1}-s'_{i,2}}{k'_{i,1}+k'_{i,2}+1}\right|\right]\\
&\qquad\leq \mathbf{Pr}[\text{extra copy of } i \text{ is in } F_1]\cdot\Theta\left(\frac{m_\ell^2}{k_\ell^2}\right)+\mathbf{Pr}[\text{extra copy of } i \text{ is in not } F_1]\cdot\Theta\left(\frac{m_\ell}{k_\ell}\right)\\
&\qquad\leq \Theta\left(\frac{k_1}{s_1+k_1}\cdot\left(\frac{s_\ell+k_\ell}{k_\ell}\right)^2+\frac{s_1}{s_1+k_1}\cdot\frac{s_\ell+k_\ell}{k_\ell}\right).
\end{aligned}$$

Note that in the above inequality ℓ can be one or two. It is not hard to show that the second term in the last line above is dominated by the first term. Now, we can put everything together and obtain the desired bound:

$$|\overline{Z}-\overline{Z}'|\leq\Theta\left(\frac{k_1}{k_1+s_1}\cdot\left(\frac{s_2+k_2}{k_2}\right)^2+\frac{k_2}{k_2+s_2}\cdot\left(\frac{s_1+k_1}{k_1}\right)^2\right)$$

and the proof is complete. □

6 Testing independence of two random variables

In this section, we provide a ξ-differentially private tester for testing independence of two random variables. The idea is to reduce the optimal non-private tester, delivered in [DK16], to a private one using the technique we explained in Section 3. We start off with defining the problem and the non-private reduction procedure, say $\mathcal{A}$, that reduces testing independence to the testing closeness of two distributions. Assume p is a distribution over $[n]\times[m]$. Without loss of generality, we assume $m\leq n$. Suppose we receive samples (x,y) from p. We say distribution p is an independent distribution, if the x's and the y's are independent from each other. The goal is to distinguish whether p is an independent distribution or is it ϵ-far from any independent distribution over $[n]\times[m]$. It is known that if p is an independent distribution, then p is equal to $p_1\times p_2$, and if p is *epsilon*-far from being independent, then p is ϵ-far from $p_1\times p_2$ where p_1 and p_2 are the marginal distributions of p. Using this fact, the non-private tester reduces the problem to testing the closeness of p and $q\coloneqq p_1\times p_2$ [BFF+01b].

Before describing the reduction procedure $\mathcal{A}$, we describe the sampling scheme of the procedure: For every sample that the algorithm needs, it draws two samples, and puts them in a *block*. We denote a block of two samples (x_1,y_1) and (x_2,y_2) by $\langle(x_1,y_1),(x_2,y_2)\rangle$. We can use the samples in block to obtain a sample from p, p_1, p_2, and q as follows. To get a sample from p, we always take the first samples, (x_1,y_1), in the block. We take x_1, y_2 as two the samples from p_1 and p_2. In addition, since x_1 and y_2 are two independent random variables, (x_1,y_2) is a sample from q. Also, we use "dot notation" to indicate an arbitrary element in the domain. For example, for a given x, $(x,.)$ is a sample that its first coordinate is x and the second coordinate can be any y in $[m]$. Similarly, we use the same notation to refer to a block, for example, for a given x and y, $\langle(x,.),(.,y)\rangle$ indicates a block that the first coordinate of the first sample is x and the second coordinate of the second sample is y, and the two other coordinates can be arbitrary elements in $[m]$ and $[n]$. Let X denotes the set of all blocks that are available to the procedure. We use f to denote a frequency of the blocks with a certain format in X., e.g., $f_{\langle(x,.),(.,.)\rangle}$ is the number of blocks in X that the first coordinate of the first sample is x. For the rest of this section, we focus on blocks and use them accordingly to extract a sample.

Here, we describe a proper procedure, $\mathcal{A}$, for reducing testing independence of p to the testing closeness of p and q. Suppose we have sample access to p, and we can draw blocks of samples from it. Procedure $\mathcal{A}$ uses the blocks for the following purposes: the *flattening samples* are used to determine the number of buckets for each domain element. They are designed to make sure that the ℓ_2-norm of q after flattening is low. Also, the *test samples* are used to generate samples from two distributions p and q, which we test their closeness. Below is how the algorithm will determine these samples. For now, assume $k^{(p_1)}$, $k^{(p_2)}$, $k^{(p)}$, $k^{(q)}$, and s are parameters that we determine later.

Flattening samples and the number of buckets: We flatten distribution p using samples from the marginal distributions p_1 and p_2, and also samples from p itself. More specifically, we draw four sets of blocks from p, namely $F^{(p_1)}, F^{(p_2)}, F^{(p)}, F^{(q)}$, which contain $\mathbf{Poi}\left(k^{(p_1)}\right)$, $\mathbf{Poi}\left(k^{(p_2)}\right)$, $\mathbf{Poi}\left(k^{(p)}\right)$, $\mathbf{Poi}\left(k^{(q)}\right)$ blocks respectively. We refer to the samples in these sets as flattening samples, and denote the collection of them by F. As we discuss earlier, we extract samples from the blocks in these sets to obtain samples from p_1, p_2, p, and q. More specifically, we use the following notation for the number of occurrences of each sample obtained from each set:

- $k_x^{(p_1)}$ denotes the number of occurrences of the blocks of the form $\langle (x,.),(.,.)\rangle$ in the flattening set $F^{(p_1)}$.
- $k_y^{(p_2)}$ denotes the number of occurrences of the blocks of the form $\langle (.,.),(.,y)\rangle$ in the flattening set $F^{(p_2)}$.
- $k_{(x,y)}^{(p)}$ denotes the number of occurrences of the blocks of the form $\langle (x,y),(.,.)\rangle$ in the flattening set $F^{(p)}$.
- $k_{(x,y)}^{(q)}$ denotes the number of occurrences of the blocks of the form $\langle (x,.),(.,y)\rangle$ in the flattening set $F^{(q)}$.

Our procedure uses $b_{(x,y)}$ many buckets for a domain element (x,y), where $b_{(x,y)}$ is defined as follows:

$$b_{(x,y)} = (k_x^{(p_1)} + 1)(k_y^{(p_2)} + 1) + k_{(x,y)}^{(p)} + k_{(x,y)}^{(q)}\,.$$

It is worth to note that $k_x^{(p_1)}$ is always determined by the first samples in the blocks, whereas $k_y^{(p_2)}$ is determined by the second samples in the blocks. Therefore, for all the x's and the y's, these quantities are independent of each other.

Test samples: To determine the test samples, we draw two sets of blocks $T^{(p)}$ and $T^{(q)}$. Each set contains $\mathbf{Poi}(s)$ many blocks. The samples in $T^{(p)}$ and $T^{(q)}$ are our test samples, and we denote the union of these two sets by T. The blocks in $T^{(p)}$ are used to obtain samples form p, and the blocks in $T^{(q)}$ are used to collect samples from q. In particular, we use the following notation for the number of occurrences of each domain element (x,y).

- $s_{(x,y)}^{(p)}$ denotes the number of occurrences of the blocks of the form $\langle (x,y),(.,.)\rangle$ in the test set $T^{(p)}$.
- $s_{(x,y)}^{(q)}$ denotes the number of occurrences of the blocks of the form $\langle (x,.),(.,y)\rangle$ in the test set $T^{(q)}$.

Now, that we showed how procedure $\mathcal{A}$ determines the number of samples, we prove it yields to a ξ-differentially private tester as well. At a high level, we first show that $\mathcal{A}$ is a proper procedure for testing independence (Section 6.1), then use our general closeness tester to achieve a ξ-differentially private tester. Since the sample complexity of the private tester depends on the sensitivity of the statistic we are using, we analyze the sensitivity of the statistic (Section 6.2). In particular, we show if the number of occurrences of certain blocks in the sample set is "as expected," then the sensitivity is low, which results in a nearly optimal sample complexity for the private tester. Next, we develop a framework to extend our results to the case where the number of occurrences of certain blocks in the sample set is *not* "as expected." Essentially, we turn the input samples sets that are "difficult" for our algorithm to test to the ones that are "easy", so that our algorithm works for any input sample set (Section 6.3). To put it all together, we have the following theorem:

Theorem 6.1. *Let p be a distribution over $[n]\times[m]$ where $n \geq m$. There exists a ξ-differentially private $(\epsilon, 2/3)$ tester for the testing independence of p that uses $\Theta(s)$ samples where s is:*

$$s = \Theta\left(\frac{n^{2/3}m^{1/3}}{\epsilon^{4/3}} + \frac{(m\,n)^{1/2}}{\epsilon^2} + \frac{(m\,n\log n)^{1/2}}{\epsilon\sqrt{\xi}} + \frac{\log n}{\epsilon^2\xi}\right).$$

Proof: At a high level, our main approach is to show procedure $\mathcal{A}$ is a proper procedure, and use the general tester to obtain a private algorithm. Our goal here is to show that the number of samples we stated in the theorem, s, is sufficient to use the general tester in Section 3. Thus, we need to upper bound n', $\mathbf{E}_F\big[\min\left(\|p^{(F)}\|_2^2, \|q^{(F)}\|_2^2\right)\big]$, and $\Delta(\overline{Z})$. We can compute the upper bound for the first two terms. However, the worse case bounds we have for $\Delta(\overline{Z})$ are too large to yield to a nearly optimal private tester. Hence, we use a tighter bound for $\Delta(\overline{Z})$ that holds for most of the input sample sets, and obtain a private tester for these *easy* sample set. Later we show for those *difficult* sample sets, we can map them an easy sample set to preserve the privacy guarantee. On the other hand, since the probability of receiving a difficult sample set is small, it does affect the accuracy guarantee substantially.

We first set up the parameters we use in the procedure $\mathcal{A}$:

$$k^{(p_2)} = m, \quad k^{(p_1)} = \min(n, n^{2/3}m^{1/3}/\epsilon^{4/3}), \quad k^{(p)} = k^{(q)} = \min(m \cdot n, s),$$

$$\text{and} \quad s = c \cdot \left(\frac{n^{2/3}m^{1/3}}{\epsilon^{4/3}} + \frac{(m\,n)^{1/2}}{\epsilon^2} + \frac{(m\,n\log n)^{1/2}}{\epsilon\sqrt{\xi}} + \frac{\log n}{\epsilon^2 \xi} \right).$$

where c is a large constant. For sufficiently large m and n, with probability 0.99 the number of blocks in each set in $\mathcal{S} = \{F^{(p_1)}, F^{(p_2)}, F^{(p)}, F^{(q)}, T^{(p)}, T^{(q)}\}$ is within a constant factor of its expectation via Lemma 5.3. Now, we have the following lemma to show $\mathcal{A}$ that we describe above is a proper procedure. We defer the proof of this lemma to Section 6.1.

Lemma 6.2. *Procedure $\mathcal{A}$ explained above is a proper procedure according to Definition 3.2 for testing independence of two random variables.*

Now, using our general private tester, and more specifically Theorem 3.3, there exists a ξ-differentially private tester for the independence property which uses the following number of test samples:

$$\Theta\left(\frac{n' \cdot \sqrt{\mathbf{E}_F\big[\min\left(\|p^{(F)}\|_2^2, \|q^{(F)}\|_2^2\right)\big]}}{\epsilon^2} + \frac{\sqrt{n'\Delta(\overline{Z})}}{\epsilon\sqrt{\xi}} \right).$$

where n' is an upper bound on the domain size of p and q after flattening, so $n' \geq \sum_{x,y} b_{(x,y)}$. Here, we show that given our parameters, we achieve the sample complexity stated in the theorem by bounding n', $\mathbf{E}_F\big[\min\left(\|p^{(F)}\|_2^2, \|q^{(F)}\|_2^2\right)\big]$, and $\Delta(\overline{Z})$ (for the easy sample sets). The expected of minimum of the ℓ_2-norm of p and q is bounded using the result in Lemma 2.6 in [DK16].

$$\mathbf{E}_{F^{(p_1)},F^{(p_2)},F^{(p)}}\Big[\min\left(\|p^{(F)}\|_2^2, \|q^{(F)}\|_2^2\right)\Big] \leq \mathbf{E}_F\Big[\|q^{(F)}\|_2^2\Big] \leq \mathbf{E}_F\left[\sum_{x=1}^{n}\sum_{y=1}^{m}\frac{q(x,y)^2}{b_{(x,y)}}\right]$$

$$\leq \sum_{x=1}^{n}\sum_{y=1}^{m}\frac{p_1(x)^2\,p_2(y)^2}{(k_x^{(p_1)}+1)\cdot(k_y^{(p_2)}+1)} \leq \|p_1^{(F^{(p_1)})}\| \cdot \|p_2^{(F^{(p_2)})}\| \leq \frac{1}{k^{(p_1)}\,k^{(p_2)}}.$$

