Optimal Differentially Private Mechanisms for Randomised Response

Naoise Holohan, Douglas J. Leith, *Senior Member, IEEE*, and 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*— Randomized response, differential privacy, local 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 [1]. The central idea is the following. Respondents are handed a *spinner* by the surveyor; they then spin the spinner in private to decide which of two questions to answer. Respondents then answer the given question truthfully with a ‘yes’ or ‘no’. For example, depending on the outcome of the random spin a respondent may answer one of the two questions:

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

The motivation for RR is that respondents are afforded *plausible deniability* as the surveyor would not know the question to which the answer refers. This should encourage respondents to engage with the survey and to answer the question truthfully. Of course, the spinner can be replaced by any appropriate randomisation device, such as coin flips, rolling dice or drawing from a pack of cards.

RR is actively used in surveying when asking questions of a sensitive nature. Examples include surveys on doping and drug use in elite athletes [2], cognitive-enhancing drug use among university students [3], faking on a CV [4], corruption [5], sexual behaviour [6], and child molestation [7].

Since its introduction, many researchers have considered the properties of the basic RR model and its extensions, and a rich body of literature now exists on RR. In particular, inefficiencies in 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 [8], the forced response model [9], Moor’s procedure [10] and two-stage RR models [11], [12]. More comprehensive lists of RR models can be found in [13] and [14].

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), [15]–[19], others remain sceptical on its advantages [20]–[22]. Public trust in RR has also been shown to be lacking in some instances [23]. When considering the effectiveness of RR and public trust in its adoption, there are two key questions to address. First, we need to understand how accurately RR approaches estimate the statistical parameters of interest. The second issue concerns characterising formally and precisely the privacy protections provided by the RR mechanism. The work of this paper is concerned with the fundamental trade-off between the performance of RR estimation mechanisms and the privacy protections they offer.

Differential privacy (DP) has emerged as a leading framework in privacy-preserving data publishing since being introduced in 2006 [24]. Mechanisms for achieving differential privacy are randomised algorithms that provide probabilistic guarantees on the privacy of individual records (corresponding to users) in a database. The core idea underlying these mechanisms is to randomly perturb the correct response to a query so that if one individual changes their record in the database, this has little effect on the (distribution of) the response from the mechanism. Succinctly put: DP mechanisms make it hard to draw inferences about individuals. In its simplest form, when parametrised by a single non-negative ϵ, differential privacy is satisfied when the likelihood of any particular output from a query on two similar datasets does not vary by more than a factor of e^{ϵ}.

Randomised response is essentially a randomised algorithm for data release and, as such, it fits within the framework of differential privacy. In particular, it is natural to quantify the privacy guarantees of RR mechanisms in terms of differential privacy. On the other hand, given that the aim of RR is to estimate unknown population parameters, the performance of RR mechanisms should be characterised in terms of how accurate the estimators built upon them are. The primary contributions of this paper are explicit characterisations of

Manuscript received December 16, 2016; revised April 20, 2017 and May 26, 2017; accepted June 3, 2017. Date of publication July 3, 2017; date of current version August 1, 2017. The work of N. Holohan was supported by the Science Foundation Ireland under Grant SFI/11/PI/1177. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Qian Wang. *(Corresponding author: Naoise Holohan.)*

N. Holohan is with IBM Research, Dublin, Ireland, and also with the School of Computer Science and Statistics, Trinity College, Dublin, Ireland (e-mail: naoise@nholohan.ie).

D. J. Leith is with the School of Computer Science and Statistics, Trinity College, Dublin, Ireland (e-mail: doug.leith@tcd.ie).

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

Digital Object Identifier 10.1109/TIFS.2017.2718487

optimal RR mechanisms subject to formal differential privacy constraints. These results allow us to enforce formal privacy guarantees when choosing parameters for a randomised response technique. 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 on the parameters of this model which ensure that the RR mechanism satisfies differential privacy. This defines a feasibility region for differentially private RR mechanisms. We explicitly characterise the minimal variance estimator for RR mechanisms in terms of the model parameters and, using this, determine the optimal differentially private RR mechanism. Our notion of optimality is based on minimising the variance of the minimal variance estimator over the differential privacy feasibility region. 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. It should be noted that other authors sometimes use alternative terminology such as pure and approximate differential privacy instead of the terms strict and relaxed. We have chosen the terms 'strict' and 'relaxed' in line with [25] in which the $\delta > 0$ case is referred to as a relaxation of ϵ-differential privacy.

