Using Randomized Response for Differential Privacy Preserving Data Collection

Yue Wang
University of North Carolina at Charlotte,USA
ywang91@uncc.edu

Xintao Wu
University of Arkansas,USA
xintaowu@uark.edu

Donghui Hu
Hefei Institute of Technology,China
hudh@hfut.edu.cn

ABSTRACT

This paper studies how to enforce differential privacy by using the randomized response in the data collection scenario. Given a client's value, the randomized algorithm executed by the client reports to the untrusted server a perturbed value. The use of randomized response in surveys enables easy estimations of accurate population statistics while preserving the privacy of the individual respondents. We compare the randomized response with the standard Laplace mechanism which is based on query-output independent adding of Laplace noise. Our research starts from the simple case with one single binary attribute and extends to the general case with multiple polychotomous attributes. We measure utility preservation in terms of the mean squared error of the estimate for various calculations including individual value estimate, proportion estimate, and various derived statistics. We theoretically derive the explicit formula of the mean squared error of various derived statistics based on the randomized response theory and prove the randomized response outperforms the Laplace mechanism. We evaluate our algorithms on YesiWell database including sensitive biomarker data and social network relationships of patients. Empirical evaluation results show effectiveness of our proposed techniques. Especially the use of the randomized response for collecting data incurs fewer utility loss than the output perturbation when the sensitivity of functions is high.

Keywords

Randomized response; differential privacy; data collection

1. INTRODUCTION

The problem of protecting individual privacy in the process of data collection, querying, mining, and release has been researched extensively. Roughly speaking, there are two scenarios in the data privacy protection. One is the privacy preserving data publishing scenario, as in which a trusted server releases datasets of individual information or answers queries on such datasets. The second one is the data collection scenario, as in which an untrusted server collects personal information from individuals.

Our paper studies how to protect privacy in the data collection scenario by using randomized response, a surveying technique for learning statistics on individuals' sensitive attribute information such as whether the survey respondent has cheated in an exam. Randomized response is purely a client-based privacy solution. It does not rely upon a trusted third-party server and puts control over data back to clients. Given a client's value x, the randomized algorithm executed by the client reports to the untrusted server a perturbed value y. The parameters of the randomized algorithm are chosen in such a way so that to limit the server's ability to learn with confidence what value x was. For example, the survey respondent can flip a biased coin, in secret, and answer the truth if it comes up head, but tell the opposite answer if it comes up tail. Using this procedure, the respondent retains confidentiality of their true value due to coin randomness.

In our analysis, we adopt the rigorous differential privacy, which was introduced by Dwork et al. [6] and has been widely studied in the data publishing or query answering scenario. Roughly speaking, differential privacy aims to ensure that the output of the algorithm does not significantly depend on any particular individual's data and ensures that an adversary should not be able to confidently infer whether a particular individual is present in a database even with access to every other entry in the database and an unbounded computational power. In the data collection scenario, the inference is in terms of the sensitive value of one individual. In particular, we study how to derive the optimal distortion matrix used in the randomized response given a differential privacy threshold.

Differential privacy of each individual value can also be achieved by using the classic Laplace mechanism [6], which is based on query-output independent adding of Laplace noise. We study the relationship between the randomized response and the Laplace mechanism and compare their performance in terms of utility preservation under the same privacy threshold. Our research starts from the simple case of data collection with one single binary attribute and extends to the general case with multiple polychotomous attributes. We evaluate utility preservation in terms of individual value estimate, proportion estimate, and various derived statistics (e.g., entropy and χ^2). Existing works on investigating the accuracy-privacy tradeoff in differential privacy often define the accuracy in terms of the variance, or magnitude expectation of the noise added to the query output [8, 14]. For example, the authors [14] studied how to optimize linear counting queries under differential privacy and defined the error as the mean squared error of query output estimates, which corresponds to the variance of the noise added to the query output to preserve differential privacy. In this paper we also measure the utility in terms of the mean squared error of the estimate when randomized response is applied. In particular, we theoretically derive the explicit formula of the mean squared error of various derived statistics based on the randomized response theory.

We conduct our empirical evaluation on a biomarker dataset and a physical activity social network extracted from from the Yesi-Well pilot study about health. We compare the performance of the randomized response and that of the Laplace mechanism and report their estimates and standard deviations. One advantage of the use of the randomized response in the data collection scenario is that the collected data can be released for as much analysis as needed without worrying further privacy disclosure. This is different from the output perturbation where each additional analysis consumes further privacy budget. Moreover, the use of the randomized response for collecting data incurs less utility loss than the output perturbation when the sensitivity of functions is high, as demonstrated in our experiment where we calculate the number of triangles in the social network while preserving differential privacy.

2. BACKGROUND

2.1 Randomized Response

Suppose there are n individual clients $C_1, ..., C_n$; each client C_i has some private value x_i regarding a sensitive attribute X. An untrusted server needs to learn certain aggregate (statistical) properties of the individual's private data. However the clients are reluctant to disclose their personal information x_i. To ensure privacy, each client C_i only sends to the server a perturbed version y_i of x_i. The server collects the perturbed information from all individuals and then recovers the statistical properties by following some reconstruction procedures.

We assume every private value x_i about an individual belongs to the same fixed domain V_X and each x_i is chosen independently at random from the same fixed probability distribution $\boldsymbol{\pi}_X$. Note that this distribution is not private and is unknown to clients. The server aims to reconstruct the distribution $\boldsymbol{\pi}_X$ or derive some statistical properties of this distribution. The independence assumption ensures that the private information x_j of all individuals C_j besides C_i tells nothing new about C_i's own private information x_i once the distribution $\boldsymbol{\pi}_X$ is learned.

To protect privacy, each individual C_i hides its own sensitive information x_i by applying a randomization algorithm. A random instance y_i is sent to the untrusted server. The domain of all possible output of y_i is denoted by V_Y. The server receives y_i from client C_i and tries to learn distribution $\boldsymbol{\pi}_X$.

2.2 Differential Privacy

Differential privacy ensures that the inclusion or exclusion of one individual's record makes no statistical difference on the output.

DEFINITION 1. *(Differential Privacy [6]) A randomized function A gives ϵ-differential privacy if for all datasets D and D' differing at most one row, and all $S \subseteq Range(A)$*

$$Pr[A(D) \in S] \leq e^{\epsilon} \cdot Pr[A(D') \in S] \quad (1)$$

The privacy parameter ϵ controls the amount by which the distributions induced by two neighboring datasets may differ (smaller values enforce a stronger privacy guarantee). A general method for computing an approximation to any function f while preserving ϵ-differential privacy is given in [6]. The mechanism for achieving differential privacy computes the sum of the true answer and random noise generated from a Laplace distribution. The magnitude of the noise distribution is determined by the sensitivity of the computation and the privacy parameter specified by the data owner.