Moreover, in Section 6.2, we provide the following bound for the sensitivity of the statistic:

Lemma 6.3. *Given that the size of all flattening and test samples are within the constant factor of their expections, the sensitivity of the statistic Z is bounded as follows:*

$$\Theta\left(\frac{s}{k^{(q)}} + \frac{s}{k^{(p)}} + \frac{s}{k^{(p)}} \cdot \frac{f_{\langle(.,b),(.,.)\rangle}}{f_{\langle(.,.),(.,b)\rangle}+1}\right)$$

To get a bound on sensitivity, for now, suppose all the input block sets X has a desired property that the ratio between $f_{\langle(.,b),(.,.)\rangle}$ and $f_{\langle(.,.),(.,b)\rangle}+1$ are bounded:

$$X \in \mathcal{X}^* := \left\{ X : \frac{f_{\langle(.,b),(.,.)\rangle}}{f_{\langle(.,.),(.,b)\rangle}+1} \leq \tau \right\}$$

where $\tau = 1200 \ln n$. Thus, using the fact that X is in $\mathcal{X}^*$, one can obtain:

$$\Delta(\overline{Z}) \leq \Theta\left(\frac{s\,\log n}{mn} + \log n\right)$$

Now, we are ready to show that $s' \leq s$ implying that we have enough samples for the ξ-private tester. It is not hard to see that we have the following bounds (up to a constant factors):

$$\begin{aligned}
s' &\leq \Theta\left(\frac{n' \cdot \sqrt{\mathbf{E}_F\left[\min\left(\|p^{(F)}\|_2^2, \|q^{(F)}\|_2^2\right)\right]}}{\epsilon^2} + \frac{\sqrt{n'\Delta(\overline{Z})}}{\epsilon\sqrt{\xi}}\right) \\
&\leq \Theta\left(\frac{m\,n}{\epsilon^2\sqrt{k^{(p_1)}\,k^{(p_2)}}} + \frac{\sqrt{m\,n\Delta(\overline{Z})}}{\epsilon\sqrt{\xi}}\right) \\
&\leq \Theta\left(\frac{m\,n}{\epsilon^2\sqrt{m\,\min(n, n^{2/3}m^{1/3}/\epsilon^{4/3})}} + \frac{\sqrt{m\,n}}{\epsilon\sqrt{\xi}} \cdot \sqrt{\frac{s\,\log n}{mn} + \log n}\right) \\
&\leq \Theta\left(\frac{n^{2/3}m^{1/3}}{\epsilon^{4/3}} + \frac{(m\,n)^{1/2}}{\epsilon^2} + \frac{\sqrt{m\,n}}{\epsilon\sqrt{\xi}} \cdot \sqrt{\frac{s\,\log n}{mn} + \log n}\right) \\
&\leq s
\end{aligned}$$

Thus, given that X is in $\mathcal{X}^*$, there exists a ξ-differentially private tester that outputs the right answer with probability 0.8. This is sufficient to show that there exists an ξ-differentially private algorithm with asymptotically the same number of samples via Lemma 6.6. □

6.1 Non-private independence tester is a proper procedure

Lemma 6.2. *Procedure $\mathcal{A}$ explained above is a proper procedure according to Definition 3.2 for testing independence of two random variables.*

Proof: Let X be the set of all blocks we received. Since the number of blocks in each of the flattening set and the test set is Poisson random variable, by Lemma 5.3, we can conclude with probability 1-0.01 we draw at most three times more samples than what is expected. Hence X is a set with a bounded number of samples.

We start proving Equation (3) by recalling a fact about the Poissonization method. The number of blocks of a certain form in one of the flattening and test sets is a Binomial random variable with the bias that is proportional to the expected size of each set. For example, if X contains t blocks of the form $\langle(x,.),(.,y)\rangle$, the number of the blocks of the form $\langle(x,.),(.,y)\rangle$ in $T^{(q)}$, namely r, is $\mathbf{Bin}(t, s/(k^{(p_1)} + k^{(p_2)} + k^{(p)} + 2s))$. Moreover, the probability of getting a block of this type is $q(x,y) = p_1(x) \cdot p_2(y)$, so t is a Poisson random variable with mean $q(x,y) \cdot (k^{(p_1)} + k^{(p_2)} + k^{(p)} + 2s)$ over the randomness of X. By Markov's inequality, with probability $1 - 1/c$, we may assume t is at most c times its expectation. As a consequence, $\mathbf{E}[r]$ is at most $c \cdot \mathbf{E}[t] \cdot s/(k^{(p_1)} + k^{(p_2)} + k^{(p)} + 2s) = c\,q(x,y)\,s$. Note that we can extend this example further to any type of block and any test or flattening sets.

Given X, for a domain element (x,y), the following holds using Jensen's inequality:

$$\begin{aligned}
\mathbf{E}_\pi\left[\frac{1}{b_{(x,y)}(X,\pi)}\right] &= \mathbf{E}_{k_x^{(p_1)}, k_y^{(p_2)}, k_{(x,y)}^{(p)}}\left[\frac{1}{b_{(x,y)}}\right] \geq \frac{1}{\mathbf{E}_{k_x^{(p_1)}, k_y^{(p_2)}, k_{(x,y)}^{(p)}}\left[b_{(x,y)}\right]} \\
&= \frac{1}{(\mathbf{E}\left[k_x^{(p_1)}\right] + 1) \cdot (\mathbf{E}\left[k_y^{(p_2)}\right] + 1) + \mathbf{E}\left[k_{(x,y)}^{(p)}\right]} \\
&\geq \frac{1}{(50\,p_1(x)\,k^{(p_1)} + 1) \cdot (50\,p_2(y)\,k^{(p_2)} + 1) + 100\,p(x,y)\,k^{(p)}} \\
&\geq \frac{1}{2500} \cdot \frac{1}{(p_1(x)\,k^{(p_1)} + 1) \cdot (p_2(y)\,k^{(p_2)} + 1) + p(x,y)\,k^{(p)}}\,.
\end{aligned}$$

where the second to last inequality holds with probability 0.95.

On the other hand, we find an upper bound of $\mathbf{E}\left[1/b_{(x,y)}\right]$ over the randomness of all variables. Note that now that X has not been observed, so the number of each block type in each set is an Poisson random variable, and it is independent from the rest. Let F denote the set of all flattening blocks. We denote $p_1(x)\,k^{(p_1)}$, $p_2(y)\,k^{(p_2)}$, and $p(x,y)\,k^{(p)}$ by λ_1, λ_2, and λ_3 respectively. The expected value of $1/b_{(x,y)}$ can be rewritten as:

$$\begin{aligned}
\mathbf{E}_F\left[\frac{1}{b_{(x,y)}}\right] &= \mathbf{E}_F\left[\frac{1}{(k_x^{(p_1)}+1)(k_y^{(p_2)}+1)+k_{(x,y)}^{(p)}}\right] \\
&\le \mathbf{E}_F\left[\min\left(\frac{1}{(k_x^{(p_1)}+1)(k_y^{(p_2)}+1)}, \frac{1}{k_{(x,y)}^{(p)}+1}\right)\right] \\
&\le \min\left(\mathbf{E}_F\left[\frac{1}{k_x^{(p_1)}+1}\right]\cdot\mathbf{E}_F\left[\frac{1}{k_y^{(p_2)}+1}\right], \mathbf{E}_F\left[\frac{1}{k_{(x,y)}^{(p)}+1}\right]\right) \\
&\le \min\left(\frac{1-e^{-\lambda_1}}{\lambda_1}\cdot\frac{1-e^{-\lambda_2}}{\lambda_2}, \frac{1-e^{-\lambda_3}}{\lambda_3}\right) \le \min\left(\frac{4}{(\lambda_1+1)(\lambda_2+1)}, \frac{2}{\lambda_3+1}\right) \\
&\le \frac{8}{(\lambda_1+1)(\lambda_2+1)+\lambda_3} \\
&= \frac{8}{(p_1(x)\,k^{(p_1)}+1)\cdot(p_2(y)\,k^{(p_2)}+1)+p(x,y)\,k^{(p)}}
\end{aligned}$$

Governed by the previous equations, we obtain:

$$\mathbf{Pr}_X\left[\mathbf{E}_\pi\left[\frac{1}{b_{(x,y)}(X,\pi)}\right] \ge \frac{1}{20000}\cdot\mathbf{E}_F\left[\frac{1}{b_{(x,y)}}\right]\right] \ge 0.96\,,$$

which is equivalent to

$$\mathbf{Pr}_X\left[\mathbf{E}_\pi\left[\frac{(p(x,y)-q(x,y))^2}{b_{(x,y)}(X,\pi)}\right] < \frac{1}{20000}\cdot\mathbf{E}_F\left[\frac{(p(x,y)-q(x,y))^2}{b_{(x,y)}}\right]\right] \le 0.04\,.$$

To prove Equation (3), we need to show that the above equation holds even for the sum of the quantities over all (x,y) with high probability. We show the claim in the following lemma. The proof is in Section 7:

Lemma 6.4. *Let $x_1, x_2, \ldots, x_n$ be n non-negative random variables. Suppose there exist two constants c and p, both at most one, such that for each random variable x_i, we have:*

$$\mathbf{Pr}[x_i < c\cdot\mathbf{E}[x_i]] \le p\,,$$

Then, one can show:

$$\mathbf{Pr}\left[\sum_{i=1}^{n} x_i < \frac{c\cdot\sum_{i=1}^{n}\mathbf{E}[x_i]}{10}\right] \le \frac{10\,p}{9}\,.$$

Now, we focus on proving Equation (4). To prove the inequality, it suffices to show that $\mathbf{E}_F\left[1/b_{(x,y)}^3\right]$ is $O\left(\mathbf{E}_F\left[1/b_{(x,y)}^2\right]^2\right)$. Again, note that we can think of $b_{(x,y)}$ to be equal to $(X+1)(Y+1)+Z$ where X, Y and Z are three Poisson random variables with means $\lambda_1 = p_1(x)k^{(p_1)}$, $\lambda_2 = p_2(y)k^{(p_2)}$, and $\lambda_3 =$ respectively. In Equation (12) we show that:

$$\mathbf{E}_{W\sim\mathbf{Poi}(\lambda)}\left[\frac{1}{W^3}\right] \le 6\cdot\left(\frac{1-e^{-\lambda}}{\lambda}\right)^2$$

Thus, we obtain an upper bound for the expected value of $1/b^3_{(x,y)}$ as follows:

$$\begin{aligned}\mathbf{E}_F\left[\frac{1}{b^3_{(x,y)}}\right] &= \mathbf{E}_{X,Y,Z}\left[\frac{1}{((X+1)\cdot(Y+1)+Z)^3}\right] \le \mathbf{E}_{X,Y,Z}\left[\min\left(\frac{1}{(X+1)^3}\cdot\frac{1}{(Y+1)^3}, \frac{1}{(Z+1)^3}\right)\right] \\ &\le \min\left(\mathbf{E}_X\left[\frac{1}{(X+1)^3}\right]\cdot\mathbf{E}_Y\left[\frac{1}{(Y+1)^3}\right], \mathbf{E}_Z\left[\frac{1}{(Z+1)^3}\right]\right) \\ &\le 36\,\min\left(\left(\frac{1-e^{-\lambda_1}}{\lambda_1}\right)^2\cdot\left(\frac{1-e^{-\lambda_2}}{\lambda_2}\right)^2, \left(\frac{1-e^{-\lambda_3}}{\lambda_3}\right)^2\right).\end{aligned}$$

Note that in the case that one of the *lambda*'s is equal to zero, one can replace $1-e^{-\lambda}/\lambda$ by one in the rest of the proof. On the other hand, we can find a lower bound for $1/b_{(x,y)}$ by Jensen's inequality:

$$\begin{aligned}\left(\mathbf{E}_F\left[\frac{1}{b_{(x,y)}}\right]\right)^2 &\ge \left(\frac{1}{\mathbf{E}_F[b_i]}\right)^2 = \left(\frac{1}{(\lambda_1+1)(\lambda_2+1)+\lambda_3}\right)^2 \\ &\ge \left(\frac{1}{2}\min\left(\frac{1}{\lambda_1+1}\cdot\frac{1}{\lambda_2+1}, \frac{1}{\lambda_3+1}\right)\right)^2 \\ &\ge \frac{1}{4}\min\left(\left(\frac{1}{\lambda_1+1}\cdot\frac{1}{\lambda_2+1}\right)^2, \left(\frac{1}{\lambda_3+1}\right)^2\right) \\ &\ge \frac{1}{64}\min\left(\left(\frac{1-e^{-\lambda_1}}{\lambda_1}\cdot\frac{1-e^{-\lambda_2}}{\lambda_2}\right)^2, \left(\frac{1-e^{-\lambda_3}}{\lambda_3}\right)^2\right)\end{aligned}$$

where the last inequality is due to the fact that $(1-e^{-t})/t$ is at most $2/(t+1)$ for a non-negative number t. Putting these two equations together, one can conclude Equation (4):

$$\begin{aligned}\mathbf{E}_F\left[\|p^{(F)}-q^{(F)}\|_4^4\right] &= \sum_{x=1}^{n}\sum_{y=1}^{m}\mathbf{E}_F\left[\frac{(p(x,y)-q(x,y))^4}{b^3_{(x,y)}}\right] \le 36\cdot 64\cdot\sum_{x=1}^{n}\sum_{y=1}^{m}\left(\mathbf{E}_F\left[\frac{(p(x,y)-q(x,y))^2}{b_{x,y}}\right]\right)^2 \\ &\le 2304\cdot\left(\sum_{x=1}^{n}\sum_{y=1}^{m}\mathbf{E}_F\left[\frac{(p(x,y)-q(x,y))^2}{b_{x,y}}\right]\right)^2 = 2304\cdot\mathbf{E}_F\left[\|p^{(F)}-q^{(F)}\|_2^2\right]^2.\end{aligned}$$

Therefore, the statement of the lemma is concluded. □

3

6.2 sensitivity of the statistic for the independence problem

In this section, we give an upper bound for the sensitivity of the independence statistic: the amount that the statistic changes if we change one sample in the input.