C. Related Work

The application of differential privacy to randomised response has been limited to date. [26] 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. In a recent revision of a paper posted to the *arXiv* repository [27], the authors have shown how to use RR to define optimal differentially private mechanisms for a broad class of private multi-party computation problems. Specifically, [27, Th. 5.1] demonstrates that an RR mechanism using a larger output space (4 values rather than 2) for each party in the computation (loosely corresponding to respondents in our setting) maximises accuracy for arbitrary accuracy functions and local and central computation. The primary focus of that paper was on the behaviour of differentially private mechanisms under composition and RR is considered as an approach to differentially private multi-party computation. In contrast, our focus is on the characterising the privacy protections offered by RR as a survey and estimation methodology and on describing optimal differentially private RR estimators.

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 [28] and [29], while differential privacy was applied to social network data in the form of graphs with randomised response in [30]. Outside the settings of randomised response and local privacy, optimal mechanisms in differential privacy have received some attention, including work on strict differential privacy [31] and relaxed differential privacy [32]. The work of [33] introduces *staircase mechanisms* which define optimal differentially private mechanisms for cost functions based on the l_1 norm and real-valued data, rather than the discrete, binary-valued data considered here. This latter result applies to strict differential privacy ($\delta = 0$) and is shown rigorously for queries taking values in $\mathbb{R}^2$; a generalisation to arbitrary dimensions is also demonstrated but is contingent on a technical conjecture being true.

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

We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ the usual 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:

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

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 describe a maximum likelihood estimator (MLE) for the mechanism and then examine its bias and error. While the proofs of the following two results are relatively straightforward, we include them here in the interests of completeness and to make the paper as self-contained as possible.

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

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

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

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

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 $\mathrm{Var}(N) = n\,\mathrm{Var}(X_i)$.

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,

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

which can be simplified to (6). □

When conducting a survey on a population, it may be necessary to calculate the margin of error of the estimate. 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 [34]. Using Chebyshev's inequality, we can calculate the margin of error of a sample 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 [34]. 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 [1] 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.

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 [24] as a means of quantifying the level of privacy achieved when publishing data via randomised algorithms or mechanisms. Using the same notation as in [35], 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 $X_{\mathbf{d}} : \Omega \to D^m$ 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. $\exists! j \in [m] : d_j \neq d'_j$) 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 ($m = 1$) (corresponding to a single respondent), 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:

$$\begin{aligned} p_{11} &\leq e^{\epsilon}(1 - p_{00}) + \delta, & \text{(11a)} \\ p_{00} &\leq e^{\epsilon}(1 - p_{11}) + \delta, \\ 1 - p_{00} &\leq e^{\epsilon} p_{11} + \delta, \\ 1 - p_{11} &\leq e^{\epsilon} p_{00} + \delta. & \text{(11b)} \end{aligned}$$

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

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

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:

$$\begin{aligned} \mathcal{R}' &= \{(p_{00}, p_{11}) \in \mathcal{R} : p_{00} + p_{11} > 1\} \\ &= \left\{ (p_{00}, p_{11}) \in \mathbb{R}^2 : \begin{array}{l} p_{00}, p_{11} \leq 1, \\ p_{00} + p_{11} > 1, \\ p_{00} \leq e^{\epsilon}(1 - p_{11}) + \delta, \\ p_{11} \leq e^{\epsilon}(1 - p_{00}) + \delta. \end{array} \right\}. \end{aligned}$$

Remark: In [27] an operational view on differential privacy in the context of hypothesis testing was presented. In particular, in Theorem 2.1 of this reference two inequalities characterising differentially private mechanisms in terms of type I and type II errors are derived. Kairouz *et al.* [27] state their result in terms of probabilities of false alarm and missed detection, denoted by P_{FA} and P_{MD} respectively. If we identify p_{01} with P_{FA} and p_{10} with P_{MD}, then we see

