RESEARCH REPORT SERIES
(Statistics #2019-08)

A Local *l*-Diversity Mechanism for Privacy Protected Categorical Data Collection

Tapan K. Nayak[1],
Xiaoyu Zhai[2]

[1]Center for Statistical Research and Methodology, U.S. Census Bureau
and Department of Statistics, George Washington University;
[2]Department of Statistics, George Washington University

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: October 22, 2019

A Local l-Diversity Mechanism for Privacy Protected Categorical Data Collection*

Tapan K. Nayak† and Xiaoyu Zhai‡

Abstract

We consider the task of protecting respondent's privacy when collecting data on categorical variables. Any mechanism for masking the true value of a respondent can be viewed as a randomized response (RR) procedure, and its prudent planning depends crucially on the given privacy criterion. We examine some existing privacy criteria and describe their drawbacks. We show that a previous notion of average security is inappropriate. Several other criteria, which simply impose upper bounds on the parity of the RR design, inflict severe data utility loss, unless the number of categories is fairly small. This applies to local differential privacy (LDP), which is a leading privacy criterion, and reveals substantial statistical inefficiency of the RAPPOR procedure, which has been in use by Google, Apple and others. We propose a new privacy procedure that is similar to l-diversity but, works locally for each respondent. The procedure is simple to implement and its privacy protection is easy to understand and communicate to survey participants. We give an unbiased estimator of the probability vector of all categories and prove its minimaxity within a class of estimators under squared error loss. We argue and believe that the new procedure offers a better privacy-utility trade-off than LDP.

Key words and Phrases: Attribute disclosure; local differential privacy; minimaxity; randomized response; RAPPOR algorithm.

*The views expressed in this article are those of the authors and not those of the U.S. Census Bureau.

†Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC 20233 and Department of Statistics, George Washington University, Washington, DC 20052.

‡Department of Statistics, George Washington University, Washington, DC 20052.

1. Introduction

The collection, analysis and sharing of data have now become central to public policy, business decisions, research and other areas. The primary goal of statistical activities is to gain knowledge, but we also need to protect respondents' privacy in order to meet privacy regulations and to uphold public trust. Public awareness and concerns about privacy have grown significantly in recent years, which have drawn considerable interest to privacy research from statisticians, computer scientists and others. Various privacy concepts, measures and methods, such as identity disclosure, differential privacy, k-anonymity, l-diversity, data swapping, cell suppression, synthetic data, randomized response, have been proposed; see Willenborg and de Waal (2001), Chen et al. (2009), Hundepool et al. (2012), Torra (2017) and Venkataramanan and Shriram (2017) for systematic discussions and references.

We should distinguish between privacy and confidentiality, which have often been used synonymously. Both are about concealing or masking information about each respondent or survey unit, so that the data may be used only to learn about the population as a whole and not about any specific unit. However, privacy appears at the time of data collection whereas confidentiality arises after data collection and they have some distinct features. Many data sets contain the *true values* or characteristics of some units, e.g., individuals, families, businesses etc. In particular, data obtained from administrative records, transactions, on-line postings, searches and many surveys contain respondents' true values. Concerns about confidentiality arise when we want to release or share such data (or even summaries), as we do not want the released data to reveal much information about any unit in the data set. Consequently, data agencies often release only high-level summary statistics or a perturbed version of the actual data.

Privacy is an individual's right to control access to his/her information, even with respect to the data collector. To protect privacy, each survey unit's true values need to be masked *locally* before reporting to the data collector. So, the data would contain only imprecise or masked

values of the units in the data set. Evidently, the masking mechanism must be chosen before the data are collected. In contrast, a method for confidentiality protection may be chosen based on the observed data. Actually, many confidentiality protection methods, such as generalization, suppression, data swapping and creating synthetic data, are data dependent. In another view, confidentiality protection involves *output perturbation* whereas privacy protection requires *input perturbation*.

One serious confidentiality breach in releasing microdata is *identity disclosure*, which occurs when a unit's records in released data can be identified by matching the values of some variables, called key variables or pseudo-identifiers, that can be easily obtained from other sources. Clearly, when a unit is identified, one can learn its values for all variables. To control identification risks, Samarati (2001) and Sweeney (2002) introduced the *k-anonymity* criterion, which requires that each record in the released data be identical to at least $k-1$ other records with respect to the key variables. So, any unit will have at least k matches by the key variables. Algorithms for achieving k-anonymity mainly use generalization and suppression of the original values. Other approaches to defining and controlling identification risks have been discussed in Bethlehem et al. (1990), Skinner and Elliott (2002), Shlomo and Skinner (2010) and Nayak et al. (2018), among others.

Machanavajjhala et al. (2007) noted that k-anonymity is inadequate for protecting against *attribute disclosure*, which occurs when a unit's true value of a sensitive variable is learned, using key variables matching. Indeed, a target unit's value of a sensitive variable (e.g., disease status) would be revealed if all of its matching units have the same sensitive variable value (e.g., cancer). For addressing this problem, Machanavajjhala et al. (2007) proposed *l-diversity*, which requires that in released data, each set of matching units with respect to the key variables must contain at least l "well represented" values for the sensitive variable. They also gave practical definitions of "well represented" and showed that l-diversity can be achieved using generalization and suppression. Domingo-Ferrer et al. (2019) presented a unified approach for achieving k-

anonymity, l-diversity and other privacy requirements.

Usually, each respondent's identity is known to the data collector. So, privacy protection mechanisms aim to control attribute disclosure for each unit assuming that its identity is already disclosed. The goal is to assure that the true values of a unit cannot be ascertained with high certainty from their masked values. Thus, a data set compiled after giving adequate privacy protection should not need confidentiality protection. For this reason, even when the true values are available, a data collector may prefer to mask those before recording. This may be particularly relevant to companies when collecting information from on-line transactions, searches, postings etc. Moreover, the masking task can be delegated to computers for both convenience and to avoid anyone looking at the original values.

In this paper, we consider privacy protection when collecting data on *categorical variables*. To describe our context, let X be a categorical survey variable or a cross-classification of several variables. Let $\mathcal{S}_X = \{c_1, \ldots, c_k\}$ be the range of X, $\pi_i = P(X = c_i), i = 1, \ldots, k$, and $\pi = (\pi_1, \ldots, \pi_k)'$, which is unknown. We want to collect data to estimate π and make other inferences about it. However, for protecting privacy we can observe only a masked version of X. Randomized response (RR) is a primary tool for privacy protection, which was first proposed by Warner (1965) for protecting respondent's privacy in interview surveys of sensitive binary variables. Subsequently, many other RR methods have been developed for applying to categorical and quantitative variables. We refer to the books Chaudhuri and Mukerjee (1988), Chaudhuri (2011) and Fox (2016) for reviews and relevant references.

In general, an RR procedure changes each true X into elements of an output space with known probabilities. Respondent's privacy is protected by revealing only the randomized response (or output). Let Z denote the output variable of an RR procedure and $\mathcal{S}_Z = \{d_1, \ldots, d_m\}$ be its range. Note that $\mathcal{S}_X$ and $\mathcal{S}_Z$ need not be the same, or even have the same cardinality. Let $p_{ij} = P(Z = d_i | X = c_j), i = 1, \ldots, m, j = 1, \ldots, k$, denote the transition probabilities that

are inherent in the RR procedure. The transition probability matrix (TPM) $P_{m\times k} = ((p_{ij}))$ characterizes an RR procedure, as it determines all of its effects on privacy and data utility. We refer to the TPM as the *RR design*. Quite importantly, any input masking procedure, locally for each unit, must have a built-in output space and transition probabilities, and thus can be viewed as an RR procedure. Consequently, the RR framework is quite general for studying and comparing all local input masking procedures. In particular, choosing a masking procedure for privacy protection boils down to choosing an RR design P.

