Optimal Differentially Private Mechanisms for Randomised Response

Naoise Holohan, Douglas J. Leith, Oliver Mason

Abstract

We examine a generalised Randomised Response (RR) technique in the context of differential privacy and examine the optimality of such mechanisms. Strict and relaxed differential privacy are considered for binary outputs. By examining the error of a statistical estimator, we present closed solutions for the optimal mechanism(s) in both cases. The optimal mechanism is also given for the specific case of the original RR technique as introduced by Warner in 1965.

Index Terms

Randomised response, randomized response, differential privacy, optimality

I. Introduction

A. Background

Stanley L. Warner first proposed the Randomised Response (RR) technique as a means to eliminate bias in surveying in 1965 [31]. Respondents would be handed a spinner by the surveyor to decide which of two questions the respondent would answer, for example,

1) Have you ever cheated on your spouse/partner?
2) Have you always been faithful to your spouse/partner?

Respondents would spin the spinner in private and answer the given question truthfully with a 'yes' or 'no'. Respondents would be afforded *plausible deniability* as the surveyor would not

Naoise Holohan and Douglas J. Leith are with the School of Computer Science and Statistics, Trinity College, Dublin 2, Ireland (e-mail: nholohan@tcd.ie; doug.leith@tcd.ie)

Oliver Mason is with the Department of Mathematics/Hamilton Institute, Maynooth University, Co. Kildare, Ireland & Lero, the Irish Software Research Centre (e-mail: oliver.mason@nuim.ie)

know the question to which the answer refers. This would encourage respondents to engage with the survey and answer the question truthfully. The spinner can be replaced by any appropriate randomisation device, such as coin flips, dice or drawing from a pack of cards.

A rich body of literature now exists on RR. The inefficiencies of Warner's original RR model have been examined by a number of authors and many new RR models have been proposed. These include the unrelated question model [13], the forced response model [2], Moor's procedure [26] and two-stage RR models [25], [24]. More comprehensive lists of RR models can be found in [21], [1].

RR is actively used in surveying when asking questions of a sensitive nature. Examples include surveys on doping and drug use in elite athletes [27], cognitive-enhancing drug use among university students [5], faking on a CV [6], corruption [11], sexual behaviour [3], and child molestation [8].

Researchers remain divided on the effectiveness of RR. While some works have shown RR to be an improvement on different survey techniques, including direct questioning (where no randomisation is involved), [29], [12], [22], [20], [28], others remain sceptical on its advantage [32], [33], [23]. Public trust in RR has also been shown to be lacking [4].

Separately, differential privacy has emerged as a model of interest in privacy-preserving data publishing since being presented in 2006 [7]. Differential privacy gives a quantitative mathematical definition to measure the level of privacy achieved in a given data release. This definition determines the amount of manipulation that needs to be applied to the data to achieve the desired level of privacy. Under differential privacy, privacy is quantified by how statistically indistinguishable the privacy-preserved outputs from two similar datasets are.

When applied to randomised response, where the output from a single individual is binary, differential privacy requires the output from any two individuals to be statistically indistinguishable, to a specified degree.

B. Our Results

In this paper we examine a generalisation of Warner's original RR technique, and establish conditions under which such a model satisfies differential privacy. By calculating the estimator of minimal variance, we determine the optimal differentially private RR mechanism. We examine strict ϵ-differential privacy and relaxed (ϵ,δ)-differential privacy. Complete solutions for the optimal mechanisms are presented for both cases. The optimal mechanism is also given for Warner's RR model satisfying (ϵ, δ)-differential privacy.

C. Related Work

The application of differential privacy to randomised response has been limited to date. [30] examined using randomised response to differentially privately collect data, although their analysis only considered strict ϵ-differential privacy and a comparison of its efficiency with respect to the Laplace mechanism, a mechanism popular in the differential privacy literature.

Randomised response has been used in conjunction with differential privacy in a more general context in the form of *local privacy*, also known as *input perturbation*. For example, extreme mechanisms for local differential privacy have been studied in [18], [16], while differential privacy was applied to social network data in the form of graphs with randomised response in [19]. Outside randomised response and local privacy, optimal mechanisms in differential privacy have received some attention, including work on strict differential privacy [10] and relaxed differential privacy [9].

D. Structure of Paper

We begin in Section II with an introduction to the Randomised Response (RR) technique, and derive the statistical estimator and associated bias and error; we also present Warner's original RR model. We introduce differential privacy in Section III and present a number of preliminary results for later use in Section IV.

The main results are given in Sections V, VI and VII, relating to strict differential privacy, relaxed differential privacy and Warner's model respectively. Concluding remarks are given in Section VIII.

II. Randomised Response

A. Introduction

We are looking to determine the proportion π of people in the population possessing a particular sensitive attribute, where possession of the attribute is binary. We conduct a survey on n individuals of the population by uniform random sampling with replacement.

A single respondent's answer $X_i \in \{0,1\}$ is a randomised version of their truthful answer $x_i \in \{0,1\}$, in order to protect their privacy. The randomised response will therefore not definitively reveal a respondent's truthful answer. By convention, a value of 1 denotes possession of the sensitive attribute, while 0 denotes that the respondent does not possess the attribute. We denote by N the number of randomised responses that return 1, hence $N = \sum_{i \in [n]} X_i$ where $[n] = [1,n] \cap \mathbb{Z}$. We are therefore looking to estimate π from $\frac{N}{n}$.

B. Generalised RR Model

In keeping with standard notation, $(\Omega, \mathcal{F}, \mathbb{P})$ denotes a probability space. $X_i : \Omega \to \{0,1\}$ is then a random variable for each $i \in [n]$, dependent on the truthful value x_i. We define the randomised response mechanism by

$$\mathbb{P}(X_i = k \mid x_i = j) = p_{jk}, \tag{1}$$

which leads us to defining the design matrix of the mechanism as follows.

Definition 1 (Design Matrix)**.** *A randomised response mechanism as defined in (1) is uniquely determined by its design matrix,*

$$P = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}.$$

For the probability mass functions of each X_i to sum to 1, we require $p_{00} + p_{01} = 1$ and $p_{10} + p_{11} = 1$. The design matrix therefore simplifies to

$$P = \begin{pmatrix} p_{00} & 1 - p_{00} \\ 1 - p_{11} & p_{11} \end{pmatrix}, \tag{2}$$

where $p_{00}, p_{11} \in [0,1]$.

As π is the true proportion of individuals in the population possessing the sensitive attribute, we can calculate the probability mass function of each X_i:

$$\begin{aligned} \mathbb{P}(X_i = 0) &= (1-\pi)p_{00} + \pi(1 - p_{11}) \\ &= p_{00} - \pi(p_{00} + p_{11} - 1), \end{aligned} \tag{3a}$$

$$\begin{aligned} \mathbb{P}(X_i = 1) &= \pi p_{11} + (1-\pi)(1 - p_{00}) \\ &= 1 - p_{00} + \pi(p_{00} + p_{11} - 1). \end{aligned} \tag{3b}$$

Remark: Direct questioning corresponds to the case where $p_{00} = p_{11} = 1$.

C. Estimator, Bias and Error

Having presented the RR mechanism previously, we now need to establish an estimator of π from the parameters of the mechanism, p_{00} and p_{11}, and from the distribution of randomised responses, namely $\frac{N}{n}$. We first establish a maximum likelihood estimator (MLE) for the mechanism and then examine its bias and error.

