Local Private Hypothesis Testing: Chi-Square Tests

Marco Gaboardi\textsuperscript{1} and Ryan Rogers\textsuperscript{2}

\textsuperscript{1}University at Buffalo, The State University of New York
\textsuperscript{2}University of Pennsylvania

September 22, 2017

Abstract

The local model for differential privacy is emerging as the reference model for practical applications collecting and sharing sensitive information while satisfying strong privacy guarantees. In the local model, there is no trusted entity which is allowed to have each individual’s raw data as is assumed in the traditional curator model for differential privacy. So, individuals’ data are usually perturbed before sharing them.

We explore the design of private hypothesis tests in the local model, where each data entry is perturbed to ensure the privacy of each participant. Specifically, we analyze locally private chi-square tests for goodness of fit and independence testing, which have been studied in the traditional, curator model for differential privacy.

1 Introduction

Hypothesis testing is a widely applied statistical tool used to test whether given models should be rejected or not based on sampled data. Hypothesis testing was initially developed for scientific and survey data, but today it is also an essential tool to test models over collections of social network, mobile, and crowdsourced data (American Statistical Association, 2014; Hunter et al., 2008; Steele et al., 2017). Collected data samples may contain highly sensitive information about the subjects, and the privacy of individuals can be compromised when the results of a data analysis are released. In work from Homer et al. (2008), it was shown that a subject in a dataset can be identified as being in the case or control group based on the aggregate statistics of a genetic-wide association study (GWAS). Privacy-risks may bring data contributors to opt out, which reduces the confidence in the data study. A way to address this concern is by developing new techniques to support privacy-preserving data analysis. Among the different approaches, differential privacy (Dwork et al., 2006b) is emerging as a viable solution: it provides strong privacy guarantees, and it allows to release accurate statistics. A standard way to achieve differential privacy is by injecting some statistical noise in the computation of the data analysis. When the noise is carefully chosen, it helps to protect the individual privacy without compromising the utility of the data analysis.

Several recent works have studied differentially private hypothesis tests that can be used in place of the standard, non-private hypothesis tests (Uhler et al., 2013; Yu et al., 2014; Sheffet, 2015; Karwa and Slavkovic, 2016; Wang et al., 2015; Gaboardi et al., 2016; Kifer and Rogers, 2017; Cai et al., 2017). These tests work in the curator model of differential privacy. In this model, the data is centrally stored and the curator carefully injects noise in the computation of the data analysis in order to meet the requirement of differential privacy.

In this work we address instead the local model of privacy, formally introduced by Raskhodnikova et al. (2008). The first differentially private algorithm called randomized response – in fact it predates the definition of differential privacy by more than 40 years – guarantees differential privacy in the local model (Warner, 1965). In this model, there is no trusted centralized entity which is responsible for the noise injection. Instead, each
individual adds enough noise to guarantee differential privacy for their own data, which provides a stronger privacy guarantee when compared to traditional differential privacy. The data analysis is then run over the collection of the individually sanitized data. The local model of differential privacy is a convenient model for several applications; for example it is used to collect statistics about the activity of the Google Chrome Web browser users (Erlingsson et al., 2014), and to collect statistics about the typing patterns of Apple’s iPhone users (Apple Press Info, 2016). These are applications which are naturally approached using the local model, rather than the centralized curator model. Despite these applications, the local model has received far less attention than the more standard centralized curator model. This is in part due to the more firm requirements imposed by this model, which make the design of effective data analysis harder.

Our main contribution is in designing chi-square hypothesis tests for the local model of differential privacy. Similar to previous works we focus on designing goodness of fit and independence hypothesis tests – other chi-square hypothesis test can be designed following similar approaches from prior work from Kifer and Rogers (2017). Each test we present is characterized by a specific mechanism used to guarantee local differential privacy. We present three different goodness of fit tests: LocalGaussGOF guarantees local (concentrated) differential privacy by adding Gaussian noise to the individual data; LocalExpGOF selects an approximate private value for each individual by using the exponential mechanism (McSherry and Talwar, 2007); LocalBitFlipGOF uses bit flipping to provide better power for higher dimensional data (Bassily and Smith, 2015). Further, we develop corresponding independence tests: LocalGaussIND, LocalExpIND, and LocalBitFlipIND.

For all these tests we study their asymptotic behavior. A desiderata for private hypothesis tests is to have a guaranteed upper bound on the probability of a false discovery (or Type I error) – rejecting a null hypothesis or model when the data was actually generated from it, and to minimize the probability of a Type II error, which is failing to reject the null hypothesis when the model is indeed false. This latter criteria corresponds to the power of the statistical test. We then present experimental results showing the power of the different tests which demonstrates that no single local differentially private algorithm is best across all data dimensions and privacy parameter regimes.

2 Related Works

There have been several works in developing private hypothesis test for categorical data, but all look at the traditional model of (concentrated) differential privacy instead of the local model, which we consider here. Motivated by the attack from Homer et al. (2008), there are several works that deal with private statistical inference with GWAS data, (Uhler et al., 2013, Yu et al., 2014, Johnson and Shmatikov, 2013). Following these works, there has also been general work in private chi-square hypothesis tests, where the main tests are for goodness of fit and independence testing, although some do extend to more general tests (Wang et al., 2015, Gaboardi et al., 2016, Kifer and Rogers, 2017, Cai et al., 2017, Kakizaki et al., 2017). There has also been work in private hypothesis testing for ordinary least squares regression (Sheffet, 2015).

There are other works that have studied statistical inference and estimators in the local model of differential privacy. Duchi et al. (2013b) focus on controlling disclosure risk in statistical estimation and inference by ensuring the analysis satisfies local differential privacy. They provide minimax convergence rates to show the tight tradeoffs between privacy and statistical efficiency, i.e. the number of samples required to give quality estimators. One of the main contributions of their work is that a generalized version of Warner’s randomized response (Warner, 1965) gives optimal sample complexity for estimating the multinomial probability vector, which we use in the hypothesis test LocalBitFlipGOF.

Kairouz et al. (2014) also considers hypothesis testing subject to local differential privacy, although they measure utility with the Kullback-Leibler divergence and does not give a decision rule, i.e. when to reject a given null hypothesis. We provide statistics whose distributions asymptotically follow a chi-square distribution.

---

1 Although in the experiments we use Laplace noise (rather than Gaussian) to ensure all our tests are locally differentially private, which requires us to use a Monte Carlo approach to estimate the statistic’s distribution.
which allows for approximating statistical p-values that can be used in a decision rule. We consider the extremal locally differentially private mechanisms given in [Kairouz et al., 2014] and empirically confirm their result that for small privacy regimes (small \( \epsilon \)) one mechanism has higher utility than other mechanisms and for large privacy regimes (large \( \epsilon \)) a different mechanism outperforms the others. However, we measure utility in terms of the power of a locally private hypothesis test subject to a given Type I error bound. Other related and notable works in the local privacy model include [Pastore and Gastpar, 2016; Kairouz et al., 2016; Ye and Barg, 2017].

3 Preliminaries

We will consider datasets \( x = (x_1, \ldots, x_n) \in X^n \) in some data universe \( X \), typically \( X = \{0,1\}^d \) where \( d \) is the dimensionality. We will first present the standard definition of differential privacy, as well as its recent variant concentrated differential privacy. We say that two datasets \( x, x' \in X^n \) are neighboring if they differ in at most one element, i.e. there exists some \( i \in [n] \) such that \( x_i \neq x'_i \) but for all \( j \neq i, x_j = x'_j \).