To discuss estimation of π based on RR data, let n denote the sample size and S_i denote the sample frequency of d_i, $i = 1, \ldots, m$. We assume random sampling, which implies that $S = (S_1, S_2, \ldots, S_m)'$ has a multinomial distribution, viz. $S \sim mult(n, \lambda)$, where

$$\lambda_{m\times 1} = P_{m\times k}\pi_{k\times 1}. \tag{1.1}$$

A common estimator of λ is $\hat{\lambda} = S/n$. When P is square and nonsingular, an estimator $\tilde{\lambda}$ (based on S) of λ gives the estimator $\tilde{\pi} = P^{-1}\tilde{\lambda}$ for π, via (1.1). For generality and relating to some recently introduced RR methods, we assume that $m \geq k$ and $rank(P) = k$. RR methods with $rank(P) < k$ are unattractive, because there the distribution of S is not identifiable with respect to π and hence π is not estimable.

As is well known, data masking reduces data utility, i.e., the scope and quality of statistical inferences that can be made from masked data are lower compared to original data. A general objective is to choose and use a masking procedure that provides desired privacy protection at a minimum loss of data utility. Alternatively, one may try to strike a good balance between privacy protection and data utility. Clearly, formalizing such ideas requires practical privacy measures. Thus, we focus this study on privacy criteria and their demands on data utility. The main contributions of this paper are twofold. (1) We examine some existing privacy criteria, including local differential privacy (LDP), and discuss their drawbacks and (2) propose a new criterion (local l-diversity) and a method for estimating π under it.

The rest of this paper is organized as follows. In the next section, we briefly review two privacy criteria, viz. strict information privacy (SIP) and ϕ-average security, that were developed taking a Bayesian approach. The LDP criterion, which has received substantial attention in recent years, is closely connected to SIP. In Section 3, we show that (i) ϕ-average security is a flawed privacy criterion and (ii) SIP and LDP are highly expensive in terms of data utility, especially for large k. In Section 4, we propose a simple privacy mechanism that provides a local l-diversity privacy protection. It generalizes the true category randomly and can be viewed as an RR procedure. We give a method of moments estimator of π, which is very easy to calculate, and prove that under squared error loss, it is minimax among all linear unbiased estimators of π. We explain that for moderate to large k, the new method is better suited than LDP or SIP based methods. Section 5 contains some concluding remarks.

2. A Brief Review of Existing Privacy Criteria

Some recently introduced privacy criteria view a privacy breach as an intruder's "too much" information gain about a respondent from his response, and develop this idea using subjective probability and the Bayes rule, to articulate an intruder's knowledge, before and after observing a response. Some relevant works of Evfimievski et al. (2003), Nayak et al. (2015), Ye and Barg (2018) and Chai and Nayak (2018) are briefly reviewed next.

Let α_i denote an intruder's prior probability that a respondent's true value of X is c_i and let $\alpha = (\alpha_1, \ldots, \alpha_k)'$. Note that α is specific to each intruder-target pair and may be different from π. For given α, the posterior probability of $X = c_j$ given the response $Z = d_i$, is

$$P_\alpha(X = c_j | Z = d_i) = \frac{P_\alpha(X = c_j, Z = d_i)}{P_\alpha(Z = d_i)} = \frac{\alpha_j p_{ij}}{\sum_{l=1}^{k} \alpha_l p_{il}}.$$

So, the prior and posterior probabilities of an event $Q \subseteq \mathcal{S}_X = \{c_1, \ldots, c_k\}$ are:

$$P_\alpha(Q) = \sum_{j: c_j \in Q} \alpha_j \quad \text{and} \quad P_\alpha(Q|d_i) = \sum_{j: c_j \in Q} P_\alpha(X = c_j | Z = d_i). \tag{2.1}$$

2.1. Strict Information Privacy and Parity Bound

One line of thought is that a privacy procedure should guarantee that $P_\alpha(Q|d_i)$ and $P_\alpha(Q)$, as defined in (2.1), would be "desirably close" *for all* Q, α and d_i. Thus, no intruder would gain "much" new information about any feature (Q) of any respondent from his response. Chai and Nayak (2018) formalized this idea generally, observing that any specification of "desirably close" is equivalent to requiring (2.2) below with two given functions h_l and h_u.

Definition 2.1. *(Chai and Nayak, 2018). Let h_l and h_u be two functions from $[0,1]$ to $[0,1]$ such that $0 \le h_l(a) \le a \le h_u(a) \le 1$ for all $0 \le a \le 1$. An RR design is said to provide strict information privacy (SIP) with respect to h_l and h_u if*

$$h_l(P_\alpha(Q)) \le P_\alpha(Q|d_i) \le h_u(P_\alpha(Q)). \tag{2.2}$$

for all $\alpha, Q \subseteq \mathcal{S}_X$ and $i = 1, \ldots, m$.

For a (generic) prior-posterior pair (p, p_*), Definition 2.1 says that p_* is "desirably close" to p if and only if $h_l(p) \le p_* \le h_u(p)$. The β-factor privacy of Nayak et al. (2015), which requires $1/\beta \le P_\alpha(Q|d_i)/P_\alpha(Q) \le \beta$, is a special case, with $h_l(p) = (1/\beta)p$ and $h_u(p) = \beta p$. Another special case is the ρ_1-to-ρ_2 privacy (Evfimievski et al., 2003), which defines privacy breaches as $P_\alpha(Q) < \rho_1$ and $P_\alpha(Q|d_i) > \rho_2$ or $P_\alpha(Q) > \rho_2$ and $P_\alpha(Q|d_i) < \rho_1$; here h_l and h_u are step functions. Chai and Nayak (2018) gave a characterization of all RR designs that satisfy SIP. It yields useful guidance on how to choose h_l and h_u in practice, and involves the following:

Definition 2.2. *(Nayak et al., 2015). The ith row parity of P is defined as*

$$\eta_i(P) = \max\left\{\frac{p_{ij}}{p_{il}} \;\middle|\; j, l = 1, \ldots, k\right\} = \frac{\max_j\{p_{ij}\}}{\min_j\{p_{ij}\}},$$

with the convention $0/0 = 1$ and $a/0 = \infty$ for any $a > 0$. The parity of P is defined as $\eta(P) = \max_i\{\eta_i(P)\}$.

Chai and Nayak (2018) proved that for given h_l and h_u, an RR design P satisfies (2.2) if and only if its parity $\eta(P)$ does not exceed a specific value $B(h_l, h_u)$ determined by h_l and h_u. In particular, necessary and sufficient conditions for P to satisfy ρ_1-to-ρ_2 and β-factor privacies are $\eta(P) \leq [\rho_2(1-\rho_1)]/[\rho_1(1-\rho_2)]$ and $\eta(P) \leq \beta$, respectively. The following criterion, which has been studied by Kairouz et al. (2016), Wang et al. (2016), Duchi et. al. (2018), Ye and Barg (2018) and others, is closely related to SIP.

