Private Distribution Testing with Heterogeneous Constraints: Your Epsilon Might Not Be Mine

Clément L. Canonne
University of Sydney, School of Computer Science, Australia

Yucheng Sun
ETH Zürich, Switzerland

Abstract

Private closeness testing asks to decide whether the underlying probability distributions of two sensitive datasets are identical or differ significantly in statistical distance, while guaranteeing (differential) privacy of the data. As in most (if not all) distribution testing questions studied under privacy constraints, however, previous work assumes that the two datasets are *equally* sensitive, i.e., must be provided the same privacy guarantees. This is often an unrealistic assumption, as different sources of data come with different privacy requirements; as a result, known closeness testing algorithms might be unnecessarily conservative, "paying" too high a privacy budget for half of the data. In this work, we initiate the study of the closeness testing problem under *heterogeneous* privacy constraints, where the two datasets come with distinct privacy requirements.

We formalize the question and provide algorithms under the three most widely used differential privacy settings, with a particular focus on the *local* and *shuffle* models of privacy; and show that one can indeed achieve better sample efficiency when taking into account the two different "epsilon" requirements.

2012 ACM Subject Classification Theory of computation → Streaming, sublinear and near linear time algorithms; Security and privacy

Keywords and phrases differential privacy, distribution testing, local privacy, shuffle privacy

Digital Object Identifier 10.4230/LIPIcs.ITCS.2024.23

Related Version *Full Version*: https://arxiv.org/abs/2309.06068

Funding *Clément L. Canonne*: Supported by an ARC DECRA (DE230101329) and an unrestricted gift from Google Research.

1 Introduction

Hypothesis testing allows a statistician, practitioner, or scientist to validate their model or to detect whether one of their assumptions is statistically improbable. One of the prototypical hypothesis testing tasks is two-sample goodness-of-fit, which asks to determine whether two unknown probability distributions are equal, based on samples from both. This task has received a lot of attention from the computer science community over the past decades as part of the broader area of *distribution testing*, where testing questions are phrased as promise problems with a particular emphasis on finite-sample guarantees and data-efficient algorithms (see, e.g., [41, 17, 19], [36, Chapter 11], and references within). In distribution testing, two-sample goodness-of-fit corresponds to *closeness testing*, where the two unknown distributions $\mathbf{p}, \mathbf{q}$ are over a discrete domain of size k. Given a distance parameter $\alpha \in (0, 1]$, one then seeks to distinguish (with high probability) between the cases $\mathbf{p} = \mathbf{q}$ and $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \alpha$ from as few samples from $\mathbf{p}, \mathbf{q}$ as possible, where d_{TV} denotes the total variation (statistical) distance. A long line of work in the distribution testing literature culminated in a full understanding of the sample complexity (i.e., the number of observations required) of this question, with respect to all parameters [15, 42, 25, 28, 23].

The recent rise of privacy concerns, and with it a focus on privacy-preserving data analysis, led researchers to consider the natural question of *private* hypothesis testing, where the samples are seen as sensitive data, and the testing algorithms must guarantee differential privacy (DP) [31], a rigorous definition of privacy which has become the *de facto* standard. Various distribution testing questions have been studied under this lens, including that of closeness testing, for which tight bounds on the sample complexity have been established in the so-called "central" model of differential privacy [10, 6, 44]. Yet, these results suffer from one major shortcoming: they assume both sets of samples (the one coming from $\mathbf{p}$, and the other from $\mathbf{q}$) to be *equally sensitive*, that is, to require the same privacy guarantees (the parameter $\varepsilon > 0$ of differential privacy, where smaller values of ε correspond to better privacy guarantees).

This assumption, while justifiable in some settings, is misguided in many others: for instance, when the two datasets come from different companies, demographic groups, or even countries subject to different legal requirements – a typical use case for closeness testing, where one seeks to check if the data from two distinct populations have similar statistical properties. Another extreme use case would be when only one of the datasets has privacy constraints, and the other is from a simulated process (e.g., when checking if the distribution of the output from a digital twin, or of synthetic data, matches that of the real world). In these cases, using the "same ε" for data from both distributions would be unnecessarily conservative, and could lead to requiring much more (costly, or hard to either gather or generate) data from one distribution than required.

To address this shortcoming, we initiate the study of closeness testing under *heterogeneous* privacy constraints, where the two datasets come with distinct privacy parameters $\varepsilon_1, \varepsilon_2 > 0$. We further formulate the question not only in the (central) DP model, but also in two others of the most widely used distributed models of privacy, the more stringent *local* DP model [38, 30] and the *shuffle* model of privacy [26, 32]. To the best of our knowledge, our work is also the first to formulate closeness testing in the two latter models, even for homogeneous privacy constraints ($\varepsilon_1 = \varepsilon_2$). We next elaborate on our results and detail our contributions.

1.1 Our Results and Contributions

Our first contribution is to formalize the question of closeness testing under heterogeneous privacy constraints in three models of differential privacy: the local model, the shuffle model, and the central model (Section 2.4). While this formalization is somewhat straightforward in the latter case, it is less so in the first two, especially in the shuffle model. Indeed, the shuffle model of privacy relies at its core on a trusted channel (the "shuffler") which randomly permutes the messages from all users, effectively anonymizing them. However, the very question of closeness testing requires the ability to distinguish between the messages from two groups of users: the ones holding samples from $\mathbf{p}$, and the ones holding samples from $\mathbf{q}$. Thus, some care has to be given in how to define the problem in the first place, putting an emphasis on what choice captures the best the possible use cases discussed earlier. Our definition for the shuffle privacy setting involves two "shufflers" (one per group of users), and the privacy guarantee applies to the shuffled output of each of them separately. This aligns with the practical setting where the two datasets come from users from two different company or entities, in which case the shuffling is performed in an early stage ("between" the users and the corresponding company), *before* being sent out in the world.

Other definitional choices could have been (1) to add to each message a (non-private) label, to specify from which population they came, and use a single random permutation for all messages; or (2) to restrict the type of permutations (no longer uniformly random) to only

shuffle the messages "within each population." While these two options are equivalent to the one we chose, they are less intuitive, possibly more cumbersome to analyze, and obscure the original motivation for the problem.

Our second contribution is to provide algorithms achieving non-trivial trade-offs between the two privacy parameters $\varepsilon_1, \varepsilon_2$ in all three models of privacy considered, showing that it *is* possible balance the different privacy requirements of the two populations to do significantly better than defaulting to $\varepsilon = \min(\varepsilon_1, \varepsilon_2)$. Moreover, our results are the first (even for homogeneous privacy constraints) for closeness testing in the local and shuffle models of privacy, and yield sample-optimal[1] (up to constant factors) bounds in the former.

In order to state our results, we first recall the distinction between *private-coin* and *public-coin* distributed protocols:[2] in the former, each user has only access to their own randomness, independent of every other user's. In the latter, however, there exists a common random seed (*in addition* to each user's personal randomness), publicly available to all parties (users, analyzer, and world) but still independent of the users' data. Thus, while in both cases the protocols are non-interactive, in public-coin protocols this common random seed can be used to achieve better accuracy, by letting the users somehow coordinate.[3] For detailed definitions of the privacy settings and types of distributed protocols, we refer the reader to Section 2.

Our first results address closeness testing under heterogeneous *local* privacy constraints, establishing a tight trade-off in all parameters.

▶ **Theorem 1** (Local Privacy, Private-Coin). *There exists a* private-coin *protocol for closeness testing which guarantees ε_1-local privacy to the n_1 users of the first group, and ε_2-local privacy to the n_2 users of the second, with $n_1 = O\left(\frac{k^{3/2}}{\varepsilon_1^2\alpha^2}\right)$ and $n_2 = O\left(\frac{k^{3/2}}{\varepsilon_2^2\alpha^2}\right)$. Moreover, this is optimal.*

▶ **Corollary 2** (Local Privacy, Public-Coin). *There exists a* public-coin *protocol for closeness testing which guarantees ε_1-local privacy to the n_1 users of the first group, and ε_2-local privacy to the n_2 users of the second, with $n_1 = O\left(\frac{k}{\varepsilon_1^2\alpha^2}\right)$ and $n_2 = O\left(\frac{k}{\varepsilon_2^2\alpha^2}\right)$. Moreover, this is optimal.*

Our next two results concern the shuffle model of privacy, with algorithms guaranteeing (approximate) differential privacy. For simplicity, we only provide here an informal statement, omitting the at most logarithmic dependence on the parameter δ and focusing on $\varepsilon_1, \varepsilon_2$. We refer the reader to Theorems 24 and 26 for the full statements.

▶ **Theorem 3** (Shuffle Privacy, Private-Coin (Informal)). *There exists a* private-coin *protocol for closeness testing which guarantees ε_1-shuffle (approximate) privacy to the n_1 users of the first group, and ε_2-shuffle (approximate) privacy to the n_2 users of the second, with*

$$n_1 = O\left(\max\left(\frac{k^{1/2}}{\alpha^2}, \frac{k^{3/4}}{\varepsilon_1\alpha}\right)\right)$$

and $n_2 = O\left(\frac{\varepsilon_1^2}{\varepsilon_2^2} n_1\right)$ (assuming without loss of generality that $\varepsilon_2 \leq \varepsilon_1$).

---

[1] The results match the lower bounds on the sample complexity of identity testing in the local model. Since one can reduce identity testing to closeness testing if assuming the ability to generate samples from known distributions "for free," the lower bounds for identity testing are also lower bounds for closeness testing in the same model.

[2] Confusingly, the "private" in "private-coin" does not refer to differential privacy, but to the fact that the randomness ("coin") of a user is hidden from all others.

[3] We note that while much of the work in the shuffle model (and, slightly less so, in the local model) does not focus on this distinction, we do so here as the availability of a common random seed is known to make a difference in related testing problems, such as uniformity and identity testing, in both the local and shuffle models of privacy [1, 11, 13, 22] as well as in other (non-private) distributed settings [3, 4].

▶ **Corollary 4** (Shuffle Privacy, Public-Coin (Informal)). *There exists a* public-coin *protocol for closeness testing which guarantees ε_1-shuffle (approximate) privacy to the n_1 users of the first group, and ε_2-shuffle (approximate) privacy to the n_2 users of the second, with*

$$n_1 = O\left(\max\left(\frac{k^{1/2}}{\alpha^2}, \frac{k^{1/2}}{\varepsilon_1\alpha}, \frac{k^{2/3}}{\varepsilon_1^{2/3}\alpha^{4/3}}\right)\right)$$

and $n_2 = O\left(\frac{\varepsilon_1^2}{\varepsilon_2^2} n_1\right)$ (assuming without loss of generality that $\varepsilon_2 \leq \varepsilon_1$).

Interestingly, specializing our results to the homogeneous privacy constraints case ($\varepsilon_1 = \varepsilon_2$), our upper bounds for local and shuffle privacy match[4] the best known algorithms for the simpler problem of *identity* testing (where one of the two distributions is fully known in advance). This shows, perhaps surprisingly, that unlike in the non-private and central DP cases, there is no sample complexity gap between closeness and identity testing.