**Definition 3.1** [Dwork et al., 2006]. An algorithm \( M : X^n \rightarrow Y \) is \((\epsilon, \delta)\)-differentially private (DP) if for all neighboring datasets \( x, x' \in X^n \) and for all outcomes \( S \subseteq Y \), we have

\[
\Pr[M(x) \in S] \leq e^{\epsilon} \Pr[M(x') \in S] + \delta.
\]

We then present the definition of zero-mean concentrated differential privacy.

**Definition 3.2** [Bun and Steinke, 2016]. An algorithm \( M : X^n \rightarrow Y \) is \( \rho \)-zero-mean concentrated differentially private (zCDP) if for all neighboring datasets \( x, x' \in X^n \), we have for all \( t > 0 \)

\[
\mathbb{E}_{y \sim M(x)} \left[ \exp \left( t \ln \left( \frac{\Pr[M(x) = y]}{\Pr[M(x') = y]} \right) \right) \right] \leq \exp \left( t^2 \rho \right).
\]

Note that in both of these privacy definitions, it is assumed that all the data is stored in a central location and the algorithm \( M \) can access all the data. Most of the work in differential privacy has been in this trusted curator model. One of the main reasons for this is that we can achieve much greater accuracy in our differentially private statistics when used in the curator setting. However, in many cases, having a trusted curator is too strong of an assumption. We then define local differential privacy, formalized by [Raskhodnikova et al., 2008] and [Dwork and Roth, 2014], which does not require the subjects to release their raw data, rather each data entry is perturbed to prevent the true entry from being stored. Thus, local differential privacy ensures a very strong privacy guarantee.

**Definition 3.3** [LR Oracle]. Given a dataset \( x \), a local randomizer oracle \( LR_x(\cdot, \cdot) \) takes as input an index \( i \in [n] \) and an \( \epsilon \)-DP algorithm \( R \), and outputs \( y \in Y \) chosen according to the distribution of \( R(x_i) \), i.e. \( LR_x(i, R) = R(x_i) \).

**Definition 3.4** [Raskhodnikova et al., 2008]. An algorithm \( M : X^n \rightarrow Y \) is \( (\epsilon, \delta) \)-local differentially private (LDP) if it accesses the input database \( x \) via the LR oracle \( LR_x \) with the following restriction: if \( LR(i, R_j) \) for \( j \in [k] \) are the \( M \)'s invocations of \( LR_x \) on index \( i \), then each \( R_j \) for \( j \in [k] \) is \( (\epsilon_j, \delta_j) \)-DP and \( \sum_{j=1}^k \epsilon_j \leq \epsilon \), \( \sum_{j=1}^k \delta_j \leq \delta \).

An easy consequence of this definition is that an algorithm which is \( (\epsilon, \delta) \)-LDP is also \( (\epsilon, \delta) \)-DP. Note that we can easily extend these definitions to include \( \rho \)-local zCDP (LzCDP) where each local randomizer is \( \rho \)-zCDP and \( \sum_{j=1}^k \rho_j \leq \rho \). We point out the following connection between LzCDP and LDP, which follows directly from results in [Bun and Steinke, 2016].

**Lemma 3.35.** If \( M : X^n \rightarrow Y \) is \((\epsilon, 0)\)-LDP then it is also \( \epsilon^2/2 \)-LzCDP. If \( M \) is \( \rho \)-LzCDP, then it is also \( \left( \rho + \sqrt{2 \rho \ln(2/\delta)} , \delta \right) \)-LDP for any \( \delta > 0 \).

Thus, in the local setting, (pure) LDP (where \( \delta = 0 \)) provides the strongest level of privacy, followed by LzCDP and then approximate-LDP (where \( \delta > 0 \)).
4 Chi-Square Hypothesis Tests

As was studied in Gaboardi et al. (2016), Wang et al. (2015), and Kifer and Rogers (2017), we will study hypothesis tests with categorical data. A null hypothesis, or model $H_0$ is how we might expect the data to be generated. The goal for hypothesis testing is to reject the null hypothesis if the data is not likely to have been generated from the given model. As is common in statistical inference, we want to design hypothesis tests to bound the probability of a false discovery (or Type I error), i.e. rejecting a null hypothesis when the data was actually generated from it, by at most some amount $\alpha$, such as 5%. However, designing tests that achieve this is easy, because we can just ignore the data and always "fail to reject" the null hypothesis, i.e. have an inconclusive test. Thus, we would additionally like to design our tests so that they can reject $H_0$ if the data was not actually generated from the given model. We then want to minimize the probability of a Type II error, which is failing to reject $H_0$ when the model is false.

For goodness of fit testing, we assume that each individual’s data $X_i$ for $i \in [n]$ is sampled i.i.d. from Multinomial$(1, p)$ where $p \in \mathbb{R}_{>0}^d$ and $p^\top \cdot 1 = 1$. The classical chi-square hypothesis test (without privacy) forms the histogram $H = (H_1, \cdots, H_d) = \sum_{i=1}^n X_i$ and computes the chi-square statistic

$$T = \sum_{j=1}^d \frac{(H_j - np_j^0)^2}{np_j^0}.$$  

The reason for using this statistic is that it converges in distribution to $\chi^2_{d-1}$ as $n \to \infty$ when $H_0 : p = p^0$ holds. Hence, we can ensure the probability of false discovery to be close to $\alpha$ as long as we only reject $H_0$ when $T > \chi^2_{d-1,1-\alpha}$ where the critical value $\chi^2_{d-1,1-\alpha}$ is defined as the following quantity $\Pr \left[ \chi^2_{d-1} > \chi^2_{d-1,1-\alpha} \right] = \alpha$.

4.1 Prior Private Chi-square Tests in the Curator Model

We now present some existing chi-square private hypothesis tests which will be useful in designing local private tests. One approach, as explored by Gaboardi et al. (2016) and Wang et al. (2015) is to add noise (Gaussian or Laplace) directly to the histogram to ensure privacy and then use the classical test statistic. Note that the resulting asymptotic distribution needs to be modified for such changes to the statistic – it is no longer a chi-square random variable. To introduce the different statistics, we will consider goodness of fit testing after adding Gaussian noise to the histogram of counts $\tilde{H} = H + N(0, \frac{1}{\rho}I_d)$, which ensures $\rho$-zCDP – prior works also considers other types of noise, e.g. Laplace. The chi-square statistic then becomes

$$\tilde{T}(\rho) = \sum_{i=1}^d \frac{(H_i + N(0, 1/\rho) - np_i^0)^2}{np_i^0}.$$  

They then show that this statistic converges in distribution to a linear combination of chi-squared random variables, when $\rho$ is also decreasing with $n$.