Theorem 1. *Let $p_{00} + p_{11} \neq 1$. Then the MLE for π of the randomised response mechanism given by (2) is*

$$\hat{\Pi}(p_{00}, p_{11}) = \frac{p_{00} - 1}{p_{00} + p_{11} - 1} + \frac{N}{(p_{00} + p_{11} - 1)n}. \tag{4}$$

Proof: Let us first index the sample so that $X_i = 1$ for each $i \leq N$, and $X_i = 0$ for each $i > N$. Then the likelihood L of the sample is

$$L = \mathbb{P}(X_i = 1)^N \mathbb{P}(X_i = 0)^{n-N}.$$

The log-likelihood is

$$\log(L) = N \log \mathbb{P}(X_i = 1) + (n - N) \log \mathbb{P}(X_i = 0),$$

whose derivatives are

$$\begin{aligned} \frac{\partial \log(L)}{\partial \pi} &= \frac{N}{\mathbb{P}(X_i = 1)} \frac{\partial \mathbb{P}(X_i = 1)}{\partial \pi} + \frac{n - N}{\mathbb{P}(X_i = 0)} \frac{\partial \mathbb{P}(X_i = 0)}{\partial \pi}, \\ \frac{\partial^2 \log(L)}{\partial \pi^2} &= -\frac{N}{\mathbb{P}(X_i = 1)^2} \left(\frac{\partial \mathbb{P}(X_i = 1)}{\partial \pi} \right)^2 - \frac{n - N}{\mathbb{P}(X_i = 0)^2} \left(\frac{\partial \mathbb{P}(X_i = 0)}{\partial \pi} \right)^2. \end{aligned}$$

We note that $\frac{\partial^2 \log(L)}{\partial \pi^2} < 0$, hence the maximum of $\log(L)$ occurs when $\frac{\partial \log(L)}{\partial \pi} = 0$. Solving for π completes the proof. ■

We note the following standard identity in probability and statistics,

$$\mathrm{Var}(Y) = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2, \tag{5}$$

for any random variable Y. We now calculate the bias and error of $\hat{\Pi}$. We use the variance of the estimator to characterise error in line with conventional practice. Similarly by convention, we characterise the bias of an estimator as its expected deviation from the quantity it is estimating (i. e. $\mathbb{E}[\hat{\Pi} - \pi]$). We remind the reader of the dependence of $\mathrm{Var}(\hat{\pi})$ on π by writing $\mathrm{Var}(\hat{\Pi}|\pi)$.

Corollary 1. *The MLE $\hat{\Pi}$ constructed in Theorem 1 is unbiased and has error*

$$\operatorname{Var}(\hat{\Pi}(p_{00}, p_{11})|\pi) = \frac{\frac{1}{4} - \left(p_{00} - \frac{1}{2} - \pi(p_{00} + p_{11} - 1)\right)^2}{(p_{00} + p_{11} - 1)^2 n}. \tag{6}$$

Proof: Since the survey we are conducting is by uniform random sampling with replacement, N is a sum of independent and identically distributed random variables. Therefore, $\mathbb{E}[N] = n\mathbb{E}[X_i]$ and $\operatorname{Var}(N) = n \operatorname{Var}(X_i)$.

Since $X_i \in \{0, 1\}$, it can be shown that $\mathbb{E}[X_i] = \mathbb{E}[X_i^2] = \mathbb{P}(X_i = 1) = 1 - p_{00} + \pi(p_{00} + p_{11} - 1)$. Hence,

$$\begin{aligned}\mathbb{E}[\hat{\Pi}] &= \frac{p_{00} - 1}{p_{00} + p_{11} - 1} + \frac{\mathbb{E}[N]}{(p_{00} + p_{11} - 1)n} \\ &= \frac{p_{00} - 1}{p_{00} + p_{11} - 1} + \frac{\mathbb{E}[X_i]}{p_{00} + p_{11} - 1} \\ &= \pi,\end{aligned}$$

and so $\hat{\Pi}$ is unbiased as claimed.

Secondly,

$$\begin{aligned}\operatorname{Var}(\hat{\Pi}|\pi) &= \frac{\operatorname{Var}(N)}{(p_{00} + p_{11} - 1)^2 n^2} \\ &= \frac{\operatorname{Var}(X_i)}{(p_{00} + p_{11} - 1)^2 n} \\ &= \frac{\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2}{(p_{00} + p_{11} - 1)^2 n} \\ &= \frac{\mathbb{P}(X_i = 1)\mathbb{P}(X_i = 0)}{(p_{00} + p_{11} - 1)^2 n},\end{aligned}$$

which can be simplified to (6). ■

When conducting a survey on a population, it is often useful and necessary to estimate the margin of error of the estimate on a sample. For a confidence level $c \in [0, 1]$, the *margin of error* of a sample is given by $\omega \geq 0$, where

$$\mathbb{P}(|\hat{\Pi} - \pi| \leq \omega) \geq c. \tag{7a}$$

In practical applications, a 95% confidence interval is typically used [17]. In the absence of any additional information on the distribution of $\hat{\Pi}$, Chebyshev's inequality can be used to derive a general, but conservative, margin of error, assuming $\hat{\Pi}$ has finite variance. In such a scenario, the margin of error of a sample is given to be 4.5σ, where the standard deviation σ is given by $\sqrt{\mathrm{Var}(\hat{\Pi}|\pi)}$, since

$$\mathbb{P}\left(|\hat{\Pi} - \pi| \leq 4.5\sqrt{\mathrm{Var}(\hat{\Pi}|\pi)}\right) \geq 0.95. \tag{7b}$$

In many practical situations, the central limit theorem is invoked to determine heuristically a margin of error. For a random variable G that is normally distributed with mean μ and variance σ^2, we have

$$\mathbb{P}(|G - \mu| \leq 1.96\sigma) \geq 0.95, \tag{7c}$$

hence 1.96σ is typically taken as the margin of error in such scenarios [17]. However, this non-rigorous approach only gives a loose representation of the margin of error, given that the guarantee of the central limit theorem only applies in the limit as the sample size n approaches infinity.

Due to this variability in defining the margin of error of a sample, we only focus on determining the error of the estimator, $\mathrm{Var}(\hat{\Pi}|\pi)$, in this paper. This error can be used to calculate the margin of error for a particular application, as outlined above.

D. Warner's RR model

Warner's model [31] is a specific case of the generalised model introduced in Section II-B. Warner proposed that surveyors would present respondents with a spinner which they would spin in private to decide which one of two questions to answer. The spinner would point to a question (e. g. "Have you ever cheated on your spouse/partner?") with probability p_w, and to the complement of that question (e. g. "Have you always been faithful to your spouse/partner?") with probability $1 - p_w$. Respondents would then be asked to answer the chosen question truthfully, but without revealing which question they were answering. As before, x_i denotes the truthful response of respondent i, while X_i denotes the randomised response, as determined by the process outlined above.