Finally, we provide a simple closeness testing algorithm for the *central* model of differential privacy:

▶ **Theorem 5** (Central Privacy). *There exists an algorithm for closeness testing which guarantees ε_1-differential privacy to the n_1 users of the first group, and ε_2-differential privacy to the n_2 users of the second, with*

$$n_1 = O\left(\max\left(\frac{k^{1/2}}{\alpha^2}, \frac{k^{1/2}}{\varepsilon_1^{1/2}\alpha}, \frac{k^{2/3}}{\alpha^{4/3}}, \frac{k^{1/3}}{\varepsilon_1^{2/3}\alpha^{4/3}}, \frac{1}{\varepsilon_1\alpha}\right)\right)$$

and $n_2 = O\left(\frac{\varepsilon_1}{\varepsilon_2} n_1\right)$ (assuming without loss of generality that $\varepsilon_2 \leq \varepsilon_1$).

Our results can be interpreted in two ways. The first focuses on privacy as a *requirement* from the two groups of users, and looks at how the costs varies among the two populations. That is, our results quantifies how much more data one needs to collect from the group with more stringent privacy requirements, compared to the group of less "privacy-demanding" users. Our upper bounds show that the overhead scales at most quadratically with the ratio of privacy parameters, i.e., as $(\varepsilon_1/\varepsilon_2)^2$ under local and shuffle privacy, and as $\varepsilon_1/\varepsilon_2$ under (central) differential privacy.

The second point of view focuses on privacy as a *promise* (or incentive) instead of a requirement: under this lens, our results show that if more users from one group are willing to participate in the analysis, or if one of the two populations is larger, then it can automatically be guaranteed better data privacy (and our algorithms provide a bound on "how much more privacy" this is).

1.2 Overview of Techniques

We now outline the main ideas behind our results, and outline some possible approaches to improve upon them.

[4] That is, exactly match the optimal sample complexity in the locally private case; and match the best known upper bounds in the shuffle private case for approximate privacy (or for pure privacy if one treats δ as a constant), and nearly match the corresponding lower bounds for shuffle DP.

1.2.1 Local privacy

The algorithm underlying Theorem 1 relies on *Hadamard response*, which was proposed in [7] as a communication-efficient mechanism under local differential privacy constraints. As shown in [1], when the privacy parameters of two groups of users are identical, this mechanism allows reducing the original closeness testing problem to testing whether the ℓ_2-distance between the mean vector of two product-Bernoulli distributions (product distributions over $\{\pm 1\}^d$) is 0 or larger than some parameter. Any sample-optimal ℓ_2-closeness testing algorithm for product-Bernoulli distributions can be used to achieve this task efficiently. Here, noise is added to samples from these two product-Bernoulli distributions to preserve privacy. When the privacy parameter becomes smaller (more privacy), the two product-Bernoulli distributions are perturbed by more noise, i.e., each attribute in the mean vector will be closer to $\frac{1}{2}$. However, when the privacy parameters are heterogeneous this reduction does not go through as is, since even in the case when $\mathbf{p} = \mathbf{q}$ the transformation will lead to two product-Bernoulli distributions with distinct mean vectors. In particular, previous ℓ_2-closeness (of the mean vectors) testing algorithm for product distributions can no longer be used as a blackbox to achieve the task.

To adapt it to the heterogeneous privacy case, we considered a test statistic for testing closeness of two product-Bernoulli distributions with different levels of noise. A side product is a new sample-optimal ℓ_2-closeness tester for product distributions, which simplifies previous algorithms (which were designed for a more general problem, either testing closeness of the product distributions in total variation distance instead of ℓ_2 distance of the mean vectors, or without the independence assumption in the soundness case).

Corollary 2 then follows from combining Theorem 1 and the *domain compression* primitive, which was proposed in [2] as a general technique to derive public-coin schemes from private-coin ones: at a high level, the idea is for the users to leverage public randomness in order to jointly hash the domain of size k into $k' \ll k$ parts, and to consider the induced probability distributions on these k' parts. This was shown to preserve the total variation distance between probability distributions up to a shrinking factor of $\sqrt{k'/k}$, effectively "replacing" α by $\alpha' \asymp \alpha\sqrt{k'/k}$. Selecting the optimal value of k' to minimize the resulting sample complexity when applying the private-coin algorithm with the new parameters k' and α' (in this case, $k' = 2$) leads to the public-coin algorithm.

1.2.2 Shuffle privacy

Turning to shuffle privacy, the algorithm behind Theorem 3 starts with the following observation, which was somewhat implicit in [22] in the context of uniformity testing: if n users get each a sample from some distribution $\mathbf{p}$ over $[k]$ and use the distributed Poisson mechanism with parameter $\frac{\mu}{n} = O\big(1/(n\varepsilon^2)\big)$ to "privately report" their data in the shuffle model, then the central server gets access to $N := n + k\mu$ i.i.d. samples from a new mixture distribution

$$\mathbf{p}' := (1-\gamma)\cdot\mathbf{p} + \gamma\cdot\mathbf{u}_k$$

where $\gamma := k\mu/N$ and $\mathbf{u}_k$ is the uniform distribution on $[k]$. (This is not totally accurate as stated, but becomes true if we replace "n users" and "N samples" by "Poisson(n) users" and Poisson(N) samples.") In our case, this means that we can obtain N_1 "ε_1-private" samples from $\mathbf{p}'$ and N_2 "ε_2-private" samples from $\mathbf{q}'$, where

$$\mathbf{p}' := (1-\gamma_1)\cdot\mathbf{p} + \gamma_1\cdot\mathbf{u}_k$$
$$\mathbf{q}' := (1-\gamma_2)\cdot\mathbf{q} + \gamma_2\cdot\mathbf{u}_k$$

where $\gamma_1 := k\mu_1/N_1 = O\Big(\frac{k}{n_1 N_1 \varepsilon_1^2}\Big)$ and $\gamma_2 := k\mu_2/N_2 = O\Big(\frac{k}{n_2 N_2 \varepsilon_2^2}\Big)$ (ignoring again, for the sake of this discussion, the dependence on the second privacy parameter δ). Now, a natural requirement is to ask that $\mathbf{p}' = \mathbf{q}'$ whenever $\mathbf{p} = \mathbf{q}$, so that our original private closeness testing task on $\mathbf{p}, \mathbf{q}$ reduces to a new one, *non-private*, on $\mathbf{p}', \mathbf{q}'$, for which we can leverage the existing results on closeness testing.

Doing so boils down to enforcing $\gamma_1 = \gamma_2$, which in turn leads to the requirement $n_2 = \Theta\big((\varepsilon_1/\varepsilon_2)^2 n_1\big)$; this also gives the "new" distance parameter $\alpha' := (1-\gamma_1)\alpha = (1-\gamma_2)\alpha$ for resulting closeness testing problem on $\mathbf{p}', \mathbf{q}'$. All that remains is to now invoke an existing and optimal (non-private) closeness testing algorithm with unequal numbers of samples (since $N_1 \neq N_2$), e.g., that of [29], and derive the resulting conditions on N_1, N_2 (and thus on n_1, n_2) this yields to establish Theorem 3.

As in the locally private case, the public-coin case (Corollary 4) then follows *via* the domain compression technique, by selecting the optimal number of parts $k' := k'(k, \alpha, \varepsilon_1, \varepsilon_2)$ to hash the domain into, to minimize the resulting sample complexity when plugging k' and $\alpha' \asymp \alpha\sqrt{k'/k}$ into Theorem 3, subject to $2 \leq k' \leq k$.

What about amplification by shuffling? We note that a natural idea to obtain a (different) private-coin shuffle private algorithm (and, via domain compression, a corresponding-public-coin one as well) would be to start from a locally private algorithm under heterogeneous privacy constraints and apply the amplification by shuffling result of [34]. This idea was used in [22] for *identity* testing (under homogeneous privacy constraints), i.e., testing whether an unknown distribution is equal to a fully known reference one. However, this approach comes with two conditions on the LDP protocol one starts with: (1) all users of the same group must use the same local randomizer, and (2) the protocol needs to work reasonably well even for large privacy parameters $\varepsilon_1, \varepsilon_2 \gg 1$, i.e., in the low-privacy regime. Unfortunately, the LDP protocol behind Theorem 1 satisfies (1) but not (2), and thus amplification by shuffling would not lead to a shuffle private algorithm with good enough sample complexity; and the LDP identity testing algorithm of [22] does not seem to generalize to closeness testing, let alone closeness testing under heterogeneous privacy constraints. We believe that obtaining a sample-optimal LDP algorithm under heterogeneous privacy constraints satisfying (1) and (2) could lead to an improvement over Theorem 3 and Corollary 4 (in terms of n_2), and leave this as an interesting future direction.

1.2.3 Central privacy

Finally, Theorem 5 follows from combining two ingredients: the first is the (sample-optimal) closeness testing algorithm of [44] for the central model of differential privacy for the equal-privacy case, which has sample complexity

$$O\left(\max\left(\frac{k^{1/2}}{\alpha^2}, \frac{k^{1/2}}{\varepsilon^{1/2}\alpha}, \frac{k^{2/3}}{\alpha^{4/3}}, \frac{k^{1/3}}{\varepsilon^{2/3}\alpha^{4/3}}, \frac{1}{\varepsilon\alpha}\right)\right).$$

This algorithm relies on adding suitably calibrated noise to the non-private, but low-sensitivity closeness testing algorithm of [28]; this low sensitivity (i.e., robust to changing any single sample) is crucial when bounding the noise required to make the algorithm differentially private, as adding Laplace noise with parameter $O(1/\varepsilon)$ *independent of* n, k, α then suffices.

A natural idea to adapt this to the heterogeneous privacy case would be, as for our shuffle privacy protocol, to first extend this algorithm to the "uneven-sample case" (i.e., $n_2 \geq n_1$) and then introduce two different levels of noise, balancing n_1, n_2 accordingly in order to achieve privacy with parameters $\varepsilon_1, \varepsilon_2$. Unfortunately, as we discuss further in Section 3.3, this approach turns out to be much trickier than expected, as the main algorithm of [28] does not appear to easily generalize to the $n_1 \neq n_2$ case. Worse, any attempt to do so (as

done in [28] for this particular algorithm, via the so-called "flattening" technique, or using the other non-private, uneven-sample closeness testing algorithms available in the literature) results in algorithms with *much* higher sensitivity, requiring too large a level of random noise in the "privatization" step and leading to vacuous sample complexity bounds.

Instead, we take recourse to a much simpler technique, *privacy amplification by subsampling* [39] (as used for a similar goal in [37]). The idea is to use the above algorithm of [44] with privacy parameter ε_1, requiring n_1 samples (where n_1 is given by the above equation) from both groups. Now, this achieves ε_1-privacy for the first group; as for the second, which requires a better privacy guarantee, we start with a larger group of n_2 users and subsample by picking uniformly at random a subgroup of n_1 users. By a standard argument, this improves the privacy guarantee from ε_1 to $\frac{n_1}{n_2}\varepsilon_1 = \varepsilon_2$, as desired.

1.3 Related Work

Uniformity, identity, and closeness testing are three of flagship (and related) questions in distribution testing, with a long history in classical statistics, where the analysis is under an asymptotic regime, as the number of samples goes to infinity. In contrast, computer scientists often study these problems under the framework of property testing, where one wishes to achieve certain accuracy with a limited number of samples (i.e., a particular focus on the "finite-sample" regime). For this regime, goodness-of-fit testing without privacy constraints has been extensively studied, with sample complexity bounds summarized in Table 1.[5] We refer the readers to [18, 12, 19] for surveys of the area.