In followup work from Kifer and Rogers (2017), they showed that modifying the chi-square statistic to account for the additional noise leads to tests with better empirical power. The projected statistic from Kifer and Rogers (2017) is the following where we use projection matrix $\Pi \overset{\text{def}}{=} (I_d - \frac{1}{d}11^\top)$

$$T_{KR}^{(n)}(\rho) = n \left( \frac{H + N(0, 1/\rho I_d)}{n} - p^0 \right)^\top \Pi \left( \text{Diag} \left(p^0 + \frac{1}{n\rho} \right) - p^0 \left(p^0\right)^\top \right)^{-1} \Pi \left( \frac{H + N(0, 1/\rho I_d)}{n} - p^0 \right).$$  

When comparing the power of our tests, we will be considering the alternate $H_1 : p = p_n^1$, where

$$p_n^1 = p^0 + \frac{\Delta}{\sqrt{n}}$$  

where $1^\top \Delta = 0$.  

4
Theorem 4.1 [Kifer and Rogers (2017)]. Under the null hypothesis $H_0: p = p^0$, the statistic $T^{(n)}_{KR}(\rho)$ given in (3) for $\rho > 0$ converges in distribution to $\chi^2_{d-1}$. Further, under the alternate hypothesis $H_1: p = p_1^n$ given in (4), the resulting asymptotic distribution is a noncentral chi-square random variable,

$$T^{(n)}_{KR}(\rho/n) \xrightarrow{D} \chi^2_{d-1} \left( \Delta^T \left( \text{Diag}(p^0) - p^0 (p^0)^T + 1/\rho I_d \right) \right)^{-1} \Delta$$

Note that the projection matrix $\Pi = (I_d - \frac{1}{d} 11^T)$ in (3) is what reduces the degrees of freedom from $d$ to $d - 1$ and gives the improvement in power from the original statistic.

5 Local Private Chi-Square Goodness of Fit Tests

We now turn to designing local private goodness of fit tests. We begin by showing how the existing statistics from the previous section can be used in the local setting and then develop new tests based on the exponential mechanism (McSherry and Talwar, 2007) and bit flipping (Bassily and Smith, 2015). Each test is locally private because it perturbs each individual’s data. Note that we did not design these local private algorithms. The novelty of this work is designing statistics based on data from these algorithms whose asymptotic distribution is chi-squared and this allows us to correctly bound the Type I error. We finish the section by empirically checking the power of each test to see which tests outperform others in different parameter regimes. We empirically show that the power of a test is directly related to the size of the noncentral parameter of the chi-square statistic under the alternate distribution.

5.1 Goodness of Fit Test with Noise Addition

In the local model we can add $Z_i \sim N \left( 0, \frac{1}{\rho} I_d \right)$ independent noise to each individual’s data $X_i$ to ensure $\rho$-LzCDP or $Z_i \sim \text{Lap} \left( \frac{2}{\rho} \right)$ independent noise to $X_i$ to ensure $\epsilon$-LDP. In either case, the resulting noisy histogram $\hat{H} = H + Z$ where $Z = \sum_i Z_i$ will have variance that scales with $n$ for fixed privacy parameters $\epsilon, \rho > 0$. Consider the case where we add Gaussian noise, which results in the following histogram, $\hat{H} = H + Z$ where $Z \sim N \left( 0, \frac{2}{\rho} I_d \right)$. Thus, we can use either statistic $\tilde{T}(\rho/n)$ or $T^{(n)}_{KR}(\rho/n)$, with the latter statistic typically having better empirical power (Kifer and Rogers, 2017). We then give our first local private hypothesis test in Algorithm 1.

Algorithm 1 LzCDP GOF Test with Gaussian Mechanism: LocalGaussGOF

Input: $x = (x_1, \cdots, x_n)$, $\rho$, $\alpha$, $H_0: p = p^0$.
Let $H = \sum_{\ell=1}^{n} x_\ell$
Set $q = T^{(n)}_{KR}(\rho/n)$ given in (3).
if $q > \chi^2_{d-1, -\alpha}$ then
    Decision $\leftarrow$ Reject.
else
    Decision $\leftarrow$ Fail to Reject.
Output: Decision

Theorem 5.1. The test LocalGaussGOF($\cdot, \rho, \alpha, p^0$) is $\rho$-LzCDP.

Proof. The proof follows from the fact that we are adding appropriately scaled noise to each individual’s data via the Gaussian mechanism and then LocalGaussGOF aggregates the privatized data, which is just a post-processing function on the privatized data.
Although we cannot guarantee the probability of a Type I error at most $\alpha$ due to the fact that we use the asymptotic distribution (as in the tests from prior work and the classical chi-square tests without privacy), we expect the Type I errors to be similar to those from the nonprivate test.

Note that the test can be modified to accommodate arbitrary noise distributions, e.g., Laplace to ensure differential privacy as was done in [Kifer and Rogers 2017]. In this case, we can use a Monte Carlo (MC) approach to estimate the critical value $\tau$ that ensures the probability of a Type I error is at most $\alpha$ if we reject $H_0$ when the statistic is larger than $\tau$. For the local setting, if each individual perturbs each coordinate by adding $\text{Lap}(2/\epsilon)$ then this will ensure our test is $\epsilon$-LDP. However, the sum of independent Laplace random variables is not Laplace, so we will need to estimate a sum of $n$ independent Laplace random variables using MC. In the experiments section we will compare the other local private tests with the version of LocalGaussGOF which uses Laplace noise and samples $m$ entries from the exact distribution under $H_0$ to find the critical value. Such a test guarantees that the probability of false discovery is at most $\alpha$ as long as $m > \lceil 1/\alpha \rceil$ (see Gaboardi et al. 2016 for more details).

5.2 Goodness of Fit Test with the Exponential Mechanism

Rather than having to add noise to each component of the original data histogram, we consider applying randomized response to obtain a LDP hypothesis test. We will use a form of the exponential mechanism (McSherry and Talwar 2007) given in Algorithm 2, which takes a single data entry from the set $\{e_1, \ldots, e_d\}$, where $e_j \in \mathbb{R}^d$ is the standard basis element with a 1 in the $j$th coordinate and is zero elsewhere, and reports the original entry with probability slightly more than uniform and otherwise reports a different element. Note that $M_{\text{EXP}}$ takes a single data entry and is $\epsilon$-differentially private.

Algorithm 2 Exponential Mechanism: $M_{\text{EXP}}$

Input: Data $x \in \{e_1, \ldots, e_d\}$, privacy parameter $\epsilon$.

Let $q(x, z) = \mathbb{I}\{x = z\}$

Select $\hat{x}$ with probability $\frac{\exp q(x, \hat{x})}{e^{\epsilon} + 1 + d}$

Output: $\hat{x}$

We have the following result when we use $M_{\text{EXP}}$ on each data entry to obtain a private histogram.

**Lemma 5.2.** If we have histogram $H = \sum_{i=1}^{n} X_i$, where $\{X_i\}_{i \sim d}$ Multinomial(1, $p$) and we write $H = \sum_{i=1}^{n} M_{\text{EXP}}(X_i, \epsilon)$ for each $i \in [n]$, then

$$H \sim \text{Multinomial}(n, \hat{p}) \quad \text{where} \quad \hat{p} = p - \frac{e^{\epsilon}}{e^{\epsilon} + d - 1} + (1 - p) \left( \frac{1}{e^{\epsilon} + d - 1} \right).$$

(5)

We can then form a chi-square statistic using the private histogram $\hat{H}$.