Let X denote a set of block that the algorithm received as the input. Assume we permute the blocks in X using a permutation π. Note that if we fix the size of each flattening set and sample set, $\hat{s}_1, \hat{s}_2$, etc., one can deterministically find $s^{(p)}_{(x,y)}, s^{(q)}_{(x,y)}$, etc.. Thus, given X, π, and sizes of sets, one can compute the following statistic:

$$Z(X,\pi) \coloneqq \sum_{x=1}^{m}\sum_{y=1}^{n}\frac{(s^{(p)}_{(x,y)}-s^{(q)}_{(x,y)})^2 - s^{(p)}_{(x,y)} - s^{(q)}_{(x,y)}}{(k^{(p_1)}_x+1)(k^{(p_2)}_y+1)+k^{(p)}_{(x,y)}+k^{(q)}_{(x,y)}}$$

We denote the average of Z over all π by $\overline{Z}(X)$. Our goal here is to calculate

$$\Delta Z = \max X, X' |\overline{Z}(X) - \overline{Z}(X')|$$

where X and X' are two neighboring data sets that they differ in exactly one element.

Through this section, we use an important property of Poissonization method: Let A and B be two sets

with $\hat{n}_1 = \mathbf{Poi}(n_1)$ and $\hat{n}_2 = \mathbf{Poi}(n_2)$ samples. Given that there are k instance of element i in A and B together, the number of instances of element i in A is a Binomial random variable: $\mathbf{Bin}(\hat{n}_1+\hat{n}_2, n_1/(n_1+n_2))$.

Lemma 6.3. *Given that the size of all flattening and test samples are within the constant factor of their expections, the sensitivity of the statistic Z is bounded as follows:*

$$\Theta\left(\frac{s}{k^{(q)}} + \frac{s}{k^{(p)}} + \frac{s}{k^{(p)}} \cdot \frac{f_{\langle(.,b),(.,.)\rangle}}{f_{\langle(.,.),(.,b)\rangle}+1}\right)$$

Proof: In this proof, we assume X and X' are fully given, and X and X' are only different in the r-th block of the samples. Note that when we permute the elements in X and X', we only permute the blocks and do not change the order of the samples within each block. The expectations in the this proof are taken over the random choice of a permutation π. As we mentioned earlier, we partition the blocks into the following sets, and the number of occurrences of each block types in each set determines the statistic:

$$\mathcal{S} = \left\{F^{(p_1)}, F^{(p_2)}, F^{(p)}, F^{(q)}, T^{(p)}, T^{(q)}\right\}$$

We can separate our calculation based on where the r-th block is:

$$\begin{aligned}
\Delta(\overline{Z}) &= |\overline{Z}(X) - \overline{Z}(X')| \\
&= \left|\mathbf{E}_\pi[Z(X,\pi) - Z(X',\pi)]\right| \leq \mathbf{E}_\pi[|Z(X,\pi) - Z(X',\pi)|] \\
&= \sum_{S\in\mathcal{S}} \mathbf{Pr}_\pi[r \in S] \cdot \left|\mathbf{E}_\pi[\,|Z(X,\pi) - Z(X',\pi)|\,|\, r \in S]\right|
\end{aligned}$$

Now, we consider each term separate.

1. **Block r is in $F^{(p_1)}$:** Suppose the types of the r-th block in X and X' are $\langle(a,.),(.,.)\rangle$ and $\langle(a',.),(.,.)\rangle$ respectively. If a and a' are equal, then the statistic will remains unchanged. Otherwise, $k_a^{(p_1)}$ and ${k'}_{a'}^{(p_1)}$ is changed by one. First, we simplify the term $|Z(X,\pi) - Z(X',\pi)|$ for a given π:

$$\begin{aligned}
|Z(X,\pi) - Z(X',\pi)| = \Bigg| &\sum_{x=1}^{n}\sum_{y=1}^{m} \frac{(s^{(p)}_{(x,y)} - s^{(q)}_{(x,y)})^2 - s^{(p)}_{(x,y)} - s^{(q)}_{(x,y)}}{(k_x^{(p_1)}+1)(k_y^{(p_2)}+1) + k^{(p)}_{(x,y)} + k^{(q)}_{(x,y)}} \\
&- \sum_{x=1}^{n}\sum_{y=1}^{m} \frac{({s'}^{(p)}_{(x,y)} - {s'}^{(q)}_{(x,y)})^2 - {s'}^{(p)}_{(x,y)} - {s'}^{(q)}_{(x,y)}}{({k'}_x^{(p_1)}+1)({k'}_y^{(p_2)}+1) + {k'}^{(p)}_{(x,y)} + {k'}^{(q)}_{(x,y)}}\Bigg| \\
\leq \sum_{x\in\{a,a'\}} \Bigg| &\sum_{y=1}^{m} \frac{(s^{(p)}_{(x,y)} - s^{(q)}_{(x,y)})^2 - s^{(p)}_{(x,y)} - s^{(q)}_{(x,y)}}{(k_x^{(p_1)}+1)(k_y^{(p_2)}+1) + k^{(p)}_{(x,y)} + k^{(q)}_{(x,y)}} \\
&- \frac{({s'}^{(p)}_{(x,y)} - {s'}^{(q)}_{(x,y)})^2 - {s'}^{(p)}_{(x,y)} - {s'}^{(q)}_{(x,y)}}{({k'}_x^{(p_1)}+1)({k'}_y^{(p_2)}+1) + {k'}^{(p)}_{(x,y)} + {k'}^{(q)}_{(x,y)}}\Bigg|
\end{aligned}$$

   For the rest of the proof, we focus on the term above when $x = a$. The other term can be upper

bounded similarly, and at the end we multiply our final bound by two.

$$
\begin{aligned}
&\left|\sum_{y=1}^{m} \frac{(s^{(p)}_{(a,y)} - s^{(q)}_{(a,y)})^2 - s^{(p)}_{(a,y)} - s^{(q)}_{(a,y)}}{(k^{(p_1)}_a + 1)(k^{(p_2)}_y + 1) + k^{(p)}_{(a,y)} + k^{(q)}_{(x,y)}} - \frac{(s'^{(p)}_{(a,y)} - s'^{(q)}_{(a,y)})^2 - s'^{(p)}_{(a,y)} - s'^{(q)}_{(a,y)}}{(k'_a(p_1) + 1)(k'^{(p_2)}_y + 1) + k'^{(p)}_{(a,y)} + k'^{(q)}_{(x,y)}}\right| \\
&\quad \le \sum_{y=1}^{m} \left| \frac{(s^{(p)}_{(a,y)} - s^{(q)}_{(a,y)})^2 - s^{(p)}_{(a,y)} - s^{(q)}_{(a,y)}}{(k'^{(p_1)}_a + 2)(k^{(p_2)}_y + 1) + k^{(p)}_{(a,y)} + k^{(q)}_{(x,y)}} - \frac{(s^{(p)}_{(a,y)} - s^{(q)}_{(a,y)})^2 - s^{(p)}_{(a,y)} - s^{(q)}_{(a,y)}}{(k'_a(p_1) + 1)(k^{(p_2)}_y + 1) + k^{(p)}_{(a,y)} + k^{(q)}_{(x,y)}}\right| \\
&\quad \le \sum_{y=1}^{m} \frac{\left(k^{(p_2)}_y + 1\right) \cdot \left|(s^{(p)}_{(a,y)} - s^{(q)}_{(a,y)})^2 - s^{(p)}_{(a,y)} - s^{(q)}_{(a,y)}\right|}{\left((k'^{(p_1)}_a + 2)(k^{(p_2)}_y + 1) + k^{(p)}_{(a,y)} + k^{(q)}_{(x,y)}\right) \cdot \left((k'_a(p_1) + 1)(k^{(p_2)}_y + 1) + k^{(p)}_{(a,y)} + k^{(q)}_{(x,y)}\right)} \\
&\quad \le \sum_{y=1}^{m} \frac{\left(k^{(p_2)}_y + 1\right) \cdot \left((s^{(p)}_{(a,y)})^2 + (s^{(q)}_{(a,y)})^2\right)}{\left((k'^{(p_1)}_a + 2)(k^{(p_2)}_y + 1) + k^{(p)}_{(a,y)} + k^{(q)}_{(x,y)}\right) \cdot \left((k'_a(p_1) + 1)(k^{(p_2)}_y + 1) + k^{(p)}_{(a,y)} + k^{(q)}_{(x,y)}\right)} \\
&\quad \le \sum_{y=1}^{m} \frac{1}{k^{(p_1)}_a + 1} \cdot \frac{(s^{(p)}_{(a,y)})^2}{k^{(p)}_{(a,y)} + 1} + \frac{1}{k^{(p_1)}_a + 1} \cdot \frac{(s^{(q)}_{(a,y)})^2}{k^{(q)}_{(a,y)} + 1}
\end{aligned}
$$

For brevity's sake, let v denote the following expectation:

$$
v \coloneqq \mathbf{E}_{\pi}\left[\sum_{y=1}^{m} \frac{1}{k^{(p_1)}_a + 1} \cdot \frac{(s^{(p)}_{(a,y)})^2}{k^{(p)}_{(a,y)} + 1} + \frac{1}{k^{(p_1)}_a + 1} \cdot \frac{(s^{(q)}_{(a,y)})^2}{k^{(q)}_{(a,y)} + 1} \middle| r \in F^{(p_1)}\right]
$$

Using the tower rule, we achieve:

$$
\begin{aligned}
v \le\ & \mathbf{E}_{F^{(p_1)}}\left[\frac{1}{(k^{(p_1)}_a + 1)} \cdot \mathbf{E}_{\pi}\left[\sum_{y=1}^{m} \frac{(s^{(p)}_{(a,y)})^2}{(k^{(p)}_{(a,y)} + 1)} \middle| r \in F^{(p_1)}, F^{(p_1)}\right] \middle| r \in F^{(p_1)}\right] \\
& + \mathbf{E}_{F^{(p_1)}}\left[\frac{1}{k^{(p_1)}_a + 1} \cdot \mathbf{E}_{\pi}\left[\sum_{y=1}^{m} \frac{(s^{(q)}_{(a,y)})^2}{k^{(q)}_{(a,y)} + 1} \middle| r \in F^{(p_1)}, F^{(p_1)}\right] \middle| r \in F^{(p_1)}\right]
\end{aligned}
$$