Table 1 Sample complexity bounds of goodness-of-fit testing without privacy constraints.

Testing	Upper bound	Lower bound
Identity testing	$O\left(\frac{k^{1/2}}{\alpha^2}\right)$	tight
Closeness testing	$O\left(\frac{k^{1/2}}{\alpha^2} + \frac{k^{2/3}}{\alpha^{4/3}}\right)$	tight

A significant body of work has considered these questions under (homogeneous) DP constraints: we list previous results in Table 2. Most relevant to our work, [44] proposed a sample-optimal closeness-testing algorithm under central differential privacy by leveraging the (non-private) test statistic of [28]. [1] gave sample-optimal identity testing algorithms under local differential privacy for both private-coin and public-coin settings. For shuffle differential privacy, [13, 22] gave identity testing algorithms with the same complexity bounds on the required number of samples (approximate DP), and [27] later provided analogous bounds in the pure privacy setting; however, the exact sample complexity of the question for homogeneous shuffle privacy (either pure or approximate) remains open.

On heterogeneous privacy

The focus of our paper is to introduce and formalize the question of heterogeneous privacy for closeness testing. However, the idea to allow for different users to have different privacy requirements has appeared in other contexts, and has been studied for different problems (with various formalizations of what allowing for "heterogeneous privacy parameters" then means). See, for instance, [37, 8, 40, 33].

[5] It is well known that uniformity testing and identity testing share the same upper bound and lower bound of sample complexity, as these two problems can be reduced to each other. Thus, we only include 'identity testing' in this table.

Table 2 Sample complexity bounds for goodness-of-fit testing under homogeneous differential privacy. (The shuffle privacy bounds from [27], marked with "*", hold for *pure* privacy, i.e., without the logarithmic factor in $1/\delta$).

	Testing question	Upper bound	Lower bound
Local	Identity, private-coin	$O\left(\frac{k^{3/2}}{\alpha^2\varepsilon^2}\right)$ [1]	tight [3]
	Identity, public-coin	$O\left(\frac{k}{\alpha^2\varepsilon^2}\right)$ [1]	tight [3, 11]
	Closeness, private-coin	No result	$\Omega\left(\frac{k^{3/2}}{\alpha^2\varepsilon^2}\right)$ [3]
	Closeness, public-coin	No result	$\Omega\left(\frac{k}{\alpha^2\varepsilon^2}\right)$ [3, 11]
Shuffle	Identity, private-coin	$O\left(\frac{k^{3/4}}{\alpha\varepsilon}\sqrt{\log(\frac{1}{\delta})} + \frac{\sqrt{k}}{\alpha^2}\right)$ [13, 22], [27]*	$\Omega\left(\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)$ [13]
	Identity, public-coin	$O\left(\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{2/3}}\log^{1/3}\frac{1}{\delta} + \frac{\sqrt{k}}{\alpha\varepsilon}\log^{1/2}\frac{1}{\delta} + \frac{\sqrt{k}}{\alpha^2}\right)$ [22], [27]*	$\Omega\left(\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)$ [13]
	Closeness (either)	No result	$\Omega\left(\frac{k^{2/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{\sqrt{k}}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)$ [13]
Central	Identity	$O\left(\frac{k^{1/2}}{\alpha^2} + \frac{k^{1/2}}{\alpha\varepsilon^{1/2}} + \frac{k^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{1}{\alpha\varepsilon}\right)$ [6]	tight [6]
	Closeness	$O\left(\frac{k^{1/2}}{\alpha^2} + \frac{k^{2/3}}{\alpha^{4/3}} + \frac{k^{1/2}}{\alpha\varepsilon^{1/2}} + \frac{k^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{1}{\alpha\varepsilon}\right)$ [44]	tight [44]

1.4 Organization of the Paper

In Section 2, we give the formal definition of our problem and introduce some necessary tools used in this paper. In Section 3, we present our algorithms and the related proofs, and give a discussion on closeness testing under homogeneous central differential privacy constraint. In Section 4, we discuss future work. All omitted proofs can be found in the full version of the paper [24].

2 Model and Preliminaries

2.1 Closeness Testing

Given sample access to two unknown distributions, closeness testing asks whether these distributions are the same, or differ significantly in terms of statistical distance:[6]

▶ **Definition 6** (Closeness testing). *Let $\mathbf{p}, \mathbf{q}$ be two unknown distributions with domain $[k]$. A closeness testing algorithm with sample complexity n takes inputs $\alpha \in (0, 1]$, a set of n i.i.d. samples from $\mathbf{p}$ and a set of n i.i.d. samples from $\mathbf{p}$ and outputs either* accept *or* reject *such that the following holds:*

- *If $\mathbf{p} = \mathbf{q}$, then the algorithm outputs* accept *with probability at least $\frac{2}{3}$;*
- *If $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \alpha$, then the algorithm outputs* reject *with probability at least $\frac{2}{3}$.*

In this definition, $\frac{2}{3}$ is just some arbitrary number picked between $\frac{1}{2}$ and 1. By a standard amplification argument any high probability $1 - \beta$ can be achieved by repeating the test independently $O(\log(1/\beta))$ times and using a majority rule.

2.2 Differential Privacy

We now recall the relevant concepts we will extensively use, starting with the definition of differential privacy.

[6] The statistical (total variation) distance between two probability distributions $\mathbf{p}, \mathbf{q}$ over the same domain $\mathcal{X}$ is defined as $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \sup_{S \subseteq \mathcal{X}} (\mathbf{p}(S) - \mathbf{q}(S)) = \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1$, where the supremum is taken over all measurable subsets.

▶ **Definition 7** (Differential privacy [31]). *A randomized algorithm $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{R}^k$ is said to be* (ε, δ)-differentially private *if for all measurable $S \subseteq \mathcal{R}^k$ and all neighboring datasets $x, y \in \mathcal{X}$:* $\Pr[\mathcal{M}(x) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(y) \in S] + \delta$, *where two datasets x, y are said to be* neighboring *if* $\mathrm{dist}(x, y) \leq 1$ *(i.e., for* $\mathrm{dist}(,)$ *being the Hamming distance, if they only differ in (at most) one entry).*

A key property of differential privacy is *immunity of post-processing.*

▶ **Lemma 8** (Immunity of post-processing). *Let $f\colon \mathcal{X}^n \to \mathcal{R}^k$ be a mapping which is (ε, δ)-differentially private. Let $g\colon \mathcal{R}^k \to \mathcal{R}^{k'}$ be any arbitrary random mapping. Then the mapping $g \circ f$ is still (ε, δ)-differentially private.*

In many cases, this randomized mapping $\mathcal{M}$ is obtained by adding random noise to some function $f(x)$ of the data, where f is the (non-private) function to be computed (such mechanisms are called *additive noise mechanisms*). The amount of noise to be added to $f(x)$ then needs to be tailored to the specific properties of f: in particular, such an important property is its ℓ_1-sensitivity.

▶ **Definition 9** (Sensitivity). *The ℓ_1-sensitivity of a function $f\colon \mathcal{N}^{|\mathcal{X}|} \to \mathcal{R}^k$ is $\triangle = \max_{x,y \in \mathcal{N}^{|\mathcal{X}|},\, \mathrm{dist}(x,y) \leq 1} \|f(x) - f(y)\|_1$. We say the function f is* $\triangle$-sensitive.

One mechanism for adding randomness using ℓ_1-sensivity as a parameter is the Poisson mechanism. We use this mechanism in the shuffle model.

▶ **Lemma 10** (Poisson mechanism [35]). *Let $f\colon \mathcal{X}^n \to \mathcal{Z}$ be a $\triangle$-sensitive function. For any $\varepsilon > 0, \delta \in (0, 1)$ and $\lambda \geq \frac{16 \log(10/\delta)}{(1 - e^{-\varepsilon/\triangle})^2} + \frac{2\triangle}{1 - e^{-\varepsilon/\triangle}}$, the randomized function $\mathcal{A}(x) = f(x) + Y$ where $x \in \mathcal{X}^n, Y \sim \mathrm{Poisson}(\lambda)$ is $(\varepsilon, \delta)-$differentially private in the central and shuffle model.*

A standard technique to amplify a differential-private algorithm is amplification by subsampling. The formal statement is as follows.

▶ **Lemma 11** (Amplification by subsampling (see, e.g.,, [39], or[14, Theorem 9])). *Let $\mathcal{A} : \mathcal{X}^{n_1} \to \mathcal{R}^k$ be a $(\varepsilon_1, \delta_1)$-differentially private mapping and $\mathcal{H}\colon \mathcal{X}^{n_2} \to \mathcal{X}^{n_1}$ be a randomized mapping which uniformly chooses n_1 elements from n_2 elements. Then, the composition of this two mappings $\mathcal{A} \circ \mathcal{H}\colon \mathcal{X}^{n_2} \to \mathcal{R}^k$ is $(\varepsilon_2, \delta_2)$-differentially private where $\varepsilon_2 = \ln(1 + \frac{n_1}{n_2}(e^{\varepsilon_1} - 1))$ and $\delta_2 = \frac{n_1}{n_2}\delta_1$. In particular, if $\varepsilon_1 \leq 1$ then $\varepsilon_2 = O\left(\frac{n_1}{n_2}\varepsilon_1\right)$.*

We conclude by recalling the privacy of Randomized Response [43], a standard mechanisms:

▶ **Fact 12** (Randomized Response). *Let $\mathcal{M}_f$ be the mechanism which takes one bit as input and flips it with probability $\frac{1}{e^{\varepsilon}+1}$. Then, $\mathcal{M}$ is $(\varepsilon, 0)-$differentially private.*

2.3 Central, local and shuffle models

In the central model of differential privacy, there is a trusted data curator who holds all the original data and guarantees its output is differentially private. A stricter model is the local model of differential privacy, where the data curator is untrusted and only receives noisy data. It is well known that the local model provides a stronger privacy guarantee, but often at the cost of utility (that is, usually leads to worse accuracy).

A third model, "between" the local and central models is the shuffle model. While a central data curator is still not trusted in this model, we allow a "shuffler" to receive messages from the users and anonymize them by applying a uniformly random permutation . The permutation is typically implemented using cryptographic primitives such as secure multi-party computation.

One advantage of shuffling is that while these cryptographic primitives are time-intensive, implementing a shuffler is simple enough and is not too time-consuming. (Instead, a fully trusted algorithm using cryptography would be too computationally intensive.)

2.4 Central, Shuffle, and Local Models for Heterogeneous Privacy and Data

Since testing involves two different distributions, it is natural to consider the case where users from different distributions have different concerns of privacy. Specifically, we want to make sure our algorithm is $(\varepsilon_1, \delta_1)$-differentially private for samples from $\mathbf{p}$ and $(\varepsilon_2, \delta_2)$-differentially private for samples from $\mathbf{q}$. We introduce the corresponding definition of the testing task below.