Warner's model corresponds to the case where $p_{00} = p_{11} = p_w$. We denote by P_w the design matrix of Warner's model, which is given by

$$P_w = \begin{pmatrix} p_w & 1 - p_w \\ 1 - p_w & p_w \end{pmatrix},$$

while the probability mass function of each X_i is defined as

$$\mathbb{P}(X_i = 0) = p_w - \pi(2p_w - 1),$$
$$\mathbb{P}(X_i = 1) = 1 - p_w + \pi(2p_w - 1).$$

Using the same unbiased MLE in (4), we denote by $\hat{\Pi}_w$ the estimator for Warner's model and, by (6), find its error to be

$$\mathrm{Var}(\hat{\Pi}_w(p_w)|\pi) = \frac{\frac{1}{4} - \left(p_w - \frac{1}{2} - \pi(2p_w - 1)\right)^2}{(2p_w - 1)^2 n}. \tag{8}$$

III. Differential Privacy

Differential privacy was first proposed by Dwork in 2006 [7] as a way to measure the level of privacy achieved when publishing data. Using the same notation as in [14], we denote by D^m the space of all m-row datasets (let D be the space of each row) and by $\mathbf{d} \in D^m$ a dataset in this space. We then denote by $X_{\mathbf{d}} : \Omega \to D^n$ a randomised version of $\mathbf{d}$.

If D is assumed to be discrete, the mechanism $X_{\mathbf{d}}$ is said to satisfy (ϵ,δ)-differential privacy if

$$\mathbb{P}(X_{\mathbf{d}} \in A) \leq e^{\epsilon}\mathbb{P}(X_{\mathbf{d}'} \in A) + \delta, \tag{9}$$

for each $\mathbf{d}, \mathbf{d}' \in D^m$ that differ in exactly one row (i.e. there exists exactly one $j \in [m]$ such that $d_j \neq d'_j$) and for each subset $A \subset D^m$.

This set-up simplifies in the case of randomised response introduced in Section II. Firstly, the datasets contain only one row ($m = 1$), and the row-space is $\{0, 1\}$. We are therefore only required to show that (9) holds for $\mathbf{d} \neq \mathbf{d}' \in \{0, 1\}$ and for $A = \{0\}, \{1\}$. Formally, (ϵ,δ)-differential privacy is satisfied if

$$\mathbb{P}(X_i = j) \leq e^{\epsilon}\mathbb{P}(X_k = j) + \delta, \tag{10}$$

for any $i, k \in [n]$ and $j \in \{0, 1\}$.

For the RR mechanism given by (2) to satisfy (ϵ, δ)-differential privacy, we require the following to hold:

$$p_{11} \leq e^{\epsilon}(1 - p_{00}) + \delta, \tag{11a}$$

$$p_{00} \leq e^{\epsilon}(1 - p_{11}) + \delta, \tag{11b}$$

$$1 - p_{00} \leq e^{\epsilon} p_{11} + \delta,$$

$$1 - p_{11} \leq e^{\epsilon} p_{00} + \delta.$$

We can now define the set of pairs (p_{00}, p_{11}) that correspond to a RR mechanism which satisfies (ϵ, δ)-differential privacy.

Definition 2 (Region of Feasibility)**.** *A RR mechanism, given by (2), satisfies (ϵ, δ)-differential privacy if $(p_{00}, p_{11}) \in \mathcal{R}$, where $\mathcal{R} \subset \mathbb{R}^2$ is defined as*

$$\mathcal{R} = \left\{ (p_{00}, p_{11}) \in \mathbb{R}^2 : \begin{array}{l} p_{00}, p_{11} \in [0, 1], \\ p_{00} \leq e^{\epsilon}(1 - p_{11}) + \delta, \\ p_{11} \leq e^{\epsilon}(1 - p_{00}) + \delta, \\ 1 - p_{11} \leq e^{\epsilon} p_{00} + \delta, \\ 1 - p_{00} \leq e^{\epsilon} p_{11} + \delta. \end{array} \right\}. \tag{12}$$

We consider the case where $p_{00} + p_{11} > 1$. Note that the estimator error, and hence the optimal mechanism, is undefined when $p_{00} + p_{11} = 1$. If $p_{00} + p_{11} < 1$, we permute all responses such that $X_i' = 1 - X_i$. This corresponds to the columns of the design matrix being swapped, giving $p_{00}' = 1 - p_{00}$ and $p_{11}' = 1 - p_{11}$, hence $p_{00}' + p_{11}' = 2 - p_{00} - p_{11} > 1$. We can therefore assume $p_{00} + p_{11} > 1$ without loss of generality.

When $p_{00} + p_{11} > 1$, we note that (i) $1 - p_{11} < p_{00} \leq e^{\epsilon}(1 - p_{11}) + \delta < e^{\epsilon} p_{00} + \delta$ and (ii) $1 - p_{00} < p_{11} \leq e^{\epsilon}(1 - p_{00}) + \delta < e^{\epsilon} p_{11} + \delta$. Hence, the region of feasibility simplifies to $\mathcal{R}'$ as follows:

$$\begin{aligned} \mathcal{R}' &= \{(p_{00}, p_{11}) \in \mathcal{R} : p_{00} + p_{11} > 1\} \\ &= \left\{ (p_{00}, p_{11}) \in \mathbb{R} : \begin{array}{l} p_{00}, p_{11} \leq 1, \\ p_{00} + p_{11} > 1, \\ p_{00} \leq e^{\epsilon}(1 - p_{11}) + \delta, \\ p_{11} \leq e^{\epsilon}(1 - p_{00}) + \delta. \end{array} \right\}. \end{aligned}$$

Furthermore, we denote by $\mathcal{R}''$ the boundary of $\mathcal{R}'$ which satisfies at least one of inequalities (11):

$$\mathcal{R}'' = \mathcal{R}' \setminus \left\{ (p_{00}, p_{11}) \in \mathbb{R} : \begin{array}{c} p_{00} < e^{\epsilon}(1 - p_{11}) + \delta, \\ p_{11} < e^{\epsilon}(1 - p_{00}) + \delta. \end{array} \right\}.$$

The set $\mathcal{R}''$ therefore consists of the union of two line segments in the unit square, where (11a) and (11b) are tight.

We are therefore looking to find the RR mechanism which minimises estimator error, while still being (ϵ, δ)-differentially private. Hence, we seek to find

$$\operatorname*{arg\,min}_{(p_{00},p_{11})\in\mathcal{R}'} \operatorname{Var}\left(\hat{\Pi}(p_{00}, p_{11})\middle|\, \pi\right). \tag{13}$$

IV. Preliminary Results

We begin by presenting two results which will be of use later in the paper. The first result concerns the non-negativity of a non-linear function on the unit square.

Lemma 1. *Let* $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *be defined by*

$$f(x, y) = 2xy - x - y + 1.$$

Then, $f(x, y) \geq 0$ *for all* $x, y \in [0, 1]$.

Furthermore,

$$\operatorname*{arg\,min}_{x,y\in[0,1]} f(x, y) = \{(0, 1), (1, 0)\}.$$

Proof: Let's first consider $\min_{x\in[0,1]} f(x, y)$:

$$\begin{aligned} \min_{x\in[0,1]} f(x, y) &= \min_{x\in[0,1]} (2xy - x) - y + 1 \\ &= \min_{x\in[0,1]} ((2y - 1)x) - y + 1 \\ &= \begin{cases} y & \text{if } y \leq \frac{1}{2}, \\ 1 - y & \text{if } y > \frac{1}{2}. \end{cases} \end{aligned} \tag{14}$$

It follows that

$$\min_{y\in[0,1]} \left(\min_{x\in[0,1]} f(x, y) \right) = 0.$$

By symmetry of f, it also follows that

$$\min_{x\in[0,1]}\left(\min_{y\in[0,1]} f(x,y)\right) = 0,$$

hence $f(x,y) \geq 0$ for all $x,y \in [0,1]$.