Let $f_{\langle (a,y),(.,.)\rangle}$, $f_{\langle (a,.),(.,y)\rangle}$, and $f_{\langle (a,.),(.,.)\rangle}$ be the numbers of blocks of the forms $\langle (a,y),(.,.)\rangle$, $\langle (a,.),(.,y)\rangle$, and $\langle (a,.),(.,.)\rangle$ in X respectively. Using Lemma 7.3, one can bound the terms inside the expectations as below:

$$
\begin{aligned}
\mathbf{E}\left[\frac{(s^{(p)}_{(a,y)})^2}{(k^{(p)}_{(a,y)} + 1)}\right] &\le \min\left(\frac{2\,(s-1)\,f_{\langle (a,y),(.,.)\rangle}}{(k^{(p)} + 1)}, 2\,f^2_{\langle (a,y),(.,.)\rangle}\right) + f_{\langle (a,y),(.,.)\rangle} \\
&\le \min\left(\left(\frac{2\,(s-1)}{(k^{(p)} + 1)} + 1\right) \cdot f_{\langle (a,y),(.,.)\rangle}, 3\,f^2_{\langle (a,y),(.,.)\rangle}\right),
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{E}\left[\frac{(s^{(q)}_{(a,y)})^2}{(k^{(q)}_{(a,y)} + 1)}\right] &\le \min\left(\frac{2\,(s-1)\,f_{\langle (a,.),(.,y)\rangle}}{(k^{(q)} + 1)}, 2\,f^2_{\langle (a,.),(.,y)\rangle}\right) + f_{\langle (a,.),(.,y)\rangle} \\
&\le \min\left(\left(\frac{2\,(s-1)}{(k^{(q)} + 1)} + 1\right) \cdot f_{\langle (a,.),(.,y)\rangle}, 3\,f^2_{\langle (a,.),(.,y)\rangle}\right).
\end{aligned}
$$

Observe that $\sum_y f_{\langle (a,y),(.,.)\rangle}$ and $\sum_y f_{\langle (a,.),(.,y)\rangle}$ are equal to $f_{\langle (a,.),(.,.)\rangle}$. Thus, using Lemma 7.4, and

Lemma 7.5, we have:

$$
\begin{aligned}
v &\leq \mathbf{E}_{F^{(p_1)}}\left[\left.\frac{1}{(k_a^{(p_1)}+1)}\right| r \in F^{(p_1)}\right] \cdot \sum_{y=1}^{m} \min\left(\left(\frac{2\,(s-1)}{(k^{(p)}+1)}+1\right)\cdot f_{\langle(a,y),(.,.)\rangle}, 3\, f^2_{\langle(a,y),(.,.)\rangle}\right) \\
&+ \mathbf{E}_{F^{(p_1)}}\left[\left.\frac{1}{k_a^{(p_1)}+1}\right| r \in F^{(p_1)}\right] \cdot \sum_{y=1}^{m} \min\left(\left(\frac{2\,(s-1)}{(k^{(q)}+1)}+1\right)\cdot f_{\langle(a,.),(.,y)\rangle}, 3\, f^2_{\langle(a,.),(.,y)\rangle}\right) \\
&\leq \min\left(1, \frac{|X|}{f_{\langle(a,.),(.,.)\rangle}\, k^{(p_1)}}\right)\cdot\left(\frac{2\,(s-1)}{(k^{(p)}+1)}+1\right)\cdot f_{\langle(a,.),(.,.)\rangle} \\
&+ \min\left(1, \frac{|X|}{f_{\langle(a,.),(.,.)\rangle}\, k^{(p_1)}}\right)\cdot\left(\frac{2\,(s-1)}{(k^{(q)}+1)}+1\right)\cdot f_{\langle(a,.),(.,.)\rangle} \\
&\leq \Theta\left(\frac{|X|}{k^{(p_1)}}\cdot\left(\frac{s}{k^{(p)}}+\frac{s}{k^{(q)}}\right)\right)
\end{aligned}
$$

Using the above calculation, it is not hard to see that the following holds

$$
\begin{aligned}
&\mathbf{Pr}_{\pi}\Big[r \in F^{(p_1)}\Big]\cdot\Big|\mathbf{E}_{\pi}\Big[\,|Z(X,\pi)-Z(X',\pi)|\,\big|\, r \in F^{(p_1)}\Big]\Big| \\
&\qquad \leq 2\cdot\frac{k^{(p_1)}}{|X|}\cdot\left(\frac{|X|}{k^{(p_1)}}\cdot\left(\frac{s}{k^{(p)}}+\frac{s}{k^{(q)}}\right)\right) \\
&\qquad \leq \Theta\left(\frac{s}{k^{(p)}}+\frac{s}{k^{(q)}}\right).
\end{aligned}
$$

2. **Block r is in $F^{(p_2)}$:** Suppose the the r-th block in X and X' are of the forms $\langle(.,.),(.,b)\rangle$ and $\langle(.,.),(.,b')\rangle$ respectively. Using the symmetry of this case and the previous case, we take the same approach.

$$
\begin{aligned}
&\mathbf{E}_\pi\left[\sum_{x=1}^{n}\frac{1}{k_b^{(p_2)}+1}\cdot\frac{(s_{(x,b)}^{(p)})^2}{k_{(x,b)}^{(p)}+1}+\frac{1}{k_b^{(p_2)}+1}\cdot\frac{(s_{(x,b)}^{(q)})^2}{k_{(x,b)}^{(q)}+1}\,\middle|\, r\in F^{(p_2)}\right]\\
&\le \mathbf{E}_{F^{(p_2)}}\left[\frac{1}{k_b^{(p_2)}+1}\cdot\mathbf{E}_\pi\left[\sum_{x=1}^{n}\frac{(s_{(x,b)}^{(p)})^2}{k_{(x,b)}^{(p)}+1}\,\middle|\, r\in F^{(p_2)},F^{(p_2)}\right]\,\middle|\, r\in F^{(p_2)}\right]\\
&+\mathbf{E}_{F^{(p_2)}}\left[\frac{1}{k_b^{(p_2)}+1}\cdot\mathbf{E}_\pi\left[\sum_{x=1}^{n}\frac{(s_{(x,b)}^{(q)})^2}{k_{(x,b)}^{(q)}+1}\,\middle|\, r\in F^{(p_2)},F^{(p_2)}\right]\,\middle|\, r\in F^{(p_2)}\right]\\
&\le \mathbf{E}_{F^{(p_2)}}\left[\frac{1}{k_b^{(p_2)}+1}\,\middle|\, r\in F^{(p_2)}\right]\cdot\sum_{x=1}^{n}\min\left(\left(\frac{2\,(s-1)}{(k^{(p)}+1)}+1\right)\cdot f_{\langle(x,b),(.,.)\rangle}\,,3\,f^2_{\langle(x,b),(.,.)\rangle}\right)\\
&+\mathbf{E}_{F^{(p_2)}}\left[\frac{1}{k_b^{(p_2)}+1}\,\middle|\, r\in F^{(p_2)}\right]\cdot\sum_{x=1}^{n}\min\left(\left(\frac{2\,(s-1)}{(k^{(q)}+1)}+1\right)\cdot f_{\langle(x,.),(.,b)\rangle}\,,3\,f^2_{\langle(x,.),(.,b)\rangle}\right)\\
&\le \min\left(1,\frac{|X|}{f_{\langle(.,.),(.,b)\rangle}\,k^{(p_2)}}\right)\cdot\left(\frac{2\,(s-1)}{(k^{(p)}+1)}+1\right)\cdot f_{\langle(.,b),(.,.)\rangle}\\
&+\min\left(1,\frac{|X|}{f_{\langle(.,.),(.,b)\rangle}\,k^{(p_2)}}\right)\cdot\left(\frac{2\,(s-1)}{(k^{(q)}+1)}+1\right)\cdot f_{\langle(.,.),(.,b)\rangle}\\
&\le \frac{|X|}{k^{(p_2)}}\cdot\left(\frac{2\,(s-1)}{(k^{(p)}+1)}+1\right)\cdot\frac{f_{\langle(.,b),(.,.)\rangle}}{f_{\langle(.,.),(.,b)\rangle}+1}+\frac{|X|}{k^{(p_2)}}\cdot\left(\frac{2\,(s-1)}{(k^{(q)}+1)}+1\right)\\
&\le \Theta\left(\frac{|X|}{k^{(p_2)}}\cdot\left(\frac{s}{k^{(p)}}\cdot\frac{f_{\langle(.,b),(.,.)\rangle}}{f_{\langle(.,.),(.,b)\rangle}+1}+\frac{s}{k^{(q)}}\right)\right)
\end{aligned}
$$

where the last inequality is due to the fact that $\min_{x>0}(\alpha/x,x)\le\sqrt{\alpha}$.

$$
\begin{aligned}
&\mathbf{Pr}_\pi\Big[r\in F^{(p_2)}\Big]\cdot\Big|\mathbf{E}_\pi\Big[\,|Z(X,\pi)-Z(X',\pi)|\,|\,r\in F^{(p_2)}\Big]\Big|\\
&\qquad\le 2\cdot\frac{k^{(p_2)}}{|X|}\cdot\Theta\left(\frac{|X|}{k^{(p_2)}}\cdot\left(\frac{s}{k^{(p)}}\cdot\frac{f_{\langle(.,b),(.,.)\rangle}}{f_{\langle(.,.),(.,b)\rangle}+1}+\frac{s}{k^{(q)}}\right)\right)\\
&\qquad\le\Theta\left(\frac{s}{k^{(p)}}\cdot\frac{f_{\langle(.,b),(.,.)\rangle}}{f_{\langle(.,.),(.,b)\rangle}+1}+\frac{s}{k^{(q)}}\right)
\end{aligned}
$$

3. **Block r is in $F^{(p)}$ or $F^{(q)}$:** Here we assume r is in $F^{(p)}$. Very similar calculation, yield the same bound if r is in $F^{(q)}$. Suppose the the r-th block in X and X' are of the forms $\langle(a,b),(.,.)\rangle$ and $\langle(a',b'),(.,.)\rangle$ respectively. Note that in this case, only two terms will be different, so we have:

$$
\begin{aligned}
|Z(X,\pi)-Z(X',\pi)|\le\sum_{\substack{(x,y)\in\\ \{(a,b),(a',b')\}}}&\left|\frac{(s_{(x,y)}^{(p)}-s_{(x,y)}^{(q)})^2-s_{(x,y)}^{(p)}-s_{(x,y)}^{(q)}}{(k_x^{(p_1)}+1)(k_y^{(p_2)}+1)+k_{(x,y)}^{(p)}+k_{(x,y)}^{(q)}}\right.\\
&\left.-\frac{(s'^{(p)}_{(x,y)}-s'^{(q)}_{(x,y)})^2-s'^{(p)}_{(x,y)}-s'^{(q)}_{(x,y)}}{(k'^{(p_1)}_x+1)(k'^{(p_2)}_y+1)+k'^{(p)}_{(x,y)}+k^{(q)}_{(x,y)}}\right|
\end{aligned}
$$

Now, we focus on the term where (x, y) is equal to (a, b). The other term can be bounded similarly.

$$\begin{aligned}
&\left| \frac{(s^{(p)}_{(a,b)} - s^{(q)}_{(a,b)})^2 - s^{(p)}_{(a,b)} - s^{(q)}_{(a,b)}}{(k^{(p_1)}_a + 1)(k^{(p_2)}_b + 1) + k^{(p)}_{(a,b)} + k^{(q)}_{(a,b)}} - \frac{(s'^{(p)}_{(a,b)} - s'^{(q)}_{(a,b)})^2 - s'^{(p)}_{(a,b)} - s'^{(q)}_{(a,b)}}{(k'^{(p_1)}_a + 1)(k'^{(p_2)}_b + 1) + k'^{(p)}_{(a,b)} + k'^{(q)}_{(a,b)}} \right| \\
&\quad = \left| \frac{(s^{(p)}_{(a,b)} - s^{(q)}_{(a,b)})^2 - s^{(p)}_{(a,b)} - s^{(q)}_{(a,b)}}{(k^{(p_1)}_a + 1)(k^{(p_2)}_b + 1) + k^{(p)}_{(a,b)} + k^{(q)}_{(a,b)}} - \frac{(s^{(p)}_{(a,b)} - s^{(q)}_{(a,b)})^2 - s^{(p)}_{(a,b)} - s^{(q)}_{(a,b)}}{(k^{(p_1)}_a + 1)(k^{(p_2)}_b + 1) + k^{(p)}_{(a,b)} + k^{(q)}_{(a,b)} + 1} \right| \\
&\quad \leq \frac{(s^{(p)}_{(a,b)})^2 + (s^{(q)}_{(a,b)})^2}{\left((k^{(p_1)}_a + 1)(k^{(p_2)}_b + 1) + k^{(p)}_{(a,b)} + k^{(q)}_{(a,b)}\right) \cdot \left((k^{(p_1)}_a + 1)(k^{(p_2)}_b + 1) + k^{(p)}_{(a,b)} + k^{(q)}_{(a,b)} + 1\right)} \\
&\quad \leq \frac{(s^{(p)}_{(a,b)})^2}{k^{(p)}_{(a,b)}\,(k^{(p)}_{(a,b)} + 1)} + \frac{(s^{(q)}_{(a,b)})^2}{k^{(q)}_a\,(k^{(q)}_a + 1)}
\end{aligned}$$