▶ **Definition 13** (Closeness testing under heterogeneous local differential privacy constraints)**.** *Let $\mathbf{p}, \mathbf{q}$ be two unknown distributions with domain $[k]$. A* closeness testing algorithm under heterogeneous local differential privacy constraints *consists of the following:*

- *n_1 randomizers $\mathcal{R}_1, ..., \mathcal{R}_{n_1} : \mathcal{X} \times \{0,1\}^r \to \mathcal{Y}$ mapping a sample drawn from $\mathbf{p}$ and a public randomness bit of length r to a privatised output.*
- *n_2 randomizers $\mathcal{R'}_1, ..., \mathcal{R'}_{n_2} : \mathcal{X'} \times \{0,1\}^r \to \mathcal{Y'}$ mapping a sample drawn from $\mathbf{q}$ and a public randomness bit of length r to a privatised output.*
- *an analyser $\mathcal{A} : \mathcal{Y}^{n_1} \times \mathcal{Y}^{n_2} \times \{0,1\}^r \to \mathcal{Z}$ mapping all privatised message and the public randomness bit to the result of analysis either* accept *or* reject *such that the following holds*
  - *If $\mathbf{p} = \mathbf{q}$, then $\mathcal{A}$ outputs* accept *with probability at least $\frac{2}{3}$*
  - *If $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \alpha$, then $\mathcal{A}$ outputs* reject *with probability at least $\frac{2}{3}$*

When $r = 0$ (no public randomness), the testing algorithm is said to belong to the private-coin local differential-private model. *Otherwise, it belongs to the* public-coin local differential-private model.

When the output $\mathcal{R}(\mathcal{X})$ of each randomizer $\mathcal{R}$ is $(\varepsilon_1, \delta_1)$-differentially private w.r.t. $\mathcal{X}$ and the output $\mathcal{R}'(\mathcal{X}')$ of each randomizer $\mathcal{R}'$ is $(\varepsilon_2, \delta_2)$-differentially private w.r.t. $\mathcal{X}'$, $\mathcal{P}$ is said to be $((\varepsilon_1, \delta_1), (\varepsilon_2, \delta_2))$-heterogeneously locally differentially private. *For simplicity, $\mathcal{P}$ is said to be* $(\varepsilon_1, \varepsilon_2)$-heterogeneously locally differentially private *when $\delta_1, \delta_2 = 0$.*

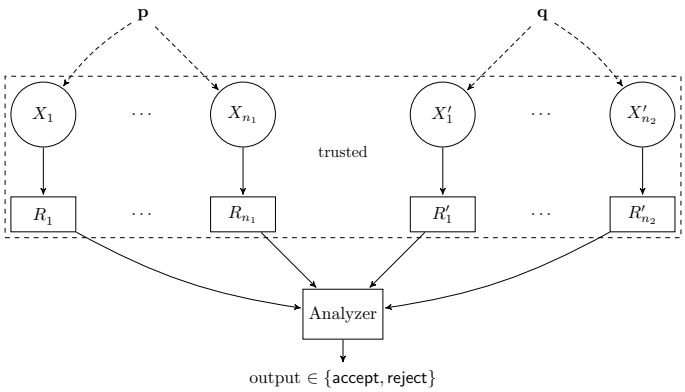

Figure 1 Closeness testing under heterogeneous *local* differential privacy constraints. The R_i's (resp. R'_i's) are the local randomizers used by the users to privatize their data prior to sending it to the analyzer.

We now provide the analogous definitions for shuffle privacy. As discussed in the introduction, this definition is not entirely straightforward, as the very definition of the closeness testing problem requires the ability to distinguish between two groups of users – those with inputs from $\mathbf{p}$, and those with inputs from $\mathbf{q}$. This goes against the objective of shuffling, and motivates the introduction of two distinct shufflers: one for each group.

▶ **Definition 14** (Closeness testing under heterogeneous shuffle differential privacy constraints)**.** *Let $\mathbf{p}, \mathbf{q}$ be two unknown distributions with domain $[k]$. A closeness testing algorithm under heterogeneous shuffle differential privacy constraints consists of the following:*

- *n_1 randomizers $\mathcal{R}_1, ..., \mathcal{R}_{n_1} : \mathcal{X} \times \{0,1\}^r \to \mathcal{Y}$ mapping a sample drawn from $\mathbf{p}$ and a public randomness bit of length r to a privatised output.*
- *n_2 randomizers $\mathcal{R}'_1, ..., \mathcal{R}'_{n_2} : \mathcal{X} \times \{0,1\}^r \to \mathcal{Y}'$ mapping a sample drawn from $\mathbf{q}$ and a public randomness bit of length r to a privatised output.*
- *A shuffler $\mathcal{S}_1 : \mathcal{Y} \to \mathcal{Y}^*$ that concatenates message vectors and then applies a uniformly random permutation to the messages.*
- *A shuffler $\mathcal{S}_2 : \mathcal{Y}' \to \mathcal{Y}'^*$ that concatenates message vectors and then applies a uniformly random permutation to the messages.*
- *an analyser $\mathcal{A} : \mathcal{Y}^* \times \mathcal{Y}'^* \times \{0,1\}^r \to \mathcal{Z}$ mapping all privatised message and the public randomness bit to the result of analysis either* accept *or* reject *such that the following holds*
  - *If $\mathbf{p} = \mathbf{q}$, then $\mathcal{A}$ outputs* accept *with probability at least $\frac{2}{3}$*
  - *If $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \alpha$, then $\mathcal{A}$ outputs* reject *with probability at least $\frac{2}{3}$*

*When the output of the shuffler $\mathcal{S}_1$ is $(\varepsilon_1, \delta_1)$-differentially private w.r.t. $\mathcal{X}$ and the output of the shuffler $\mathcal{S}_2$ is $(\varepsilon_2, \delta_2)$-differentially private w.r.t. $\mathcal{X}'$, $\mathcal{P}$ is said to be $((\varepsilon_1, \delta_1), (\varepsilon_2, \delta_2))$-*heterogeneously shuffle differentially private*. For simplicity, $\mathcal{P}$ is said to be $(\varepsilon_1, \varepsilon_2)$-*heterogeneously shuffle differentially private *when $\delta_1, \delta_2 = 0$.*

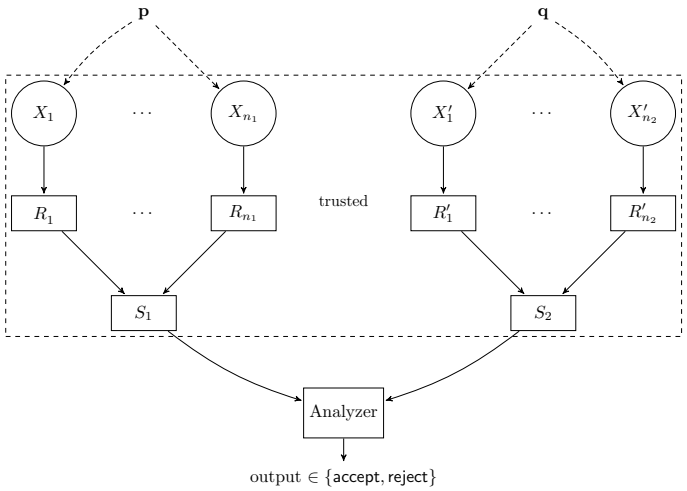

Figure 2 Closeness testing under heterogeneous *shuffle* differential privacy constraints. Here, S_1 and S_2 are the shufflers for the two groups, and the R_i's (resp. R'_i's) are the randomizers used by the users, prior to the shuffling.

Finally, we conclude with the definition of the task under (central) differential privacy:

▶ **Definition 15** (Closeness testing under heterogeneous central differential privacy constraints)**.** *Let* $\mathbf{p}, \mathbf{q}$ *be two unknown distributions with domain* $[k]$*. A closeness testing algorithm under central local differential privacy constraints consists of the following:*

- *an analyser* $\mathcal{A} : \mathcal{X} \times \mathcal{X}' \times \{0,1\}^r \to \mathcal{Z}$ *mapping a sample vector* $\mathcal{X}$ *from* $\mathbf{p}$ *and a sample vector* $\mathcal{X}'$ *from* $\mathbf{q}$ *to the result of analysis either* accept *or* reject *such that the following holds*
  - *If* $\mathbf{p} = \mathbf{q}$*, then* $\mathcal{A}$ *outputs* accept *with probability at least* $\frac{2}{3}$
  - *If* $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \alpha$*, then* $\mathcal{A}$ *outputs* reject *with probability at least* $\frac{2}{3}$

When the output $\mathcal{A}(\mathcal{X})$ *of the analyser* $\mathcal{A}$ *is* $(\varepsilon_1, \delta_1)$*-differentially private w.r.t.* $\mathcal{X}$ *and* $(\varepsilon_2, \delta_2)$*-differentially private w.r.t.* $\mathcal{X}'$*,* $\mathcal{P}$ *is said to be* $((\varepsilon_1, \delta_1), (\varepsilon_2, \delta_2))$-heterogeneously centrally differentially private*. For simplicity,* $\mathcal{P}$ *is said to be* $(\varepsilon_1, \varepsilon_2)$-heterogeneously centrally differentially private *when* $\delta_1, \delta_2 = 0$*.*

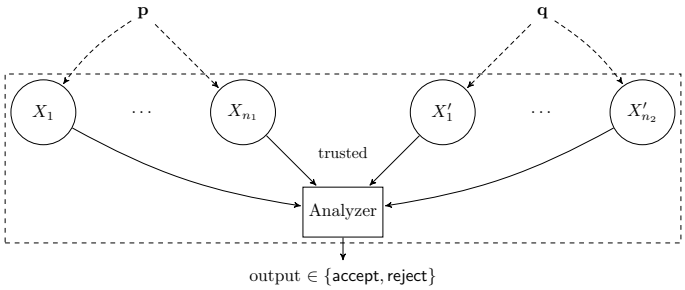

Figure 3 Closeness testing under heterogeneous *central* differential privacy constraints.

▶ Remark 16 (Using the right tool for the job). While the shuffle model of privacy is very appealing due to its balance between privacy guarantees and utility, and stems from practical considerations [26, 16], we emphasize that it might not be a "silver bullet" for *every* scenario. We believe that the use of two shufflers for some goodness-of-fit questions is natural, both conceptually and in practice; however, there may be settings where implementing instead the *central* privacy algorithms with secure multiparty computation (MPC) may be easier.

2.5 Domain compression

Finally, to obtain public-coin protocols from private-coin ones, we will rely on the following *domain compression* result, a hashing-type technique that allows to trade domain size for distance parameter in distribution testing:

▶ **Lemma 17** (Domain Compression [2])**.** *There exist absolute constants* $c_1, c_2 > 0$ *such that the following holds. For any* $2 \leq k' \leq k$ *and any distributions* $\mathbf{p}, \mathbf{q}$ *over* $[k]$*,*

$$\Pr_{\Pi}\left[\mathrm{d}_{\mathrm{TV}}(\mathbf{p}_\Pi, \mathbf{q}_\Pi) \geq c_1\sqrt{\frac{k'}{k}}\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q})\right] \geq c_2\,,$$

where $\Pi = (\Pi_1, \ldots \Pi_{k'})$ *is a uniformly random partition of* $[k]$ *in* k' *subsets, and* $\mathbf{p}_\Pi$ *denotes the probability distribution on* $[k']$ *induced by* $\mathbf{p}$ *and* Π *via* $\mathbf{p}_\Pi(i) = \mathbf{p}(\Pi_i)$*.*

Notation

Throughout, we write $a_n \gtrsim b_n$ (resp. $a_n \lesssim b_n$) to denote the existence of an absolute constant $C > 0$ such that $a_n \leq C \cdot b_n$ (resp. $a_n \geq C \cdot b_n$) for all n; and use $a_n \asymp b_n$ when both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. Besides this, we use the standard $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ notation. Hereafter, we identify a probability distribution $\mathbf{p}$ over a discrete domain $\mathcal{X}$ with its probability mass function (pmf), writing $\mathbf{p}(x)$ for $\Pr_{X\sim\mathbf{p}}[\,X = x\,]$; and for a subset S of the domain, write $\mathbf{p}(S) = \sum_{x\in S} \mathbf{p}(x)$.