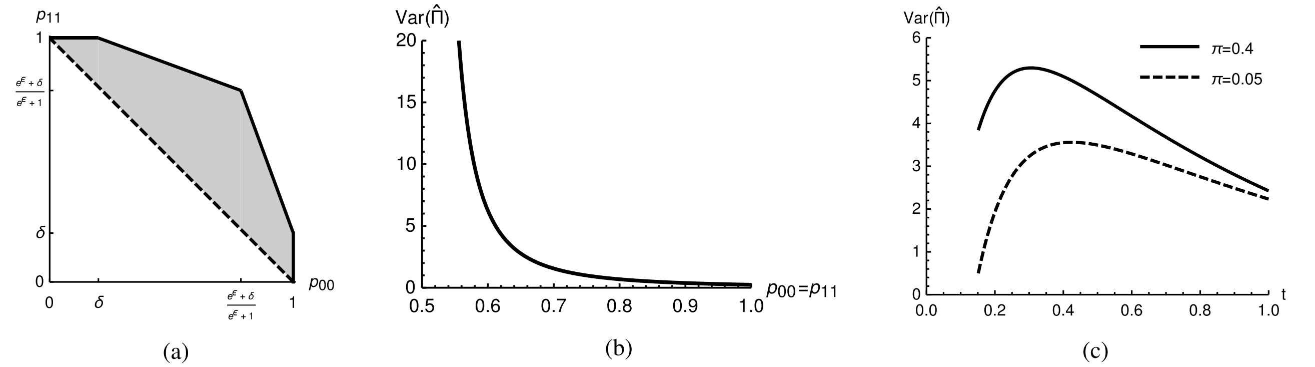


Fig. 1. The problem being considered in this paper is outlined in the above plots. We are seeking to minimise the estimator error, $\mathrm{Var}(\hat{\Pi})$ on the region of feasibility $\mathcal{R}'$, shown in (a). The estimator error is shown to be decreasing on the diagonal centerline, $p_{00} = p_{11}$, in (b). However, on the boundary $p_{11} = e^{\epsilon}\ (1 - p_{00}) + \delta$ (which is shown later in Lemma 2 to contain the optimum), (c) shows that the optimal mechanism depends on π.

that the region $\mathcal{R}'$ given above is identical to the feasibility region described in [27].

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

$$\mathcal{R}'' = \mathcal{R}' \setminus \left\{ (p_{00}, p_{11}) \in \mathbb{R}^2 : \begin{array}{l} p_{00} < e^{\epsilon}(1 - p_{11}) + \delta, \\ p_{11} < e^{\epsilon}(1 - p_{00}) + \delta. \end{array} \right\}.$$

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

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

Figure 1 gives a graphical illustration of the central question being considered in this paper.

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,

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

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 additionaly that when $\pi \in\ (0, 1)$, optimal mechanisms only occur on $\mathcal{R}''$.

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

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

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

Firstly, after some rearranging/manipulation,

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

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 $\frac{\partial\, \mathrm{Var}(\hat{\Pi}|\pi)}{\partial p_{11}} \leq 0$.

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 $\frac{\partial\, \mathrm{Var}(\hat{\Pi}|\pi)}{\partial p_{11}} = 0$ only when $\pi = 0$ and $p_{00} = 1$. Equivalently,

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

Secondly, after some rearranging/manipulation,

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

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 $\frac{\partial\, \mathrm{Var}(\hat{\Pi}|\pi)}{\partial p_{00}} \leq 0$.

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.

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

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,

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

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 inequalities (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.

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

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

Using the results of Section IV, we can now establish results on the optimal randomised response mechanism for differential privacy. We begin in this section by examining strict ϵ-differential privacy, where $\delta = 0$. In Section VI we consider relaxed (ϵ,$\delta > 0$)-differential privacy. We then briefly consider the Warner RR mechanism, where $p_{00} = p_{11}$, in Section VII.

Using Lemma 2 of the previous section, we establish the optimal RR mechanism for ϵ-differential privacy in the following result.