Now, using Lemma 7.5, we bound the expected value of the above quantity from above:

$$\begin{aligned}
&\mathbf{E}_\pi\left[\frac{(s^{(p)}_{(a,b)})^2}{k^{(p)}_{(a,b)}\,(k^{(p)}_{(a,b)} + 1)} + \frac{(s^{(q)}_{(a,b)})^2}{k^{(q)}_a\,(k^{(q)}_a + 1)} \,\middle|\, r \in F^{(p)} \right] \\
&\qquad \leq f^2_{\langle (a,b),(.,.)\rangle} \cdot \mathbf{E}_\pi\left[\frac{1}{k^{(p)}_{(a,b)}\,(k^{(p)}_{(a,b)} + 1)} \,\middle|\, r \in F^{(p)} \right] \\
&\qquad + f^2_{\langle (a,.),(.,b)\rangle} \cdot \mathbf{E}_\pi\left[\frac{1}{k^{(q)}_a\,(k^{(q)}_a + 1)} \,\middle|\, r \in F^{(p)} \right] \\
&\qquad \leq \frac{|X|(|X|+1)}{k^{(p)}\,(k^{(p)}+1)} + \frac{|X|(|X|+1)}{k^{(q)}\,(k^{(q)}+1)}
\end{aligned}$$

Using the above equation, and the fact that $|X| = \Theta(s)$, it is not hard to see that

$$\begin{aligned}
&\mathbf{Pr}_\pi\Big[r \in F^{(p)} \Big] \cdot \Big| \mathbf{E}_\pi\Big[|Z(X,\pi) - Z(X',\pi)| \,\Big|\, r \in F^{(p)} \Big] \Big| \\
&\qquad \leq 2 \cdot \frac{k^{(p)}}{|X|} \cdot \Theta\left(\frac{|X|(|X|+1)}{k^{(p)}\,(k^{(p)}+1)} + \frac{|X|(|X|+1)}{k^{(q)}\,(k^{(q)}+1)} \right) \\
&\qquad \leq \Theta\left(\frac{s}{k^{(p)}} + \frac{s}{k^{(q)}} \right)
\end{aligned}$$

Note that the factor of two in the above inequality, comes from including the symmetric term for (a', b').

4. **Block r is in $T^{(p)}$ or $T^{(q)}$:** Suppose the the r-th block in X and X' are of the forms $\langle (a,b),(.,.)\rangle$ and $\langle (a',b'),(.,.)\rangle$ respectively. Note that in this case, only two terms will be different for the two datasets, so we have:

$$\begin{aligned}
|Z(X,\pi) - Z(X',\pi)| \leq \sum_{\substack{(x,y)\in \\ \{(a,b),(a',b')\}}} &\left| \frac{(s^{(p)}_{(x,y)} - s^{(q)}_{(x,y)})^2 - s^{(p)}_{(x,y)} - s^{(q)}_{(x,y)}}{(k^{(p_1)}_x + 1)(k^{(p_2)}_y + 1) + k^{(p)}_{(x,y)} + k^{(q)}_{(x,y)}} \right. \\
&\left. - \frac{(s'^{(p)}_{(x,y)} - s'^{(q)}_{(x,y)})^2 - s'^{(p)}_{(x,y)} - s'^{(q)}_{(x,y)}}{(k'^{(p_1)}_x + 1)(k'^{(p_2)}_y + 1) + k'^{(p)}_{(x,y)} + k'^{(q)}_{(x,y)}} \right|
\end{aligned}$$

Below, we assume r is in $T^{(p)}$. However, the calculation will be the same if r was in $T^(q)$. Now, we focus on the term where (x,y) is equal to (a,b). The other term can be bounded similarly.

$$\begin{aligned}
&\left|\frac{(s^{(p)}_{(a,b)}-s^{(q)}_{(a,b)})^2-s^{(p)}_{(a,b)}-s^{(q)}_{(a,b)}}{(k^{(p_1)}_a+1)(k^{(p_2)}_b+1)+k^{(p)}_{(a,b)}+k^{(q)}_{(a,b)}}-\frac{(s'^{(p)}_{(a,b)}-s'^{(q)}_{(a,b)})^2-s'^{(p)}_{(a,b)}-s'^{(q)}_{(a,b)}}{(k'^{(p_1)}_a+1)(k'^{(p_2)}_b+1)+k'^{(p)}_{(a,b)}+k'^{(q)}_{(a,b)}}\right| \\
&\quad=\left|\frac{(s^{(p)}_{(a,b)}-s^{(q)}_{(a,b)})^2-s^{(p)}_{(a,b)}-s^{(q)}_{(a,b)}}{(k^{(p_1)}_a+1)(k^{(p_2)}_b+1)+k^{(p)}_{(a,b)}+k^{(q)}_{(a,b)}}-\frac{(s^{(p)}_{(a,b)}-1-s^{(q)}_{(a,b)})^2-(s^{(p)}_{(a,b)}-1)-s^{(q)}_{(a,b)}}{(k^{(p_1)}_a+1)(k^{(p_2)}_b+1)+k^{(p)}_{(a,b)}+k^{(q)}_{(a,b)}}\right| \\
&\quad\le\frac{2s^{(p)}_{(a,b)}+2s^{(q)}_{(a,b)}}{(k^{(p_1)}_a+1)(k^{(p_2)}_b+1)+k^{(p)}_{(a,b)}+k^{(q)}_{(a,b)}}\le\frac{2s^{(p)}_{(a,b)}}{k^{(p)}_{(a,b)}+1}+\frac{2s^{(q)}_{(a,b)}}{k^{(q)}_a+1}
\end{aligned}$$

Now, using Lemma 7.4 , we bound the expected value of the above quantity from above:

$$\begin{aligned}
&\mathbf{E}_\pi\left[\frac{2s^{(p)}_{(a,b)}}{k^{(p)}_{(a,b)}+1}+\frac{2s^{(q)}_{(a,b)}}{k^{(q)}_a+1}\,\middle|\,r\in F^{(p)}\right] \\
&\qquad\le 2\,f_{\langle(a,b),(.,.)\rangle}\cdot\mathbf{E}_\pi\left[\frac{1}{k^{(p)}_{(a,b)}+1}\,\middle|\,r\in F^{(p)}\right] \\
&\qquad+2\,f_{\langle(a,.),(.,b)\rangle}\cdot\mathbf{E}_\pi\left[\frac{1}{k^{(q)}_a+1}\,\middle|\,r\in F^{(p)}\right] \\
&\qquad\le\frac{|X|}{k^{(p)}+1}+\frac{|X|}{k^{(q)}+1}
\end{aligned}$$

Using the above equation, and the fact that $|X|=\Theta(s)$, it is not hard to see that

$$\begin{aligned}
&\mathbf{Pr}_\pi\left[r\in F^{(p)}\right]\cdot\left|\mathbf{E}_\pi\left[\,|Z(X,\pi)-Z(X',\pi)|\,|\,r\in F^{(p)}\right]\right| \\
&\qquad\le 2\cdot\frac{k^{(p)}}{|X|}\cdot\Theta\left(\frac{|X|}{k^{(p)}+1}+\frac{|X|}{k^{(q)}+1}\right)\le\Theta\left(1\right)
\end{aligned}$$

Note that the factor of two in the above inequality, comes from including the symmetric term for (a',b').

Putting all the terms computed above together, we obtain:

$$\begin{aligned}
|\overline{Z}(X)-\overline{Z}(X')|&=\sum_{S\in\mathcal{S}}\mathbf{Pr}_\pi[r\in S]\cdot\left|\mathbf{E}_\pi[\,|Z(X,\pi)-Z(X',\pi)|\,|\,r\in S]\right| \\
&\le\Theta\left(\frac{s}{k^{(q)}}+\frac{s}{k^{(p)}}+\frac{s}{k^{(p)}}\cdot\frac{f_{\langle(.,b),(.,.)\rangle}}{f_{\langle(.,.),(.,b)\rangle}+1}\right)
\end{aligned}$$

□

6.3 Stretching the domain of a private algorithm

In this section, we investigate whether we can extend the domain of a differentially private tester under certain conditions. We start off by defining the domains. The input of a differentially private tester is a sample set from a universe Ω. Suppose we have a dataset X of $2s$ samples from $[n]$ that are arranged in two rows each of size s (namely top and bottom rows). Let the domain of a differential algorithm, denoted by $\mathcal{X}$, be the set of all such pairs of rows, namely $[n]^{2s}$. We denote the frequency of an element $i\in[n]$ in the

top row by $t_i(X)$, and in the bottom row by $b_i(X)$. A desired property of a dataset is that the ratio of the frequencies in the row is bounded by a fixed parameter $A \geq 2$. More precisely, we define the subset of $\mathcal{X}$ which contains the data sets with this property as:

$$\mathcal{X}^* = \left\{ X \,\middle|\, \forall i \in [n] : \frac{t_i(X)}{b_i(X)+1} \leq A \right\}$$

Let $\mathcal{A}$ be a tester that receives a set of samples, X, as its input, and outputs $\mathcal{A}(X)$ which is known to be correct with probability at least $1-\delta$. Suppose $\mathcal{A}$ is ξ-differentially private when X is in $\mathcal{X}^*$. Our goal here is to design a $\Theta(\xi)$-differentially private algorithm, namely $\mathcal{B}$, that takes $X \in \mathcal{X}$ as its input, and outputs $\mathcal{B}(X)$ which is incorrect with probability slightly larger than δ.

At a high level, we implement $\mathcal{B}$ by using $\mathcal{A}$ as a blackbox as follows: We first look at the input $X \in \mathcal{X}$. If X is already in $\mathcal{X}^*$, we output $\mathcal{A}(X)$. Otherwise, if X is in $\mathcal{X} \setminus \mathcal{X}^*$, we pass X through a "filter" and turn it into another dataset Y which is in $\mathcal{X}^*$. Then, we output $\mathcal{A}(Y)$.

To show that $\mathcal{B}$ is the desired algorithm, we have few challenges: (1) we need to show the mapping does not affect the correctness probability by too much. (2) $\mathcal{B}$ is $\Theta(\xi)$-differentially private although its input may be from $\mathcal{X} \setminus \mathcal{X}^*$. Overcoming the second challenge is closely related to the design of the mapping. If two datasets have Hamming distance one, then we need to make sure they will remain "close". In the following section, we explain the mapping, and in the next section, we prove that $\mathcal{B}$ is a ξ-differentially private algorithm with large correctness probability.

6.4 Mapping datasets in $\mathcal{X} \setminus \mathcal{X}^*$ to datasets in $\mathcal{X}^*$

In this section, we provide a randomized mapping that takes $X \in \mathcal{X}$ as the input, and maps it to randomly selected Y in $\mathcal{X}^*$ with two important properties stated in the following Lemma:

Lemma 6.5. *There exists a randomized mapping that takes $X, X' \in \mathcal{X}$ and maps them to $Y, Y' \in \mathcal{X}^*$ respectively with the following property:*

- *If X is in $\mathcal{X}^*$, then it will always be mapped to itself. : $Y = X$.*
- *If the Hamming distance between X and X' is one, then there exists a coupling $\mathcal{C}$ between the random outputs of the mapping, Y and Y', where for any (Y, Y') drawn from $\mathcal{C}$, the Hamming distance between Y and Y' is at most a constant $c = 4$ (independent of A).*

Proof: The main idea is to decrease the ratio $t_i(X)/(b_i(X)+1)$ by replacing a subset of samples in the bottom row with the copies of i to decrease the ratio without introducing new elements that violate the ratio condition. For a dataset X, we look at each element $i \in [n]$, and see how many copies of i are needed to "fix" the ratio. It is not hard to see that if for each element i, $r_i(X)$ many copies is sufficient where $r_i(X)$ is defined as below:

$$r_i(X) \coloneqq \max\left(\left\lceil \frac{t_i(X)}{A} \right\rceil - b_i(X) - 1,\, 0 \right).$$

Let R be a multiset that contains $r_i(X)$ copies of i. We find $|R|$ slots in the bottom row, and replace the samples in those slots with an element in R. If we carefully select the slots and do not replace any copy of i in the bottom row, the new ratio will be: $t_i(X)/(b_i(X)+r_i(X)+1)$ which is at most A. Now, we focus on finding the slots in the bottom row. We can select a slot containing an instance of i, only if the replacement of i does not increase the ratio of the frequencies above A. For an element i, we may remove at most $s_i(X)$ samples where

$$s_i(X) \coloneqq \max\left(b_i(X) + 1 - \left\lceil \frac{t_i(X)}{A} \right\rceil,\, 0 \right).$$

For each element i, we mark $s_i(X)$ many slots which contains copy of i in the bottom row as "available" preferring the slots with the smaller index. Observe that we always have at least $|R|$ many slots since A is

at least two:

$$|R| = \sum_{i=1}^{n} r_i(X) \le \sum_{i=1}^{n} \frac{t_i(X)}{A} \le \frac{s}{A} \le (A-1)\frac{s}{A} = s - \sum_{i=1}^{n} \frac{t_i(X)}{A} \le s - \sum_{i=1}^{n} \left(\left\lceil \frac{t_i(X)}{A} \right\rceil - 1 \right)$$
$$\le s - \sum_{i=1}^{n} b_i(X) - s_i(X) = \sum_{i=1}^{n} s_i(X)$$

We choose the first $|R|$ available slots (i.e. with the smaller indices), and replace the bottom samples in them by the samples in R randomly. After the replacements, it is clear that we did not remove a sample where its ratio could go above A, and we fixed all those elements with the ratio above A as well. Thus, the dataset we get after this process is surely in $\mathcal{X}^*$. Furthermore, if X is already in $\mathcal{X}^*$, then R is an empty set, and the mapping does not change it, so $Y = X$.