3 Our Algorithms

We now provide the details of our algorithms, starting with those under (heterogeneous) local privacy.

3.1 Under Local Privacy

3.1.1 Private-coin protocol

Our algorithm for testing closeness under heterogeneous local differential privacy constraints is based on the Hadamard Response mechanism [5]; we recall one of its key properties below.

▶ **Theorem 18** ([5]). *Let H^K be the $K \times K$ Hadamard matrix where $K = 2^{\lfloor \log_2(k+1)\rfloor}$, which is the smallest power of two larger than k. Let C_j be the locations of 1s in the jth column of H^K where $j \in [K]$. For any distribution $\mathbf{p}$, let $\mathbf{p}(C_j)$ be the probability that a sample from $\mathbf{p}$ falls in set C_j. Then we have*

$$\sum_{j=1}^{k} (\mathbf{p}(C_j) - \mathbf{q}(C_j))^2 = \frac{K}{4}\|\mathbf{p} - \mathbf{q}\|_2^2.$$

Recall that identity and closeness testing fix a distance α, and test whether two distributions $\mathbf{p}, \mathbf{q}$ are the same or the total variation distance between two distributions is larger than α. By using the Cauchy–Schwarz inequality, we have $\sqrt{k}\|\mathbf{p} - \mathbf{q}\|_2 \geq \|\mathbf{p} - \mathbf{q}\|_1$ for any two distributions $\mathbf{p}$, $\mathbf{q} \in [k]$. Thus, if $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1 > \alpha$ then we must have $\frac{4}{K}\sum_{j=1}^{k}(\mathbf{p}(C_j) - \mathbf{q}(C_j))^2 = \|\mathbf{p} - \mathbf{q}\|_2^2 \geq (\frac{1}{\sqrt{k}}\|\mathbf{p} - \mathbf{q}\|_1)^2 > \frac{4}{k}\alpha^2$. Since $K \geq k$, that implies

$$\sum_{j=1}^{k} (\mathbf{p}(C_j) - \mathbf{q}(C_j))^2 > \alpha^2. \tag{1}$$

Otherwise, if $\mathbf{p} = \mathbf{q}$, we must have

$$\sum_{j=1}^{k} (\mathbf{p}(C_j) - \mathbf{q}(C_j))^2 = \frac{K}{4}\|\mathbf{p} - \mathbf{q}\|_2^2 = 0. \tag{2}$$

Motivated by this observation, the identity testing algorithm in [1] divides users from distribution $\mathbf{p}$ into K disjoint groups with equal size. It then assigns the users in the jth group to a set C_j. Each user generates a 1-bit message, indicating whether the data of the user belongs to C_j. To make the output of each user differentially private, each user needs to add some noise to their output. Specifically, each user flips the 1 bit of message with a certain probability $\frac{1}{e^{\varepsilon}+1}$ and sends it using Randomized Response Fact 12.

Recall that the message sent by each user is only one bit, and the messages sent in the same group follows the same Bernoulli distribution. For users from distribution $\mathbf{p}$, by taking

the message from one user from every set, we can obtain a sample from a product-Bernoulli distribution P with length K. Let $\mu(P)$ be the mean of the product-Bernoulli distribution, we have

$$\mu(P)_j = \frac{e^\varepsilon}{e^\varepsilon+1}\sum_{x\in C_j}\mathbf{p}(x) + \frac{1}{e^\varepsilon+1}\sum_{x\notin C_j}\mathbf{p}(x) = \frac{e^\varepsilon-1}{e^\varepsilon+1}\mathbf{p}(C_j) + \frac{1}{e^\varepsilon+1} \tag{3}$$

Since our task is closeness testing instead of identity testing, we also need to perform the same operations for users from $\mathbf{q}$ and obtain samples from another product-Bernoulli distribution Q. Similarly, we use $\mu(Q)$ to denote the mean of the product-Bernoulli distribution Q.

There is a very intuitive understanding of Equation (3). When $\varepsilon \to \infty$ (no privacy), $\mu(P)_j \to \mathbf{p}(C_j)$. When $\varepsilon \to 0$ (no accuracy), $\mu(P)_j$ converges to $\frac{1}{2}$. Moreover, if users from two distributions are using the same parameter $(\varepsilon, 0)$ for differential privacy and $\mathrm{d_{TV}}(\mathbf{p},\mathbf{q}) \geq \alpha$, we have $\|\mu(P)-\mu(Q)\|_2 > \frac{e^\varepsilon-1}{e^\varepsilon+1}\alpha$ by combining Equations (1) and (3).

That implies we can test whether $\mathbf{p}=\mathbf{q}$ or $d_{TV}(\mathbf{p},\mathbf{q}) > \alpha$ by *non-privately* testing whether $\mu(P)=\mu(Q)$ or $\|\mu(P)-\mu(Q)\|_2 > \frac{e^\varepsilon-1}{e^\varepsilon+1}\alpha.$, using any sample-optimal ℓ_2-testing algorithm for testing identity and closeness of product distributions (e.g., [21]). This leads to the optimal sample complexity for identity testing under LDP constraints, as shown in [1].

If we want to use the same privatizing method with heterogeneous constraints, however, we can no longer simply use a ℓ_2 closeness testing algorithm for product distributions as outlined above: indeed, our differential privacy constraints for the two distributions $\mathbf{p},\mathbf{q}$ are not the same. By Fact 12 we need to flip the bits of message from two distributions with probabilities $\frac{1}{e^{\varepsilon_1}+1}$ and $\frac{1}{e^{\varepsilon_1}+1}$ respectively, if we want messages from $\mathbf{p}$ to be $(\varepsilon_1,0)$-differentially private and messages from $\mathbf{q}$ to be $(\varepsilon_2,0)$-differentially private. But that implies $\mu(P)$ and $\mu(Q)$ will not be the same even if $\mathbf{p}=\mathbf{q}$. Thus, we need to have a different algorithm for testing closeness between $\mathbf{p},\mathbf{q}$ given $\mu(P),\mu(Q)$. To design this algorithm, we first provide an algorithm for closeness testing of product distributions tailored to our purpose, which is simpler than the (more general) algorithm in [21]. The idea of this testing algorithm is inspired by the test statistic in [20], which we can simplify as we do not need, in our case, to deal with arbitrary covariances. While the end guarantees are not new, we provide this slightly simpler algorithm for completeness.

Algorithm 1 Testing closeness of two product distributions P, Q.

Require: Two groups of samples $X^{(1)},...,X^{(n)},X'^{(1)},...,X'^{(n)}$ from the $d-$dimensional product distribution P and two groups of samples $Y^{(1)},...,Y^{(n)},Y'^{(1)},...,Y'^{(n)}$ from the $d-$dimensional product distribution Q, where $P,Q\in\{-1,1\}^d$ and $n=\frac{100\sqrt{d}}{\alpha^2}$

1: Calculate $\hat{X},\hat{X}',\hat{Y},\hat{Y}'$, which are mean vectors of X,X',Y,Y' respectively.
2: Define $Z_1 = \langle \hat{X}-\hat{Y}, \hat{X}'-\hat{Y}'\rangle$.
3: **if** $Z_1 \leq \frac{1}{2}\alpha^2$ **then**
4: return accept.
5: **else**
6: return reject.

▶ **Lemma 19.** *Assume we can draw samples from two d-dimensional product-Bernoulli distributions $P,Q\in\{0,1\}^d$. Given a distance parameter $\alpha>0$, Algorithm 1 is a sample-optimal algorithm which distinguishes between $P=Q$ and $\|\mu(P)-\mu(Q)\|_2>\alpha$ with probability at least $\frac{2}{3}$ using $O\left(\frac{\sqrt{d}}{\alpha^2}\right)$ samples.*

The proof can be found in the full version.

We then show that Algorithm 1, with some modification, can deal with two groups of samples under different differential privacy constraints.

▶ **Theorem 20.** *Assume we can draw samples with noise from two d-dimensional product distributions $P, Q \in \{0,1\}^d$, where each coordinate in samples from P, Q is flipped with probability $\frac{1}{e^{\varepsilon_1}+1}$, $\frac{1}{e^{\varepsilon_2}+1}$ respectively. Given privacy parameters $\varepsilon_1, \varepsilon_2 \in (0,1]$ and a distance parameter α, there exist an sample-optimal algorithm which uses $\frac{\sqrt{d}}{\alpha^2\varepsilon_1^2}$ samples from P and $\frac{\sqrt{d}}{\alpha^2\varepsilon_2^2}$ samples from Q, and distinguish between $\|\mu(P)-\mu(Q)\|_2 = 0$ and $\|\mu(P)-\mu(Q)\|_2 > \alpha$ with high probability.*

In the interest of space and clarity of exposition, the proof of this theorem is deferred to the full version.

Finally, we can claim that we have an sample-optimal algorithm for testing closeness of two distributions under heterogeneous local privacy constraints.

▶ **Theorem 21.** *For every $k \geq 0$ and $\varepsilon_1, \varepsilon_2 \in (0,1]$, there exist a private-coin protocol for (k,α)-closeness testing between two unknown distributions $\mathbf{p}$, $\mathbf{q}$ using $O\left(\frac{k^{3/2}}{\alpha^2\varepsilon_1^2}\right)$ samples from $\mathbf{p}$ and $O\left(\frac{k^{3/2}}{\alpha^2\varepsilon_2^2}\right)$ samples from $\mathbf{q}$, as this protocol is $(\varepsilon_1, 0)-$LDP for samples in $\mathbf{p}$ and $(\varepsilon_2, 0)-$LDP private for samples in $\mathbf{q}$ respectively.*

Proof. We claim Algorithm 2 below satisfies our demands. Its correctness directly follows from Theorems 18 and 20, Equation (1), , and Fact 12. ◀

Algorithm 2 Closeness testing under heterogeneous local differential privacy constraints.

Require: Privacy parameters $\varepsilon_1, \varepsilon_2$, a distance parameter α, n_1 users from unknown distribution $\mathbf{p}$ and n_2 users from unknown distribution $\mathbf{q}$.