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 $\operatorname{Var}(\hat{\Pi}(p_{00}, p_{11})|\pi)$ 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

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

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

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

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

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

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

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

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

We have now established the optimal mechanism for ϵ-differential privacy. Now, let's consider the case of (ϵ, δ)-differential privacy, where $\delta \in (0, 1]$ is non-zero.

As before, we first parametrise $\mathcal{R}''$. 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 $t = 1$ corresponds to 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 an interior 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

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

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,

$$\begin{aligned} r_\delta(t_0(\epsilon, \delta)) &= 1, \\ s_\delta(t_0(\epsilon, \delta)) &= \delta. \end{aligned}$$

By definition, it follows that $(1, \delta) \in \mathcal{R}' \cup \partial\mathcal{R}'$, and since $p_{00} \le 1$ 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 $t_0(\epsilon, \delta) < t \le 1$, 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 establishes 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 $\pi \le \frac{1}{2}$.

Theorem 3: Let $\delta > 0$ and $0 < \pi \le \frac{1}{2}$, 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),

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

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 $\delta > 0$,

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

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

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

We are therefore seeking to determine the sign of

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

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, $\operatorname{Var}(\hat{\Pi}(r_\delta(t_0), s_\delta(t_0))|\pi) < \operatorname{Var}(\hat{\Pi}(r_\delta(1), s_\delta(1))|\pi)$ 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 [36, Th. 10] 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 $\operatorname{Var}(\hat{\Pi}(p_{00},p_{11})|\pi)$ 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),

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

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

Proof: The result follows from Theorem 3 since

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

C. Discussion

We note that the result of Theorem 3 depends on knowing the true value of π to determine the optimal mechanism to estimate π. Clearly, this has the potential to limit the applicability of the result and appears to involve circularity. However, in many practical scenarios, an approximate range for π is likely to be known in advance and this will often be sufficient to allow us to determine the optimal mechanism using Theorem 3, given ϵ and δ.

The approximate range for π will impact on how Theorem 3 is used in practice. To see how this would work, suppose we know a lower bound l and an upper bound u for π so that $l \leq \pi \leq u$. The theorem describes explicitly the optimal mechanisms for all values of the privacy parameters ϵ,δ satisfying $g(\epsilon,\delta) > u$ and $g(\epsilon,\delta) < l$. One way of viewing the result is thus the following. Once bounds l, u are known for π, we can select a desired privacy level (specified by ϵ,δ) satisfying either $g(\epsilon,\delta) > u$ or $g(\epsilon,\delta) < l$ and then determine the optimal mechanism for this level of privacy. For example, if δ is small, i. e. 10^{-5}, then the optimal mechanism will be the ϵ-differential privacy optimal mechanism.

Alternatively, the privacy parameters ϵ,δ may be fixed independently of π and its bounds l, u (e. g. if specified by law). In such circumstances, when $g(\epsilon,\delta) < l$ or $g(\epsilon,\delta) > u$, the optimal mechanism can easily be determined. If $g(\epsilon,\delta) \in [l,u]$, we cannot be certain of the optimal mechanism using Theorem 3. One heuristic approach which could be applied in this scenario is to approximate π, using the midpoint $\frac{l+u}{2}$ for instance, to select the optimal mechanism according to the theorem. Determining precisely the performance gap resulting from an incorrect decision made in this way is an interesting topic for future work.

Example 1 and Figure 2 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 $\operatorname{Var}(\hat{\Pi}(r_\delta(1), s_\delta(1))|\pi) = 2.372$ and $\operatorname{Var}(\hat{\Pi}(r_\delta(t_0), s_\delta(t_0))|\pi) = 2.438$.