DEFINITION 2. *(Global Sensitivity [6]) The global sensitivity of a function $f : D^n \to \boldsymbol{R}^d$,*

$$GS_f(D) := \max_{D,D' s.t. D' \in \Gamma(D)} ||f(D) - f(D')||_1 \quad (2)$$

THEOREM 1. *(Laplace Mechanism [6]) An algorithm A takes as input a dataset D, and some $\epsilon > 0$, a query Q with computing function $f : D^n \to \boldsymbol{R}^d$, and outputs*

$$A(D) = f(D) + (Y_1, ..., Y_d) \quad (3)$$

where the Y_i are drawn i.i.d from $Lap(GS_f(D)/\epsilon)$. The Algorithm satisfies ϵ-differential privacy.

3. BINARY ATTRIBUTE

Suppose there are n individuals $C_1, ..., C_n$ and each individual C_i has a private binary value $x_i \in \{0, 1\}$ regarding a sensitive binary attribute X. To ensure privacy, each individual C_i sends to the untrusted server a modified version y_i of x_i. Using the randomized response, the server can collect perturbed data from individuals.

3.1 Randomized Response

A randomized response scheme on a binary attribute X follows a 2 × 2 design matrix (also called distortion matrix):

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} \quad (4)$$

where $p_{uv} = P[y_i = u|x_i = v]$ $(u, v \in \{0, 1\})$ denotes the probability that the random output is u when the real attribute value x_i for C_i is v; here $p_{uv} \in (0, 1)$. In the design matrix, the sum of probabilities of each column is 1.

In this section, we focus on two types of classic queries in the data collection scenario.

- Q1: what is the probability of correctly estimating x_i of individual C_i corresponding to the sensitive binary attribute X?
- Q2: what is the proportion of $X = 1$ $(X = 0)$?

For Q1, the original value $x_i = v(\in \{0, 1\})$ is outputted as $y_i = u(\in \{0, 1\})$ with probability p_{uv} from the design matrix $\mathbf{P}$ in Equation 4. Let $\hat{x}_i$ denote the reconstructed variable of x_i and $Pr(x_i = v \to \hat{x}_i = v)$ denote the probability of correctly reconstructing the individual's value as v from the perturbed data, given that the original value x_i is v where $v \in \{0, 1\}$. This reconstruction probability implies how much information is preserved in the randomization process.

$$Pr(x_i = v \to \hat{x}_i = v) = \sum_{u=0,1} P(y_i = u|x_i = v)P(\hat{x}_i = v|y_i = u) \quad (5)$$

Q2 aims to learn the population distribution based on the collected randomized dataset. We use π_0 (π_1) to denote the true proportion of value 0 (1) to be estimated in the original population. The observed proportion of value 0 (1) in the collected dataset is denoted as $\lambda_0(\lambda_1)$. We denote the unbiased estimator for π_0, π_1 respectively as $\hat{\pi}_0, \hat{\pi}_1$.

LEMMA 1. *(Chapter 1.2 [3]) Given the design matrix $\mathbf{P}$ and the observed proportion of value b $(\in \{0, 1\})$ in randomized dataset $\hat{D}_{rr}$, an unbiased estimator of the fraction of records whose attribute value is b is*

$$\hat{\pi}_b = \frac{p_{bb} - 1}{2p_{bb} - 1} + \frac{\lambda_b}{2p_{bb} - 1}, \quad (6)$$

where $p_{bb} \neq 0.5$ and $0 < p_{bb} < 1$. Since the observed number of records whose attribute value equals b follows binomial distribution, the variance of $\hat{\pi}_b$ is

$$var(\hat{\pi}_b) = \frac{\hat{\pi}_b(1-\hat{\pi}_b)}{n-1} + \frac{1}{n-1}[\frac{1}{16(p_{bb}-0.5)^2} - \frac{1}{4}] \qquad (7)$$

which is the expected error for the estimator $\hat{\pi}_b$.

3.2 Randomized Response vs. Laplace Mechanism

The values in each row u ($u \in \{0,1\}$) of the design matrix denote the probability that the random output is u. For example, p_{00} (p_{01}) denotes the distortion probability that the random output value is 0 when the real individual value is 0 (1). Without loss of generality, we assume the randomized response still favors the true value, i.e., $p_{00}, p_{11} > 0.5$. Differential privacy requires that $p_{00}/p_{01} \leq e^{\epsilon}$. Thus we show how the randomized response will achieve differential privacy in the following result. In addition, we also give the form of the design matrix that is expected to achieve the optimal utility while satisfying the given ϵ-differential privacy.

RESULT 1. *For a given differential privacy parameter ϵ, the randomized response scheme following the design matrix $\mathbf{P}$ in Equation 4 satisfies ϵ-differential privacy if $\max\{\frac{p_{00}}{p_{01}}, \frac{p_{11}}{p_{10}}\} \leq e^{\epsilon}$.*

In order to maximize $p_{00} + p_{11}$ while satisfying ϵ-differential privacy, the design matrix should have the following pattern,

$$\mathbf{P}_{rr} = \begin{pmatrix} \frac{e^{\epsilon}}{1+e^{\epsilon}} & \frac{1}{1+e^{\epsilon}} \\ \frac{1}{1+e^{\epsilon}} & \frac{e^{\epsilon}}{1+e^{\epsilon}} \end{pmatrix} \qquad (8)$$

PROOF. Assume $\frac{p_{00}}{p_{01}} = p, \frac{p_{11}}{p_{10}} = q$. In order to satisfy ϵ-differential privacy, we have $1 < p \leq e^{\epsilon}$ and $1 < q \leq e^{\epsilon}$. In this case, the distortion matrix will have the general form:

$$\mathbf{P}_{rr} = \begin{pmatrix} \frac{p(q-1)}{pq-1} & \frac{q-1}{pq-1} \\ \frac{p-1}{pq-1} & \frac{(p-1)q}{pq-1} \end{pmatrix}$$

We denote

$$func(p,q) = \mathbf{P}_{rr}(1,1) + \mathbf{P}_{rr}(2,2) = \frac{p(q-1)}{pq-1} + \frac{(p-1)q}{pq-1}.$$