Now, we focus on the proof of the existence of the coupling. Let S be the indices of the $|R|$ available slots we select. First note that we consider all the elements in R to be distinct (even though they might be different copies of the same sample, we can index $r_i(X)$ copies of i by $1, 2, \ldots, r_i(X)$). Thus, there are $|R|!$ for assigning the samples in R into the slots in S, and each assigning has probability $1/|R|!$. Suppose two datasets, X and X', differ in exactly one sample: X has an extra copy of i, and X' instead has an extra copy of j. Also, let R' and S' be the equivalents of R and S respectively for X'. Clearly, we have $|R| = |S|$, and $|R'| = |S'|$. This discrepancy between X and X' happens in either on the top row or the bottom row. Since the frequency of i and j changes by at most one, $r_i(X)$, $s_i(X)$, $r_j(X)$, and $s_j(X)$ will change by at most. Without loss of generality, if we consider all possible cases, it is not hard to see that one of the two following cases happens:

Case 1: R and R' has the same size, and $|R \cap R'|$ and $|S \cap S'|$ is at least $|R| - 1$. It is not hard to see that there is a bijection between Y and Y'. Assume there exists a set of replacement that turns X into Y. We construct the corresponding Y' accordingly. We start off with X'. We apply the same set of replacements with only two exceptions: Suppose we want to replace the sample in the slot ℓ with k according to the original set of replacement, then we see if k is not in R', we carry on the replacement with $k' = R' \setminus k$. Also, if the ℓ is not in S', we will choose slot $\ell' = S' \setminus S$, pick the slot ℓ' for the replacement. After performing all the replacement we get Y' which has Hamming distance at most four to Y. It is not hard to see that we can map Y' to Y similarly, so there exists a matching between the Y's, and the Y''s. We define the coupling $\mathcal{C}$ to be a probability distribution over $\mathcal{X}^* \times \mathcal{X}^*$, where the probability of (Y, Y') according to the above definition is $1/|R|!$, and it is zero for the rest of the pairs.

Case 2: R and S have one extra member: $R' = R \cup \{k\}$, and $S' = S \cup \{\ell\}$. Assume there exists a set of replacements that turns X into Y. We construct $|R| + 1$ sets of replacements that turn X' into $Y'_1, Y'_2, \ldots, Y'_{|R|+1}$. We start off with X'. We choose one of the replacement in the set which replaces the sample in slot ℓ' by k'. Then, we perform all the replacements on X' except the one that is left out. Now, we do the following: We replace the sample in slot ℓ by k' and the sample in slot ℓ' by k. Clearly, we found an assignment between R' and S', so we construct $Y'_1, \ldots, Y'_{|R|}$. We also perform all the replacement in the set, and in addition to that, we replace the sample in slot ℓ by k to obtain $Y'_{|R|+1}$. It is not hard to see that given Y', we can construct Y as well, so there is a matching between Y and the Y'_t's. Also, Y and the Y''s have a Hamming distance of at most three. Now, we define the coupling $\mathcal{C}$. We set the probability of the pairs (Y, Y'_t) to be $1/(|R| + 1)!$ for $t = 1, \ldots, |R| + 1$. It is clear that each Y appears with probability $(|R| + 1)/(|R| + 1)! = 1/|R|!$. Thus, the desired coupling exists.

Note that in both case, there exists a coupling $\mathcal{C}$ such that each pair drawn from $\mathcal{C}$ have a Hamming distance of at most four. Hence the proof is complete. □

6.5 Proving privacy guarantee after extending the domain

As we describe $\mathcal{B}$ at a high level before, now we formally described it in Algorithm 2. Below, we formally show that the algorithm is differentially private as well.

Algorithm 2 A private procedure for extending the domain

procedure PRIVATE TESTER(X, A)
  $R, S \leftarrow \emptyset$.
  for $i = 1, 2, \ldots, n$ **do**
    if $r_i(X) \geq 1$ **then**
      $R \leftarrow R \cup \{r_i(X) \text{ copies of } i\}$
    if $s_i(X) \geq 1$ **then**
      $S_i \leftarrow$ Set of the smallest $s_i(X)$ indices of the entries in the bottom row of X which contains i.
      $S \leftarrow S \cup S_i$.
  $S \leftarrow |R|$ smallest element in S.
  for each $k \in R$ **do**
    $\ell \leftarrow$ a random element in S.
    $S \leftarrow S \setminus \{\ell\}$.
    $X-\text{bottom}(\ell) \leftarrow k$.
  Output $\mathcal{A}(X)$.

Lemma 6.6. *Assume $\mathcal{A}$ is a $\xi/4$-differentially private algorithm over $\mathcal{X}^*$ with parameter $A \geq 12 \ln n/\delta'$ that output the correct answer with probability at least $1 - \delta$. Algorithm 2 is a ξ-differentially private algorithm over $\mathcal{X}$. which outputs the correct answer with probability at least $1 - \delta - \delta'$.*

Proof: First, we claim that the algorithm changes X with probability at most δ'. Assume s is a Poisson random variable with parameter λ, and let X be the set of $2\,s$ samples from a distribution p. Using Poissonization method, we can think of $t_i(X)$ and $b_i(X)$ as two Poisson random variables with mean $\lambda_i := p(i) \cdot \lambda$. Now, we bound the probability that $t_i(X)/(b_i(X) + 1)$ become larger than one. If λ_i is zero, then $t_i(X)$ and $b_i(X)$ must be zero, so the ratio is below A. Let $B = b_i(X)/\lambda_i$. We consider the following cases for λ_i:

Case 1: $\lambda_i \leq A/2$. By the concentration of a Poisson random variables, we have the following:

$$\begin{aligned}\mathbf{Pr}\left[\frac{t_i(X)}{b_i(X)+1} \geq A\right] &\leq \mathbf{Pr}[t_i(X) - \lambda_i \geq A - \lambda_i] \leq \exp\left(-\frac{(A-\lambda_i)^2}{2\,A}\right) \\ &\leq \exp\left(-\frac{A}{8}\right)\end{aligned}$$

Case 2: $\lambda_i > A/2$. Clearly, we have:

$$\mathbf{Pr}\left[\frac{t_i(X)}{b_i(X)+1} \geq A\right] \leq \mathbf{Pr}[t_i(X) \geq A \cdot b_i(X) + A] = \mathbf{Pr}[t_i(X) \geq A \cdot B \cdot \lambda_i + A]$$

Now, if $A \cdot B \geq 2$, we obtain:

$$\begin{aligned}\mathbf{Pr}\left[\frac{t_i(X)}{b_i(X)+1} \geq A\right] &\leq \mathbf{Pr}[t_i(X) - \lambda_i \geq \lambda_i + A] \leq \exp\left(-\frac{(A+\lambda_i)^2}{2\,(A+2\,\lambda_i)}\right) \\ &\leq \exp\left(-\frac{\lambda_i^2}{6\,\lambda_i}\right) \leq \exp\left(-\frac{A}{12}\right)\end{aligned}$$

If $A \cdot B = A\, b_i(X)/\lambda_i < 2$, it means that $b_i(X)$is at most $2\lambda_i/A$. Thus, we have:

$$\begin{aligned}\mathbf{Pr}\left[\frac{t_i(X)}{b_i(X)+1} \geq A\right] &\leq \mathbf{Pr}\left[b_i(X) \leq \frac{2\,\lambda_i}{A}\right] = \mathbf{Pr}\left[\lambda_i - b_i(X) \geq \frac{(A-2)\cdot\lambda_i}{A}\right] \\ &\leq \exp\left(-\frac{(A-2)^2\,\lambda_i^2}{2\,A^2\,\left(\frac{2\,A-2}{A}\cdot\lambda_i\right)}\right) = \exp\left(-\frac{(A-2)^2}{2\,A\,(2\,A-2)}\cdot\lambda_i\right) \\ &\leq \exp\left(-\frac{\lambda_i}{6}\right) \leq \exp\left(-\frac{A}{12}\right)\end{aligned}$$

where the second to last inequality is true when $A \geq 10$.

In all of the cases above, The probability that the ratio associated with element i goes above A is at most $\exp(-A/12) \leq \delta'/n$. By union bound, the probability of having at least one i with the ratio above A is at most δ'. Observe that if all the ratios are below A, all the $r_i(X)$'s will be zero. Thus, the algorithm does not change X with probability $1-\delta'$. Also, if $\mathcal{A}$ outputs the correct answer with probability at least $1-\delta$, then $\mathcal{B}$ outputs the correct answer with probability at least $1-\delta-\delta'$.

Now, we show that $\mathcal{B}$ is private. In Lemma 6.5, we show our mapping has the following property: Let X and X' in $\mathcal{X}$ be two datasets with Hamming distance at most one. Let Y and Y' be the randomized datasets that X and X' are mapped to. There exists a coupling $\mathcal{C}$ between Y and Y' where the Hamming distance between any (Y, Y') with non-zero probability in $\mathcal{C}$ is at most four. The existence of the coupling and the fact that $\mathcal{A}$ is an $\xi/4$ private algorithm help us to prove the privacy guarantee for $\mathcal{B}$. Let O be an arbitrary output for $\mathcal{B}$. In the context of our paper O can be accept or reject. Below, we show the probability of outputting O on two neighboring dataset X and X' with Hamming distance one, is the same up to a e^{ξ} factor.

$$\begin{aligned}\mathbf{Pr}[\mathcal{B}(X)=O] &= \sum_{Y} \mathbf{Pr}[\mathcal{A}(Y)=O]\cdot\mathbf{Pr}[X \text{ is mapped to } Y] \\ &= \sum_{Y,Y'} \mathbf{Pr}[\mathcal{A}(Y)=O]\cdot\mathcal{C}(Y,Y') \\ &\leq \sum_{Y,Y'} e^{(\xi/c)\cdot|Y-Y'|}\mathbf{Pr}[\mathcal{A}(Y')=O]\cdot\mathcal{C}(Y,Y') \\ &\leq \sum_{Y,Y'} e^{\xi}\mathbf{Pr}[\mathcal{A}(Y')=O]\cdot\mathcal{C}(Y,Y') \\ &\leq \sum_{Y'} e^{\xi}\mathbf{Pr}[\mathcal{A}(Y')=O]\cdot\mathbf{Pr}[X' \text{ is mapped to } Y'] \\ &= e^{\xi}\mathbf{Pr}[\mathcal{B}(X')=O]\end{aligned}$$

Therefore, $\mathcal{B}$ is ξ-private on $\mathcal{X}$. □

7 Proof of the Lemmas

7.1 Proof of Lemma 3.1

Lemma 3.1. *Let $s_{i,1}$ (similarly $s_{i,2}$) be the number of occurrences of element i in the sample sets of p (q). Assume b_i is the number of buckets assigned to element i. Let $v_{i,j,1}$ (similarly $v_{i,j,2}$) be the number of occurrences of bucket (i,j) in the sample set from $p^{(F)}$ ($q^{(F)}$). Then, we have:*

$$\mathbf{E}_r\left[\sum_{j=1}^{b_i}(v_{i,j,1}-v_{i,j,2})^2 - v_{i,j,1} - v_{i,j,2}\,\middle|\, b_i, s_{i,1}, s_{i,2}\right] = \frac{(s_{i,1}-s_{i,2})^2 - s_{i,1} - s_{i,2}}{b_i},$$

where the expectation is taken over all random assignments of the samples to the buckets.