1: Define $C_j = \{i \in [K] : H_{ij}^{(K)} = 1\}$, $j \in [K]$.
2: n_1 users from $\mathbf{p}$ and n_2 users from $\mathbf{q}$ are divided into K disjoint subgroups of equal size separately. Users in the jth group generate a bit of message 1 or 0 depending on whether their data is in the set C_J.
3: Users from $\mathbf{p}$ flip their one bit of message with probability $\frac{1}{e^{\varepsilon_1}+1}$, and users from $\mathbf{q}$ flip their one bit of message with probability $\frac{1}{e^{\varepsilon_2}+1}$. Then users send their message to the analyser.
4: For users from $\mathbf{p}$, by taking one user from each block and viewing the resulting collection of messages as a length-K binary vector, the analyser gets n_1/K independent samples of a product-Bernoulli distribution $P \in \{0,1\}^K$.
5: The analyser does the same thing for users from $\mathbf{q}$ and gets n_2/K independent samples of a product-Bernoulli distribution $Q \in \{0,1\}^K$.
6: The analyser calculates Z_2 as defined as follows:

$$Z_2 = \left\langle a \cdot (\hat{X} - \tfrac{1}{e^{\varepsilon_1}+1}) - b \cdot (\hat{Y} - \tfrac{1}{e^{\varepsilon_2}+1}),\ a \cdot (\hat{X}' - \tfrac{1}{e^{\varepsilon_1}+1}) - b \cdot (\hat{Y}' - \tfrac{1}{e^{\varepsilon_2}+1}) \right\rangle \tag{4}$$

where $a := \frac{e^{\varepsilon_1}+1}{e^{\varepsilon_1}-1}$, $b := \frac{e^{\varepsilon_2}+1}{e^{\varepsilon_2}-1}$.
7: **if** $Z_2 \leq \frac{1}{2}\alpha^2$ **then**
8: return accept.
9: **else**
10: return reject.

▶ Remark 22. The sample complexity of this algorithm matches the known lower bound of sample complexity of locally private identity testing, and thus this algorithm is sample-optimal.

3.1.2 Public-coin protocol

By combining the domain compression technique stated in Lemma 17 with Theorem 21, we are then able to obtain a sample-efficient public-coin algorithm:

▶ **Theorem 23.** *For every $k \geq 0$ and $\varepsilon_1, \varepsilon_2 \in (0, 1]$, there exist a public-coin protocol for (k, α)-closeness testing between two unknown distributions $\mathbf{p}$, $\mathbf{q}$ using $O\left(\frac{k}{\alpha^2 \varepsilon_1^2}\right)$ samples from $\mathbf{p}$ and $O\left(\frac{k}{\alpha^2 \varepsilon_2^2}\right)$ samples from $\mathbf{q}$, as this protocol is $(\varepsilon_1, 0)-$LDP for samples in $\mathbf{p}$ and $(\varepsilon_2, 0)-$LDP private for samples in $\mathbf{q}$ respectively.*

Proof. Recall that by using the domain compression, we are able to compress the size of domain to L while the total variation distance between any two distributions are preserved with high probability. Specifically, if we set the size of the compressed domain to be a constant $L = c_1$, we have a $(c_1, \sqrt{\frac{c_1 c_2}{k}}, \beta)$-domain compression such that for all $\mathbf{p}, \mathbf{q}$ with domain sizes k and total variation distance $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \alpha$, the mapping satisfies

$$\Pr[\mathrm{d}_{\mathrm{TV}}(\mathbf{p}^{\psi_U}, \mathbf{q}^{\psi_U}) \geq \sqrt{\frac{c_1 c_2}{k}}\alpha] > 1 - \beta.$$

Also, $\mathbf{p}^{\psi_U} = \mathbf{q}^{\psi_U}$ when $\mathbf{p} = \mathbf{q}$.

From Theorem 21 we know that there is an algorithm testing whether $\mathbf{p}^{\psi_U} = \mathbf{q}^{\psi_U}$ or $\Pr[\mathrm{d}_{\mathrm{TV}}(\mathbf{p}^{\psi_U}, \mathbf{q}^{\psi_U}) > \sqrt{\frac{c_1 c_2}{k}}\alpha]$ using $O\left(\frac{c_1^{3/2}}{\sqrt{\frac{c_1 c_2}{k}}^2 \alpha^2 \varepsilon_1^2}\right)$ samples from $\mathbf{p}$ and $O\left(\frac{c_1^{3/2}}{\sqrt{\frac{c_1 c_2}{k}}^2}\right)$ from $\mathbf{q}$ as this algorithm is $(\varepsilon_1, 0)-$LDP for samples in $\mathbf{p}$ and $(\varepsilon_2, 0)-$LDP private for samples in $\mathbf{q}$. By standard probability amplification techniques, we can decrease the error probability by a constant factor by increasing the number of samples by a constant factor, to achieve any desired constant error probability β_0. That implies the probability of this algorithm outputting the correct answer is at least $(1 - c_3)(1 - \beta_0)$. By choosing the right constants, the probability can be made larger than $\frac{2}{3}$. ◀

3.2 Under Shuffle Privacy

We now turn to our algorithms under the less stringent shuffle privacy model.

3.2.1 Private-coin protocol

For closeness testing under heterogeneous shuffle differential privacy constraints, we propose an algorithm based on Poisson mechanism, whose guarantees are stated in Lemma 10. (I.e., each user will add Poisson noise to their data to preserve their privacy in the shuffled model.) Since $1 - e^{-x} \geq x/2$ for any $x \in [0, 1]$, we only need to set $\mu = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$ when $\varepsilon \in [0, 1], \triangle \geq 1$, in this lemma.

We now claim their is a sample-efficient algorithm for closeness testing under heterogeneous shuffle differential privacy constraints, as stated in Theorem 24.

▶ **Theorem 24.** *For every $k \geq 0$, $\varepsilon_1, \varepsilon_2 \in (0, 1]$ (w.l.o.g. $\varepsilon_1 \geq \varepsilon_2$), and $\delta \in (0, 1]$, there exists a private-coin protocol for (k, α) closeness testing between two unknown distributions $\mathbf{p}$, $\mathbf{q}$ using*

$$n_1 = O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{k^{3/4}\sqrt{\log(1/\delta)}}{\alpha \varepsilon_1} + \min\left(\frac{\varepsilon_1^2 \varepsilon_2^2}{\alpha^4 \log^2(1/\delta)}, \frac{k^{2/3}}{\alpha^{4/3}}\left(\frac{\varepsilon_2}{\varepsilon_1}\right)^{2/3}\right)\right)$$

samples from $\mathbf{p}$ *and* $n_2 = (\frac{\varepsilon_1}{\varepsilon_2})^2 n_1$ *samples from* $\mathbf{q}$*; and this protocol is* $(\varepsilon_1, \delta)-$*shuffle differentially private for samples in* $\mathbf{p}$ *and* $(\varepsilon_2, \delta)-$*shuffle differentially private for samples in* $\mathbf{q}$ *respectively.*

▶ Remark 25. When $\varepsilon_1 = \varepsilon_2$ and ignoring δ, our algorithm retrieves the best known upper bound even for identity testing in the shuffle differential privacy model.

Due to space constraints, the proof of Theorem 24 is deferred to the full version.

3.2.2 Public-coin protocol

By leveraging again the domain compression technique of Lemma 17 with Theorem 24, we get the following public-coin sample complexity:

▶ **Theorem 26.** *For every* $k \geq 0$ *and* $\varepsilon_1, \varepsilon_2 \in (0, 1]$*, and* $\delta \in (0, 1]$*, there exist a private-coin protocol for* (k, α) *closeness testing between two unknown distributions* $\mathbf{p}$, $\mathbf{q}$ *using*

$$n_1 = O\left(\frac{\sqrt{k}}{\alpha^2} + \frac{k^{2/3}}{\alpha^{4/3}\varepsilon_1^{2/3}} \log^{1/3} \frac{1}{\delta} + \frac{\sqrt{k}}{\alpha\varepsilon_1}\sqrt{\log \frac{1}{\delta}}\right)$$

samples from $\mathbf{p}$ *and* $n_2 = O\left(\left(\frac{\varepsilon_1}{\varepsilon_2}\right)^2 n_1\right)$ *samples from* $\mathbf{q}$*, as this protocol is* $(\varepsilon_1, \delta)-$*shuffle differentially private for samples in* $\mathbf{p}$ *and* $(\varepsilon_2, \delta)-$*shuffle differentially private for samples in* $\mathbf{q}$ *respectively.*

Proof. We also use domain compression technique, and the procedure is the same as that in the proof of Theorem 23. The only thing we need to do is to choose an appropriate size of the compressed domain. I.e., we need to choose $2 \leq L \leq k$ such that the following is minimised:

$$n_1 \gtrsim \frac{\sqrt{L}}{\left(\frac{\sqrt{L}\alpha}{\sqrt{k}}\right)^2} + \frac{L^{3/4}\sqrt{\mu_1}}{\left(\frac{\sqrt{L}\alpha}{\sqrt{k}}\right)} + \min\left(\frac{1}{\left(\frac{\sqrt{L}\alpha}{\sqrt{k}}\right)^4 \mu_1\mu_2}, \frac{L^{2/3}}{\left(\frac{\sqrt{L}\alpha}{\sqrt{k}}\right)^{4/3}}\left(\frac{\mu_1}{\mu_2}\right)^{1/3}\right),$$

i.e.

$$n_1 \gtrsim \frac{k}{\sqrt{L}\alpha^2} + \frac{L^{1/4}\sqrt{k}\sqrt{\mu_1}}{\alpha} + \min\left(\frac{k^2}{L^2\alpha^4\mu_1\mu_2}, \frac{k^{2/3}}{\alpha^{4/3}}\left(\frac{\mu_1}{\mu_2}\right)^{1/3}\right).$$

If we set L to minimise the sum of the two terms, that is, $L := \min\left(k, \max\left(2, \frac{k^{2/3}}{\alpha^{4/3}\mu_1^{2/3}}\right)\right)$, we get

$$n_1 \gtrsim \frac{\sqrt{k}}{\alpha^2} + \frac{k^{2/3}}{\alpha^{4/3}}\mu_1^{1/3} + \frac{\sqrt{k}}{\alpha}\sqrt{\mu_1} + \min\left(\max\left(\frac{k^2}{\alpha^4\mu_1\mu_2}, \frac{1}{\alpha^4\mu_1\mu_2}, \frac{k^{2/3}}{\alpha^{4/3}\mu_1^{2/3}\mu_2}\right), \frac{k^{2/3}}{\alpha^{4/3}}\left(\frac{\mu_1}{\mu_2}\right)^{1/3}\right)$$

i.e.

$$n_1 \gtrsim \frac{\sqrt{k}}{\alpha^2} + \frac{k^{2/3}}{\alpha^{4/3}}\mu_1^{1/3} + \frac{\sqrt{k}}{\alpha}\sqrt{\mu_1} + \min\left(\frac{k^2}{\alpha^4\mu_1\mu_2}, \frac{k^{2/3}}{\alpha^{4/3}}\left(\frac{\mu_1}{\mu_2}\right)^{1/3}\right) \tag{5}$$

Since we have $\min\left(\frac{k^2}{\alpha^4\mu_1\mu_2}, \frac{k^{2/3}}{\alpha^{4/3}}\left(\frac{\mu_1}{\mu_2}\right)^{1/3}\right) \leq \frac{k^{2/3}}{\alpha^{4/3}}\left(\frac{\mu_1}{\mu_2}\right)^{1/3} \leq \frac{k^{2/3}}{\alpha^{4/3}}\mu_1^{1/3}$, we can remove the last term in Equation (5). Thus, the sample complexity is

$$n_1 \gtrsim \frac{\sqrt{k}}{\alpha^2} + \frac{k^{2/3}}{\alpha^{4/3}\varepsilon_1^{2/3}} \log^{1/3} \frac{1}{\delta} + \frac{\sqrt{k}}{\alpha\varepsilon_1}\sqrt{\log \frac{1}{\delta}},$$

$$n_2 = \left(\frac{\varepsilon_1}{\varepsilon_2}\right)^2 n_1.$$

This concludes the proof. ◀

3.3 Under Central Privacy?

We now recall our result for closeness testing in the heterogeneous central privacy model, Theorem 5.