Since $\frac{\partial func}{\partial p} = \frac{(q-1)^2}{(pq-1)^2} > 0$ and $\frac{\partial func}{\partial q} = \frac{(p-1)^2}{(pq-1)^2} > 0$ when $p, q \in (1, e^{\epsilon}]$, thus $func$ will achieve its maximum value if and only if $p = q = e^{\epsilon}$. In this way, we get the form in Equation 8. □

Similarly, individual C_i can achieve differential privacy by using the Laplace mechanism. The Laplace mechanism first adds a random noise generated from the Laplace distribution with parameter $\frac{1}{\epsilon}$ (with a given ϵ and the global sensitivity of 1) to the true answer x_i. Since the output should be a binary value, we postprocess the perturbed result by outputting 0 if the perturbed value is less than c and outputting 1 otherwise, shown in Equation 9.

$$y_i = \begin{cases} 0;\ \text{if } x_i + Lap(1/\epsilon) < c \\ 1;\ \text{if } x_i + Lap(1/\epsilon) \geq c \end{cases} \qquad (9)$$

The probability of $y_i = 0$ is $F_{x_i,1/\epsilon}(c)$ and the probability of $y_i = 1$ is $1 - F_{x_i,1/\epsilon}(c)$ where $F_{\mu,b} = \frac{1}{2} + \frac{1}{2}sgn(x-\mu)(1 - e^{(-\frac{|x-\mu|}{b})})$ denotes the cumulative distribution function of Laplace distribution $Lap(\mu, b)$ with the location parameter μ and the scale parameter b (and with the mean μ and variance $2b^2$). Thus we can map the Laplace mechanism to the randomized response with the design matrix as

$$\mathbf{P}_{lm} = \begin{pmatrix} F_{0,1/\epsilon}(c) & F_{1,1/\epsilon}(c) \\ 1 - F_{0,1/\epsilon}(c) & 1 - F_{1,1/\epsilon}(c) \end{pmatrix} \qquad (10)$$

For a given ϵ, the perturbed result of Laplace mechanism satisfies differential privacy because the postprocessing process does not consume any privacy budget. Thus we have the following result indicating the Laplace mechanism with the postprocessing satisfies ϵ-differential privacy. We also show that the best postprocessing strategy is to set $c = 0.5$ for Equation 9.

RESULT 2. *For a given differential privacy parameter ϵ, the Laplace mechanism based scheme with the postprocessing strategy following Equation 9 satisfies ϵ-differential privacy.*

The corresponding design matrix should have the following form,

$$\mathbf{P}_{lm} = \begin{pmatrix} 1 - \frac{1}{2}e^{-\frac{\epsilon}{2}} & \frac{1}{2}e^{-\frac{\epsilon}{2}} \\ \frac{1}{2}e^{-\frac{\epsilon}{2}} & 1 - \frac{1}{2}e^{-\frac{\epsilon}{2}} \end{pmatrix}, \qquad (11)$$

in order to maximize $p_{00} + p_{11}$ while satisfying ϵ-differential privacy.

PROOF. With the assumption that we need to preserve the real value with probability greater than 0.5, c in Equation 9 is in the range [0,1]. We have

$$\mathbf{P}_{lm} = \begin{pmatrix} 1 - \frac{1}{2}e^{-c\epsilon} & \frac{1}{2}e^{-(1-c)\epsilon} \\ \frac{1}{2}e^{-c\epsilon} & 1 - \frac{1}{2}e^{-(1-c)\epsilon} \end{pmatrix}.$$

We denote

$$func(c) = \mathbf{P}_{lm}(1,1) + \mathbf{P}_{lm}(2,2) = 1 - \frac{1}{2}e^{-c\epsilon} + 1 - \frac{1}{2}e^{-(1-c)\epsilon},$$

where $c \in [0,1]$. Since

$$\frac{\partial func}{\partial c} = \frac{\epsilon}{2}(e^{-c\epsilon} - e^{-(1-c)\epsilon}),$$

we have

$$\frac{\partial func}{\partial c} \begin{cases} > 0, & \text{when } c \in [0, 0.5) \\ < 0, & \text{when } c \in (0.5, 1]. \end{cases}$$

Thus the maximum value for $func(c)$ is achieved when $c = 0.5$. □

3.3 Utility Comparison

In this paper we measure the utility in terms of the mean squared error of the estimate for x_i given a randomized mechanism A.

$$\textsf{ERROR}_{\textsf{A}}(\hat{x}_i) = \mathbb{E}[(\hat{x}_i - x_i)^2] \qquad (12)$$

Note that by replacing $P(y_i = u | x_i = v)$ in Equation 5 with values in $\mathbf{P}_{rr}$ of the randomized response and those in $\mathbf{P}_{lm}$ of the Laplace mechanism, we can calculate the estimates $\hat{x}$ respectively. We can then compare the utility of the randomized response with that of the Laplace mechanism based on Equation 12.

Intuitively, under the same privacy standard, the mechanism with larger diagonal elements in the corresponding design matrix tends to achieve better utility. The diagonal elements in Equation 8 are larger than those in Equation 11. Based on such intuition, we can prove that the randomized response actually can achieve better utility than the classic Laplace mechanism in the scenario of binary data collection.

THEOREM 2. *Given ϵ, for the randomized response scheme with $\mathbf{P}_{rr}$ and the Laplace mechanism based on $\mathbf{P}_{lm}$, we have $\textsf{ERROR}_{\textsf{rr}}(\hat{x}_i) \leq \textsf{ERROR}_{\textsf{lm}}(\hat{x}_i)$.*

PROOF. We have $(\hat{x}_i - x_i)^2 = 0$ with probability $Pr(x_i = v \rightarrow \hat{x}_i = v)$ in Equation 5; and $(\hat{x}_i - x_i)^2 = 1$ with probability $1 - Pr(x_i = v \rightarrow \hat{x}_i = v)$. So

$$\begin{aligned}&\textsc{Error}_{\mathsf{A}}(\hat{x}_i)\\ &= 0 \times Pr(x_i = v \rightarrow \hat{x}_i = v) + 1 \times (1 - Pr(x_i = v \rightarrow \hat{x}_i = v))\\ &= 1 - Pr(x_i = v \rightarrow \hat{x}_i = v).\end{aligned}$$

Without loss of generality, assume $v = 1$, we denote the prior probability of $x_i = 1$ as π_1. According to Bayes's theorem, we have

$$\textsc{Error}_{\mathsf{A}}(\hat{x}_i) = 1 - \left(\frac{p_{11}^2\pi_1}{p_{11}\pi_1 + p_{10}(1-\pi_1)} + \frac{p_{01}^2\pi_1}{p_{01}\pi_1 + p_{00}(1-\pi_1)}\right).$$