**Theorem 5.3.** Let $H \sim \text{Multinomial}(n, p)$ and $\hat{H}$ be given in Lemma 5.2 with privacy parameter $\epsilon > 0$. Under the null hypothesis $H_0 : p = p^0$, we have for $\hat{p} = \frac{1}{e^{\epsilon} + d - 1} \left( e^{\epsilon} p^0 + (1 - p^0) \right)$, and

$$T_{\text{Exp}}^{(n)}(\epsilon) \overset{D}{\rightarrow} \chi_{d-1}^2 \sum_{j=1}^{d} \frac{\Delta_j^2}{\hat{p}_j^2}.$$  

Further, with alternate $H_1 : p = p_1$ given in (4), the resulting asymptotic distribution is the following,

$$T_{\text{Exp}}^{(n)}(\epsilon) \overset{D}{\rightarrow} \chi_{d-1}^2 \left( \frac{e^{\epsilon} - 1}{e^{\epsilon} + d - 1} \right)^2 \sum_{j=1}^{d} \frac{\Delta_j^2}{\hat{p}_j^2}.$$  

We point out that $M_{\text{EXP}}$ is $\epsilon$-differentially private, whereas the traditional exponential mechanism tells us that it is 2$\epsilon$-differentially private. The savings in the factor of 2 is the result of the normalizing constant not being affected by the input data.
Proof. This follows the same analysis as in the classical result for the asymptotic distribution of the statistic in \cite{1} (see Ferguson \cite{1996} for a reference). To compute the noncentral parameter, we note that the expected difference between $\hat{H}$ and the null hypothesis $\hat{p}^0$ is $\left(\frac{e^{1/d} - 1}{e^{1/d} + 1}\right)^2 \Delta$. Thus, the noncentral parameter is $\left(\frac{e^{1/d} - 1}{e^{1/d} + 1}\right)^2 \Delta^2 \Delta$

We then base our LDP goodness of fit test on this result to obtain the correct threshold (or critical value) to reject the null hypothesis based on a chi-square distribution. The test is presented in Algorithm 3

Algorithm 3 Local DP GOF Test with Exponential Mechanism: LocalExpGOF

Input: $x = (x_1, \cdots, x_n)$, $\epsilon$, $\alpha$, $H_0 : p = p^0$.
Let $\hat{p}^0 = \frac{1}{n} \left(\frac{e^{1/d} - 1}{e^{1/d} + 1}\right) (e^{i\epsilon} p^0 + (1 - p^0))$.
Let $H = \sum_{i=1}^n M_{\text{Exp}} (x_i, \epsilon)$.
Set $q = \sum_{j=1}^d \left(\frac{h_i - np_i^0}{np_i^0}\right)^2$ if $q > \chi^2_{d-1, 1-\alpha}$ then
Decision $\leftarrow$ Reject.
else
Decision $\leftarrow$ Fail to Reject.
Output: Decision

Theorem 5.4. The test LocalExpGOF($\cdot, \epsilon, \alpha, p^0$) is $\epsilon$-LDP.

Proof. The proof follows from the fact that we use $M_{\text{Exp}}$ for each individual’s data and then LocalExpGOF aggregates the privatized data, which is just a post-processing function.

5.3 Goodness of Fit Test with Bit Flipping

Note that the noncentral parameter in Theorem 5.3 goes to zero as $d$ grows large due to the coefficient being $\left(\frac{e^{1/d} - 1}{e^{1/d} + 1}\right)^2$. Thus, for larger dimensional data the exponential mechanism cannot reject a false null hypothesis.

We next consider another differentially private algorithm $M : \{e_1, \cdots, e_d\} \rightarrow \{0, 1\}^d$, given in Algorithm 4 used in \cite{Bassily and Smith (2015)} that flips each bit with some biased probability.

Algorithm 4 Local DP GOF Test with Bit Flipping: $M_{\text{bit}}$

Input: Data $x \in \{e_1, \cdots, e_d\}$, privacy parameter $\epsilon$.
for $j \in [d]$ do
Set $z_j = x_j$ with probability $\frac{e^{2/\epsilon}}{e^{2/\epsilon} + 1}$, otherwise $z_j = (1 - x_j)$.
Output: $z$

Theorem 5.5. The algorithm $M_{\text{bit}} : \{e_1, \cdots, e_d\} \rightarrow \{0, 1\}^d$ is $\epsilon$-DP.

We then want to form a statistic based on the output $z \in \{0, 1\}^d$ that is asymptotically distributed as a chi-square under the null hypothesis. We defer the proof to the appendix.

\footnote{Special thanks to Adam Smith for recommending to use this particular algorithm.}
The histogram $\tilde{T}$ is distributed as a chi-square matrix $\Sigma$.

Further, we can then design a hypothesis test based on the outputs from Algorithm 5.

We now show that the statistic in (8) is asymptotically distributed as $\chi^2_d$.

Proof. The proof follows from the fact that we are using Algorithm 4 which is $\epsilon$-DP on each person’s data, and then LocalBitFlipGOF aggregates the resulting output vector.

We now show that the statistic in (8) is asymptotically distributed as $\chi^2_d$.

Theorem 5.8. If the null hypothesis $H_0 : p = p^0$ holds, then the statistic $T_{\text{BitFlip}}^{(n)}(\epsilon)$ is asymptotically distributed as a chi-square $T_{\text{BitFlip}}^{(n)}(\epsilon) \xrightarrow{D} \chi^2_d$. Further, if we consider the alternate hypothesis $H_1 : p = p^1$ given in (4), then $T_{\text{BitFlip}}^{(n)}(\epsilon) \xrightarrow{D} \frac{\chi^2_d}{\epsilon^2 + 1} \cdot \Delta^\top \Sigma(p^0)^{-1} \Delta$.

Proof. We first note that $\Sigma(p^0)$ has eigenvector 1 with eigenvalue $\frac{\epsilon^2}{\epsilon^2 + 1}$. Hence, we can project out this eigenvector with $\Pi$, as was done for the projected statistic in Kifer and Rogers (2017) to get that the stated statistic is asymptotically distributed as $\chi^2_d$. Further the noncentral parameter under the alternate hypothesis $H_1$ follows the same analysis in Kifer and Rogers (2017).

Algorithm 5 Local DP GOF Test with Bit Flipping: LocalBitFlipGOF

Input: $x = (x_1, \ldots, x_n)$, $\epsilon$, $\alpha$, $H_0 : p = p^0$.

Let $\tilde{H} = \sum_{i=1}^n M_{\text{bit}}(x_i, \epsilon)$.

Set $q = T_{\text{BitFlip}}^{(n)}(\epsilon)$

if $q > \chi^2_{d-1,1-\alpha}$ then
  Decision $\leftarrow$ Reject.
else
  Decision $\leftarrow$ Fail to Reject.

Output: Decision.