We note that $f(1,0) = f(0,1) = 0$, and by (14) we see that these values uniquely minimise $f(x,y)$ for all $x,y \in [0,1]$. ■

In the second result of this section we prove that an optimal mechanism exists on $\mathcal{R}''$ (i. e. on the boundary of $\mathcal{R}'$ where at least one of inequalities (11) is tight), and additionally that when $\pi \in (0,1)$, optimal mechanisms only occur on $\mathcal{R}''$.

Lemma 2. *Let* $p_{00} + p_{11} > 1$. *Then there exists* $(p_{00}^*, p_{11}^*) \in \arg\min_{\mathcal{R}'} \operatorname{Var}(\hat{\Pi}|\pi)$ *such that* $(p_{00}^*, p_{11}^*) \in \mathcal{R}''$.

Furthermore, when $0 < \pi < 1$, $\arg\min_{\mathcal{R}'} \operatorname{Var}(\hat{\Pi}|\pi) \subseteq \mathcal{R}''$.

Proof: Let's consider $\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{00}}$ and $\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{11}}$.

Firstly, after some rearranging/manipulation,

$$\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{11}} = -\frac{2p_{00}(1-p_{00})(1-\pi) + \pi(2p_{00}p_{11} - p_{00} - p_{11} + 1)}{(p_{00} + p_{11} - 1)^3 n}.$$

By Lemma 1, we know that $2p_{00}p_{11} - p_{00} - p_{11} + 1 \geq 0$, and since $p_{00} + p_{11} - 1 > 0$ by hypothesis, we conclude that $\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{11}} \leq 0$.

We further note that $2p_{00}p_{11} - p_{00} - p_{11} + 1 > 0$ by Lemma 1, since the assumption that $p_{00} + p_{11} > 1$ means $p_{00}, p_{11} > 0$. Hence $\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{11}} = 0$ only when $\pi = 0$ and $p_{00} = 1$. Equivalently,

$$\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{11}} < 0 \text{ when } \pi > 0 \text{ or } p_{00} < 1. \tag{15}$$

Secondly, after some rearranging/manipulation,

$$\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{00}} = -\frac{(2p_{00}p_{11} - p_{00} - p_{11} + 1)(1-\pi) + 2p_{11}\pi(1-p_{11})}{(p_{00} + p_{11} - 1)^3 n}.$$

Since, by assumption, we have $2p_{00}p_{11} - p_{00} - p_{11} + 1 \geq 0$ and since $p_{11} \in [0,1]$, we see that $\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{00}} \leq 0$.

Similar to the reasoning above, since $2p_{00}p_{11} - p_{00} - p_{11} + 1 > 0$ and $p_{11} > 0$, $\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{00}} = 0$ only when $\pi = 1$ and $p_{11} = 1$. Equivalently,

$$\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{00}} < 0 \text{ when } \pi < 1 \text{ or } p_{11} < 1. \tag{16}$$

Since $\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{00}} \leq 0$ and $\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{11}} \leq 0$, there exists a mechanism on the boundary of $\mathcal{R}'$ which minimises the estimator error, i. e.

$$\partial\mathcal{R}' \cap \left(\underset{(p_{00},p_{11})\in\mathcal{R}'}{\arg\min} \operatorname{Var}(\hat{\Pi}(p_{00},p_{11})|\pi) \right) \neq \emptyset. \tag{17}$$

However, if $0 < \pi < 1$, we see from (15) and (16) that $\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{00}} < 0$ and $\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{11}} < 0$. Hence,

$$\underset{(p_{00},p_{11})\in\mathcal{R}'}{\arg\min} \operatorname{Var}(\hat{\Pi}(p_{00},p_{11})|\pi) \subseteq \partial\mathcal{R}', \tag{18}$$

i. e. the optimal mechanisms *only* occur on the boundary of $\mathcal{R}'$.

Finally, suppose $(p_{00},p_{11}) \in \partial\mathcal{R}'$, but neither of the inequalities in (11) are tight. Then there exist $\Delta_0, \Delta_1 \geq 0$, $\Delta_0 + \Delta_1 > 0$ where $(p_{00}+\Delta_0, p_{11}+\Delta_1) \in \partial\mathcal{R}'$, but because $\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{00}} \leq 0$ and $\frac{\partial \operatorname{Var}(\hat{\Pi}|\pi)}{\partial p_{11}} \leq 0$, then $\operatorname{Var}(\hat{\Pi}(p_{00},p_{11})|\pi) \geq \operatorname{Var}(\hat{\Pi}(p_{00}+\Delta_0,p_{11}+\Delta_1)|\pi)$. Hence minimal error is achieved when at least one of the inequalities (11) is tight, i. e.

$$\underset{(p_{00},p_{11})\in\mathcal{R}'}{\arg\min} \operatorname{Var}(\hat{\Pi}(p_{00},p_{11})|\pi) \subseteq \mathcal{R}''.$$

■

For the remainder of this paper, we assume $\pi \in (0,1)$. Note that the results on optimal mechanisms still hold for $\pi \in [0,1]$, however these optima may not be unique.

V. Optimal Mechanism for ϵ-Differential Privacy

We have already established that the parameters for the optimal (ϵ,δ)-differentially private mechanism lie on $\mathcal{R}''$. We now examine the case of ϵ-differential privacy, where $\delta = 0$, with the additional assumption that $\epsilon > 0$.

Theorem 2. *Let $\pi \in (0,1)$, $p_{00}+p_{11} > 1$ and $\epsilon > 0$. The ϵ-differentially private RR mechanism which minimises estimator error is given by the design matrix*

$$P_\epsilon = \begin{pmatrix} \frac{e^\epsilon}{e^\epsilon+1} & \frac{1}{e^\epsilon+1} \\ \frac{1}{e^\epsilon+1} & \frac{e^\epsilon}{e^\epsilon+1} \end{pmatrix}.$$

Proof: By Lemma 2, we know that the parameters (p_{00},p_{11}) of the optimal mechanism exist on the boundary of $\mathcal{R}'$, with at least one of the inequalities (11) tight. We now separately consider the cases where (11a) and (11b) are tight. By hypothesis, $\delta = 0$ and $\epsilon \neq 0$.

1) (11a) tight: $p_{11} = e^{\epsilon}(1 - p_{00})$, constrained by $p_{11} \geq 0$ and $p_{00} \leq e^{\epsilon}(1 - p_{11})$. By (11b) and since $p_{00} = 1 - e^{-\epsilon}p_{11}$, we have

$$\begin{aligned} e^{\epsilon}p_{11} &\leq e^{\epsilon} - p_{00} \\ &= e^{\epsilon} - (1 - e^{-\epsilon}p_{11}) \\ &= e^{\epsilon} - 1 + e^{-\epsilon}p_{11}, \end{aligned}$$

which we rewrite as

$$p_{11}(e^{\epsilon} - e^{-\epsilon}) \leq e^{\epsilon} - 1,$$

and noting that $e^{2\epsilon} - 1 = (e^{\epsilon} - 1)(e^{\epsilon} + 1)$, we see that

$$\begin{aligned} p_{11} &\leq \frac{e^{\epsilon} - 1}{e^{-\epsilon}(e^{2\epsilon} - 1)} \\ &= \frac{e^{\epsilon}}{e^{\epsilon} + 1}. \end{aligned}$$