For $\textsc{Error}_{\mathsf{lm}}(\hat{x}_i)$, we have $p_{11}^{lm} = p_{00}^{lm} = 1 - \frac{1}{2}e^{-0.5\epsilon}$. For $\textsc{Error}_{\mathsf{rr}}(\hat{x}_i)$, we have $p_{11}^{rr} = p_{00}^{rr} = \frac{e^\epsilon}{e^\epsilon+1}$. Now we are to prove: for $\pi_1 \in [0, 1]$,

$$func(\pi_1) = \textsc{Error}_{\mathsf{lm}}(\hat{x}_i) - \textsc{Error}_{\mathsf{rr}}(\hat{x}_i) \geq 0.$$

For a given $\epsilon > 0$, since $func(\pi_1)$ is continual and it has only two roots for the parameter range $\pi_1 \in [0, 1]$. Respectively they are $\pi_1 = 0$ and $\pi_1 = 1$. It indicates that all $\pi_1 \in (0, 1)$, the output of $func(\pi_1)$ has the same sign. Thus, we only need to prove for one specific π_1, say $\pi_1 = 0.5$, that $func(\pi_1) > 0$. Then the same result holds for all $\pi_1 \in (0, 1)$.
In this case, we have

$$\textsc{Error}_{\mathsf{A}}(\hat{x}_i) = 1 - (p_{11}^2 + (1 - p_{11}^2)).$$

Since $p_{11}^{rr} > p_{11}^{lm} > 0.5$ for all $\epsilon > 0$, we have $func(0.5) > 0$. Thus for all $\pi_1 \in (0, 1)$ we have $func(\pi_1) > 0$. The same idea can be applied for the situation of $v = 0$. So for $\pi_1 \in [0, 1]$, we have $\textsc{Error}_{\mathsf{rr}}(\hat{x}_i) \leq \textsc{Error}_{\mathsf{lm}}(\hat{x}_i)$. □

Using either the randomized response or the Laplace mechanism based approach, the server can collect private data from individuals. Both collected datasets satisfy ϵ-differential privacy (rather than $n\epsilon$-differential privacy) according to the independence assumption. Formally, $\hat{D}_{rr}$ denotes the dataset generated by the randomized response following the design matrix $\mathbf{P}_{rr}$ as in Equation 8. Similarly, $\hat{D}_{lm}$ denotes the dataset generated by the Laplace mechanism. We define the expected error of the estimator $\hat{\pi}_b$ as its variance, $\textsc{Error}_{\mathsf{A}}(\hat{\pi}_b) = var(\hat{\pi}_b)$.

THEOREM 3. *Given ϵ, for the randomized response scheme with $\mathbf{P}_{rr}$ and the Laplace mechanism based on $\mathbf{P}_{lm}$, we have* $\textsc{Error}_{\mathsf{rr}}(\hat{\pi}_b) \leq \textsc{Error}_{\mathsf{lm}}(\hat{\pi}_b)$.

PROOF. From Equation 7, we see comparing the utility of estimation from $\hat{D}_{rr}$ and $\hat{D}_{lm}$ relies only on p_{bb} in the distortion matrix $\mathbf{P}$. For the Laplace mechanism, we have $p_{00}^{lm} = p_{11}^{lm} = 1 - \frac{1}{2}e^{-\frac{\epsilon}{2}}$; For the randomized response, we have $p_{11}^{rr} = p_{00}^{rr} = \frac{e^\epsilon}{e^\epsilon+1}$. Since $p_{11}^{rr} > p_{11}^{lm} > 0.5$ for all $\epsilon > 0$, we have $\textsc{Error}_{\mathsf{rr}}(\hat{\pi}_b) \leq \textsc{Error}_{\mathsf{lm}}(\hat{\pi}_b)$, according to Equation 7. □

4. POLYCHOTOMOUS ATTRIBUTE

In the previous section, we compared the Laplace mechanism and the randomized response approach in collecting information about one private binary attribute. In this section, we extend to a sensitive polychotomous attribute with $t(t \geq 2)$ mutually exclusive and exhaustive classes. Due to space limits, we skip all proofs of results in this section. Refer to [19] for proof details.

4.1 Randomized Response

The corresponding unknown proportions to be estimated are denoted as $\pi_1, ..., \pi_t$. The randomization device is such that an individual belonging to the vth category($v = 1, ..., t$) reports a random value u ($u = 1, ..., t$) with probability p_{uv} and $\Sigma_{u=1}^{t} p_{uv} = 1$ for all $v = 1, ..., t$.

The matrix $\mathbf{P} = \{p_{uv}\}$ is called the design matrix, where the sum of each column in $\mathbf{P}$ is 1.

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1v} & \cdots & p_{1t} \\ p_{21} & p_{22} & \cdots & p_{2v} & \cdots & p_{2t} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{u1} & p_{u2} & \cdots & p_{uv} & \cdots & p_{ut} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{t1} & p_{t2} & \cdots & p_{tv} & \cdots & p_{tt} \end{pmatrix} \tag{13}$$

Similarly we have two types of classic queries in the data collection scenario.

- Q1: what is the probability of correctly estimating x_i of individual C_i corresponding to the sensitive attribute X?
- Q2: what is the proportion of $X = 1, \cdots, t$?

Let $Pr(x_i = v \rightarrow \hat{x}_i = v)$ denote the probability of correctly reconstructing the individual's value as v from the perturbed data, given that the original value x_i is v where $v \in \{1, \cdots, t\}$. This reconstruction probability implies how much information is preserved in the randomization process.

$$\begin{aligned}&Pr(x_i = v \rightarrow \hat{x}_i = v) =\\ &\sum_{u=1}^{t} P(y_i = u|x_i = v)P(\hat{x}_i = v|y_i = u)\end{aligned} \tag{14}$$