Definition 2.3. *An RR design provides ϵ-local differential privacy (ϵ-LDP), for $\epsilon > 0$, if*

$$\sup_{B \subseteq \mathcal{S}_Z} \sup_{c_i, c_j \in \mathcal{S}_X} \frac{P(Z \in B | X = c_i)}{P(Z \in B | X = c_j)} \leq e^{\epsilon}.$$

Chai and Nayak (2018) showed that an RR procedure provides ϵ-LDP if and only if

$$\frac{P_\alpha(Q)}{1 + (\gamma - 1)(1 - P_\alpha(Q))} \leq P_\alpha(Q|d_i) \leq \frac{\gamma P_\alpha(Q)}{1 + (\gamma - 1)P_\alpha(Q)} \tag{2.3}$$

for all α, Q and d_i, where $\gamma = e^{\epsilon}$, i.e., the procedure provides SIP with h_l and h_u defined as the lower and upper bounds in (2.3), respectively. It also follows that an RR design P provides ϵ-LDP if and only if $\eta(P) \leq \gamma = e^{\epsilon}$. An important conclusion is that satisfying ϵ-LDP or SIP, including its special cases, all reduce to imposing an upper bound on the RR design's parity.

2.2. Privacy as Average Information Gain

Boreale and Paolini (2015) introduced a "worst-case breach" criterion, which is essentially the same as β-factor privacy. The phrase "worst-case" refers to the requirement that (2.2) must hold, with $h_l(p) = (1/\beta)p$ and $h_u(p) = \beta p$, *for all possible* responses $d_i, i = 1, \ldots, m$, irrespective of their probabilities, which are $P_\alpha(Z = d_i) = \sum_{l=1}^{k} \alpha_l p_{il}, i = 1, \ldots, m$. They also proposed an "average-case breach" criterion, taking the response probabilities into account. They considered the scenario where an intruder uses Z to predict whether $X \in Q$ or $X \in Q^c$, for some $Q \subseteq \mathcal{S}_X$. Under 0 – 1 loss, an optimum rule declares $X \in Q$ if $P_\alpha(Q|d_i) \geq 1/2$, and otherwise $X \in Q^c$. For

this rule, the probability of a correct prediction is

$$G_\alpha(Q|Z) = \sum_{i=1}^{m} \max\{P_\alpha(Q|d_i), P_\alpha(Q^c|d_i)\} P_\alpha(Z = d_i)$$
$$= \sum_{i=1}^{m} \max\{P_\alpha(Q \cap d_i), P_\alpha(Q^c \cap d_i)\}$$

and the Bayes risk is $1 - G_\alpha(Q|Z)$. Similarly, the correct prediction probability when only the prior (and not Z) is used is $G_\alpha(Q) = \max\{P_\alpha(Q), P_\alpha(Q^c)\}$. It can be seen that $0.5 \leq G_\alpha(Q) \leq G_\alpha(Q|Z) \leq 1$ and so, the ratio $G_\alpha(Q|Z)/G_\alpha(Q)$ must be between 1 and 2.

Definition 2.4. *Boreale and Paolini (2015). An RR procedure P permits an average-case breach at level $\phi \in (1, 2)$ if there exists some $Q \subseteq \mathcal{S}_X$ and a prior α such that $G_\alpha(Q|Z)/G_\alpha(Q) > \phi$. Moreover, P is said to be ϕ-average secure if it does not allow any average-case breach at level ϕ.*

Actually, they used a negative log scale and connected the correct prediction probabilities to Renyi's min-entropy. Huang and Du (2008) also considered $G_\alpha(Q|Z)$, but evidently with $\alpha = \pi$. Boreale and Paolini (2015) also proved the following:

Theorem 2.1. *For an RR design $P_{m\times k}$, let $\vec{p}_i$ denote its ith column and let*

$$l_1(P) = \max_{i,j} \|\vec{p}_i - \vec{p}_j\|_1 = \max_{i,j} \sum_{u=1}^{m} |p_{ui} - p_{uj}|.$$

Then, P is ϕ-average secure if and only if $l_1(P) \leq 2(\phi - 1)$.

3. Disadvantages of Average Security and Parity Bounding

3.1. Drawbacks of Average Security

The average security criterion in Definition 2.4 might seem intuitively reasonable, but it is ill-suited for assuring privacy. To illustrate this, we consider $k = 2$ and the following two design

matrices:

$$P_1 = \begin{matrix} d_1 \\ d_2 \\ d_3 \end{matrix} \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \qquad \text{and} \qquad P_2 = \begin{matrix} d_1 \\ d_2 \end{matrix} \begin{bmatrix} 0.7 & 0.2 \\ 0.3 & 0.8 \end{bmatrix}.$$

Note that $l_1(P_1) = l_1(P_2) = 1.0$ and by Theorem 2.1, both P_1 and P_2 are ϕ-average secure at $\phi = 1.5$. However, under P_1, the responses d_1 and d_2 (corresponding to the first two rows of P_1) reveal that X is c_1 and c_2, respectively. Thus, P_1 discloses the true category of each respondent with probability 0.5, which is highly unsatisfactory. Evidently, P_1 is unacceptable for privacy protection and P_2 is much better, but the average-case breach criterion cannot recognize that or any difference between P_1 and P_2.

Similarly, for general ϕ and $k = 2$, the design

$$P_0 = \begin{matrix} d_1 \\ d_2 \\ d_3 \end{matrix} \begin{bmatrix} \phi - 1 & 0 \\ 0 & \phi - 1 \\ 2 - \phi & 2 - \phi \end{bmatrix} \tag{3.1}$$

is ϕ-average secure, but it discloses the true X category when the response is d_1 or d_2. In contrast, $P(Z = d_3|X = x)$ does not depend on x. Consequently, d_3 does not contain any information about π and under it, all posterior probabilities are the same as the corresponding prior probabilities. Thus, P_0, reveals the true category of each respondent with probability $\phi - 1$ and hides it fully with probability $2 - \phi$, which would be unsatisfactory privacy protection unless ϕ is very close to 1. But, when ϕ is close to 1, a large proportion (viz. $2 - \phi$) of the observations (all d_3 values) would be totally noninformative about π (and wasted). Also, as we discuss below, P_0 yields the best data utility among all ϕ-average secure designs. Thus, the average security criterion leads to a dire trade-off between privacy and data utility.

The following adaptation of Blackwell's (1951, 1953) approach to comparing statistical experiments gives a strong criterion for comparing data utility of RR designs.

Definition 3.1. *An RR design $P_{m\times k}$ is said to be at least as informative (or good) as another design $A_{r\times k}$, to be denoted $P \succeq A$, if there exists a transition probability matrix $C_{r\times m}$ such that $A = CP$.*

From Blackwell's works, it follows that $P \succeq A$ implies that under any loss function, for any inference rule δ based on the data from A, there exists a rule δ_* based on P such that the risk function of δ_* does not exceed that of δ. So, the assertion "at least as informative" holds in a very broad sense. In an intuitive way, $A = CP$ means that applying A is equivalent to further randomizing by C the responses obtained using P, and the second RR (by C) should inflict further data utility loss. Logically, we should say P and A are equivalent (denoted $P \sim A$) if both $P \succeq A$ and $A \succeq P$ hold, and P is better than A (written $P \succ A$) if $P \succeq A$ but $A \not\succeq P$. Also, P is said to be admissible if there exists no A such that $A \succ P$. For $k = 2$, the following result asserts optimal data utility of P_0 among all RR designs that are ϕ-average secure.