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

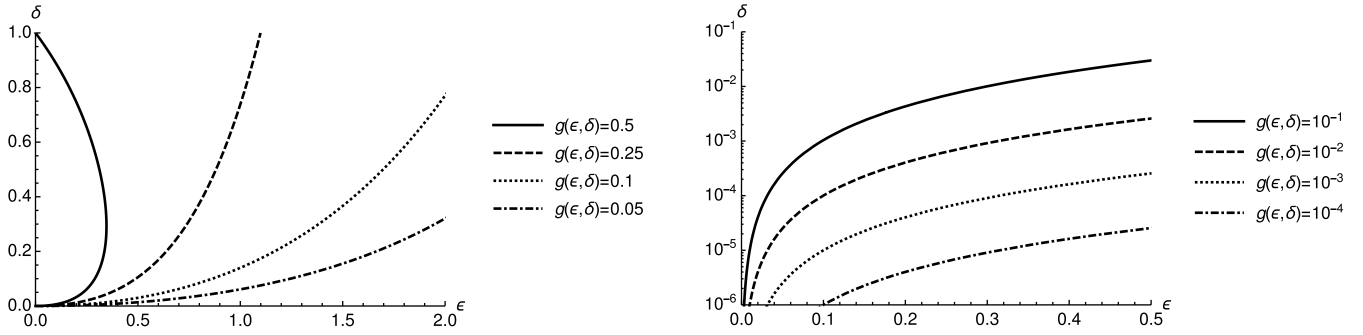
$$\begin{pmatrix} 1 & 0 \\ 1-\delta & \delta \end{pmatrix}.$$

Again, this can be verified by noting that $\operatorname{Var}(\hat{\Pi}(r_\delta(1), s_\delta(1))|\pi) = 0.385$ and $\operatorname{Var}(\hat{\Pi}(r_\delta(t_0), s_\delta(t_0))|\pi) = 0.24$.

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 $\operatorname{Var}(\hat{\Pi}(s_\delta(1), r_\delta(1))|\pi) = 0.854$ and $\operatorname{Var}(\hat{\Pi}(s_\delta(t_0), r_\delta(t_0))|\pi) = 0.143$. Note also that



(a)

(b)

Fig. 2. Linear and log contour plots of various level sets of $g(\epsilon,\delta)$ for various ranges of ϵ and δ. Given π, ϵ and δ, these level sets can be used to determine the optimal (ϵ,δ)-differentially private RR mechanism. (a) $\epsilon \in [0, 2]$, $\delta \in [0, 1]$; (b) $\epsilon \in \left[0, \frac{1}{2}\right]$, $\delta \in \left[10^{-6}, 10^{-1}\right]$.

$\mathrm{Var}(\hat{\Pi}(r_\delta(0), s_\delta(0))|\pi) = 1.911$, 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 $\mathrm{Var}(\hat{\Pi}(r_\delta(1), s_\delta(1))|\pi) = \mathrm{Var}(\hat{\Pi}(r_\delta(t_0), s_\delta(t_0))|\pi) = \frac{15}{16}$.

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,

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

Proof: By (8), we note that

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

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

$$\arg\min_{p_w \in \mathcal{R}_w} \mathrm{Var}(\hat{\Pi}_w(p_w)|\pi) = \max\ (\mathcal{R}_w),$$

and the result follows. □

VIII. Conclusion

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. The results here all concern the simple RR model in which responses are binary-valued. For numerous practical applications, an extension to the case of multi-valued responses will be needed. Such an extension would first require a characterisation of the minimal variance estimator for such mechanisms. Results such as those presented in [35] for discrete, categorical data are then likely to be useful in developing appropriate extensions of the work here. This is an ongoing line of work and the authors hope to be able to report more general results in the near future.

References

[1] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. Amer. Statist. Assoc.*, vol. 60, no. 309, pp. 63–69, 1965.

[2] H. Striegel, R. Ulrich, and P. Simon, "Randomized response estimates for doping and illicit drug use in elite athletes," *Drug Alcohol Dependence*, vol. 106, nos. 2–3, pp. 230–232, 2010. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0376871609003202

[3] P. Dietz, H. Striegel, A. G. Franke, K. Lieb, P. Simon, and R. Ulrich, "Randomized response estimates for the 12-month prevalence of cognitive-enhancing drug use in university students," *Pharmacotherapy, J. Human Pharmacol. Drug Therapy*, vol. 33, no. 1, pp. 44–50, 2013. [Online]. Available: http://dx.doi.org/10.1002/phar.1166

[4] J. J. Donovan, S. A. Dwight, and G. M. Hurtz, "An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique," *Human Perform.*, vol. 16, no. 1, pp. 81–106, 2003.