The probability λ_u of the (randomized) response u is given by

$$\lambda_u = \Sigma_{v=1}^{t} p_{uv}\pi_v \qquad (u = 1, ..., t) \tag{15}$$

Defining $\boldsymbol{\lambda}$= $(\lambda_1, ..., \lambda_t)'$, $\boldsymbol{\pi} = (\pi_1, ..., \pi_t)'$, we obtain in matrix notation

$$\boldsymbol{\lambda} = \mathbf{P}\boldsymbol{\pi} \tag{16}$$

LEMMA 2. *(Chapter 3.3 [3]) With a simple random sample with replacement of size n, let $\hat{\boldsymbol{\lambda}}$ be the vector of sample proportions corresponding to $\boldsymbol{\lambda}$. Then assuming the nonsingularity of the design matrix $\mathbf{P}$, an unbiased estimator of $\boldsymbol{\pi}$ emerges as*

$$\hat{\boldsymbol{\pi}} = \mathbf{P}^{-1}\hat{\boldsymbol{\lambda}}. \tag{17}$$

An unbiased estimator of the dispersion matrix is given by

$$disp(\hat{\boldsymbol{\pi}}) = (n-1)^{-1}\mathbf{P}^{-1}(\hat{\lambda}^\delta - \hat{\boldsymbol{\lambda}}\hat{\boldsymbol{\lambda}}')\mathbf{P}'^{-1}, \tag{18}$$

where $\hat{\lambda}^\delta$ is a diagonal matrix with the same diagonal elements as those of $\hat{\boldsymbol{\lambda}}$

4.2 Randomized Response vs. Laplace Mechanism

Similar as the binary case, the values in each row u ($u \in \{1, 2, ..., t\}$ of the design matrix denote the probability that the random output is u. Differential privacy requires that the maximum value difference in each row is bounded by e^ϵ. Thus we have the following result.

RESULT 3. *The randomized response is ϵ-differentially private if $\epsilon \geq ln \max_{u=1..t} \frac{\max_{v=1..t} p_{uv}}{\min_{v=1..t} p_{uv}}$.*

In order to maximize the sum of the diagonal elements,the design matrix for randomized response $\mathbf{P}_{rr} = \{p_{uv}\}$ should be in the following form,

$$p_{uv} = \begin{cases} \frac{e^{\epsilon}}{t-1+e^{\epsilon}};\ \textit{if}\ u = v \\ \frac{1}{t-1+e^{\epsilon}};\ \textit{if}\ u \neq v \end{cases} \tag{19}$$

In other words, in the optimal form of the design matrix, all diagonal entries are set as $\frac{e^{\epsilon}}{t-1+e^{\epsilon}}$ and all off-diagonal entries are set as $\frac{1}{t-1+e^{\epsilon}}$. We can also achieve differential privacy by adding Laplace noise. The global sensitivity is $t-1$. So the Laplace noise is generated from the distribution $Lap(\frac{t-1}{\epsilon})$. Because the perturbed outputs are numerical, we postprocess to map them to an index value from 1 to t as shown in Equation 20.

$$y_i = \begin{cases} 1;\ \text{if}\ x_i + Lap((t-1)/\lambda) \in (-\infty, c_1] \\ 2;\ \text{if}\ x_i + Lap((t-1)/\lambda) \in (c_1, c_2] \\ \ldots \\ u;\ \text{if}\ x_i + Lap((t-1)/\lambda) \in (c_{u-1}, c_u] \\ \ldots \\ t;\ \text{if}\ x_i + Lap((t-1)/\lambda) \in (c_{t-1}, \infty) \end{cases} \tag{20}$$

where c_u is in the range $[u, u+1]$.

Note that in this scenario, the strategy of perturbation by Laplace mechanism is also a special case of the randomized response strategy. We give the form of the corresponding design matrix in Equation 21. The following result shows such Laplace mechanism with postprocessing satisfies ϵ-differential privacy. We also give the best postprocessing strategy with the corresponding design matrix.

RESULT 4. *The Laplace mechanism of adding random noise from distribution $Lap(\frac{t-1}{\epsilon})$, with postprocessing strategy following Equation 20 is ϵ- differentially private.*

In order to maximize the sum of the diagonal elements in the corresponding design matrix for Laplace mechanism, $\mathbf{P}_{lm}$, we have $c_u = u + 0.5$ for $u \in \{1, 2, ..t-1\}$ in Equation 20. The corresponding design matrix $\mathbf{P}_{lm} = \{p_{uv}\}$ has the following form,

$$p_{uv} = \begin{cases} F_{v,\frac{\epsilon}{t-1}}(1.5);\ \textit{if}\ u = 1 \\ 1 - F_{v,\frac{\epsilon}{t-1}}(t-0.5);\ \textit{if}\ u = t \\ F_{v,\frac{\epsilon}{t-1}}(u+0.5) - F_{v,\frac{\epsilon}{t-1}}(u-0.5);\ \textit{otherwise} \end{cases} \tag{21}$$

where $F_{v,\frac{\epsilon}{t-1}}$ is the cumulative distribution function of Laplace distribution with mean value of v ($v \in \{1, 2, ..., t\}$), variance of $2\lambda^2$ and $\lambda = (t-1)/\epsilon$.

4.3 Utility Comparison

Intuitively, the utility depends on the diagonal elements in the design matrix. The Laplace mechanism based approach degrades the utility by favoring the values near the correct value. The sum of diagonal elements in $\mathbf{P}_{lm}$ is actually smaller than that in $\mathbf{P}_{rr}$ (for details see the proof of Theorem 4). In consistence with the binary case, the randomized response achieves better utility than the classic Laplace mechanism in data collection scenario.

THEOREM 4. *Given ϵ, for the randomized response scheme with $\mathbf{P}_{rr}$ and the Laplace mechanism based on $\mathbf{P}_{lm}$, we have $\text{ERROR}_{\text{rr}}(\hat{x}_i) \leq \text{ERROR}_{\text{lm}}(\hat{x}_i)$.*

Similarly as the binary case, we define the expected error of the estimator $\hat{\pi}_v$ for the proportion of category v ($v \in \{1, 2, ..., t\}$) as its variance, the diagonal element in the unbiased estimate of dispersion matrix $disp(\hat{\boldsymbol{\pi}})$ following the randomized mechanism A. We have $\text{ERROR}_{\text{A}}(\hat{\pi}_v) = disp(\hat{\boldsymbol{\pi}})_{vv}$ where $v \in \{1, 2, ..., t\}$.

However, it is intractable to directly prove that the randomized response strategy following the design matrix in Equation 19 could achieve lower expected error of the estimator $\hat{\pi}_v$ than the Laplace mechanism based approach following Equation 21 does. Intuitively we can see that the Laplace mechanism based approach will degrade the utility by favoring the values near the correct value. As shown in the proof of Theorem 4, the sum of the diagonal elements in $\mathbf{P}_{lm}$ is smaller than that in $\mathbf{P}_{rr}$, which indicates that the estimation based on the randomized response mechanism following $\mathbf{P}_{rr}$ is expected to achieve smaller error than that based on the Laplace mechanism following $\mathbf{P}_{lm}$.