Theorem 5.7. The test LocalBitFlipGOF($\cdot, \epsilon, \alpha, p^0$) is $\epsilon$-LDP.
We now compare the noncentral parameters of the three local private tests we presented in Algorithms [1][3] and [4]. We consider the null hypothesis $p^0 = (1/d, \cdots, 1/d)$ for $d > 2$, and alternate $H_1: p = p^0 + \Delta$, as in [4]. In this case, we can easily compare the various noncentral parameters for various privacy parameters and dimensions $d$. In Figure 1, we give the coefficient to the term $\Delta^T\Delta$ in the noncentral parameter of the asymptotic distribution for each local private test presented thus far. The larger this coefficient is, the better the power will be for any alternate $\Delta$ vector. Note that in LocalGaussGOF, we set $\rho = \epsilon^2/8$ which makes the variance the same as for a random variable distributed as $\text{Lap}(2/\epsilon)$ for an $\epsilon$-DP guarantee – recall that LocalGaussGOF does not satisfy $\epsilon$-DP for any $\epsilon > 0$. We give results for $\epsilon \in \{1, 2, 3, 4\}$ which are all in the range of privacy parameters that have been considered in actual locally differentially private algorithms used in practice. From the plots, we see how LocalExpGOF may outperform LocalBitFlipGOF depending on the privacy parameter and dimension of the data. We can use these plots to determine which test to use given a privacy budget $\epsilon$ and dimension of data $d$. When $H_0$ is not uniform, we can use the noncentral parameters given for each test to determine which test has the largest noncentral parameter for a particular privacy budget $\epsilon$.

5.5 Empirical Results

We then empirically compare the power between LocalGaussGOF in Algorithm [1][3], LocalExpGOF in Algorithm [4], and LocalBitFlipGOF in Algorithm [5]. Note that rather than use Gaussian noise in Algorithm [1][3] we use Laplace noise with scale $2/\epsilon$, and we then use an MC approach where we sample $m$ points from the exact distribution of the statistic under $H_0$. We will denote the resulting test as LocalLapGOF, This will ensure that all the tests we use are $\epsilon$-LDP. In our experiments we fix $m = 999$, $\alpha = 0.05$ and $\epsilon \in \{1, 2, 4\}$. We then consider null hypotheses of the form $p^0 = (1/d, 1/d, \cdots, 1/d)$ and alternate $H_1: p = p^0 + \eta(1, -1, \cdots, 1, -1)$ for some $\eta > 0$. In Figure 2 we plot the number of times our tests correctly rejects the null hypothesis in 1000 independent trials for various sample sizes $n$ and privacy parameters $\epsilon$.

From Figure 2 we can see that the test statistics that have the largest noncentral parameter for a particular dimension $d$ and privacy parameter $\epsilon$ will have the best empirical power. When $d = 4$, we see that LocalExpGOF performs the best. However, for $d = 40$ it is not so clear cut. When $\epsilon = 4$, we can see that LocalExpGOF does the best, but then when $\epsilon = 2$, LocalBitFlipGOF does best. Thus, the best Local DP Goodness of Fit test depends on the noncentral parameter, which is a function of $\epsilon$, the null hypothesis $p^0$, and alternate $p = p^0 + \Delta$. Note that the worst local DP test also depends on the privacy parameter and the dimension $d$.

Based on our empirical results, we see that no single locally private test is best for all data dimensions. Knowing the corresponding noncentral parameter for a given problem is useful in determining which tests to

\footnote{In [Erlingsson et al. (2014)], we know that Google uses $\epsilon = \ln(3) \approx 1.1$ in RAPPOR. We are also aware of a Twitter post by Aleksandra Korolova’s that Apple uses $\epsilon = 1, 4$.}
What makes this test difficult is that the analyst does not know the data distribution. We consider two multinomial random variables \( \pi^{(1)} \) and \( \pi^{(2)} \) for \( \pi^{(1)} \in \mathbb{R}^r \) and \( \pi^{(2)} \in \mathbb{R}^c \) and no component of \( \pi^{(1)} \) or \( \pi^{(2)} \) is zero. Without loss of generality, we will consider an individual to be in one of \( r \) groups who reports a data record that is in one of \( c \) categories. The collected data consists of \( n \) joint outcomes \( H \) whose \((i, j)\)th coordinate is \( H_{i,j} = \sum_{\ell=1}^{n} 1 \{ U_{\ell,i} = 1 \ \& \ V_{\ell,j} = 1 \} \). Note that \( H \) is then the contingency table over the joint outcomes.

Under the null hypothesis of independence between \( \{ U_{\ell} \}_{\ell=1}^{n} \) and \( \{ V_{\ell} \}_{\ell=1}^{n} \), we have

\[
H \sim \text{Multinomial} \left( n, p(\pi^{(1)}, \pi^{(2)}) \right) \quad \text{where} \quad p(\pi^{(1)}, \pi^{(2)}) = \pi^{(1)} \left( \pi^{(2)} \right)^\top
\]

What makes this test difficult is that the analyst does not know the data distribution \( p(\pi^{(1)}, \pi^{(2)}) \) and so cannot simply plug it into the chi-square statistic. Rather, we use the data to estimate the best guess for the unknown probability distribution that satisfies the null hypothesis.

Note that without privacy, each individual \( \ell \in [n] \) is reporting a \( r \times c \) matrix \( X_{\ell} \) which would be 1 in exactly one location. Thus we can alternatively write the contingency table as \( H = \sum_{\ell=1}^{n} X_{\ell} \). We then use the three local private algorithms we presented earlier for noise addition, exponential mechanism, and bit flipping to see how we can form a private chi-square statistic for independence testing. We want to be able to ensure the privacy of both the group and the category that each individual belongs to.
6.1 Independence Test with Noise Addition

For noise addition, we will have the following contingency table $H + Z$ (treating $H$ as an $r \times c$ vector) where $Z = N(0, \frac{2}{\epsilon})$ for $\rho$-LzCDP or $Z$ is the sum of $n$ independent Laplace random vectors with scale parameter $2/\epsilon$ in each coordinate for $\epsilon$-LDP, although we will only consider the Gaussian noise case. We can then use the same statistic presented in Kifer and Rogers (2017) with a rescaling of the variance in the noise and flattening the contingency table as a vector. For convenience, we write the marginals as $H_{i,} = \sum_{j=1}^{c} H_{i,j}$ and similarly for $H_{,j}$. Further, we will write $\hat{n} = n + \sum_{i,j} Z_{i,j}$ in the following test statistic.

Let $H_{i,} = \sum_{j=1}^{c} H_{i,j}$ and similarly for $H_{,j}$. Further, we will write $\hat{n} = n + \sum_{i,j} Z_{i,j}$ in the following test statistic.

$$\hat{\pi}^{(1)} = \left( \frac{1}{\hat{n}} (H_{i,} + Z_{i,}) : i \in [r] \right), \quad \hat{\pi}^{(2)} = \left( \frac{1}{\hat{n}} (H_{,j} + Z_{,j}) : j \in [c] \right)$$

$$\hat{M} = \text{Diag} \left( \hat{p}(\hat{\pi}^{(1)}, \hat{\pi}^{(2)}) \right) - \hat{p}(\hat{\pi}^{(1)}, \hat{\pi}^{(2)}) \left( \hat{p}(\hat{\pi}^{(1)}, \hat{\pi}^{(2)}) \right)^{\top} + \frac{1}{\rho} I_{rc}$$

$$\hat{T}_{KR}^{(n)} \left( \rho/n; (\hat{\pi}^{(1)}, \hat{\pi}^{(2)}) \right) = \frac{1}{\hat{n}} \left( H + Z - np \left( \hat{\pi}^{(1)}, \hat{\pi}^{(2)} \right) \right)^{\top} \Pi \left( \hat{M} \right)^{-1} \Pi \left( H + Z - np \left( \hat{\pi}^{(1)}, \hat{\pi}^{(2)} \right) \right)$$ (10)

We use this statistic because it is asymptotically distributed as a chi-square random variable given the null hypothesis holds, which follows directly from the general chi-square theory presented in Kifer and Rogers (2017).