Proof: Observe that given b_i, $s_{i,1}$, and $s_{i,2}$, the number of instances of element i in each bucket, $v_{i,j,1}$, is random variables drawn from a binomial distribution $\mathbf{Bin}(s_{i,1}, 1/b_i)$. Similarly, $v_{i,j,2}$ is drawn from $\mathbf{Bin}(s_{i,2}, 1/b_i)$. Thus, we have:

$$\mathbf{E}\big[v_{i,j,1}\big] = \frac{s_{i,1}}{b_i}\ , \quad \mathbf{E}\big[v_{i,j,1}^2\big] = \mathbf{Var}\big[v_{i,j,1}\big] + \mathbf{E}\big[v_{i,j,1}\big]^2 = s_{i,1}\cdot\left(1-\frac{1}{b_i}\right)\cdot\frac{1}{b_i} + \frac{s_{i,1}^2}{b_i^2} = \frac{s_{i,1}}{b_i} + \frac{s_{i,1}^2 - s_{i,1}}{b_i^2}\ ,$$

$$\mathbf{E}\big[v_{i,j,2}\big] = \frac{s_{i,2}}{b_i}\ , \quad \mathbf{E}\big[v_{i,j,2}^2\big] = \mathbf{Var}\big[v_{i,j,2}\big] + \mathbf{E}\big[v_{i,j,2}\big]^2 = s_{i,2}\cdot\left(1-\frac{1}{b_i}\right)\cdot\frac{1}{b_i} + \frac{s_{i,2}^2}{b_i^2} = \frac{s_{i,2}}{b_i} + \frac{s_{i,2}^2 - s_{i,2}}{b_i^2}\ .$$

Since $v_{i,j,1}$ is independent from $v_{i,j,2}$, then we have:

$$\begin{aligned}
&\mathbf{E}\left[\sum_{j=1}^{b_i}(v_{i,j,1}-v_{i,j,2})^2 - v_{i,j,1} - v_{i,j,2}\,\middle|\, b_i, s_{i,1}, s_{i,2}\right] \\
&= \sum_{j=1}^{b_i}\mathbf{E}\big[(v_{i,j,1}-v_{i,j,2})^2 - v_{i,j,1} - v_{i,j,2}|b_i, s_{i,1}, s_{i,2}\big] \\
&= b_i\cdot\big(\mathbf{E}\big[u_{i,1}^2 + v_{i,1}^2 - 2\cdot v_{i,j,1}\cdot v_{i,j,1} - v_{i,j,1} - v_{i,j,2}|b_i, s_{i,1}, s_{i,2}\big]\big) \\
&= b_i\cdot\left(\frac{s_{i,1}^2 - s_{i,1}}{b_i^2} + \frac{s_{i,2}^2 - s_{i,2}}{b_i^2} - 2\cdot\frac{s_{i,1}}{b_i}\cdot\frac{s_{i,2}}{b_i}\right) \\
&= \frac{(s_{i,1}-s_{i,2})^2 - s_{i,1} - s_{i,2}}{b_i}\ .
\end{aligned}$$

which completes the proof. □

Lemma 5.3. *Assume random variable x is drawn from $\mathbf{Poi}(\lambda)$. If λ is at least $1.5\cdot\ln(1/c)$, then the probability of x being larger than $3\,\lambda$ is at most $1-c$.*

Proof: We use the tail bound for the Poisson distribution we have:

$$\mathbf{Pr}_x[x \geq \lambda + 2\lambda] \leq \exp\left(-\frac{(2\,\lambda)^2}{2\cdot(2+1)\cdot\lambda}\right) \leq \exp(-2\,\lambda/3) \leq c\,.$$

Thus, the proof is complete. □

Lemma 5.4. *Assume we have n independent random variables $x_1, x_2, \ldots, x_n$ in the range $[0, +\infty)$. Suppose each x_i is at least A_i with probability $p \geq 0.95$ where A_i is a fixed number. Then, with probability at least 0.9, $\sum_{i=1}^n x_i$ is at least $0.1\sum_{i=1}^n A_i$.*

Proof: We define another set of random variables, y_i's, as follows:

$$y_i = \begin{cases} A_i & \text{with probability p} \\ 0 & \text{with probability 1-p} \end{cases}$$

Clearly, the expected value of $\sum_{i=1}^n y_i$ is $p\cdot\sum_{i=1}^n A_i$. Note that we can see y_i as A_i multiplied by a Bernoulli random variable with bias p. Thus, the variance of $\sum_{i=1}^n y_i$ is:

$$\mathbf{Var}\left[\sum_{i=1}^n y_i\right] = \sum_{i=1}^n \mathbf{Var}[y_i] = \sum_{i=1}^n A_i^2\mathbf{Var}[\mathbf{Ber}(p)] = p(1-p)\cdot\sum_{i=1}^n A_i^2\,.$$

Now, by Chebyshev's inequality, we can bound the probability of being far from their expectation:

$$\begin{aligned}\mathbf{Pr}\left[\sum_{i=1}^{n} y_i \le 0.1 \cdot \mathbf{E}\left[\sum_{i=1}^{n} y_i\right]\right] &\le \mathbf{Pr}\left[\left|\sum_{i=1}^{n} y_i - \mathbf{E}\left[\sum_{i=1}^{n} y_i\right]\right| \ge 0.9 \cdot \mathbf{E}\left[\sum_{i=1}^{n} y_i\right]\right] \\ &\le \frac{\mathbf{Var}[\sum_{i=1}^{n} y_i]}{0.9^2 \cdot \mathbf{E}[\sum_{i=1}^{n} y_i]^2} \le \frac{p(1-p)\sum_{i=1}^{n} A_i^2}{0.9^2\, p^2 \cdot (\sum_{i=1}^{n} A_i)^2} \le \frac{p(1-p)}{0.9^2\, p^2} \le 0.1\,.\end{aligned}$$

Observe that y_i's are defined such that the probability of $\sum_{i=1}^{n} x_i > a$ for any number a is at least the probability of $\sum_{i=1}^{n} y_i > a$. Thus, we have:

$$\mathbf{Pr}\left[\sum_{i=1}^{n} x_i \ge 0.1 \sum_{i=1}^{n} A_i\right] \ge \mathbf{Pr}\left[\sum_{i=1}^{n} y_i \ge 0.1 \sum_{i=1}^{n} A_i\right] \ge 0.9\,.$$

Hence, the proof is complete. □

Lemma 6.4. *Let $x_1, x_2, \ldots, x_n$ be n non-negative random variables. Suppose there exist two constants c and p, both at most one, such that for each random variable x_i, we have:*

$$\mathbf{Pr}[x_i < c \cdot \mathbf{E}[x_i]] \le p\,,$$

Then, one can show:

$$\mathbf{Pr}\left[\sum_{i=1}^{n} x_i < \frac{c \cdot \sum_{i=1}^{n} \mathbf{E}[x_i]}{10}\right] \le \frac{10\,p}{9}\,.$$

Proof: At a high level, we expect each random variable x_i to "contribute" to the sum of x_i's by $\mathbf{E}[x_i]$. If a random variable x_i is at least $c\,\mathbf{E}[x_i]$, it is contributing "enough" to the sum. While otherwise, the sum "misses" a contribution of amount $\mathbf{E}[x_i]$. The main idea is to show that total amount that the sum misses is not too large.

More formally, for each i, we define an auxiliary random variables y_i as below. Roughly speaking y_i indicates how much the sum is missing due to a low x_i:

$$y_i = \begin{cases} \mathbf{E}[x_i] & \text{if } x_i < c \cdot \mathbf{E}[x_i] \\ 0 & \text{otherwise} \end{cases}$$

First, we claim that the sum of y_i's is not too large since we have:

$$\mathbf{E}\left[\sum_{i=1}^{n} y_i\right] = \sum_{i=1}^{n} \mathbf{E}[x_i] \cdot \mathbf{Pr}[x_i < c \cdot \mathbf{E}[x_i]] \le p \cdot \sum_{i=1}^{n} \mathbf{E}[x_i]\,.$$

Using Markov's inequality, the sum of y_i's cannot be larger than $0.9 \cdot \sum_{i=1}^{n} \mathbf{E}[x_i]$ with probability more than $10\,p/9$. Hence, with probability $1 - 10\,p/9$, we may assume $\sum_{i=1}^{n} y_i$ is at most $0.9 \cdot \sum_{i=1}^{n} \mathbf{E}[x_i]$.

Now, we show that the sum of x_i's cannot be too small when the sum of y_i's is less than $0.9 \cdot \sum_{i=1}^{n} \mathbf{E}[x_i]$. To see this, let I be the set of indices i for which $x_i \ge c \cdot \mathbf{E}[x_i]$. Then, one can obtain:

$$\begin{aligned}\sum_{i=1}^{n} x_i \ge \sum_{i \in I} x_i \ge c \cdot \sum_{i \in I} \mathbf{E}[x_i] &= c \cdot \left(\sum_{i=1}^{n} \mathbf{E}[x_i] - \sum_{i \notin I} \mathbf{E}[x_i]\right) \\ = c \cdot \left(\sum_{i=1}^{n} \mathbf{E}[x_i] - \sum_{i \notin I} y_i\right) &= c \cdot \left(\sum_{i=1}^{n} \mathbf{E}[x_i] - \sum_{i=1}^{n} y_i\right) \\ \ge \frac{c \cdot \sum_{i=1}^{n} \mathbf{E}[x_i]}{10}\,,\end{aligned}$$

which concludes the lemma. □

Lemma 7.1. *Assume x is binomial random variable with n trials and bias p. Then, the following is true.*

$$\mathbf{E}_x\Big[\frac{1}{x+1}\Big] \leq \min\Big(\frac{1}{p\cdot(n+1)}, 1\Big)$$

Proof:

$$\begin{aligned}
\mathbf{E}_x\Big[\frac{1}{x+1}\Big] &= \frac{1}{p\cdot(n+1)}\sum_{x=0}^{n}\frac{n+1}{x+1}\binom{n}{x}p^{x+1}(1-p)^{n-x} \\
&= \frac{1}{p\cdot(n+1)}\sum_{y=1}^{n}\binom{n+1}{y}p^{y}(1-p)^{(n+1)-y} = \frac{1-(1-p)^{n+1}}{p\cdot(n+1)} \\
&\leq \min\Big(\frac{1}{p\cdot(n+1)}, 1\Big)
\end{aligned}$$

□

Lemma 7.2. *Assume x is binomial random variable with n trials and bias p. Then, the following is true.*

$$\mathbf{E}_x\Big[\frac{1}{(x+2)(x+1)}\Big] \leq \min\Big(\frac{1}{p^2\cdot(n+1)(n+2)}, 1\Big)$$

Proof:

$$\begin{aligned}
\mathbf{E}_x\Big[\frac{1}{(x+2)(x+1)}\Big] &= \frac{1}{p^2\cdot(n+1)(n+2)}\sum_{x=0}^{n}\frac{(n+2)(n+1)}{(x+2)(x+1)}\binom{n}{x}p^{x+2}(1-p)^{n-x} \\
&= \frac{1}{p^2\cdot(n+1)(n+2)}\sum_{y=2}^{n}\binom{n+2}{y}p^{y}(1-p)^{(n+1)-y} \\
&= \frac{1-(1-p)^{n+2}-(n+2)\,p(1-p)^{n+1}}{p^2\cdot(n+1)(n+2)} \\
&\leq \min\Big(\frac{1}{p^2\cdot(n+1)(n+2)}, 1\Big)
\end{aligned}$$

□

Lemma 7.3. *Suppose we have a bin with m balls where exactly t of them are red. We draw balls from the bin without replacement. Let X be the number of red balls in the first s trials and let Y be the number of red balls in the next k trials. Then, we have:*

$$\mathbf{E}\Big[\frac{X^2}{Y+1}\Big] \leq \min\Big(\frac{2\,(s-1)t}{(k+1)}, 2\,t^2\Big) + t\,.$$

Proof: We write the expectation explicitly:

$$\begin{aligned}
\mathbf{E}\left[\frac{X^2}{Y+1}\right] &\leq \sum_a \sum_b \frac{a^2}{b+1} \cdot \mathbf{Pr}[X=a] \cdot \mathbf{Pr}[Y=b] = \sum_a \sum_b \frac{a^2}{b+1} \cdot \frac{\binom{s}{a}\binom{k}{b}\binom{m-s-k}{t-a-b}}{\binom{m}{t}} \\
&= \sum_{a\geq 2} \sum_b \frac{2\,a\,(a-1)}{b+1} \frac{\binom{s}{a}\binom{k}{b}\binom{m-s-k}{t-a-b}}{\binom{m}{t}} + s \cdot \sum_b \frac{1}{b+1} \frac{\binom{k}{b}\binom{m-s-k}{t-1-b}}{\binom{m}{t}} \\
&= \sum_{a\geq 2} \sum_b \frac{2\,a\,(a-1)}{b+1} \frac{\binom{s}{a}\binom{k}{b}\binom{m-s-k}{t-a-b}}{\binom{m}{t}} + s \cdot \sum_b \frac{1}{b+1} \frac{\binom{k}{b}\binom{m-s-k}{t-1-b}}{\binom{m}{t}} \\
&= \frac{2\,s(s-1)\,t}{(k+1)\,m} \sum_{a\geq 2} \sum_b \frac{\binom{s-2}{a-2}\binom{k+1}{b+1}\binom{m-s-k}{t-a-b}}{\binom{m-1}{t-1}} + \frac{s}{k+1} \cdot \sum_b \frac{\binom{k+1}{b+1}\binom{m-s-k}{t-1-b}}{\binom{m}{t}}
\end{aligned}$$