We are therefore considering $\mathrm{Var}(\hat{\Pi}(p_{00}, p_{11})|\pi)$ on the line $p_{00} = 1 - e^{-\epsilon}p_{11}$ for $0 \leq p_{11} \leq \frac{e^{\epsilon}}{e^{\epsilon}+1}$. We parametrise this line as follows, where $0 < t \leq 1$, $p_{00} = r(t)$ and $p_{11} = s(t)$ (we require $t > 0$ since $p_{00} + p_{11} > 1$):

$$\begin{aligned} r(t) &= (1 - t) + \frac{e^{\epsilon}}{1 + e^{\epsilon}}t = 1 - e^{-\epsilon}s(t), \\ s(t) &= \frac{e^{\epsilon}}{1 + e^{\epsilon}}t. \end{aligned} \tag{19}$$

For simplicity, we let $\hat{\Pi}(r(t), s(t)) = \hat{\Pi}_1(t)$. After some manipulation, we see that

$$\frac{\partial \, \mathrm{Var}(\hat{\Pi}_1(t)|\pi)}{\partial t} = -\frac{(1 + e^{\epsilon})(1 + \pi(e^{\epsilon} - 1))}{(e^{\epsilon} - 1)^2 t^2 n},$$

and noting that $e^{\epsilon} > 1$, we see that $\frac{\partial \, \mathrm{Var}(\hat{\Pi}_1(t)|\pi)}{\partial t} < 0$. Hence,

$$\operatorname*{arg\,min}_{t \in (0,1]} \mathrm{Var}(\hat{\Pi}_1(t)|\pi) = \{1\}. \tag{20}$$

2) (11b) tight: By symmetry of the equations (11), we simply let $p_{00} = s(t)$ and $p_{11} = r(t)$. By examining (3) and (6), we see that

$$\mathrm{Var}(\hat{\Pi}(p_{00}, p_{11})|1 - \pi) = \mathrm{Var}(\hat{\Pi}(p_{11}, p_{00})|\pi),$$

and by letting $\hat{\Pi}(s(t), r(t)) = \hat{\Pi}_2(t)$, we get

$$\frac{\partial \, \mathrm{Var}(\hat{\Pi}_2(t)|\pi)}{\partial t} = -\frac{(1 + e^{\epsilon})(1 + (1 - \pi)(e^{\epsilon} - 1)}{(e^{\epsilon} - 1)^2 t^2 n}.$$

Again it follows that $\frac{\partial \operatorname{Var}(\hat{\Pi}_2(t)|\pi)}{\partial t} < 0$, and so

$$\operatorname*{arg\,min}_{t\in(0,1]} \operatorname{Var}(\hat{\Pi}_2(t)|\pi) = \{1\}. \tag{21}$$

By (18), (20) and (21), we can now conclude that

$$\operatorname*{arg\,min}_{(p_{00},p_{11})\in\mathcal{R}'} \operatorname{Var}(\hat{\Pi}(p_{00},p_{11})|\pi) = \left\{\left(\frac{e^\epsilon}{e^\epsilon+1}, \frac{e^\epsilon}{e^\epsilon+1}\right)\right\},$$

and so the result follows. ■

Remark: When $\epsilon = 0$, all rows of the design matrix must be identical, i. e. $p_{00} = 1 - p_{11}$ and $p_{11} = 1 - p_{00}$. This gives $p_{00} + p_{11} = 1$, leading to an unbounded estimator error (6). In practical terms, 0-differential privacy enforces the same output distribution for every respondent, hence nothing meaningful can be learned.

VI. Optimal Mechanism for (ϵ, δ)-Differential Privacy

Let's now consider the optimal mechanism for (ϵ, δ)-differential privacy. We parametrise $\mathcal{R}''$ as follows. If we let

$$\begin{aligned} r_\delta(t) &= \left(1 + e^{-\epsilon}\delta\right)(1-t) + \frac{e^\epsilon + \delta}{e^\epsilon + 1} t, \\ &= 1 - e^{-\epsilon}(s_\delta(t) - \delta), \\ s_\delta(t) &= \frac{e^\epsilon + \delta}{e^\epsilon + 1} t, \end{aligned} \tag{22}$$

for $t \in [0,1]$, then the boundary where (11a) holds is parametrised by $p_{00} = r_\delta(t)$ and $p_{11} = s_\delta(t)$; by symmetry, the boundary where (11b) holds is parametrised by $p_{00} = s_\delta(t)$ and $p_{11} = r_\delta(t)$.

We note that $t = 1$ denotes an extreme point of $\mathcal{R}'$ (and $\mathcal{R}''$), the point at which both inequalities (11) are tight. Here $p_{00} = p_{11} = r_\delta(1) = s_\delta(1) = \frac{e^\epsilon+\delta}{e^\epsilon+1}$.

A. *Preliminary Lemmas*

Before proceeding to the main result of this section, we first present a collection of lemmas for later use. The first result states that the minimal variance of $\hat{\Pi}$ on $\mathcal{R}''$ will occur at one of its extreme points (i. e. at one of the endpoints of the two line segments which comprise $\mathcal{R}''$).

Lemma 3. *Let r_δ and s_δ be given by (22), let $\delta > 0$ and let $a \le b \in [0,1]$. Then,*

$$\operatorname*{arg\,min}_{t\in[a,b]} \operatorname{Var}(\hat{\Pi}(r_\delta(t), s_\delta(t))|\pi) \subseteq \{a, b\}.$$

Proof: For simplicity, we denote $\hat{\Pi}(r_\delta(t), s_\delta(t))$ by $\hat{\Pi}_{1,\delta}(t)$.

By some manipulation, it can be shown that the numerator of $\frac{\partial \operatorname{Var}(\hat{\Pi}_{1,\delta}(t)|\pi)}{\partial t}$ is linear in t, hence it has at most one root at

$$t = \frac{\delta(1+e^\epsilon)(2e^\epsilon + 2\delta - 1 - \pi(e^\epsilon + 2\delta - 1))}{(e^\epsilon + \delta)(e^\epsilon + 2\delta - 1)(1 + (e^\epsilon - 1)\pi)}.$$

By substitution, we find that

$$\frac{\partial^2 \operatorname{Var}(\hat{\Pi}_{1,\delta}(t)|\pi)}{\partial t^2} = -\frac{(e^\epsilon + \delta)^2(e^\epsilon + 2\delta - 1)^4(1 + (e^\epsilon - 1)\pi)^4}{8e^{2\epsilon}\delta^3(e^\epsilon + \delta - 1)^3(1 + e^\epsilon)^2 n},$$

when $\frac{\partial \operatorname{Var}(\hat{\Pi}_{1,\delta}(t)|\pi)}{\partial t} = 0$. By inspection, and since $\delta > 0$, we see that $\frac{\partial^2 \operatorname{Var}(\hat{\Pi}_{1,\delta}(t)|\pi)}{\partial t^2} < 0$ when $\frac{\partial \operatorname{Var}(\hat{\Pi}_{1,\delta}(t)|\pi)}{\partial t} = 0$, and so this point is the maximum of $\operatorname{Var}(\hat{\Pi}_{1,\delta}(t)|\pi)$. Hence, the minimum of $\operatorname{Var}(\hat{\Pi}_{1,\delta}(t)|\pi)$ cannot occur at a mid-point of an interval. The result follows. ■

We next show that the error of $\hat{\Pi}$ along the boundary constrained by (11a) is uniformly greater than along the boundary constrained by (11b) when $\pi \le \frac{1}{2}$.

Lemma 4. *Let r_δ and s_δ be given by (22) and let $\delta > 0$. Then, when $\pi \le \frac{1}{2}$,*

$$\operatorname{Var}(\hat{\Pi}(r_\delta(t), s_\delta(t))|\pi) \le \operatorname{Var}(\hat{\Pi}(s_\delta(t), r_\delta(t))|\pi),$$

for $t \in [0,1]$.