▶ **Theorem 27** (Central Privacy, restated). *There exists an algorithm for closeness testing which guarantees ε_1-differential privacy to the n_1 users of the first group, and ε_2-differential privacy to the n_2 users of the second, with*

$$n_1 = O\left(\max\left(\frac{k^{1/2}}{\alpha^2}, \frac{k^{1/2}}{\varepsilon_1^{1/2}\alpha}, \frac{k^{2/3}}{\alpha^{4/3}}, \frac{k^{1/3}}{\varepsilon_1^{2/3}\alpha^{4/3}}, \frac{1}{\varepsilon_1\alpha}\right)\right)$$

and $n_2 = O\left(\frac{\varepsilon_1}{\varepsilon_2} n_1\right)$ (assuming without loss of generality that $\varepsilon_2 \leq \varepsilon_1$).

As outlined in Section 1.2, this follows straightforwardly by combining the result from [44] and privacy amplification by subsampling Lemma 11.

This begs the question of whether this simple approach can be improved upon. We discuss below some other natural approaches, and why they failed or did not pan out.

3.3.1 Second idea: find a test statistics with heterogeneous sensitivity

Designing a sample-efficient closeness testing algorithm under heterogeneous central differential privacy constraints is much more challenging than that under local and shuffle constraints. Our first attempt was to generalize previous work. As mentioned in the background section, a sample-optimal closeness testing algorithm under central differential privacy constraints was proposed in [44] by using the test statistic in [28]. To be more specific, that test statistic is as follows. Suppose we take two sets of n i.i.d. samples from both $\mathbf{p}$ and $\mathbf{q}$, and let X_i, X_i', Y_i, Y_i' be the number of occurrences of i in those four sets respectively for $i \in [k]$ where k is the size of the domain of $\mathbf{p}$, $\mathbf{q}$. Then the test statistic Z is defined as

$$Z := |X_i - Y_i| + |X_i' - Y_i'| - |X_i - X_i'| - |Y_i - Y_i'| \tag{6}$$

In [44], the algorithm shifts Z to get a new test statistic

$$Z' = (Z - C_1\sqrt{n} - \frac{C_2}{\varepsilon})/2 \tag{7}$$

Then, it uses a sigmoid function to map $\varepsilon Z'$ to $(0, 1)$ and then draws a Bernoulli random variable using this value as a parameter. Finally, the algorithm outputs accept or reject depending on whether the value of this random variable is 1 or 0. Roughly speaking, the idea is that Z will either be close to 0 or greater than $\sqrt{n}$, depending on whether $\mathbf{p} = \mathbf{q}$ or $\mathrm{d}_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) > \alpha$; and therefore after shifting, Z' will be either $< -C_2/\varepsilon$ or $> C_2/\varepsilon$ (with high probability). Using the sigmoid function on $\varepsilon Z'$ maps this to a Bernoulli with bias either $1/2 - \Omega(1)$ or $1/2 + \Omega(1)$, which allows to distinguish the two cases while satisfying ε-DP (since changing one sample will change the value of Z by at most 2, as we will see below, and thus of $\varepsilon Z'$ by at most ε.

If we view Z as a function of $\mathbf{X}, \mathbf{X}', \mathbf{Y}, \mathbf{Y}'$ respectively, the ℓ_1-sensitivity of Z must be smaller than or equal to 2. To see why, let us w.l.o.g. consider one sample in the first set taken from $\mathbf{p}$. The change of that sample will only add 1 to X_i and decrease 1 from X_j for some $i, j \in [k]$ where $i \neq j$. Thus, Z will only by increased by 1 or decreased by 1 when that sample changes. By using a similar statement, one can show that the sensitivity of $\varepsilon Z'$ is not larger than ε. Combined with the fact $e^{-|\gamma|} \leq \sigma(x + \gamma)\sigma(x) \leq e^{|\gamma|}$, the author of [44] showed the algorithm is $(\varepsilon, 0)$-differential private in the central model.

Now, we consider how to generalize their algorithm. One natural idea is to adapt the test statistic to differentiate its sensitivity for samples from $\mathbf{p}$ and samples from $\mathbf{q}$. An example would be the following statistic:

$$Z = |\frac{X_i}{n} - \frac{Y_i}{m}| + |\frac{X_i'}{n} - \frac{Y_i'}{m}| - |\frac{X_i}{n} - \frac{X_i'}{n}| - |\frac{Y_i}{m} - \frac{Y_i'}{m}|. \tag{8}$$

However, that test static is not easy to analyze as previous approaches do not naturally extend to this. Specifically, in [28], the analysis of the test static defined in Equation (6) relied on two things: (1) the samples are drawn using the "Poissonisation trick" (X_i, X_i', Y_i, Y_i' are Poisson random variables), and (2) clever use of the additive property of Poisson distributions. Thus, we cannot use their proof technique for our proposed test statistic as a scaled Poisson (such as X_i/n) is no longer Poisson, and the different scalings prevent their approach from going through. To overcome that difficulty, we found a simpler and more general method to analyze the static defined in Equation (6). It relies on an identity (Zolotarev identity) relating the expectation of the absolute value of any random variable to the integral of its characteristic function:

$$\mathbb{E}|X| = \frac{2}{\pi}\int_0^\infty \frac{1 - \Re(\mathbb{E}e^{itX})}{t^2}\,dt$$

We have written the complete analysis in [23]. However, we were still not able to find an appropriate tester even after finding the new analysis method. However, even this new analysis did not allow us to establish the desired properties of the testers we considered (such as the one in Equation (8)), and it is unclear whether they would actually work. We conjecture so, and proving this would be an interesting (and non-trivial) future direction.

3.3.2 Third idea: using a different privatizing method

In [44], the closeness testing algorithm firstly calculates the test statistic using samples drawn from $\mathbf{p}$ and $\mathbf{q}$, then privatizes the test static. Instead, we want to privatize the histograms drawn from $\mathbf{p}$ and $\mathbf{q}$ respectively, and then use the privatized histogram to calculate a test statistic.

However, we do not want to use a continuous noise such as the Laplace noise. If we use one of those continuous distributions for the noise, then analysing the statistic will become very hard. (E.g. the distribution of sums of Poisson/multinomial and Laplace can be very complicated.) For example, we can consider using the Skellam mechanism. (A Skellam distribution is the same as the distribution of the difference between two Poisson random variables.) The correctness of the Skellam mechanism directly follows from the correctness of the Poisson mechanism.

▶ **Lemma 28** (Skellam noise). *Let $f\colon X^n \to \mathcal{Z}$ be a $\triangle$-sensitive function. For any $\varepsilon > 0, \delta \in (0,1)$ and $\lambda \geq \frac{16\log(10/\delta)}{(1-e^{-\varepsilon/\triangle})^2} + \frac{2\triangle}{1-e^{-\varepsilon/\triangle}}$, the randomized function $\mathcal{A}(X^n) = f(x) + Y_1 - Y_2$ where $Y_1, Y_2 \sim \mathrm{Poisson}(\lambda)$ is $(\varepsilon,\delta)-$differentially private in the shuffle model.*

Proof. Since the Poisson mechanism is correct, the randomized function f is already $(\varepsilon,\delta)-$differentially private after adding Y_1. Then it should still be $(\varepsilon,\delta)-$differentially private after adding Y_2 due to the immunity of post-processing of differential privacy. ◀

Assuming we are dealing with homogeneous privacy constraints and want our output to be $(\varepsilon,\delta)-$differentially private, we would set the parameter μ of the Skellam mechanism to be $\Theta(\log(1/\delta)/\varepsilon^2)$. With some transformation, our test statistic could be written in this form:

$$\tilde{Z} = \sum_{i=1}^{k} (|X_i + a_i - Y_i - c_i| + |X_i' + b_i - Y_i' - d_i| - |X_i + a_i' - X_i' - b_i'| - |Y_i + c_i' - Y_i' - d_i'|)$$

where $a_1, b_i, c_i, d_i, a_1', b_i', c_i', d_i' \sim \text{Poisson}(2\mu)$ due to the additivity property of the Poisson distribution. By using the additivity property of the Poisson distribution again, it could be rewritten as

$$\tilde{Z} = \sum_{i=1}^{k} (|A_i - B_i| + |A_i' - B_i'| - |A_i - A_i'| - |B_i - B_i'|)$$

where $A_i, A_i' \sim \text{Poisson}(n\mathbf{p}_i + 2\mu)$, $B_i, B_i' \sim \text{Poisson}(n\mathbf{q}_i + 2\mu)$. Since $n\mathbf{p}_i + 2\mu = (n + 2k\mu)\frac{n\mathbf{p}_i+2\mu}{n+2k\mu}$ and $\sum_{i=1}^{k} \frac{n\mathbf{p}_i+2\mu}{n+2k\mu} = 1$, using this test statistic is then equivalent to using the test statistic in [28] by taking $n + 2k\mu$ samples from another distribution $\mathbf{p}'$ and $n + 2k\mu$ samples from another distribution $\mathbf{q}'$, where $\mathbf{p}_i' = \frac{n\mathbf{p}_i+2\mu}{n+2k\mu}, \mathbf{q}_i' = \frac{n\mathbf{q}_i+2\mu}{n+2k\mu}$. Then, directly from the analysis in $\frac{np_i+2\mu}{n+2k\mu}$ we have

$$\mathbb{E}[\tilde{Z}]^2 \gtrsim \frac{n^3\alpha^4}{k(1 + k\mu/n)}$$

To bound the variance of the test statistic, we will use the Efron–Stein inequality. Our test statistic could be seen as the sum of $12k$ independent Poisson random variables. For each $i \in [k]$, there are 2 Poisson($n\mathbf{p}_i$) random variables, 2 Poisson($n\mathbf{q}_i$) random variables, 8 Poisson(2μ) random variables, and they are independent of each other. Then by using the Efron–Stein inequality, we get

$$\text{Var}(\tilde{Z}) \leq \frac{1}{2} \sum_{i=1}^{12k} \mathbb{E}\left[(\tilde{Z} - \tilde{Z}_i')^2\right]$$

where the ith independent random variable in $\tilde{Z}$ is different from that in $\tilde{Z}_i'$, and other pairs of random variables are the same, respectively. This allows us to bound the variance as

$$\text{Var}(\tilde{Z}) \leq \frac{1}{2} \sum_{i=1}^{k} 2 \cdot 2np_i + 2 \cdot 2nq_i + 8 \cdot 2 \cdot 2\mu = O(n + k\mu).$$

Then, for our tester to work, we need the "signal" to be larger than the "noise", i.e., we want $\mathbb{E}[\tilde{Z}]^2 \gg \text{Var}(\tilde{Z})$. When $n \geq k\mu$, this implies $\frac{n^3\alpha^4}{k} \gtrsim n$, and thus $n = \Omega\left(\frac{k^{1/2}}{\alpha^2}\right)$. Otherwise, we have $\frac{n^3\alpha^4}{k(1+k\mu/n)} \gtrsim k\mu$, and thus $n^3 \gtrsim \frac{k^3\mu}{n\alpha^4} \gtrsim \frac{k^3}{n\varepsilon^3\alpha^4}$, which implies $n = \Omega\left(\frac{k^{3/2}}{\varepsilon^{3/2}\alpha^2}\right)$. Combining the results, we get the overall requirement that, for this tester to work, n must be $\Omega\left(\frac{k^{1/2}}{\alpha^2} + \frac{k^{3/4}}{\varepsilon\alpha}\sqrt{\log(1/\delta)}\right)$.