**Theorem 6.1.** Under the null hypothesis that $U$ and $V$ are independent, then we have as $n \to \infty$

$$\min_{\pi^{(1)}, \pi^{(2)}} \left\{ \hat{T}_{KR}^{(n)} \left( \rho/n; (\pi^{(1)}, \pi^{(2)}) \right) \right\} \xrightarrow{D} \chi_{(r-1)(c-1)}^{2}.$$

We present the test in Algorithm 6 for Gaussian noise which uses the statistic in (10).

**Algorithm 6** LzCDP Independence Test with Gaussian Mechanism: LocalGaussIND

**Input:** $(x_1, \ldots, x_n), \rho, \alpha$.

Let $H = \sum_{\ell=1}^{n} x_{\ell}$

Set $q = \min_{\pi^{(1)}, \pi^{(2)}} \left\{ \hat{T}_{KR}^{(n)} \left( \rho/n; (\pi^{(1)}, \pi^{(2)}) \right) \right\}$ given in (10).

if $q > \chi_{(r-1)(c-1),1-\alpha}^{2}$ then

Decision ← Reject.

else

Decision ← Fail to Reject.

**Output:** Decision

We then have the following result which follows from the privacy analysis from before.

**Theorem 6.2.** LocalGaussIND is $\rho$-LzCDP.

6.2 Independence Test with the Exponential Mechanism

Next we want to design an independence test when the data is generated from $M_{\text{EXP}}$ given in Algorithm 2. In this case our contingency table can be written as $H \sim \text{Multinomial}(n, \hat{p}(\pi^{(1)}, \pi^{(2)}))$ where we use both [4] and [5] to get

$$\hat{p}(\pi^{(1)}, \pi^{(2)}) = p(\pi^{(1)}, \pi^{(2)}) \left( \frac{e^{\epsilon}}{e^{\epsilon} + rc - 1} \right) + \left( 1 - p(\pi^{(1)}, \pi^{(2)}) \right) \left( \frac{1}{e^{\epsilon} + rc - 1} \right)$$ (11)

We will follow a common rule of thumb for small sample sizes, so that if $n\hat{\pi}^{(1)} (\hat{\pi}^{(2)})^{\top} \leq 5$ for any cell, then we simply fail to reject $H_0$. 

11
We then obtain an estimate for the unknown parameters by minimizing over \( \pi^{(1)}, \pi^{(2)} \) the chi-square statistic in (6) where we replace \( \hat{P}^0 \) with \( \hat{P}(\pi^{(1)}, \pi^{(2)}) \). Thus, the resulting statistic becomes,

\[
\hat{T}^{(n)}(\epsilon) = \sum_{i,j} \left( \frac{\hat{H}_{i,j} - n\hat{p}_{i,j}(\pi^{(1)}, \pi^{(2)})}{n\hat{p}_{i,j}(\pi^{(1)}, \pi^{(2)})} \right)^2
\]

(12)

We can then use classical results from minimum chi-square theory (Ferguson, 1996) or recent results from Kifer and Rogers (2017) to prove the following

\textbf{Theorem 6.3.} Under the null hypothesis that \( U \) and \( V \) are independent, then we have as \( n \to \infty \)

\[
\hat{T}^{(n)}(\epsilon) \overset{D}{\to} \chi^2_{(r-1)(c-1)}.
\]

We then use this result to design our private chi-square test for independence.

\textbf{Algorithm 7} LDP Independence Test with Exponential Mechanism: \( \text{LocalExpIND} \)

\begin{itemize}
  \item **Input:** \( x = (x_1, \ldots, x_n) \), \( \epsilon, \alpha \), \( H_0 : \hat{p} = \hat{p}^0 \).
  \item Let \( \hat{p}(\pi^{(1)}, \pi^{(2)}) = \frac{1}{e^{\epsilon^2/2} + 1} \left( e^{\epsilon^2/2} + 1 \right) p(\pi^{(1)}, \pi^{(2)}) + (1 - p(\pi^{(1)}, \pi^{(2)})) \).
  \item Let \( \hat{H} = \sum_{i=1}^{n} \mathcal{M}_{\text{Exp}}(x_i, \epsilon) \).
  \item Set \( q = \hat{T}^{(n)}(\epsilon) \) from (12)
  \item if \( q > \chi^2_{(r-1)(c-1)} \) then
    \item Decision \( \leftarrow \) Reject.
  \item else
    \item Decision \( \leftarrow \) Fail to Reject.
\end{itemize}

\textbf{Output:} Decision

We then have the following result which again follows from the privacy analysis from before.

\textbf{Theorem 6.4.} \( \text{LocalExpIND} \) is \( \epsilon \)-LDP.

### 6.3 Independence Test with Bit Flipping

Lastly, we design an independence test when the data is reported via \( \mathcal{M}_{\text{bit}} \) in Algorithm 4. Assuming that \( H = \sum_{\ell=1}^{n} X_\ell \sim \text{Multinomial} (n, p(\pi^{(1)}, \pi^{(2)})) \), then we know that replacing \( \hat{p}^0 \) with \( p(\pi^{(1)}, \pi^{(2)}) \) in Section 5.3 gives us the following for the covariance matrix \( \Sigma(\cdot) \) given in (7)

\[
\sqrt{n} \left( \frac{\hat{H}}{n} - \left[ \frac{(e^{\epsilon^2/2} - 1)}{e^{\epsilon^2/2} + 1} p(\pi^{(1)}, \pi^{(2)}) + \frac{1}{e^{\epsilon^2/2} + 1} \right] \right) \overset{D}{\to} \mathcal{N} \left( 0, \Sigma(\hat{p}(\pi^{(1)}, \pi^{(2)})) \right)
\]

(13)

\[\text{Again, we will follow a common rule of thumb for small sample sizes, so that if } n\pi^{(1)}(\pi^{(2)})^\top \leq 5 \text{ for any cell, then we simply fail to reject } H_0.\]
Following the analysis of [Kifer and Rogers 2017], we start with a rough estimate for the unknown parameters which converges in probability to the true estimates, so we use:

\[
\bar{\pi}(1) = \left( e^{\epsilon^2/2} + 1 \right) \left( \frac{H_i}{n} - \frac{c}{e^{\epsilon^2/2} + 1} : i \in [r] \right)
\]

\[
\bar{\pi}(2) = \left( e^{\epsilon^2/2} + 1 \right) \left( \frac{H_j}{n} - \frac{r}{e^{\epsilon^2/2} + 1} : j \in [c] \right).
\]

(14)

We then give the resulting statistic, parameterized by the unknown parameters \( \bar{\pi}(\ell) \), for \( \ell \in \{1, 2\} \).

\[
\tilde{M} = \left( e^{\epsilon^2/2} - 1 \right)^2 \left[ \text{Diag} \left( p \left( \bar{\pi}(1), \bar{\pi}(2) \right) \right) - p \left( \bar{\pi}(1), \bar{\pi}(2) \right) \left( p \left( \bar{\pi}(1), \bar{\pi}(2) \right) \right)^T \right] + \frac{e^{\epsilon^2/2}}{(e^{\epsilon^2/2} + 1)^2} I_{rc}
\]

\[
\tilde{T}_{\text{BitFlip}}^{(n)} \left( \epsilon; (\bar{\pi}(1), \bar{\pi}(2)) \right) = \frac{1}{n} \left( \tilde{H} - np \left( \bar{\pi}(1), \bar{\pi}(2) \right) \right)^T \left( \tilde{M} \right)^{-1} \left( \tilde{H} - np \left( \bar{\pi}(1), \bar{\pi}(2) \right) \right)
\]

(15)

Again, using the results from [Kifer and Rogers 2017], we know that minimizing \( \tilde{T}_{\text{BitFlip}}^{(n)} \left( \epsilon; (\bar{\pi}(1), \bar{\pi}(2)) \right) \) over \( (\bar{\pi}(1), \bar{\pi}(2)) \) results in a statistic that is distributed as a chi-square random variable.