5. ACCURACY ANALYSIS OF RANDOMIZED DATASET

5.1 Multiple Attributes

To be consistent with notations, we denote the set of variables by $\mathcal{X} = \{X_1, \cdots, X_s\}$. Note that, for ease of presentation, we use the terms "attribute" and "variable" interchangeably. Each variable X_u has d_u mutually exclusive and exhaustive categories. We use $i_u = 1, \cdots, d_u$ to denote the index of its categories. For each data record, we apply the randomized response model independently on each sensitive variable X_u using different settings of distortion.

Formally, let $\pi_{i_1,\cdots,i_s}$ denote the true proportion corresponding to the categorical combination of s variables $(X_{1i_1}, \cdots, X_{si_s})$ in the original data, where $i_u = 1, \cdots, d_u$ ($u = 1, \cdots, s$), and X_{1i_1} denotes the i_1th category of attribute X_1. Let $\boldsymbol{\pi}$ be a vector with elements $\pi_{i_1,\cdots,i_s}$ arranged in a fixed order. The combination vector corresponds to a fixed order of cell entries in the contingency table formed by these s variables. Similarly, we denote $\lambda_{i_1,\cdots,i_s}$ as the expected proportion in the randomized data.

For the case of s multi-variables, we denote $\lambda_{\mu_1,\cdots,\mu_s}$ as the expected probability of getting a response $(X_{1\mu_1}, \cdots, X_{s\mu_s})$ and $\boldsymbol{\lambda}$ the vector with elements $\lambda_{\mu_1,\cdots,\mu_s}$ arranged in a fixed order. For example, given a dataset with two variables, *Gender* with domain values {male, female} and *Race* with domain values {black, white,asian}, we have $d_1 = 2$ and $d_2 = 3$. The vector $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{23})'$ corresponds to a fixed order of cell entries π_{ij} in the 2×3 contingency table. π_{12} denotes the proportion of records with *male* and *white*.

Let $P = P_1 \times \cdots \times P_s$, we can obtain

$$\boldsymbol{\lambda} = P\boldsymbol{\pi} = (P_1 \times \cdots \times P_s)\boldsymbol{\pi} \tag{22}$$

where × stands for the Kronecker product [1].

The original database $\mathcal{D}$ is changed to $\mathcal{D}_{rr}$ after randomization. An unbiased estimate of $\boldsymbol{\pi}$ based on one given realization $\mathcal{D}_{rr}$ follows as

$$\hat{\boldsymbol{\pi}} = P^{-1}\hat{\boldsymbol{\lambda}} = (P_1^{-1} \times \cdots \times P_s^{-1})\hat{\boldsymbol{\lambda}} \tag{23}$$

where $\hat{\boldsymbol{\lambda}}$ is the vector of proportions calculated from $\mathcal{D}_{rr}$ corresponding to $\boldsymbol{\lambda}$ and P_u^{-1} denotes the inverse of the matrix P_u.

5.2 Variance of Derived Measure

Many measures (including entropy, mutual information, Pearson Correlation, G^2-likelihood) can be expressed as one derived random variable (or function) from the observed variable $\boldsymbol{\pi}$. Similarly,

[1] Kronecker product is an operation on two matrices, an m-by-n matrix A and a p-by-q matrix B, resulting in the mp-by-nq block matrix

its estimate from the randomized data can be considered as another derived random variable from the input variable $\hat{\boldsymbol{\pi}}$. One natural question is how to calculate the variance of those estimates. In the following, we introduce the use of the delta method [11] to derive the variance of variours measures.

Let Z be a random variable derived from the observed random variables X_i ($i = 1, \cdots, k$): $Z = g(X_1, X_2, ..., X_k)$. According to the delta method, a Taylor approximation of the variance of a function with multiple variables can be expanded as

$$var\{g(X_1, X_2, ..., X_k)\} = \sum_{i=1}^{k}\{g_i'(\theta)\}^2 var(X_i) + \sum\sum_{i\neq j=1}^{k} g_i'(\theta)g_j'(\theta)cov(X_i, X_j) + o(n^{-r}) \tag{24}$$

where θ_i is the mean of x_i, $g_i'(\theta)$ is the $\frac{\partial g(X_1,X_2,...,X_k)}{\partial X_i}$ evaluated at $\theta_1, \theta_2, \cdots, \theta_k$.

We use the entropy function as an example. The entropy function from information theory is defined as follows:

$$H(X) = -\sum_{j \in Range(X)} \pi_j log_2 \pi_j \tag{25}$$

We can estimate the entropy of the discrete random variable X with possible values $\{1, 2, .., t\}$ in the original dataset using the estimator of the distribution $\boldsymbol{\pi}$ and the estimator of the dispersion matrix $disp(\boldsymbol{\pi})$ calculated following Equations 17 and 18.

RESULT 5. *The variance of the estimated entropy can be computed following Equation 24 where* $i, j \in \{1, 2, ..., t\}$, $k = t$, $X_i = \hat{\pi}_i$, $\frac{\partial g(X_1,X_2,...,X_k)}{\partial X_i} = log_2\hat{\pi}_i + \frac{1}{ln2}$ *and* $var(X_i) = disp(\hat{\boldsymbol{\pi}})(i,i)$, $cov_{i\neq j}(X_i, X_j) = disp(\hat{\boldsymbol{\pi}})(i,j)$.

Different from the entropy which involves only one variable, some measures such as chi-square statistics involve multiple variables.

$$\hat{\chi}^2 = n\sum_i\sum_j \frac{\{\pi_{ij} - \pi_{i+}\pi_{+j}\}^2}{\pi_{i+}\pi_{+j}} \tag{26}$$

It is easy to see $\hat{\chi}^2$ can be considered as one derived variable from the observed elements $\hat{\pi}_{X_1 \cdots X_s}$ and the marginal totals of the contingency table. Following the same delta method, we can derive its variance.

6. EMPIRICAL EVALUATION

6.1 YesiWell Data

We conduct our empirical evaluation using the real dataset collected from the YesiWell pilot study. The study was conducted in 2010-2011 as the collaboration among several health laboratories and universities to help people maintain active lifestyles and lose weight. Data gained from this study includes information of various domains such as biomarker, biometrics, social activities.

We conduct experiments on a chosen table which contains 248 individuals' biomarker information. In particular, we focus on two sensitive attributes: LDL cholesterol (LDL) with six domain levels and Total cholesterol (TC) with three domain levels. Under each differential privacy threshold ϵ, we compare the performance of the randomized response (with the corresponding derived design matrix $\mathbf{P}_{rr}$) and that of the Laplace mechanism (with the corresponding derived design matrix $\mathbf{P}_{lm}$) from the utility preservation perspective. We focus on proportion estimates of categories based on LDL levels, the derived entropy of LDL, and the χ^2 statistics of LDL and TC. For each statistics, we report their estimate values and derive standard deviations for two strategies: randomized response and Laplacian mechanism.