This bound is not optimal, but it is not too bad. However, it cannot be generalized to the heterogeneous setting because the privatizing procedure is, in fact, not separated (because it can be viewed as adding noise to the difference of histograms taken from two distributions). We also considered other methods of privatizing but were not able to develop an algorithm with good sample complexity due to the time frame of this thesis.

3.3.3 Fourth idea: use "flattening samples" and "testing samples"

In [29], the authors proposed a new approach for sample-optimally closeness testing. They firstly use some "flattening samples" from one distribution to "flatten" the domain in order to decrease the expected value of ℓ_2-norm. Then, they use a standard ℓ_2 tester to achieve ℓ_1 closeness testing. It is natural to develop a close testing algorithm using different numbers of samples from two distributions, which is proposed in that paper. However, it is hard to make that algorithm differential private because the sensitivity of the "flattening samples" can be prohibitively large (as large as the number of samples).

It is worth noting that the authors of [9] developed a method for making the algorithm in [29] deferentially private. They considered all permutations of "flattening samples" and "testing samples" in the calculation of the test statistic to decrease the sensitivity of "flattening samples." However, we were not able to generalize their proof to our needs, and it is unclear whether this technique could yield a sample-efficient tester for our problem.

4 Future Work

Our results raise several future directions. First, it is not clear how to achieve pure privacy for the shuffle model, which is interesting since pure privacy is a stricter privacy guarantee. Second, while we are able to obtain sample-optimal algorithms for the local model tight bounds for the central and shuffle model remain unknown. Third, since our algorithms only work for the high privacy regime where $\varepsilon \in (0, 1]$, it will be interesting to determine algorithms for the low privacy regime. Note that a low-privacy algorithm for the local model should directly lead to a high-privacy algorithm for the shuffle model, *via* amplification by shuffling. Last, it would be interesting to consider a mixed privacy guarantee, e.g., where one group of users works a local model, while the other group relies on the shuffle model.

References

1 Jayadev Acharya, Clément L. Canonne, Cody Freitag, Ziteng Sun, and Himanshu Tyagi. Inference under information constraints III: local privacy constraints. *IEEE J. Sel. Areas Inf. Theory*, 2(1):253–267, 2021.

2 Jayadev Acharya, Clément L. Canonne, Yanjun Han, Ziteng Sun, and Himanshu Tyagi. Domain compression and its application to randomness-optimal distributed goodness-of-fit. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, volume 125 of *Proceedings of Machine Learning Research*, pages 3–40. PMLR, 2020. URL: `http://proceedings.mlr.press/v125/acharya20a.html`.

3 Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints I: Lower bounds from chi-square contraction. *IEEE Trans. Inform. Theory*, 66(12):7835–7855, 2020. Preprint available at arXiv:abs/1812.11476. `doi:10.1109/TIT.2020.3028440`.

4 Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Inference under information constraints II: Communication constraints and shared randomness. *IEEE Transactions on Information Theory*, 2020. In press. Preprint available at arXiv:abs/1804.06952. `doi:10.1109/TIT.2020.3028439`.

5 Jayadev Acharya and Ziteng Sun. Communication complexity in locally private distribution estimation and heavy hitters. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 51–60. PMLR, 2019. URL: `http://proceedings.mlr.press/v97/acharya19c.html`.

6 Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private testing of identity and closeness of discrete distributions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6878–6891. Curran Associates, Inc., 2018. URL: http://papers.nips.cc/paper/7920-differentially-private-testing-of-identity-and-closeness-of-discrete-distributions.pdf.

7 Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 1120–1129. PMLR, 2019. URL: http://proceedings.mlr.press/v89/acharya19a.html.

8 Mohammad Alaggan, Sébastien Gambs, and Anne-Marie Kermarrec. Heterogeneous differential privacy. *J. Priv. Confidentiality*, 7(2), 2016. doi:10.29012/jpc.v7i2.652.

9 Maryam Aliakbarpour, Ilias Diakonikolas, Daniel Kane, and Ronitt Rubinfeld. Private testing of distributions via sample permutations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 10878–10889. Curran Associates, Inc., 2019. URL: http://papers.nips.cc/paper/9270-private-testing-of-distributions-via-sample-permutations.pdf.

10 Maryam Aliakbarpour, Ilias Diakonikolas, and Ronitt Rubinfeld. Differentially private identity and equivalence testing of discrete distributions. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 169–178, Stockholmsmässan, Stockholm Sweden, 10–15 July 2018. PMLR. URL: http://proceedings.mlr.press/v80/aliakbarpour18a.html.

11 Kareem Amin, Matthew Joseph, and Jieming Mao. Pan-private uniformity testing. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 183–218. PMLR, 09–12 July 2020. URL: http://proceedings.mlr.press/v125/amin20a.html.

12 Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics*, 12(2):727–749, 2018. doi:10.1214/18-AOAS1155SF.

13 Victor Balcer, Albert Cheu, Matthew Joseph, and Jieming Mao. Connecting robust shuffle privacy and pan-privacy. *CoRR*, abs/2004.09481, 2020. arXiv:2004.09481.

14 Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *NeurIPS*, pages 6280–6290, 2018.

15 Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *41st Annual Symposium on Foundations of Computer Science, FOCS 2000*, pages 189–197, 2000.

16 Andrea Bittau, Úlfar Erlingsson, Petros Maniatis, Ilya Mironov, Ananth Raghunathan, David Lie, Mitch Rudominer, Ushasree Kode, Julien Tinnés, and Bernhard Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *SOSP*, pages 441–459. ACM, 2017.

17 Clément L. Canonne. A Survey on Distribution Testing: your data is Big. But is it Blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, April 2015. URL: http://eccc.hpi-web.de/report/2015/063.

18 Clément L. Canonne. *A Survey on Distribution Testing: Your Data is Big. But is it Blue?* Number 9 in Graduate Surveys. Theory of Computing Library, 2020. doi:10.4086/toc.gs.2020.009.

19 Clément L. Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Found. Trends Commun. Inf. Theory*, 19(6):1032–1198, 2022. Also available at https://ccanonne.github.io/survey-topics-dt.html. doi:10.1561/0100000114.

20 Clément L. Canonne, Xi Chen, Gautam Kamath, Amit Levi, and Erik Waingarten. Random restrictions of high dimensional distributions and uniformity testing with subcube conditioning. In Dániel Marx, editor, *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms, SODA 2021, Virtual Conference, January 10 - 13, 2021*, pages 321–336. SIAM, 2021. doi:10.1137/1.9781611976465.21.

21 Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing bayesian networks. *IEEE Trans. Inf. Theory*, 66(5):3132–3170, 2020. doi:10.1109/TIT.2020.2971625.

22 Clément L. Canonne and Hongyi Lyu. Uniformity testing in the shuffle model: Simpler, better, faster. In *SOSA*, pages 182–202. SIAM, 2022.

23 Clément L. Canonne and Yucheng Sun. Optimal closeness testing of discrete distributions made (complex) simple. *CoRR*, abs/2204.12640, 2022.

24 Clément L. Canonne and Yucheng Sun. Private distribution testing with heterogeneous constraints: Your epsilon might not be mine. *CoRR*, abs/2309.06068, 2023. doi:10.48550/arXiv.2309.06068.

25 Siu-on Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of SODA*, pages 1193–1203, 2014.

26 Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *Advances in cryptology—EUROCRYPT 2019. Part I*, volume 11476 of *Lecture Notes in Comput. Sci.*, pages 375–403. Springer, Cham, 2019. doi:10.1007/978-3-030-17653-2_13.

27 Albert Cheu and Chao Yan. Pure differential privacy from secure intermediaries. *CoRR*, abs/2112.10032, 2021.

28 Ilias Diakonikolas, Themis Gouleakis, Daniel M. Kane, John Peebles, and Eric Price. Optimal testing of discrete distributions with high probability. In *STOC '21—Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 542–555. ACM, New York, 2021. doi:10.1145/3406325.3450997.

29 Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *57th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2016*. IEEE Computer Society, 2016.

30 John C. Duchi and Martin J. Wainwright. Distance-based and continuum Fano inequalities with applications to statistical estimation. *arXiv*, abs/1311.2669, 2013. arXiv:1311.2669.

31 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography*, volume 3876 of *Lecture Notes in Comput. Sci.*, pages 265–284. Springer, Berlin, 2006.

32 Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: from local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, Philadelphia, PA, 2019. doi:10.1137/1.9781611975482.151.

33 Alireza Fallah, Ali Makhdoumi, Azarakhsh Malekian, and Asuman E. Ozdaglar. Optimal and differentially private data acquisition: Central and local mechanisms. In *EC*, page 1141. ACM, 2022.

34 Vitaly Feldman, Audra McMillan, and Kunal Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *FOCS*, pages 954–964. IEEE, 2021.

35 Badih Ghazi, Ravi Kumar, Pasin Manurangsi, and Rasmus Pagh. Private counting from anonymous messages: Near-optimal accuracy with vanishing communication overhead. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3505–3514. PMLR, 2020. URL: http://proceedings.mlr.press/v119/ghazi20a.html.

36 Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017. URL: http://www.wisdom.weizmann.ac.il/~oded/pt-intro.html.

37 Zach Jorgensen, Ting Yu, and Graham Cormode. Conservative or liberal? personalized differential privacy. In Johannes Gehrke, Wolfgang Lehner, Kyuseok Shim, Sang Kyun Cha, and Guy M. Lohman, editors, *31st IEEE International Conference on Data Engineering, ICDE 2015, Seoul, South Korea, April 13-17, 2015*, pages 1023–1034. IEEE Computer Society, 2015. doi:10.1109/ICDE.2015.7113353.

38 Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826, 2011.

39 Ninghui Li, Wahbeh H. Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In Heung Youl Youm and Yoojae Won, editors, *7th ACM Symposium on Information, Compuer and Communications Security, ASIACCS '12, Seoul, Korea, May 2-4, 2012*, pages 32–33. ACM, 2012. doi:10.1145/2414456.2414474.

40 Ben Niu, Yahong Chen, Boyang Wang, Zhibo Wang, Fenghua Li, and Jin Cao. Adapdp: Adaptive personalized differential privacy. In *INFOCOM*, pages 1–10. IEEE, 2021.

41 Ronitt Rubinfeld. Taming big probability distributions. *XRDS: Crossroads, The ACM Magazine for Students*, 19(1):24, September 2012. doi:10.1145/2331042.2331052.

42 Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.

43 Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965. doi:10.1080/01621459.1965.10480775.

44 Huanyu Zhang. *Statistical Inference in the Differential Privacy Model*. PhD thesis, Cornell University, 2021.