**Theorem 6.5.** Under the null hypothesis that \( U \) and \( V \) are independent, then we have as \( n \to \infty \)

\[
\min_{\bar{\pi}(1), \bar{\pi}(2)} \left\{ \tilde{T}_{\text{BitFlip}}^{(n)} \left( \epsilon; (\bar{\pi}(1), \bar{\pi}(2)) \right) \right\} \xrightarrow{D} \chi^2_{(r-1)(c-1)}.
\]

We present the test in Algorithm 8.

**Algorithm 8** LDP Independence Test with Bit Flipping: LocalBitFlipIND

**Input:** \( (x_1, \cdots, x_n), \epsilon, \alpha. \)

Let \( \tilde{H} = \sum_{i=1}^{n} \mathcal{M}_{\text{bit}}(x_i, \epsilon). \)

Set \( q = \min_{\bar{\pi}(1), \bar{\pi}(2)} \left\{ \tilde{T}_{\text{BitFlip}}^{(n)} \left( \epsilon; (\bar{\pi}(1), \bar{\pi}(2)) \right) \right\} \) given in (15).

if \( q > \chi^2_{(r-1)(c-1), 1-\alpha} \) then

Decision \( \leftarrow \) Reject.

else

Decision \( \leftarrow \) Fail to Reject.

**Output:** Decision

We then have the following result which follows from the privacy analysis from before.

**Theorem 6.6.** LocalBitFlipIND is \( \epsilon \)-LDP.

7 Conclusion

We have designed several hypothesis tests, each depending on different local differentially private algorithms: LocalGaussGOF (LocalLapGOF), LocalExpGOF, and LocalBitFlipGOF as well as their corresponding independence tests LocalGaussIND, LocalExpIND, and LocalBitFlipIND. This required constructing different statistics so that the resulting distribution after injecting noise into the data in order to satisfy privacy

7Once again, we will follow a common rule of thumb for small sample sizes, so that if \( n \bar{\pi}(1) \left( \bar{\pi}(2) \right)^T \leq 5 \) for any cell, then we simply fail to reject \( H_0 \).
could be closely approximated with a chi-square distribution. Hence, we designed rules for when a null hypothesis $H_0$ should be rejected while satisfying some bound $\alpha$ on Type I error. Further, we showed that each statistic has a noncentral chi-square distribution when the data is drawn from some alternate hypothesis $H_1$. Depending on the form of the alternate probability distribution, the dimension of the data, and the privacy parameter, either LocalExpGOF or LocalBitFlipGOF gave better power. This corroborates the results from Kairouz et al. (2014) who showed that in hypothesis testing, different privacy regimes have different optimal local differentially private mechanisms, although utility in their work was in terms of KL divergence. Our results show that the power of the test is directly related to the noncentral parameter of the test statistic that is used. This requires the data analyst to carefully consider alternate hypotheses, as well as the data dimension and privacy parameter for a particular test and then see which test statistic results in the largest noncentral parameter.

We focused primarily on goodness of fit testing, where the null hypothesis is a single probability distribution. We further developed local private independence tests which resulted in using previous general chi-square theory presented in Kifer and Rogers (2017). This basic framework can be used to develop other chi-square hypothesis tests where the null hypothesis is not a single parameter. We hope that this will lead to future work on designing local differentially private hypothesis tests beyond chi-square testing.
References

American Statistical Association. Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society, 2014. URL http://www.amstat.org/policy/pdfs/BigDataStatisticsJune2014.pdf

Apple Press Info. Apple previews ios 10, the biggest ios release ever, 2016. URL https://www.apple.com/pr/library/2016/06/13Apple-Previews-iOS-10-The-Biggest-iOS-Release-Ever.html

Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing, STOC ’15, pages 127–135, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3536-2. doi: 10.1145/2746539.2746632. URL http://doi.acm.org/10.1145/2746539.2746632

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In Theory of Cryptography - 14th International Conference, TCC 2016-B, Beijing, China, October 31 - November 3, 2016, Proceedings, Part I, pages 635–658, 2016. doi: 10.1007/978-3-662-53641-4_24. URL http://dx.doi.org/10.1007/978-3-662-53641-4_24

Bryan Cai, Constantinos Daskalakis, and Gautam Kamath. Priv’it: Private and sample efficient identity testing. In International Conference on Machine Learning, 2017.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In 54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA, pages 429–438, 2013a. doi: 10.1109/FOCS.2013.53. URL https://doi.org/10.1109/FOCS.2013.53

John C. Duchi, Martin J. Wainwright, and Michael I. Jordan. Local privacy and minimax bounds: Sharp rates for probability estimation. In Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States., pages 1529–1537, 2013b. URL http://papers.nips.cc/paper/5013-local-privacy-and-minimax-bounds-sharp-rates-for-probability-estimation

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3–4):211–407, 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In Advances in Cryptology - EUROCRYPT 2006, 25th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28 - June 1, 2006, Proceedings, pages 486–503, 2006a. doi: 10.1007/11761679_29.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In In Proceedings of the 3rd Theory of Cryptography Conference, pages 265–284. Springer, 2006b.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, CCS ’14, pages 1054–1067, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2957-6. doi: 10.1145/2660267.2660348. URL http://doi.acm.org/10.1145/2660267.2660348

T.S. Ferguson. A Course in Large Sample Theory. Chapman & Hall Texts in Statistical Science Series. Taylor & Francis, 1996. ISBN 9780412043710. URL https://books.google.com/books?id=DDh_DiTz9agC

Marco Gaboardi, Hyun Woo Lim, Ryan Rogers, and Salil P. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML’16, pages 2111–2120. JMLR.org, 2016. URL http://dl.acm.org/citation.cfm?id=3045390.3045613
Nils Homer, Szabolcs Szelingo, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density sup genotyping microarrays. *PLoS Genet.*, 4(8), 08 2008.

David R. Hunter, Steven M. Goodreau, and Mark S. Handcock. Goodness of fit of social network models. *Journal of the American Statistical Association*, 103:248–258, 2008. URL [http://EconPapers.repec.org/RePEc:bes:jnlasa:v:103:y:2008:m:march:p:248-258](http://EconPapers.repec.org/RePEc:bes:jnlasa:v:103:y:2008:m:march:p:248-258).

Aaron Johnson and Vitaly Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’13, pages 1079–1087, New York, NY, USA, 2013. ACM.