[5] D. W. Gingerich, "Understanding off-the-books politics: Conducting inference on the determinants of sensitive behavior with randomized response surveys," *Political Anal.*, vol. 18, no. 3, pp. 349–380, 2010. [Online]. Available: http://pan.oxfordjournals.org/content/18/3/349.abstract

[6] X. Chen, Q. Du, Z. Jin, T. Xu, J. Shi, and G. Gao, "The randomized response technique application in the survey of homosexual commercial sex among men in Beijing," *Iranian J. Public Health*, vol. 43, no. 4, pp. 416–422, Apr. 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4433722/

[7] D. Finkelhor and I. A. Lewis, "An epidemiologic approach to the study of child molestationa," *Ann. New York Acad. Sci.*, vol. 528, no. 1, pp. 64–78, 1988. [Online]. Available: http://dx.doi.org/10.1111/j.1749-6632.1988.tb50852.x

[8] B. G. Greenberg, A.-L. A. Abul-Ela, W. R. Simmons, and D. G. Horvitz, "The unrelated question randomized response model: Theoretical framework," *J. Amer. Statist. Assoc.*, vol. 64, no. 326, pp. 520–539, 1969. [Online]. Available: http://www.jstor.org/stable/2283636

[9] R. F. Boruch, "Assuring confidentiality of responses in social research: A note on strategies," *Amer. Sociol.*, vol. 6, no. 4, pp. 308–311, 1971. [Online]. Available: http://www.jstor.org/stable/27701807

[10] J. J. A. Moors, "Optimization of the unrelated question randomized response model," *J. Amer. Statist. Assoc.*, vol. 66, no. 335, pp. 627–629, 1971. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482320

[11] N. S. Mangat and R. Singh, "An alternative randomized response procedure," *Biometrika*, vol. 77, no. 2, pp. 439–442, 1990.

[12] N. S. Mangat, "An improved randomized response strategy," *J. Roy. Statist. Soc. Ser. B, Methodol.*, vol. 56, no. 1, pp. 93–95, 1994. [Online]. Available: http://www.jstor.org/stable/2346030

[13] I. Krumpal, "Determinants of social desirability bias in sensitive surveys: A literature review," *Quality Quantity*, vol. 47, no. 4, pp. 2025–2047, 2013. [Online]. Available: http://dx.doi.org/10.1007/s11135-011-9640-9

[14] G. Blair, K. Imai, and Y.-Y. Zhou, "Design and analysis of the randomized response technique," *J. Amer. Statist. Assoc.*, vol. 110, no. 511, pp. 1304–1319, 2015. [Online]. Available: http://dx.doi.org/10.1080/01621459.2015.1050028

[15] P. G. M. van der Heijden, G. van Gils, J. Bouts, and J. J. Hox, "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," *Sociol. Methods Res.*, vol. 28, no. 4, pp. 505–537, 2000. [Online]. Available: http://smr.sagepub.com/content/28/4/505.abstract

[16] M. S. Goodstadt and V. Gruson, "The randomized response technique: A test on drug use," *J. Amer. Statist. Assoc.*, vol. 70, no. 352, pp. 814–818, 1975. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/01621459.1975.10480307

[17] D. K. Lara, J. Strickler, C. D. Olavarrieta, and C. Ellertson, "Measuring induced abortion in Mexico: A comparison of four methodologies," *Sociol. Methods Res.*, vol. 32, no. 4, pp. 529–558, 2004. [Online]. Available: http://smr.sagepub.com/content/32/4/529.abstract

[18] I. Krumpal, "Estimating the prevalence of xenophobia and anti-Semitism in Germany: A comparison of randomized response and direct questioning," *Soc. Sci. Res.*, vol. 41, no. 6, pp. 1387–1403, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0049089X12001172

[19] P. E. Tracy and J. A. Fox, "The validity of randomized response for sensitive measurements," *Amer. Sociol. Rev.*, vol. 46, no. 2, pp. 187–200, 1981. [Online]. Available: http://www.jstor.org/stable/2094978