Theorem 3.1. *Within $\mathcal{A}_\phi = \{P_{m\times 2} : m \geq 2, l_1(P) \leq 2(\phi - 1)\}$, with $0 < \phi < 1$, only P_0 in (3.1) is admissible. Thus, P_0 is the most informative design among all ϕ-average secure designs.*

A proof of the theorem is given in the appendix. The drawbacks of ϕ-average security discussed above hold for all $k \geq 2$. In particular, it can be seen that for any $k \geq 2$, the design

$$P_{(k+1)\times k} = \begin{pmatrix} (\phi - 1)I_k \\ \hline (2 - \phi)1_k' \end{pmatrix}$$

is ϕ-average secure and also admissible. But, it discloses the true category of each respondent with probability $\phi - 1$. We may note that for $k > 2$, other admissible designs exist. We can prove that a ϕ-secure design P is admissible if and only if in each row of P, all nonzero elements are equal.

3.2. Cost of Parity Bounds

Here, we examine the effects of privacy demands on estimating π under squared error loss. Consider an RR design P and an estimator $\hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_k)'$ of π, derived under P and multinomial sampling. Then, under squared error loss, the risk function of $(P, \hat{\pi})$, normalized for sample size for convenience, is

$$R(P, \hat{\pi}; \pi) = n \times E_{P,\pi}\Big[\|\hat{\pi} - \pi\|^2\Big] = n \times E_{P,\pi}\Big[\sum_{i=1}^{k}(\hat{\pi}_i - \pi_i)^2\Big], \tag{3.2}$$

where the expectation is with respect to both sampling and randomization. It is well known that when $P = I$ (i.e., no randomization is used), $\hat{\pi}_0 = S/n$ is the best unbiased estimator of π; recall that S denotes the frequency vector from RR data. Also, $R(I, \hat{\pi}_0; \pi) = \sum_{i=1}^{k} \pi_i(1 - \pi_i)$. So, for assessing the data utility cost of an RR design P (and an estimator $\hat{\pi}$) we compare $R(P, \hat{\pi}; \pi)$ with $R(I, \hat{\pi}_0; \pi)$.

For given γ, let $\mathcal{C}_\gamma = \{P : \eta(P) \leq \gamma\}$, the class of all P that satisfy the parity constraint $\eta(P) \leq \gamma$. Chai and Nayak (2018) give a full characterization of $\mathcal{C}_\gamma$. Generally, $\mathcal{C}_\gamma$ contains many designs with different effects on data utility. So, we assess the cost of the privacy requirement $\eta(P) \leq \gamma$ by the cost of an optimal design in $\mathcal{C}_\gamma$ and an optimal estimator (inspired by Chai and Nayak, 2019). Specifically, for each P, we consider only unbiased estimators of π that are linear in S. This is equivalent to considering estimators of the form $\hat{\pi} = LS/n$, where $L_{k\times m}$ is a fixed matrix satisfying $LP = I$. With this restriction, a design and estimator pair $(P_*, \tilde{\pi} = L_*S/n)$ is minimax under (3.2) if $P_* \in \mathcal{C}_\gamma, P_*L_* = I$ and

$$\sup_{\pi} E_{P_*,\pi}\Big[\|\frac{L_*S}{n} - \pi\|^2\Big] = \inf_{P\in\mathcal{C}_\gamma} \inf_{L:LP=I} \sup_{\pi} E_{P,\pi}\Big[\|\frac{LS}{n} - \pi\|^2\Big].$$

Chai and Nayak (2019) solved this problem and derived the minimax pair $(P_*, \tilde{\pi})$. From their work, it also follows that the risk function of $(P_*, \tilde{\pi})$ is

$$R(P_*, \tilde{\pi}; \pi) = \Big[\frac{(k-1)^2}{f(q) - k} + \frac{1}{k} - 1\Big] + \sum_{i=1}^{k} \pi_i(1 - \pi_i), \tag{3.3}$$

where the function f and the quantity q are as defined in Chai and Nayak (2019). Actually, q is an integer valued function of k and γ, and q is either $\lfloor \frac{k}{1+\gamma} \rfloor$ or $\lceil \frac{k}{1+\gamma} \rceil$. Because of unbiasedness, $R(P_*, \tilde{\pi}; \pi)$ is also $n \times \text{tr}[V(\tilde{\pi})] = n \times \sum V(\tilde{\pi}_i)$. The last term of (3.3) is $R(I, \hat{\pi}_0; \pi)$; it reflects the *sampling variability* (with no randomization) and will be denoted $V_S(\pi)$. The term [ ] in (3.3), to be denoted V_R, is the *added variance* due to RR. Interestingly, note that V_R is independent of π, unlike the sampling variance $V_S(\pi)$.

	k								
γ	2	3	5	10	20	50	100	150	200
1.1	220	640	1441.333	3571.2	7959.1	21129.05	43125.92	65124.08	87121.7
1.5	12	32	76	193.5	432.25	1151.5	2351.25	3551.167	4751.125
2	4	10	25.333	64.2857	143.6484	383.2656	783.1343	1183.06	1583.067
3	1.5	3.5	9	23.7857	53.2	143.1798	293.04	443.0614	593.02
5	0.625	1.375	3.25	9.3515	21.70	59.0807	121.5399	184.015	246.5202
10	0.2469	0.5185	1.1358	3.1111	7.9883	22.7995	47.4115	72.1117	96.7882
20	0.1108	0.2271	0.4765	1.1967	3.0526	9.7502	20.7440	31.8096	42.9131
30	0.0713	0.1451	0.2996	0.7277	1.7621	5.9575	13.0144	20.1314	27.2974
50	0.0416	0.0841	0.1716	0.4048	0.9338	3.0204	7.1749	11.3367	15.5002
80	0.0256	0.0516	0.1045	0.2423	0.5419	1.6331	4.0926	6.6071	9.2567
100	0.0204	0.0410	0.0828	0.1910	0.4226	1.2399	3.0101	5.1851	7.0862

Table 1: The added variance (V_R) due to RR for the minimax method.

In Table 1, we report the values of V_R for some k and γ. The values of k are given in the second row and the γ values are shown in the first column. For each (k, γ) pair, the value of V_R is reported in the corresponding cell. As expected, V_R increases with k and it decreases as γ increases. However, our main point, elaborated below, is that unless k is small and γ is large, the added variance (V_R) is quite large relative to sampling variance $V_S(\pi)$.