Conversely, if $\pi \ge \frac{1}{2}$, then

$$\operatorname{Var}(\hat{\Pi}(r_\delta(t), s_\delta(t))|\pi) \ge \operatorname{Var}(\hat{\Pi}(s_\delta(t), r_\delta(t))|\pi),$$

for $t \in [0,1]$.

Proof: After manipulation of the terms, we can show that

$$\operatorname{Var}(\hat{\Pi}(r_\delta(t), s_\delta(t))|\pi) - \operatorname{Var}(\hat{\Pi}(s_\delta(t), r_\delta(t))|\pi) = -\frac{(e^\epsilon + 1)(e^\epsilon + \delta)(1 - 2\pi)(1 - t)}{(e^\epsilon(e^\epsilon - 1)t + \delta(1 - t + e^\epsilon(1 + t)))n}.$$

We see that $1 - 2\pi \ge 0$ when $\pi \le \frac{1}{2}$, and $1 - 2\pi \le 0$ when $\pi \ge \frac{1}{2}$, and, since $t \in [0,1]$ and $\delta > 0$, the result follows. ■

Finally, we present $t_0(\epsilon, \delta)$ as the t-value which gives the endpoints of the line segments of $\mathcal{R}''$ at the boundary of the unit square.

Lemma 5. *Define $t_0 : \mathbb{R} \times \mathbb{R} \to [0,1]$ by*

$$t_0(\epsilon, \delta) = \frac{\delta(e^\epsilon + 1)}{e^\epsilon + \delta},$$

then,

$$(r_\delta(t_0(\epsilon,\delta)), s_\delta(t_0(\epsilon,\delta))) \in \partial\mathcal{R}'.$$

Proof: By explicit calculation,

$$r_\delta(t_0(\epsilon,\delta)) = 1,$$
$$s_\delta(t_0(\epsilon,\delta)) = \delta.$$

By definition, it follows that $(1,\delta) \in \mathcal{R}' \cup \partial\mathcal{R}'$, and since $p_{00} \leq 1$ is a boundary of $\{(p_{00}, p_{11}) \in \mathcal{R}'\}$, it follows that $(1,\delta) \in \partial\mathcal{R}'$. ■

Remark: When $\delta = 0$, $(r_\delta(t_0(\epsilon,\delta)), s_\delta(t_0(\epsilon,\delta))) \notin \mathcal{R}'$, since we require $r_\delta + s_\delta > 1$.

Remark: By linearity, it follows that $(r_\delta(t), s_\delta(t)) \in \mathcal{R}'$ for all $t_0(\epsilon,\delta) < t \leq 1$, and that $(r_\delta(t), s_\delta(t)) \notin \mathcal{R}'$ when $t < t_0(\epsilon,\delta)$.

B. Main Result

We now present the main results of this paper, which establish the optimal (ϵ,δ)-differentially private RR mechanism(s). The following results assume $\delta > 0$; the optimal mechanism when $\delta = 0$ was presented in Theorem 2. Note that we continue to assume $\pi \in (0,1)$ to ensure uniqueness of the optima.

The following theorem establishes the optimal RR mechanism(s) when $\pi \leq \frac{1}{2}$.

Theorem 3. *Let $\delta > 0$ and $0 < \pi \leq \frac{1}{2}$, and define $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by*

$$g(\epsilon,\delta) = \frac{\delta(e^\epsilon + \delta)}{(e^\epsilon + 2\delta - 1)^2}. \tag{23}$$

Then, for r_δ and s_δ given by (22),

$$\underset{(p_{00},p_{11})\in\mathcal{R}'}{\arg\min}\ \mathrm{Var}(\hat{\Pi}(p_{00},p_{11})|\pi) = \begin{cases} \{(r_\delta(t_0), s_\delta(t_0))\}, & \text{if } g(\epsilon,\delta) > \pi, \\ \{(r_\delta(1), s_\delta(1))\}, & \text{if } g(\epsilon,\delta) < \pi, \\ \{(r_\delta(t_0), s_\delta(t_0)), (r_\delta(1), s_\delta(1))\}, & \text{if } g(\epsilon,\delta) = \pi. \end{cases}$$

where $t_0 = t_0(\epsilon,\delta)$.

Proof: By Lemmas 2, 3 and 4, we know that when $0 < \pi \leq \frac{1}{2}$ and $\delta > 0$,

$$\underset{(p_{00},p_{11})\in\mathcal{R}'}{\arg\min}\ \mathrm{Var}(\hat{\Pi}(p_{00},p_{11})|\pi) \subseteq \{(r_\delta(t_0), s_\delta(t_0)), (r_\delta(1), s_\delta(1))\}.$$

We are therefore considering two candidate points, which can be shown to resolve to

$$r_\delta(t_0) = 1, \qquad s_\delta(t_0) = \delta,$$

$$r_\delta(1) = \frac{e^\epsilon + \delta}{e^\epsilon + 1}, \qquad s_\delta(1) = \frac{e^\epsilon + \delta}{e^\epsilon + 1}.$$

We are therefore seeking to determine the sign of

$$\mathrm{Var}(\hat{\Pi}(1, \delta)|\pi) - \mathrm{Var}\left(\hat{\Pi} \left(\frac{e^\epsilon + \delta}{e^\epsilon + 1}, \frac{e^\epsilon + \delta}{e^\epsilon + 1} \right) \middle| \pi \right). \tag{24}$$

After some manipulation, we can show that (24) simplifies to

$$\frac{(1-\delta)(\pi(e^\epsilon + 2\delta - 1) - \delta(e^\epsilon + \delta))}{\delta(e^\epsilon + 2\delta - 1)^2 n},$$

and we note that its denominator is strictly positive since $\delta > 0$. Note additionally that (24) simplifies to zero when $\delta = 1$, which is trivial since $r_1(t_0) = s_1(t_0) = r_1(1) = s_1(1) = 1$.

The sign of (24) is therefore determined by the sign of $\pi(e^\epsilon + 2\delta - 1) - \delta(e^\epsilon + \delta)$, which gives $g(\epsilon, \delta)$ when solved for π. Hence, $\mathrm{Var}(\hat{\Pi}(r_\delta(t_0), s_\delta(t_0))|\pi) < \mathrm{Var}(\hat{\Pi}(r_\delta(1), s_\delta(1))|\pi)$ when $g(\epsilon, \delta) > \pi$. The other results follow similarly. ■

Remark: When $g(\epsilon, \delta) \leq \pi$, the optimal mechanism corresponds with that established for ϵ-differential privacy on RR (with an added dependence for δ) and also with the optimal mechanism established in Theorem 10 of [15] for mechanisms on categorical data. However, when $g(\epsilon, \delta) > \pi$, the optimal mechanism is one which we have not encountered previously.

The next corollary establishes the optimal mechanism(s) when $\pi \geq \frac{1}{2}$, and follows from Theorem 3 by the symmetry of $\mathrm{Var}(\hat{\Pi}(p_{00}, p_{11})|\pi)$ in p_{00} and p_{11}.