We also conduct experiments on the YesiWell physical activity social network which contains 185 users and 684 interactions. Each interaction, represented as an edge between two user nodes, is considered sensitive in our context. We study how to enforce edge differential privacy in our social network, i.e., the inclusion or exclusion of a link between two individuals from the graph makes no statistical difference to the results found. We focus on two classic graph features: the degree sequence $D = \{d_i\}$ where each entry represents the degree of node i, and the number of triangle sequence $N_\Delta = \{N_\Delta(i)\}$ where each entry represents the number of triangles involving node i. We compare the performance of the randomized response and that of the Laplace mechanism and report their estimates and standard deviations for the above two graph statistics.

In addition to our above study in the data collection scenario, we also compare our randomized response with two mechanisms, Laplacian mechanism and smooth sensitivity, in the data query answering scenario where the trusted server keeps all unperturbed values and returns differential privacy preserving query answers.

6.2 Proportion Estimate

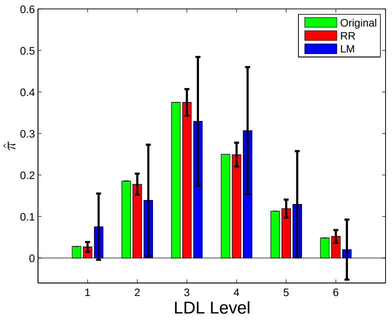

Figure 1: Estimation of the distribution of each LDL level when $\epsilon = 5$

Figure 1 shows the estimation result for each of six LDL levels when $\epsilon = 5$. For each level, the green bar represents the original proportion value, the red bar shows the estimated proportion value from the randomized response, and the blue bar shows the estimated proportion value from the Laplace mechanism. For each estimate, we also report its standard deviation. We can easily observe that the randomized response achieves better utility preservation for each level (with more accurate estimate and smaller standard deviation) than the Laplace mechanism.

Figure 2 shows the estimation results in terms of the average mean squared error of proportion estimates of the six different levels between the randomized response and the Laplace mechanism given varying ϵ values. We can easily observe the averaged estimation error of the randomized response is two-three orders lower than that of the Laplace mechanism and the randomized response shows more superiority than the Laplace mechanism when ϵ is small.

6.3 Derived Measures

We calculate the estimates of the entropy of the LDL. Figure 3 shows the estimation results of the calculated entropy values from

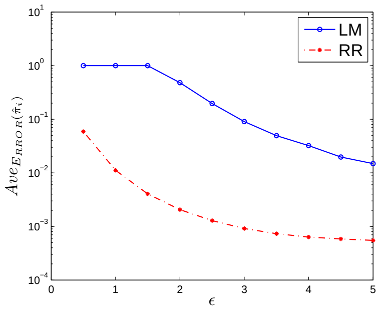

Figure 2: Average mean squared error for estimation of the distribution of the six different LDL levels vs. varying ϵ

two approaches with varying ϵ. We can see that the red line (corresponding to the randomized response) is more close to the green line (corresponding to the real entropy value) than the blue line (corresponding to the Laplace mechanism). The bar values (corresponding to their standard deviation values) also clearly show the superiority of the randomize response.

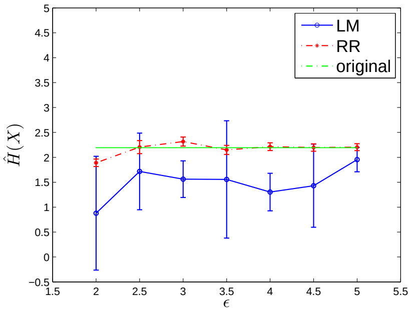

Figure 3: Estimation of the entropy of LDL vs. varying ϵ

We calculate the estimates of the chi-square statistics between the LDL and TC. Figure 4 shows the estimation results of the χ^2 statistics from two approaches with varying ϵ. We can see that the red line (corresponding to the randomized response) generally lies more close to the green line (corresponding to the real entropy value) than the blue line (corresponding to the Laplace mechanism) with varying ϵ values. The randomized response also has much smaller standard deviation values than the Laplace mechanism, which also indicates better utility preservation.

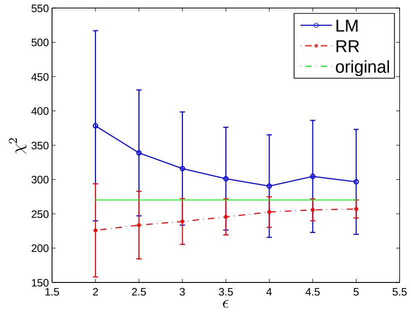

Figure 4: Estimation of χ^2 between LDL and TC vs. varying ϵ

6.4 Graph Statistics

The graph of the YesiWell social network contains 185 nodes and 684 edges. In the data collection scenario, the untrusted server collects the link relationship information from users. The link relationship between two users is sensitive and should be protected. The collected social network data with n users and m relationships can be represented as an adjacency matrix $A_{n\times n}$ with $2m$ non-zero entries where $A_{ij} = 1$ denotes the presence of an relationship between user i and user j, and $A_{ij} = 0$ otherwise. In our setting, for A_{ij}, the client C_i applies the randomized response (or the Laplace mechanism) to send the server a randomized output $Y_{ij} \in 0, 1$. After collecting all randomized relationships, the server then applies the reconstruction process and generates one instance of the social network with $2m$ non-zero entries (denoted as $\hat{A}$). The generated graph instance satisfies ϵ differential privacy and can be released for any analysis. In this experiment, we conduct performance comparison between the randomized response and the Laplace mechanism using two graph features, the degree sequence $D = \{d_i\}$ and the number of triangle sequence $N_\Delta = \{N_\Delta(i)\}$. Figure 5 and Figure 6 show comparison results in terms of the degree sequence and the number of triangle sequence respectively.