Recall that the sampling variance depends on π, whereas the added variance is independent

of π. So, the ratio $V_R/V_S(\pi)$, which is of natural interest, depends on π. However, it can be seen that $0 \leq V_S(\pi) \leq 1 - 1/k < 1$ for all π. Consequently, $V_R/V_S(\pi) > V_R$ and $V_R/V_S(\pi)$ can be arbitrarily large depending on $V_S(\pi)$. Consider, for example, the case of $\gamma = 20$, which we believe gives only light privacy protection. There, for $k = 20$, $V_R(= 3.0526)$ is more than three times $V_S(\pi)$, and for $k = 100$, $V_R > (20.74)V_S(\pi)$. As another example, consider $k = 5$ (a small value) and $\pi = (.4, .2, .2, .1, .1)$, which yields $V_S(\pi) = 0.74$. Here, for $\gamma = 20$, $V_R = 0.4765$ and so, $V_R = (0.644)V_S(\pi)$ and $[V_R + V_S(\pi)]/V_S(\pi) = 1.644$. Thus, providing SIP with $\gamma = 20$ (or equivalently, LDP at $\epsilon = 2.996$) increases the overall variance by 64.4% (compared to no RR and no privacy). It can be seen that for $k = 5$, SIP at $\gamma = 10$ (providing moderate privacy) increases the variance by 153.49%.

In summary, LDP (or SIP) is suitable only for small k, because for reasonable values of γ (say ≤ 20), data utility costs are substantial even for moderate k. We might mention that a large sample size mitigates high added variance and yields a small $\text{tr}[Var(\tilde{\pi})]$. Naturally, with a (very) large sample size it is possible to give high privacy protection and yet estimate π accurately. So, LDP might be useful when data from many thousands or millions of transactions, searches etc. are captured.

4. A Local l-Diversity Method

4.1. The Randomization Mechanism

We propose a new procedure that is easy to implement and offers a better privacy-utility trade-off than SIP (or LDP) for moderate to large k. Our main motivation comes from the practice of collecting data on quantitative variables in measurement classes, to protect privacy. A direct approach to adopting that idea for categorical variables is collecting data after merging categories, i.e., coarsening (or generalizing) the response categories. Common methods for providing k-

anonymity and l-diversity do that (but after data collection). One drawback of that approach is that the probabilities of the merged categories cannot be estimated from the resulting data. To avoid this problem, we propose to randomly coarsen the true category of each respondent, as described next.

Our basic idea is to generalize each respondent's category to a superset of l categories consisting of the true category and additional $l-1$ randomly selected categories. The randomization mechanism and statistical inferences can be discussed conveniently using indicator vectors. Represent the categorical variable X with a row vector $\vec{X} = (X_1, \ldots, X_k)$, where $X_i = 1$ if the true category is c_i and otherwise $X_i = 0$. Thus, $\vec{X}$ is an indicator vector for the true category. We represent a random generalization of X by a row vector $\vec{Z} = (Z_1, \ldots, Z_k)$. The conversion from $\vec{X}$ to $\vec{Z}$ is done as follows. For a response $\vec{X}$ with $X_i = 1$, we (i) set $Z_i = 1$, (ii) randomly select $l-1$ of the remaining components of $\vec{Z}$ and set them to 1 and (iii) assign 0 to all other components. Thus, each $\vec{Z}$ contains exactly l ones and $k-l$ zeros.

The preceding mechanism is equivalent to an RR procedure. Its output range $\mathcal{S}_Z$ consists of all k dimensional row vectors, each having 1 in l components and 0 in the rest. Clearly, $\mathcal{S}_Z$ contains $\binom{k}{l}$ elements. Let $a_l = 1/\binom{k-1}{l-1}$ and $E_{[i]} = \{\vec{z} = (z_1, \ldots, z_k) \in \mathcal{S}_Z : z_i = 1\}$. Then, it can be seen easily that

$$P(\vec{Z} = \vec{z}|X = c_i) = \begin{cases} a_l & \text{if } \vec{z} \in E_{[i]} \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

The resultant TPM is of order $\binom{k}{l} \times k$ and in its each row, l values are a_l and $(k-l)$ are 0.

Clearly, a randomized response $\vec{z}$ simply reveals that the respondent's true category is one of l categories. To be specific, for a given $\vec{z} = (z_1, \ldots, z_k)$, let $B_{\vec{z}} = \{c_i \in \mathcal{S}_X : z_i = 1\}$. Then, $\vec{z}$ reveals that the respondent's true category is in $B_{\vec{z}}$ and all are equally plausible, as $P(\vec{Z} = \vec{z}|X = c_i)$ is a constant (a_l) for all $i \in B_{\vec{z}}$. It follows easily that an intruder with prior α will have the following

posterior probabilities:

$$P(X = c_i|\vec{Z} = \vec{z}) = \begin{cases} \frac{\alpha_i}{P_\alpha(B_{\vec{z}})} & \text{if } c_i \in B_{\vec{z}} \\ 0 & \text{otherwise} \end{cases}$$

for $i = 1, \ldots, k$. Thus, from $\vec{Z} = \vec{z}$, an intruder only learns that $X \in B_{\vec{z}}$ and updates his probabilities by simply normalizing his prior over $B_{\vec{z}}$.

We regard our procedure as a *local* l-diversity mechanism, noting that it's connection to l-diversity is similar to the connection between local differential privacy and differential privacy. However, certain differences between l-diversity and our approach should be noted. Fundamentally, we address privacy protection whereas l-diversity is about data confidentiality. We collect responses after masking whereas for l-diversity, the data agency perturbs the true data, after collected, using a data dependent mechanism. Also, l-diversity depends on the choice of key variables, but that is irrelevant to us. Both, our mechanism and its privacy implications are local to each respondent, independent of the values of other units.

We should mention that while $\vec{z}$ only reveals that the true category is in $B_{\vec{z}}$, an intruder with strong prior information may learn a lot from it. For example, if an intruder knows (a priori) for a respondent that all but one category in $B_{\vec{z}}$ are impossible, then he would learn the respondent's true category with certainty from $\vec{z}$. This applies also to l-diversity, as Li et al. (2007) noted. However, such situations are unlikely for large l. Thus, to mitigate effects of strong priors on privacy protection, we suggest to use reasonably large values for l, depending on the sensitivity of X and the value of k.

4.2. Statistical Estimation

With vector representation of the responses, our RR procedure yields an $n \times k$ data matrix $\mathcal{Z} = ((z_{ij}))$, with each row showing one respondent's data. A method of moments estimator of π can be obtained easily as follows. Let $V' = (V_1, \ldots, V_k)$ denote the column sums of $\mathcal{Z}$. Note that

V is the sum of a random sample of n vectors drawn from the distribution of $\vec{Z}$. It can be seen that for $i = 1, \ldots, n$,

$$E[\frac{V_i}{n}] = E[Z_i] = P(Z_i = 1) = \pi_i + (1 - \pi_i)\Big(\frac{l-1}{k-1}\Big) = \zeta_i, \text{ say.} \tag{4.2}$$

Using (4.2), we obtain the following method of moments estimators:

$$\hat{\pi}_{i*} = \Big(\frac{k-1}{k-l}\Big)\frac{V_i}{n} - \frac{l-1}{k-l}, i = 1, \ldots, k. \tag{4.3}$$

Let $\hat{\pi}_* = (\hat{\pi}_{1*}, \ldots, \hat{\pi}_{k*})'$. Then, it can be seen that $\hat{\pi}_*$ is an unbiased estimator of π and $V(\hat{\pi}_*) = [(k-1)/n(k-l)]^2\Sigma$, where $\Sigma = ((\sigma_{ij}))$ denotes $V(\vec{Z})$. Furthermore, $\sigma_{ii} = \zeta_i(1 - \zeta_i)$ and for $i \neq j$, $\sigma_{ij} = \zeta_{ij} - \zeta_i\zeta_j$, where

$$\begin{aligned}\zeta_{ij} = P(Z_i = Z_j = 1) &= \binom{k-2}{l-2}a_l\big(\pi_i + \pi_j\big) + \binom{k-3}{l-3}a_l\big(1 - \pi_i - \pi_j\big) \\ &= \frac{(l-1)(k-l)}{(k-1)(k-2)}\Big[\pi_i + \pi_j + \frac{l-2}{k-l}\Big].\end{aligned}$$

As in Section 1, let S denote the $\binom{k}{l}$ dimensional vector of frequencies of the elements in $\mathcal{S}_Z$. Then, with λ defined as in (1.1), a natural estimator of λ is $\hat{\lambda} = S/n$. In view of (1.1), one may consider linear functions of $\hat{\lambda}$ (or equivalently of S) for estimating π. In the appendix, we prove the following optimality property of our method of moments estimator.