Corollary 2. *Let $\delta > 0$ and $\frac{1}{2} \leq \pi < 1$. Then, for r_δ and s_δ given by (22) and g given by (23),*

$$\operatorname*{arg\,min}_{(p_{00}, p_{11}) \in \mathcal{R}'} \mathrm{Var}(\hat{\Pi}(p_{00}, p_{11})|\pi) = \begin{cases} \{(s_\delta(t_0), r_\delta(t_0))\}, & \text{if } g(\epsilon, \delta) > 1 - \pi, \\ \{(s_\delta(1), r_\delta(1))\}, & \text{if } g(\epsilon, \delta) < 1 - \pi, \\ \{(s_\delta(t_0), r_\delta(t_0)), (s_\delta(1), r_\delta(1))\}, & \text{if } g(\epsilon, \delta) = 1 - \pi, \end{cases}$$

where $t_0 = t_0(\epsilon, \delta)$.

Proof: The result follows from Theorem 3 since

$$\mathrm{Var}(\hat{\Pi}(p_{00}, p_{11})|\pi) = \mathrm{Var}(\hat{\Pi}(p_{11}, p_{00})|1 - \pi).$$

■

Example 1 and Figure 1 illustrate the conclusion of Theorem 3.

Example 1. Consider Theorem 3 and Corollary 2 for various values of ϵ, δ and π. For simplicity, in each of these examples we set $n = 1$.

1) $\epsilon = \frac{1}{2}$, $\delta = \frac{1}{10}$, $\pi = \frac{1}{4}$: In this case, we have $g(\epsilon, \delta) = 0.243 < \pi$. Hence, the design matrix of the optimal mechanism is denoted by
$$\begin{pmatrix} \frac{e^\epsilon+\delta}{e^\epsilon+1} & \frac{1-\delta}{e^\epsilon+1} \\ \frac{1-\delta}{e^\epsilon+1} & \frac{e^\epsilon+\delta}{e^\epsilon+1} \end{pmatrix}.$$
This can be verified by noting that $\mathrm{Var}(\hat{\Pi}(r_\delta(1), s_\delta(1))|\pi) = 2.372$ and $\mathrm{Var}(\hat{\Pi}(r_\delta(t_0), s_\delta(t_0))|\pi) = 2.438$.

2) $\epsilon = 1$, $\delta = \frac{2}{5}$, $\pi = \frac{1}{10}$: In this case, $g(\epsilon, \delta) = 0.197 > \pi$. Hence, the design matrix of the optimal mechanism is denoted by
$$\begin{pmatrix} 1 & 0 \\ 1-\delta & \delta \end{pmatrix}.$$
Again, this can be verified by noting that $\mathrm{Var}(\hat{\Pi}(r_\delta(1), s_\delta(1))|\pi) = 0.385$ and $\mathrm{Var}(\hat{\Pi}(r_\delta(t_0), s_\delta(t_0))|\pi) = 0.24$.

3) $\epsilon = \frac{1}{2}$, $\delta = \frac{1}{3}$, $\pi = \frac{9}{10}$: Since $\pi \geq \frac{1}{2}$, we use Corollary 2 for this example. We note that $g(\epsilon, \delta) = 0.382 > 1 - \pi$. Hence, the design matrix of the optimal mechanism is denoted by
$$\begin{pmatrix} \delta & 1-\delta \\ 0 & 1 \end{pmatrix}.$$
We see that $\mathrm{Var}(\hat{\Pi}(s_\delta(1), r_\delta(1))|\pi) = 0.854$ and $\mathrm{Var}(\hat{\Pi}(s_\delta(t_0), r_\delta(t_0))|\pi) = 0.143$. Note also that $\mathrm{Var}(\hat{\Pi}(r_\delta(0), s_\delta(0))|\pi) = 1.911$, corresponding with the conclusion of Lemma 4

4) $\epsilon = \ln(2), \delta = \frac{1}{4}, \pi = \frac{1}{4}$: In this case, we have $g(\epsilon, \delta) = \frac{1}{4} = \pi$, hence there are two optimal mechanisms,
$$\begin{pmatrix} \frac{e^\epsilon+\delta}{e^\epsilon+1} & \frac{1-\delta}{e^\epsilon+1} \\ \frac{1-\delta}{e^\epsilon+1} & \frac{e^\epsilon+\delta}{e^\epsilon+1} \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1-\delta & \delta \end{pmatrix}.$$
This can be verified by noting that $\mathrm{Var}(\hat{\Pi}(r_\delta(1), s_\delta(1))|\pi) = \mathrm{Var}(\hat{\Pi}(r_\delta(t_0), s_\delta(t_0))|\pi) = \frac{15}{16}$.

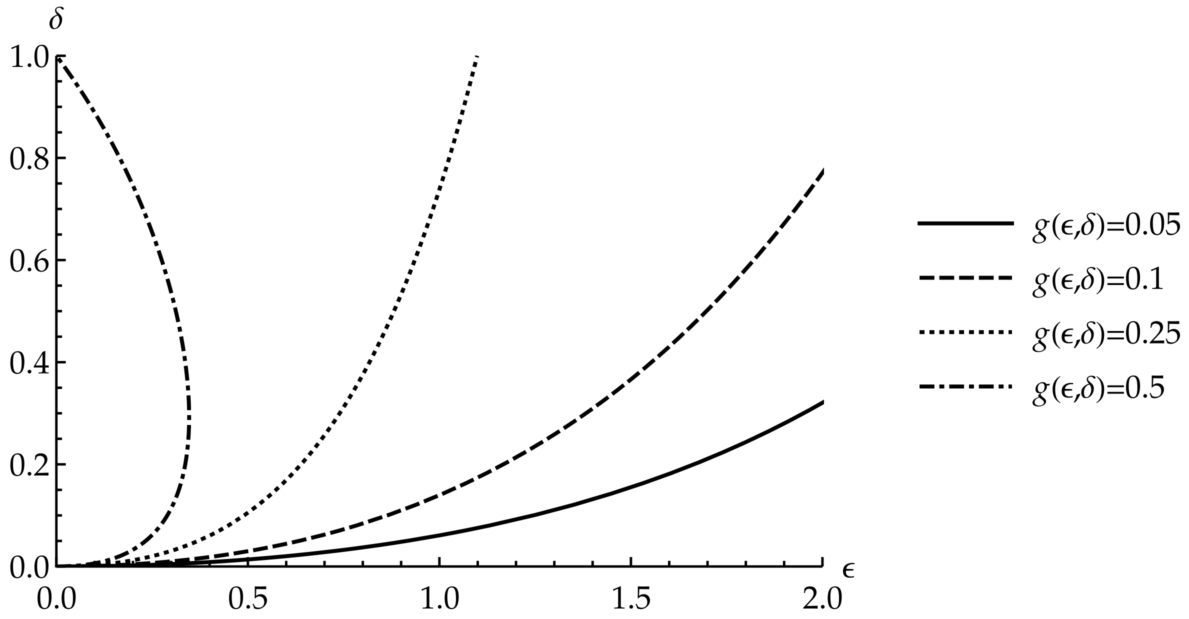

Fig. 1. A contour plot of various level sets of $g(\epsilon, \delta)$. Given π, ϵ and δ, these level sets can be used to determine the optimal (ϵ, δ)-differentially private RR mechanism.

VII. Optimal Warner Mechanism for (ϵ, δ)-Differential Privacy

In the final result of this paper, we examine the optimal mechanism for Warner's RR mechanism. We recall that Warner's mechanism imposed the additional constraint that $p_{00} = p_{11} = p_w$, so the design matrix becomes

$$\begin{pmatrix} p_w & 1 - p_w \\ 1 - p_w & p_w \end{pmatrix}.$$

The error of such a mechanism is only a function of p_w and the population proportion π, as shown in (8).