[20] B. L. Williams and H. Suen, "A methodological comparison of survey techniques in obtaining self-reports of condom-related behaviors," *Psychol. Rep.*, vol. 75, no. 3, pp. 1531–1537, 1994. [Online]. Available: http://prx.sagepub.com/content/75/3_suppl/1531.abstract

[21] F. Wolter and P. Preisendörfer, "Asking sensitive questions: An evaluation of the randomized response technique versus direct questioning using individual validation data," *Sociol. Methods Res.*, vol. 42, no. 3, pp. 321–353, 2013. [Online]. Available: http://smr.sagepub.com/content/42/3/321.abstract

[22] E. R. Larkins, E. C. Hume, and B. S. Garcha, "The validity of the randomized response method in tax ethics research," *J. Appl. Bus. Res.*, vol. 13, no. 3, pp. 25–32, 1997. [Online]. Available: http://elib.tcd.ie/login?url=http://search.proquest.com/docview/227596424?accountid=14404

[23] E. Coutts and B. Jann, "Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT)," *Sociol. Methods Res.*, vol. 40, no. 1, pp. 169–193, 2011. [Online]. Available: http://smr.sagepub.com/content/40/1/169.abstract

[24] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*. Berlin, Germany: Springer, 2006, pp. 1–12.

[25] C. Dwork and G. N. Rothblum. (Mar. 2016). "Concentrated differential privacy." [Online]. Available: http://arxiv.org/abs/1603.01887

[26] Y. Wang, X. Wu, and D. Hu, "Using randomized response for differential privacy preserving data collection," Faculty Comput. Sci. Comput. Eng., Univ. Arkansas, Fayetteville, AR, USA, Tech. Rep. DPL-2014-003, 2014.

[27] P. Kairouz, S. Oh, and P. Viswanath. (Nov. 2013). "The composition theorem for differential privacy." [Online]. Available: https://arxiv.org/abs/1311.0776

[28] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in *Advances in Neural Information Processing Systems*, vol. 27. Red Hook, NY, USA: Curran Associates, Inc., 2014, pp. 2879–2887. [Online]. Available: http://papers.nips.cc/paper/5392-extremal-mechanisms-for-local-differential-privacy.pdf

[29] N. Holohan, D. J. Leith, and O. Mason. (May 2016). "Extreme points of the local differential privacy polytope." [Online]. Available: https://arxiv.org/abs/1605.05510

[30] V. Karwa, A. B. Slavković, and P. Krivitsky, *Differentially Private Exponential Random Graphs*. Cham, Switzerland: Springer, 2014, pp. 143–155. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11257-2_12

[31] Q. Geng and P. Viswanath, "The optimal mechanism in differential privacy," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2014, pp. 2371–2375.

[32] Q. Geng and P. Viswanath, "The optimal mechanism in (ϵ,δ)-differential privacy," *CoRR*, vol. abs/1305.1330, May 2013. [Online]. Available: http://arxiv.org/abs/1305.1330

[33] Q. Geng, P. Kairouz, S. Oh, and P. Viswanath, "The staircase mechanism in differential privacy," *IEEE J. Sel. Topics Signal Process.*, vol. 9, no. 7, pp. 1176–1184, Oct. 2015.

[34] S. Jackman, "Pooling the polls over an election campaign," *Austral. J. Political Sci.*, vol. 40, no. 4, pp. 499–517, 2005.

[35] N. Holohan, D. J. Leith, and O. Mason, "Differential privacy in metric spaces: Numerical, categorical and functional data under the one roof," *Inf. Sci.*, vol. 305, pp. 256–268, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025515000596

[36] N. Holohan, D. J. Leith, and O. Mason, "Differentially private response mechanisms on categorical data," *Discrete Appl. Math.*, vol. 211, pp. 86–98, Oct. 2016.

Naoise Holohan, photograph and biography not available at the time of publication.

Douglas J. Leith, photograph and biography not available at the time of publication.

Oliver Mason, photograph and biography not available at the time of publication.