Theorem 4.1. *Among all unbiased estimators of π that are linear in S, the estimator given by (4.3) is minimax under squared error loss.*

Using the above expressions and standard algebra we find that

$$nV(\hat{\pi}_{i*}) = \big(\frac{l-1}{k-l}\big)(1 - \pi_i) + \pi_i(1 - \pi_i).$$

and

$$R(P_{k,l}, \hat{\pi}_*; \pi) = n \times \mathrm{tr}\big(V(\hat{\pi}_*)\big) = \frac{(k-1)(l-1)}{(k-l)} + \sum_{i=1}^{k}\pi_i(1 - \pi_i), \tag{4.4}$$

where $P_{k,l}$ denotes the TPM of our RR mechanism. The first term on the right side of (4.4), to be denoted V_+^*, is the added variance due to our random generalization. Note that V_+^* is independent of π and it increases with l and decreases to $(l-1)$ as k increases to ∞. Naturally, as l increases, both privacy protection and added variance increase.

4.3. Comparison with Other Criteria

A natural question is: which of the two approaches – local l-diversity and parity bounding – we should use in practice? Recall that LDP, ρ_1-to-ρ_2 privacy, β-factor privacy and SIP all result in imposing an upper bound on the parity of the RR design P. To help answer the question, we compare their privacy protection aspects at equal added variance (a measure of data utility). Specifically, for a given locally l-diverse mechanism, we take the strongest parity (or LDP) protective mechanism, in the minimax sense of Chai and Nayak (2019), with the same added variance and compare their privacy protection characteristics.

For given k and l, let $P_{k,l}$ denote the locally l-diverse design. From (4.4), the added variance of $P_{k,l}$ is $(k-1)(l-1)/(k-l)$. To find the minimax design $P_{k,l}^*$ with the same added variance, we calculate γ and q such that the term [ ] in (3.3) equals $(k-1)(l-1)/(k-l)$. This gives us the smallest γ for which the parity condition $\eta(P) \leq \gamma$ can be satisfied with the same added variance. It specifies the largest privacy protection, in the sense of SIP or LDP, that can be afforded with that added variance. We may remark that the facts that the added variance terms in (3.3) and (4.4) are independent of π enable us to hold them at the same level easily. In Table 2, for some values of k and l, we report the corresponding values of γ and q, in parentheses.

For clarity of comparison, we briefly review certain features of the minimax mechanism of Chai and Nayak (2019). For given k and γ, they first calculate an optimum integer value q. Then, using vector representations as discussed earlier, from a true value $\vec{x} = (x_1, \ldots, x_k)$ a response $\vec{z} = (z_1, \ldots, z_k)$ is generated as follows. Suppose $x_i = 1$. Then, they set $z_i = 1$ with probability

$p = (q\gamma)/(q\gamma + k - q)$; otherwise $z_i = 0$. If $z_i = 1$, they randomly choose $q-1$ other components of z and set those to 1. On the other hand, if $z_i = 0$, q of the remaining components of $\vec{z}$ are selected at random and set to 1. In both cases, the other $(k - q)$ components are assigned 0. Similar to our procedure, the response vector $\vec{Z}$ contains exactly q ones and $k - q$ zeros. One important difference is that in our case, $B_{\vec{z}}$ (the set of categories for which $z_i = 1$) includes the true category but in their method, $B_{\vec{z}}$ contains the true category with probability p (as given above).

	k						
l	10	15	20	50	100	200	500
5	6(2)	10.11(1)	14.19(1)	38.50(1)	78.96(1)	159.87(1)	402.58(1)
10	NA	3.73(3)	5.83(3)	18.02(3)	38.23(3)	78.60(3)	199.71(3)
15	NA	NA	3.00(5)	11.25(4)	24.63(4)	51.34(4)	131.44(4)
20	NA	NA	NA	7.88(6)	17.96(5)	37.98(5)	97.99(5)
25	NA	NA	NA	5.83(7)	13.94(7)	30.01(6)	78.04(6)
30	NA	NA	NA	4.44(9)	11.25(8)	24.63(8)	64.69(8)

Table 2: Best privacy protecting minimax designs at l-diversity's added variance.

To discuss the values in Table 2, take for example, $k = 20$ and $l = 5$ and the associated locally l-diverse design $P_{20,5}$. Table 2 shows that the corresponding minimax design $P^*_{20,5}$ has $\gamma = 14.19$ and $q = 1$. Here, the two competing RR designs, $P_{20,5}$ and $P^*_{20,5}$, have the same data utility (added variance). So, to choose between the two, we may ask: Privacy characteristics of which one are easier to comprehend and communicate? Which of the two designs is likely to be more comforting to respondents? We believe that empirical work is needed to find out respondents' preferences, but we also note a few points below.

In terms of SIP, the minimax design $P^*_{20,5}$ (with $\gamma = 14.19$) assures each respondent that for all intruders, events (or properties) of interest and RR outputs, the ratio of posterior to

prior probabilities will be between 0.0705 (=1/14.19) and 14.19. Alternatively, it assures LDP at $\epsilon = 2.6525\,(= \ln 14.19)$. We speculate that in practice, it would be difficult to communicate these assurances (especially the LDP) correctly and effectively to many respondents (not having much background in probability). Indeed, LDP and SIP are mathematically rigorous, but are also highly technical and difficult for the public to grasp. Also, it is not even clear that real life intruders revise their opinion probabilistically and using the Bayes rule (which requires a full elicitation of the intruder's prior distribution over $\mathcal{S}_X$—not an easy task).

The RR mechanisms of minimax designs, discussed above, give practical information about $P^*_{20,5}$. Here, as $q = 1$, each respondent reports exactly one category. Furthermore, $p = (q\gamma)/(q\gamma + k - q) = 0.4275$ and so, each respondent reports his true category with probability 0.4275; otherwise, he randomly selects and reports one of the other 19 categories. Thus, under $P^*_{20,5}$, an intruder (or data collector) can correctly predict a respondent's true category with probability 0.4275. In contrast, the locally 5-diverse design $P_{20,5}$ asks each respondent to report a set of 5 (out of 20) categories, one of which must be his true category (the other 4 are added randomly). Here, an intruder with no prior knowledge can correctly guess a respondent's true category (from the RR output) with probability 0.2 (generally, $1/l$ for $P_{k,l}$), which is much smaller than 0.4275. Thus, we suspect that respondents will largely prefer the local l-diversity mechanism.