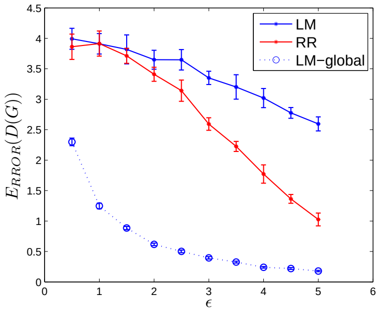

Figure 5: Average entrywise error of the degree sequence vs. varying ϵ

Figure 5 shows the average entrywise error of the degree sequence calculated from different approaches with varying ϵ. In the figure, we denote the Laplace mechanism as LM and the randomized response as RR, each of which uses its randomized graph topology respectively. We also report the comparison with the output perturbation method, LM-global, which adds the Laplace noise directly to the query output. Note that LM-global is used in the data query answering scenario where the server is assumed to have all the true unperturbed data. However, any differential privacy preservation query consumes a separate privacy budget. On the contrary, the randomized data collected from LM and RR can be released for any analysis with the same privacy threshold. We can observe in Figure 5 that the randomized response achieves better utility preservation than the Laplace method in the data collection scenario and the LM-global incurs less estimation error than both RR and LM (due to its small global sensitivity value $GS_D = 2$).

Figure 6 shows the average entrywise error of the number of triangle sequence calculated from different approaches with varying ϵ. Note that the global sensitivity of N_Δ is $3(n-2)$. We denote the approach of directly adding the Laplace noise based on the global sensitivity as LM-global. We denote the approach of adding the Laplace noise based on the smooth sensitivity [16] as LM-smooth. As above, we denote the Laplace mechanism in our data collection scenario as LM and the randomized response as RR. We can observe that the average entrywise error of RR and LM

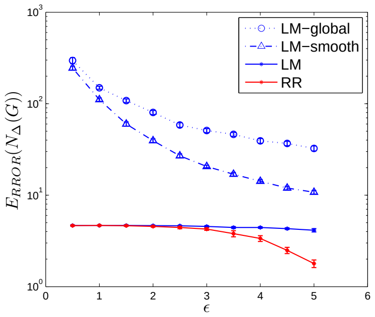

Figure 6: Average entrywise error of the number of triangle sequence vs. varying ϵ; LM-global (LM-smooth) denotes the global sensitivity based Laplace (the smooth sensitivity based) mechanism.

is lower than that either of LM-global or LM-smooth, indicating the local differential privacy preserving data collection could be a better choice than output perturbation for queries or analysis with very large sensitive values. It is unsurprise that RR achieves the best utility preservation.

7. RELATED WORK

Randomized response techniques have been extensively investigated in statistics (e.g., see a book [3]). Previous work on privacy preservation using the randomized response model mainly focused on evaluating the trade-off between privacy preservation and utility loss of the reconstructed data (e.g., [1, 18]). Some research studied the problem of determining the optimal distortion parameters to achieve good performance (e.g., [9]). The authors in [7] first presented the notation of privacy breaches based on *amplification* where it provides guarantee limits on privacy breaches without any knowledge of the distribution of original data.

Differential privacy research has been significantly studied from the theoretical perspective, e.g., [4, 12], and the application perspective, e.g., [15, 20]. The mechanisms of achieving differential privacy mainly include the classic approach of adding Laplacian noise [6], the exponential mechanism based on the smooth sensitivity [15], and the functional perturbation approach [4]. Most of the above works focused on the data publishing scenario.

Local differential privacy was formally proposed in [5, 10] as a strong measure of privacy under the data collection scenario, where individual clients are willing to share their data but are concerned about revealing sensitive information. The authors studied the problem of utility maximization under local differential privacy and developed a family of extremal mechanisms called the staircase mechanisms and showed that two simple staircase mechanisms (the binary and randomized response mechanisms) are optimal in the high and low privacy regimes. In [10], the author mainly studies the tradeoff between local privacy and utility in hypothesis testing. In [5], the authors studied the tradeoff between privacy guarantees and the utility of mean estimation in location.

8. FUTURE WORK

In this paper, we measure the utility preservation in terms of the variance. Several theoretical works on the privacy mechanism design (e.g., [5]) proposed the use of a general utility-maximization framework under differential privacy where the utility function can be a general function depending on the noise added to the query output. We will explore the use of the general function to measure the utility.

Acknowledgments

The authors would like to thank anonymous reviewers for their valuable comments and suggestions. This work was supported in part by U.S. National Science Foundation (DGE-1523115 and IIS-1502273) and U.S. National Institute of Health (1R01GM103309).

9. REFERENCES

[1] S. Agrawal and J. R. Haritsa. A framework for high-accuracy privacy-preserving mining. In *ICDE*, pages 193–204, 2005.

[2] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *PODS*, pages 273–282. ACM, 2007.

[3] A. Chaudhuri and R. Mukerjee. *Randomized response: Theory and techniques*. Marcel Dekker New York, 1988.

[4] K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *NIPS*, pages 289–296. Citeseer, 2008.

[5] J. Duchi, M. Jordan, and M. Wainwright. Local privacy and statistical minimax rates. In *FOCS*, pages 429–438, Oct 2013.

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography*, pages 265–284, 2006.

[7] A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *PODS*, pages 211–222. ACM, 2003.

[8] M. Hardt and K. Talwar. On the geometry of differential privacy. In *STOC*, pages 705–714, 2010, ACM.

[9] Z. Huang and W. Du. Optrr: Optimizing randomized response schemes for privacy-preserving data mining. In *ICDE*, pages 705–714, 2008.

[10] P. Kairouz, S. Oh, and P. Viswanath. Extremal mechanisms for local differential privacy. *CoRR*, 2014.

[11] M. G. Kendall and A. Stuart. The advanced theory of statistics, vol. 2: Hafner. *New York*, page 133, 1969.

[12] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *SIGMOD*, pages 193–204, 2011.

[13] J. Lee and C. Clifton. Differential identifiability. In *KDD*, pages 1041–1049, 2012.

[14] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. Optimizing linear counting queries under differential privacy. In *PODS*, pages 123–134, ACM.

[15] F. McSherry and I. Mironov. Differentially Private Recommender Systems. In *KDD*. ACM, 2009.

[16] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, pages 75–84. ACM, 2007.

[17] V. Rastogi, M. Hay, G. Miklau, and D. Suciu. Relationship privacy: Output perturbation for queries with joins. In *PODS*, pages 107–116. ACM, 2009.

[18] S. J. Rizvi and J. R. Haritsa. Maintaining data privacy in association rule mining. In *VLDB*, 2002.

[19] Y. Wang, X. Wu, and D. Hu. Using Randomized Response for Differential Privacy Preserving Data Collection. In *Technical Report, DPL-2014-003, University of Arkansas*, 2014.

[20] X. Xiao, G. Wang, and J. Gehrke. Differential privacy via wavelet transforms. In *ICDE*, pages 225–236. IEEE, 2010.