As before, we require $2p_w > 1$. Our region of feasibility is therefore

$$\mathcal{R}_w = \left(\frac{1}{2}, \frac{e^\epsilon + \delta}{e^\epsilon + 1}\right].$$

Theorem 4. *Consider Warner's RR mechanism as presented in Section II-D. Then,*

$$\underset{p_w \in \mathcal{R}_w}{\arg\min} \operatorname{Var}(\hat{\Pi}_w(p_w)|\pi) = \left\{\frac{e^\epsilon + \delta}{e^\epsilon + 1}\right\}.$$

Proof: By (8), we note that

$$\frac{\partial \operatorname{Var}(\hat{\Pi}_w(p_w)|\pi)}{\partial p_w} = \frac{1}{(1-2p_w)^3 n},$$

hence $\frac{\partial \operatorname{Var}(\hat{\Pi}_w(p_w)|\pi)}{\partial p_w} < 0$ when $p_w > \frac{1}{2}$. Therefore,

$$\operatorname*{arg\,min}_{p_w \in \mathcal{R}_w} \operatorname{Var}(\hat{\Pi}_w(p_w)|\pi) = \max(\mathcal{R}_w),$$

and the result follows. ∎

VIII. Conclusions

We have presented the optimal differentially private RR mechanisms with respect to a maximum likelihood estimator, where both strict and relaxed differential privacy were considered. For a given desired level of privacy, as determined by ϵ and δ, we presented a method to quickly determine the optimal mechanism. This will allow for the optimal implementation of differential privacy in any randomised response survey.

Acknowledgement

The first named author was supported by the Science Foundation Ireland grant SFI/11/PI/1177.

References

[1] Blair, G., Imai, K., and Zhou, Y.-Y. Design and analysis of the randomized response technique. *Journal of the American Statistical Association 110*, 511 (2015), 1304–1319.

[2] Boruch, R. F. Assuring confidentiality of responses in social research: A note on strategies. *The American Sociologist 6*, 4 (1971), 308–311.

[3] Chen, X., Du, Q., Jin, Z., Xu, T., Shi, J., and Gao, G. The randomized response technique application in the survey of homosexual commercial sex among men in Beijing. *Iran J Public Health 43*, 4 (Apr 2014), 416–422. 26005651[pmid].

[4] Coutts, E., and Jann, B. Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods & Research 40*, 1 (2011), 169–193.

[5] Dietz, P., Striegel, H., Franke, A. G., Lieb, K., Simon, P., and Ulrich, R. Randomized response estimates for the 12-month prevalence of cognitive-enhancing drug use in university students. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy 33*, 1 (2013), 44–50.

[6] Donovan, J. J., Dwight, S. A., and Hurtz, G. M. An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique. *Human Performance 16*, 1 (2003), 81–106.

[7] Dwork, C. Differential privacy. In *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 1–12.

[8] FINKELHOR, D., AND LEWIS, I. A. An epidemiologic approach to the study of child molestationa. *Annals of the New York Academy of Sciences 528*, 1 (1988), 64–78.

[9] GENG, Q., AND VISWANATH, P. The optimal mechanism in (ϵ,δ)-differential privacy. *CoRR abs/1305.1330* (2013).

[10] GENG, Q., AND VISWANATH, P. The optimal mechanism in differential privacy. In *Information Theory (ISIT), 2014 IEEE International Symposium on* (2014), IEEE, pp. 2371–2375.

[11] GINGERICH, D. W. Understanding off-the-books politics: Conducting inference on the determinants of sensitive behavior with randomized response surveys. *Political Analysis 18*, 3 (2010), 349–380.

[12] GOODSTADT, M. S., AND GRUSON, V. The randomized response technique: A test on drug use. *Journal of the American Statistical Association 70*, 352 (1975), 814–818.

[13] GREENBERG, B. G., ABUL-ELA, A.-L. A., SIMMONS, W. R., AND HORVITZ, D. G. The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association 64*, 326 (1969), 520–539.

[14] HOLOHAN, N., LEITH, D. J., AND MASON, O. Differential privacy in metric spaces: Numerical, categorical and functional data under the one roof. *Information Sciences 305* (2015), 256–268.

[15] HOLOHAN, N., LEITH, D. J., AND MASON, O. Differentially private response mechanisms on categorical data. *Discrete Applied Mathematics 211* (2016), 86–98.

[16] HOLOHAN, N., LEITH, D. J., AND MASON, O. Extreme Points of the Local Differential Privacy Polytope. *ArXiv e-prints* (May 2016).

[17] JACKMAN, S. Pooling the polls over an election campaign. *Australian Journal of Political Science 40*, 4 (2005), 499–517.

[18] KAIROUZ, P., OH, S., AND VISWANATH, P. Extremal mechanisms for local differential privacy. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2879–2887.

[19] KARWA, V., SLAVKOVIĆ, A. B., AND KRIVITSKY, P. *Differentially Private Exponential Random Graphs*. Springer International Publishing, Cham, 2014, pp. 143–155.

[20] KRUMPAL, I. Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning. *Social Science Research 41*, 6 (2012), 1387–1403.

[21] KRUMPAL, I. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity 47*, 4 (2013), 2025–2047.

[22] LARA, D., STRICKLER, J., OLAVARRIETA, C. D., AND ELLERTSON, C. Measuring induced abortion in Mexico: A comparison of four methodologies. *Sociological Methods & Research 32*, 4 (2004), 529–558.

[23] LARKINS, E. R., HUME, E. C., AND GARCHA, B. S. The validity of the randomized response method in tax ethics research. *Journal of Applied Business Research 13*, 3 (1997), 25–32.

[24] MANGAT, N. S. An improved randomized response strategy. *Journal of the Royal Statistical Society. Series B (Methodological) 56*, 1 (1994), 93–95.

[25] MANGAT, N. S., AND SINGH, R. An alternative randomized response procedure. *Biometrika 77*, 2 (1990), 439–442.

[26] MOORS, J. J. A. Optimization of the unrelated question randomized response model. *Journal of the American Statistical Association 66*, 335 (1971), 627–629.

[27] STRIEGEL, H., ULRICH, R., AND SIMON, P. Randomized response estimates for doping and illicit drug use in elite athletes. *Drug and Alcohol Dependence 106*, 2–3 (2010), 230–232.

[28] TRACY, P. E., AND FOX, J. A. The validity of randomized response for sensitive measurements. *American Sociological Review 46*, 2 (1981), 187–200.

[29] VAN DER HEIJDEN, P. G. M., VAN GILS, G., BOUTS, J., AND HOX, J. J. A comparison of randomized response, computer-assisted self-interview, and face-to-face direct questioning: Eliciting sensitive information in the context of welfare and unemployment benefit. *Sociological Methods & Research 28*, 4 (2000), 505–537.

[30] WANG, Y., WU, X., AND HU, D. Using randomized response for differential privacy preserving data collection. Tech. rep., Technical Report, DPL-2014-003, University of Arkansas, 2014.

[31] WARNER, S. L. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association 60*, 309 (1965), 63–69.

[32] WILLIAMS, B. L., AND SUEN, H. A methodological comparison of survey techniques in obtaining self-reports of condom-related behaviors. *Psychological Reports 75*, 3 suppl (1994), 1531–1537.

[33] WOLTER, F., AND PREISENDRFER, P. Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data. *Sociological Methods & Research 42*, 3 (2013), 321–353.