As another example, for $k = 50$ and $l = 10$, Table 2 gives $\gamma = 18.02$ and $q = 3$. We also calculate that $\epsilon = 2.8915$ and $p = 0.5349$. Under local l-diversity (with $l = 10$) and with no prior knowledge, a respondent's true category can be guessed correctly with probability 0.1. Under the corresponding minimax design, each respondent reports three categories as follows. With probability 0.5349 each respondent reports his true category plus two randomly selected categories. Otherwise, i.e., with probability 0.4651, he reports three randomly selected categories, excluding his true category. Here, a respondent might observe that by randomly selecting one of the three reported categories, intruders will be able to correctly guess his true response with

probability $(1/3)(0.5349) = 0.1783$. As this is noticeably larger than 0.1, respondents are likely to favor local l-diversity.

5. Discussion

Privacy criteria are very important as they are central to developing and choosing privacy mechanisms. In this paper, we examined several existing criteria. We found that the average security criterion has a serious flaw, which makes it unsatisfactory. To examine LDP and SIP (and its special cases), we took a common approach using the fact that each of those simply imposes an upper bound on the parity of the design matrix. We showed that when the number of categories is moderate to large, the condition $\eta(P) \leq \gamma$, with reasonably small γ (to give modest privacy), induces high data utility loss, even when an optimal design is used to satisfy the parity constraint. In particular, this implies that LDP, which has received significant attention in recent years (and seems to be the leading privacy criterion), may be suitable only for small k.

The LDP is used most prominently in the RAPPOR procedure (see Erlingsson et al., 2014; and Fanti et al., 2016), which is a significant privacy mechanism that is being used notably by Google and Apple for Internet data collection. It was developed to satisfy LDP. Previously, Chai and Nayak (2019) showed that basic RAPPOR is substantially inefficient (and does greater harm to data utility) compared to their minimax procedure. So, our findings about the minimax procedure's data utility cost indicate that in many applications, RAPPOR would produce data with only little statistical information (and practical value).

The proposed l-diversity procedure has certain attractive features. First, its response randomization can be implemented easily. Actually, it is easier than RAPPOR. Second, its privacy protection is easy to grasp and communicate to the public. Simply, each respondent reveals l categories, of which one is his true category. Third, statistical estimation is simple and effective. The data can be presented nicely as an $n \times k$ matrix and the computation of $\hat{\pi}_*$, in (4.3), is very

simple. Also, the estimator is unbiased and minimax in a specific sense. As the privacy concept and its impacts vary widely depending on contexts, various privacy criteria and protection methods are needed to handle practical problems. We hope that our findings and the proposed l-diversity method will be useful in practice and stimulate further research.

6. Appendix

A. Proof of Theorem 3.1.

Proof. Chai and Nayak (2018) showed that if P has rows that are proportional and $\tilde{P}$ is obtained from P by merging its proportional rows, then $P \sim \tilde{P}$. They also argued that one should merge all proportional rows of an RR design, if any. It can also be verified directly that $l_1(P) = l_1(\tilde{P})$. So, it suffices to consider only the designs $P_{m\times 2} \in \mathcal{A}_\phi$ that have no propositional rows.

Next, we prove that if $P_{m\times 2}$ has a row $(c \;\; d)$ with $0 < c, d < 1$ and $c \neq d$, then P is inadmissible. Without loss of generality, suppose that the first row of P is $(c \;\; d)$ with $0 < d < c < 1$ and let P_* denote the matrix of the remaining rows of P. Now, consider

$$P_1 = \left(\begin{array}{cc} c-d & 0 \\ d & d \\ \hline & P_* \end{array}\right)$$

and note that $P = (e_1 \mid I)P_1$, where e_1 is a column vector whose first element is 1 and the rest are 0. So, $P_1 \succeq P$. Next, to see that $P \not\succeq P_1$, suppose there exists a transition probability matrix $C_{(m+1)\times m} = ((c_{ij}))$ such that $P_1 = CP$. Then, each row of P_1 is a weighted combination of the rows of P, with all weights being in $[0, 1]$. The first row of P_1 can satisfy these conditions only if P_* has a row $(a, 0)$, with $a \geq c - d$. Without loss of generality, suppose that the first row of P_* (i.e., the second row of P) is $(a, 0)$. Then, for $P_1 = CP$ to hold, both the first and third rows of P_1 must be reconstructed from the second row of P. So, we must have $c_{12} = (c - d)/a$ and $c_{32} = 1$,

in which case the sum of the second column of C would exceed 1, contradicting the assumption that C is a TPM. Thus, we can conclude that $P_1 \succ P$.

The preceding discussions imply that any admissible $P \in \mathcal{A}_\phi$ with no proportional rows must be of the form

$$P_a = \begin{bmatrix} a & 0 \\ 0 & a \\ 1-a & 1-a \end{bmatrix}$$

with $a \leq \phi - 1$ (to be ϕ-average secure). But, it can be seen, using arguments similar to those used above, that if $a < \phi - 1$, then $P_0 \succ P_a$. Thus, within $\mathcal{A}_\phi$ only P_0 is admissible, and hence it is the best design. □

A. Proof of Theorem 4.1.

Proof. To prove the theorem, we use some results and ideas from Chai and Nayak (2019). Label the elements of $\mathcal{S}_Z$ using the vectors $\vec{z}$ as described in Section 4.1. Then, the transition probabilities for our RR mechanism are as in (4.1). Let $P_{k,l} = ((p_{ij}))$ denote the transition probability matrix, with some ordering of all possible $\vec{z}$. Note that $P_{k,l}$ is of order $m \times k$, with $m = \binom{k}{l}$. Let $\lambda = P_{k,l}\pi = (\lambda_1, \ldots, \lambda_m)'$ and $\hat{\lambda} = S/n$. From Chai and Nayak (2019) it follows that for any fixed $L_{k\times m}$, $L\hat{\lambda}$ is an unbiased estimator of π if and only if $LP_{k,l} = I$, and the following:

Lemma 6.1. *Consider any $L_{k\times m}$ such that $LP_{k,l} = I$. Let $R(P_{k,l}, L; \pi) = nE_{P_{k,l},\pi}\|L\hat{\lambda} - \pi\|^2$ denote the risk function of $L\hat{\lambda}$ for estimating π under squared error loss and let D_λ denote the diagonal matrix with elements $\lambda_1, \ldots, \lambda_m$. Then,*

$$\begin{aligned} R(P_{k,l}, L; \pi) &= n[tr(V_{P_{k,l},\pi}(L\hat{\lambda}))] = tr[L(D_\lambda - \lambda\lambda')L'] \\ &= tr(LD_\lambda L') - \sum_{i=1}^{k} \pi_i^2 \qquad (6.1) \\ &\geq tr(P'_{k,l}D_\lambda^{-1}P_{k,l})^{-1} - \sum_{i=1}^{k} \pi_i^2 \qquad (6.2) \end{aligned}$$

and the lower bound in (6.2) is attained when

$$L = (P'_{k,l}D_\lambda^{-1}P_{k,l})^{-1}P'_{k,l}D_\lambda^{-1} = L_*(\pi), \ \textit{say.} \tag{6.3}$$