We define the two sums in the last line as A and B:

$$A \coloneqq \sum_{a\geq 2} \sum_b \frac{\binom{s-2}{a-2}\binom{k+1}{b+1}\binom{m-s-k}{t-a-b}}{\binom{m-1}{t-1}}, \qquad B \coloneqq \sum_b \frac{\binom{k+1}{b+1}\binom{m-s-k}{t-1-b}}{\binom{m}{t}}.$$

We claim A and B are two probabilities of the following randomized processes, so we can bound them. Suppose we have an urn with $m-1$ balls, $t-1$ of them are red. A is the probability that we get at least one red ball if we draw $k+1$ balls from the bin without replacement. Let Z be the number of red balls we draw after $k+1$ draws. Using Markov's inequality, we get:

$$A = \mathbf{Pr}[Z \geq 1] \leq \min(1, \mathbf{E}[Z]) \leq \min\left(1, \frac{(t-1)\cdot(k+1)}{(m-1)}\right)$$

Furthermore, we can define B as the following probability: Assume we have an urn of m balls including t red balls. If we draw $(s-1)+(k+1)$ balls from the urn without replacement. B is the probability that non of the $s-1$ draws are red, and there is at least one red draw in the next $k+1$ draws. This is clearly smaller than the probability of seeing at least one red ball in the $k+1$ draws. Thus, similar to the above, we have:

$$B \leq \min\left(1, \frac{t\cdot(k+1)}{m}\right).$$

Now, putting all these together, and using the fact that $s \leq m$, we obtain:

$$\mathbf{E}\left[\frac{X^2}{Y+1}\right] \leq \min\left(\frac{2\,(s-1)t}{(k+1)}, 2\,t^2\right) + t\,.$$

□

Lemma 7.4. *Assume X is a random variable drawn from $\mathbf{HG}(m,t,k)$, then*

$$\mathbf{E}_X\left[\frac{1}{(X+1)}\right] \leq \min\left(1, \frac{(m+1)}{(t+1)(k+1)}\right).$$

Proof: Clearly, the expectation cannot be larger than one since $X \geq 0$. For the other term, by the definition,

we can achieve the following bound:

$$\begin{aligned}
\mathbf{E}_x\left[\frac{1}{x+1}\right] &= \sum_{x=\max(0,k-(m-t))}^{\min(t,k)} \mathbf{HG}(x;m,t,k)\cdot\frac{1}{x+1} = \sum_x \frac{\binom{t}{x}\binom{m-t}{k-x}}{\binom{m}{k}}\cdot\frac{1}{x+1} \\
&= \sum_x \frac{m+1}{(t+1)(k+1)}\frac{\frac{t+1}{x+1}\binom{t}{x}\binom{m-t}{k-x}}{\frac{m+1}{k+1}\binom{m}{k}} = \sum_x \frac{m+1}{(t+1)(k+1)}\cdot\frac{\binom{t+1}{x+1}\binom{(m+1)-(t+1)}{(k+1)-(x+1)}}{\binom{m+1}{k+1}} \\
&\leq \frac{m+1}{(t+1)(k+1)}\cdot\sum_x \mathbf{HG}(x+1;m+1,t+1,k+1) \leq \frac{m+1}{(t+1)(k+1)}
\end{aligned}$$

where the last line is true because the sum of the probabilities according to a distribution is at most one. □

Lemma 7.5. *Assume X is a random variable drawn from $\mathbf{HG}(m,t,k)$, then*

$$\mathbf{E}_X\left[\frac{1}{(X+2)(X+1)}\right] \leq \min\left(1, \frac{(m+2)(m+1)}{(t+2)(t+1)(k+2)(k+1)}\right).$$

Proof: Clearly, the expectation cannot be larger than one since $X \geq 0$. For the other term, by the definition, we can achieve the following bound:

$$\begin{aligned}
\mathbf{E}_X\left[\frac{1}{(X+2)(X+1)}\right] &= \sum_X \mathbf{HG}(X;m,t,k)\cdot\frac{1}{(X+2)(X+1)} \\
&= \sum_X \frac{\binom{t}{X}\binom{m-t}{k-X}}{\binom{m}{k}}\cdot\frac{1}{(X+2)(X+1)} \\
&= \frac{(m+2)(m+1)}{(t+2)(t+1)(k+2)(k+1)}\cdot\sum_X \frac{\binom{t+2}{x+2}\binom{(m+2)-(t+2)}{(k+2)-(x+2)}}{\binom{m+2}{k+2}} \\
&= \frac{(m+2)(m+1)}{(t+2)(t+1)(k+2)(k+1)}\sum_x \mathbf{HG}(x;m+2,t+2k+2) \\
&\leq \frac{(m+2)(m+1)}{(t+2)(t+1)(k+2)(k+1)}
\end{aligned}$$

where the last line is true, because the sum of the probabilities in a distribution is at most one. □

Acknowledgments

MA is supported by funds from the MIT-IBM Watson AI Lab (Agreement No. W1771646), the NSF grants IIS-1741137, and CCF-1733808. ID is supported by NSF Award CCF-1652862 (CAREER), NSF AiTF award CCF-1733796 and a Sloan Research Fellowship. DK is supported by NSF Award CCF-1553288 (CAREER) and a Sloan Research Fellowship. RR is supported by funds from the MIT-IBM Watson AI Lab (Agreement No. W1771646) the NSF grants CCF-1650733, CCF-1733808, IIS-1741137 and CCF-1740751.

References

[ACFT19] J. Acharya, C. Canonne, C. Freitag, and H. Tyagi. Test without trust: Optimal locally private distribution testing. In *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2067–2076. PMLR, 2019.

[ADK15] J. Acharya, C. Daskalakis, and G. Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pages 3591–3599, 2015.

[ADR18] M. Aliakbarpour, I. Diakonikolas, and R. Rubinfeld. Differentially private identity and equivalence testing of discrete distributions. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pages 169–178, 2018.

[AJOS14] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh. Sublinear algorithms for outlier detection and generalized closeness testing. In *2014 IEEE International Symposium on Information Theory*, pages 3200–3204, 2014.

[ASZ18] J. Acharya, Z. Sun, and H. Zhang. Differentially private testing of identity and closeness of discrete distributions. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018*, pages 6879–6891, 2018.

[BC17] T. Batu and C. L. Canonne. Generalized uniformity testing. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017*, pages 880–889, 2017.

[BFF+01a] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *Proc. 42nd IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001.

[BFF+01b] T. Batu, E. Fisher, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variabls for indenpence and identity. In *Proc. 42nd Annual IEEE Symposium on Foundations of Computer Science*, pages 442–451, 2001.

[BFR+00] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing that distributions are close. In *IEEE Symposium on Foundations of Computer Science*, pages 259–269, 2000.

[BFR+13] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. Testing closeness of discrete distributions. *J. ACM*, 60(1):4, 2013.

[BKR04] T. Batu, R. Kumar, and R. Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *ACM Symposium on Theory of Computing*, pages 381–390, 2004.

[BV15] B. B. Bhattacharya and G. Valiant. Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pages 2611–2619, 2015.

[Can15] C. L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015.

[CDGR16] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld. Testing shape restrictions of discrete distributions. In *33rd Symposium on Theoretical Aspects of Computer Science, STACS 2016*, pages 25:1–25:14, 2016.

[CDK17] B. Cai, C. Daskalakis, and G. Kamath. Priv'it: Private and sample efficient identity testing. In *International Conference on Machine Learning, ICML*, pages 635–644, 2017.

[CDKS17a] C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart. Testing bayesian networks. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 370–448, 2017.

[CDKS17b] C. L. Canonne, I. Diakonikolas, D. M. Kane, and A. Stewart. Testing conditional independence of discrete distributions. *CoRR*, abs/1711.11560, 2017. In STOC'18.

[CDS17] C. L. Canonne, I. Diakonikolas, and A. Stewart. Fourier-based testing for families of distributions. *CoRR*, abs/1706.05738, 2017. In NeurIPS 2018.

[CDVV14] S. Chan, I. Diakonikolas, P. Valiant, and G. Valiant. Optimal algorithms for testing closeness of discrete distributions. In *SODA*, pages 1193–1203, 2014.

[DDS+13] C. Daskalakis, I. Diakonikolas, R. Servedio, G. Valiant, and P. Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *SODA*, pages 1833–1852, 2013.

[DGPP16] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. Collision-based testers are optimal for uniformity and closeness. *Electronic Colloquium on Computational Complexity (ECCC)*, 23:178, 2016.

[DGPP17] I. Diakonikolas, T. Gouleakis, J. Peebles, and E. Price. Sample-optimal identity testing with high probability. *CoRR*, abs/1708.02728, 2017. In ICALP 2018.

[DHS15] I. Diakonikolas, M. Hardt, and L. Schmidt. Differentially private learning of structured discrete distributions. In *Conference on Neural Information Processing Systems, NIPS*, pages 2566–2574, 2015.

[DK16] I. Diakonikolas and D. M. Kane. A new approach for testing properties of discrete distributions. In *IEEE Symposium on Foundations of Computer Science, FOCS*, pages 685–694, 2016. Full version available at abs/1601.05557.

[DKN15a] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015*, pages 1183–1202, 2015.

[DKN15b] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Testing identity of structured distributions. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015*, pages 1841–1854, 2015.

[DKN17] I. Diakonikolas, D. M. Kane, and V. Nikishkin. Near-optimal closeness testing of discrete histogram distributions. In *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017*, pages 8:1–8:15, 2017.

[DKP19] I. Diakonikolas, D. M. Kane, and J. Peebles. Testing identity of multidimensional histograms. In *Conference on Learning Theory, COLT 2019*, pages 1107–1131, 2019.

[DKS18] I. Diakonikolas, D. M. Kane, and A. Stewart. Sharp bounds for generalized uniformity testing. In *Advances in Neural Information Processing Systems 31, NeurIPS 2018*, pages 6204–6213, 2018.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, pages 265–284, Berlin, Heidelberg, 2006. Springer-Verlag.

[DP17] C. Daskalakis and Q. Pan. Square Hellinger subadditivity for Bayesian networks and its applications to identity testing. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 697–703, 2017.

[DR14a] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[DR14b] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[Dwo09] C. Dwork. The differential privacy frontier (extended abstract). In *TCC*, pages 496–502, 2009.

[GLRV16] M. Gaboardi, H. W. Lim, R. M. Rogers, and S. P. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *International Conference on Machine Learning, ICML*, pages 2111–2120, 2016.

[Gol17] O. Goldreich. *Introduction to Property Testing*. Forthcoming, 2017.

[GR18] M. Gaboardi and R. Rogers. Local private hypothesis testing: Chi-square tests. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pages 1612–1621, 2018.

[KFS17] K. Kakizaki, K. Fukuchi, and J. Sakuma. Differentially private chi-squared test by unit circle mechanism. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1761–1770. PMLR, 2017.

[KR17] D. Kifer and R. Rogers. A new class of private chi-square hypothesis tests. In *International Conference on Artificial Intelligence and Statistics, AISTATS*, pages 991–1000, 2017.

[LRR11] R. Levi, D. Ron, and R. Rubinfeld. Testing properties of collections of distributions. In *ICS*, pages 179–194, 2011.

[NP33] J. Neyman and E. S. Pearson. 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.

[Pan08] L. Paninski. A coincidence-based test for uniformity given very sparsely-sampled discrete data. *IEEE Transactions on Information Theory*, 54:4750–4755, 2008.

[Pea00] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

[Rub12] R. Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

[She18] O. Sheffet. Locally private hypothesis testing. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, pages 4612–4621, 2018.

[VV14] G. Valiant and P. Valiant. An automatic inequality prover and instance optimal identity testing. In *FOCS*, 2014.

[WLK15] Y. Wang, J. Lee, and D. Kifer. Revisiting differentially private hypothesis tests for categorical data. *CoRR*, abs/1511.03376, 2015.