Peter Kairouz, Sewoong Oh, and Pranod Viswanath. Extremal mechanisms for local differential privacy. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2879–2887. Curran Associates, Inc., 2014. URL [http://papers.nips.cc/paper/5392-extremal-mechanisms-for-local-differential-privacy.pdf](http://papers.nips.cc/paper/5392-extremal-mechanisms-for-local-differential-privacy.pdf).

Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016*, New York City, NY, USA, June 19-24, 2016, pages 2436–2444, 2016. URL [http://jmlr.org/proceedings/papers/v48/kairouz16.html](http://jmlr.org/proceedings/papers/v48/kairouz16.html).

Kazuya Kakizaki, Kazuto Fukuchi, and Jun Sakuma. Differentially private chi-squared test by unit circle mechanism. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1761–1770, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL [http://proceedings.mlr.press/v70/kakizaki17a.html](http://proceedings.mlr.press/v70/kakizaki17a.html).

Vishesh Karwa and Aleksandra Slavković. Inference using noisy degrees: Differentially private β-model and synthetic graphs. *Ann. Statist.*, 44(P1):87–112, 02 2016.

Daniel Kifer and Ryan Rogers. A New Class of Private Chi-Square Hypothesis Tests. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 991–1000, Fort Lauderdale, FL, USA, 20–22 Apr 2017. PMLR. URL [http://proceedings.mlr.press/v54/rogers17a.html](http://proceedings.mlr.press/v54/rogers17a.html).

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, Providence, RI, October 2007. IEEE. URL [https://www.microsoft.com/en-us/research/publication/mechanism-design-via-differential-privacy/](https://www.microsoft.com/en-us/research/publication/mechanism-design-via-differential-privacy/).

A. Pastore and M. Gastpar. Locally differentially-private distribution estimation. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 2694–2698, July 2016. doi: 10.1109/ISIT.2016.7541788.

Sofya Raskhodnikova, Adam Smith, Homin K. Lee, Kobbi Nissim, and Shiva Prasad Kasiviswanathan. What can we learn privately? *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 00: 531–540, 2008. ISSN 0272-5428. doi: doi.ieeecomputersociety.org/10.1109/FOCS.2008.27.

Or Sheffet. Differentially private least squares: Estimation, confidence and rejecting the null hypothesis. *arXiv preprint arXiv:1507.02482*, 2015.

Jessica E. Steele, Pål Roe Sundsøy, Carla Pezzulo, Victor A. Alegana, Tomas J. Bird, Joshua Blumenstock, Johannes Bjelland, Keith Engo-Monsen, Yves-Alexandre de Montjoye, Asif M. Iqbal, Khandakar N. Hadiuzzaman, Xin Lu, Erik Wetter, Andrew J. Tatem, and Linus Bengtsson. Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127), 2017. ISSN 1742-5689. doi: 10.1098/rsif.2016.0690. URL [http://rsif.royalsocietypublishing.org/content/14/127/20160690](http://rsif.royalsocietypublishing.org/content/14/127/20160690).
Caroline Uhler, Aleksandra Slavkovic, and Stephen E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. *Journal of Privacy and Confidentiality*, 5(1), 2013.

Yue Wang, Jaewoo Lee, and Daniel Kifer. Differentially private hypothesis testing, revisited. *arXiv preprint arXiv:1511.03376*, 2015.

Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60:63–69, 1965.

M. Ye and A. Barg. Optimal schemes for discrete distribution estimation under local differential privacy. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 759–763, June 2017. doi: 10.1109/ISIT.2017.8006630.

Fei Yu, Stephen E. Fienberg, Aleksandra B. Slavkovic, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50:133–141, 2014.
A Omitted Proofs

Proof of Lemma 5.6. This follows from the central limit theorem. We first compute the expected value

\[ \frac{1}{n} E \left[ \tilde{H} \right] = \frac{e^{\epsilon/2}}{e^{\epsilon/2} + 1} p + \frac{1}{e^{\epsilon/2} + 1} (1 - p). \]

In order to compute the covariance matrix, we consider the diagonal term \((j,j)\)

\[ E \left[ (\tilde{H}_j)^2 \right] = \frac{e^{\epsilon/2}}{e^{\epsilon/2} + 1} p_j + \frac{1}{e^{\epsilon/2} + 1} (1 - p_j) \]

Next we compute the off diagonal term \((j,k)\)

\[ E \left[ \tilde{H}_j \tilde{H}_k \right] = \left( \frac{e^{\epsilon/2}}{e^{\epsilon/2} + 1} \right)^2 \Pr [X_j = 1 \& X_k = 1] \]

\[ + \frac{e^{\epsilon/2}}{(e^{\epsilon/2} + 1)^2} (\Pr [X_j = 1 \& X_k = 0] + \Pr [X_j = 0 \& X_k = 1]) \]

\[ + \frac{1}{(e^{\epsilon/2} + 1)^2} \Pr [X_j = 0 \& X_k = 0] \]

\[ = \left( \frac{1}{e^{\epsilon/2} + 1} \right)^2 \left[ e^{\epsilon/2} (p_j + p_k) + (1 - p_j - p_k) \right] \]

Before we construct the covariance matrix, we simplify a few terms

\[ E \left[ (\tilde{H}_j)^2 \right] - E \left[ \tilde{H}_j \right]^2 \]

\[ = \left( \frac{1}{e^{\epsilon/2} + 1} \right)^2 \left[ (e^{\epsilon/2} - 1) \left( e^{\epsilon/2} + 1 \right) p_j + 1 + e^{\epsilon/2} - \left( e^{\epsilon/2} - 1 \right) p_j + 1 \right]^2 \]

\[ = \left( \frac{1}{e^{\epsilon/2} + 1} \right)^2 \left[ \left( e^{\epsilon/2} - 1 \right) \left( e^{\epsilon/2} + 1 - 2 \right) p_j - (e^{\epsilon/2} - 1)^2 (p_j)^2 + e^{\epsilon/2} \right] \]

\[ = \frac{1}{(e^{\epsilon/2} + 1)^2} \left[ (e^{\epsilon/2} - 1)^2 p_j (1 - p_j) + e^{\epsilon/2} \right] \]

Further, we have

\[ E \left[ \tilde{H}_j \tilde{H}_k \right] - E \left[ \tilde{H}_j \right] E \left[ \tilde{H}_k \right] \]

\[ = \left( \frac{1}{e^{\epsilon/2} + 1} \right)^2 \left[ \left( e^{\epsilon/2} - 1 \right) (p_j + p_k) + 1 - \left( (e^{\epsilon/2} - 1) p_j + 1 \right) \left( (e^{\epsilon/2} - 1) p_k + 1 \right) \right] \]

\[ = - \left( \frac{e^{\epsilon/2} - 1}{e^{\epsilon/2} + 1} \right)^2 p_j p_k \]

Putting this together, the covariance matrix can then be written as

\[ \Sigma(p) = E \left[ \tilde{H} \left( \tilde{H} \right)^T \right] - E \left[ \tilde{H} \right] \left( E \left[ \tilde{H} \right] \right)^T \]

\[ = \left( \frac{e^{\epsilon/2} - 1}{e^{\epsilon/2} + 1} \right)^2 \left[ \text{Diag} \ (p) - p \ (p)^T \right] + \frac{e^{\epsilon/2}}{(e^{\epsilon/2} + 1)^2} I_d \]

\[ \square \]