Note that the optimum L in (6.3) depends on π via D_λ, which shows that a uniformly minimum risk linear unbiased estimator π does not exist. Let $\pi_u = (1/k, \ldots, 1/k)' = \frac{1}{k}1_k$, $L_u = L_*(\pi_u)$ and $\tilde{\pi} = L_u S/n$. It is easy to see, using (6.3), that $L_u P_{k,l} = I$ and thus, $\tilde{\pi}$ is an unbiased estimator of π. Recall that in each row of $P_{k,l}$, l values are $a_l\,[= 1/\binom{k-1}{l-1}]$ and the rest are 0. So, when $\pi = \pi_u$,

$$\lambda = P_{k,l}\pi_u = \frac{1}{k}P_{k,l}1_k = \frac{1}{k}la_l 1_m = \frac{1}{m}1_m,$$

which implies that $D_\lambda = \frac{1}{m}I$ and hence $L_u = (P'_{k,l}P_{k,l})^{-1}P'_{k,l}$. Let $P'_{k,l}P_{k,l} = H = ((h_{ij}))$. Then, $h_{ii} = \|i\text{th column of } P_{k,l}\|^2$ and for $i \neq j$, h_{ij} is the inner product of the ith and jth columns of $P_{k,l}$. Note that in each column of $P_{k,l}$, a_l values are a_l and the rest are zero and any two columns of $P_{k,l}$ have the value a_l in $\binom{k-2}{l-2}$ common rows. From these, it follows that

$$P'_{k,l}P_{k,l} = \frac{a_l}{k-1}\big[(k-l)I_k + (l-1)1_k 1'_k\big]$$

and

$$\begin{aligned} L_u = (P'_{k,l}P_{k,l})^{-1}P'_{k,l} &= \frac{1}{a_l(k-l)}\Big[(k-1)I_k - (1-\frac{1}{l})1_k 1'_k\Big]P'_{k,l} \\ &= \frac{1}{k-l}\Big[(\frac{k-1}{a_l})P'_{k,l} - (l-1)1_k 1'_m\Big]. \end{aligned} \tag{6.4}$$

Next, we note that $\mathrm{tr}(L_u D_\lambda L'_u)$ does not depend on π. To see that, let $((g_{ij})) = G = L'_u L_u$. Since the rows of $P_{k,l}$ are permutations of each other, the same holds true for L'_u, in view of (6.4). So, g_{ii} is a constant, say g_0 for all i. Now,

$$tr(L_u D_\lambda L'_u) = tr(D_\lambda L'_u L_u) = tr(D_\lambda G) = \sum_{i=1}^{m}\lambda_i g_{ii} = g_0\sum_{i=1}^{m}\lambda_i = g_0,$$

which is independent of π. This and (6.1) and the fact that $\sum \pi_i^2$ is minimum when $\pi = \pi_u$ now give us

$$\sup_\pi R(P_{k,l}, L_u; \pi) = tr(L_u D_\lambda L'_u) - \inf_\pi \sum_{i=1}^{k}\pi_i^2 = R(P_{k,l}, L_u; \pi_u).$$

Then,

$$\inf_L \sup_\pi R(P_{k,l}, L; \pi) \geq \inf_L R(P_{k,l}, L; \pi_u) = R(P_{k,l}, L_u; \pi_u) = \sup_\pi R(P_{k,l}, L_u; \pi),$$

where the first "=" follows from $L_u = L_*(\pi_u)$. In conclusion we have the following:

Lemma 6.2. *Under $P_{k,l}$, the estimator $\tilde{\pi} = L_u S/n$ is minimax among all linear unbiased estimators of π.*

We complete the proof of Theorem 4.1 by showing that $\hat{\pi}_*$ in (4.3) and $\tilde{\pi}$ (in Lemma 6.2) are the same. Using (6.4), we obtain

$$\frac{L_u S}{n} = \frac{k-1}{n(k-l)a_l} P'_{k,l} S - \Big(\frac{l-1}{k-l}\Big) 1_k. \tag{6.5}$$

The jth element of $P'_{k,l}S$ is the inner product of S and the jth column of $P_{k,l}$. Recall from (4.1) that $p_{ij} = a_l$ if $\vec{z} \in E_{[j]} = \{\vec{z} = (z_1, \ldots, z_k) \in \mathcal{S}_Z : z_j = 1\}$ and 0 otherwise. So, the inner product that yields the jth element of $P'_{k,l}S$ is a_l times the total number of responses with $z_j = 1$ or $a_l V_j$. Now, it can be seen easily that (6.5) reduces to the method of moments estimator $\hat{\pi}_*$. □

Acknowledgment. We sincerely thank Eric Slud for giving us several suggestions, which helped us to improve the paper substantially.

References

[1] Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association*, 85, 38-45.

[2] Blackwell, D. (1951). Comparison of experiments. In *Proceedings of Second Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley, pp. 93-102.

[3] Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24, 265-272.

[4] Boreale, M., and Paolini, M. (2015). Worst-and average-case privacy breaches in randomization mechanisms. *Theoretical Computer Science*, 597, 40-61.

[5] Chai, J., and Nayak, T.K. (2018). A criterion for privacy protection in data collection and its attainment via randomized response procedures. *Electronic Journal of Statistics*, 12, 4264-4287.

[6] Chai, J., and Nayak, T.K. (2019). Minimax randomized response methods for providing local differential privacy. Research Report Series, Statistics #2019-04, Center for Statistical Research & Methodology, U.S. Census Bureau.
https://www.census.gov/srd/papers/pdf/RRS2019-04.pdf

[7] Chaudhuri, A. (2011). *Randomized Response and Indirect Questioning Techniques in Surveys*. Boca Raton, FL, CRC Press.

[8] Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*. New York, Marcel Dekker.

[9] Chen, B-C., Kifer, D., LeFevre, K. and Machanavajjhala, A. (2009) Privacy-preserving data publishing. *Foundations and Trends in Databases*, 2, 1-167.

[10] Domingo-Ferrer, J., Soria-Comas, J. and Mulero-Vellido, R. (2019). Steered microaggregation as a unified primitive to anonymize data sets and data streams. *IEEE Transactions on Information Forensics and Security*, 14, 3298-3311

[11] Duchi, J.C., Jordan, M.I., and Wainwright, M.J. (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113, 182-201.

[12] Erlingsson, U., Pihur, V., and Korolova, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054-1067.

[13] Evfimievski, A., Gehrke, J. and Srikant, R. (2003). Limiting privacy breaches in privacy preserving data mining. *In Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 211-222.

[14] Fanti, G., Pihur, V. and lfar Erlingsson, U. (2016). Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. *Proceedings on Privacy Enhancing Technologies*, 3, 41-61

[15] Fox, J.A. (2016). *Randomized Response and Related Methods: Surveying Sensitive Data.* Thousand Oaks, CA, Sage Publications.

[16] Huang, Z. and Du, W. (2008). OptRR: Optimizing randomized response schemes for privacy-preserving data mining. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*, pp. 705-714.

[17] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K. and de Wolf, P.-P. (2012). *Statistical Disclosure Control.* New York: John Wiley & Sons.

[18] Kairouz, P., Oh, S., and Viswanath, P. (2016). Extremal mechanisms for local differential privacy. *Journal of Machine Learning Research*, 17, 1-51.

[19] Li, N., Li, T. and Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In *Proceedings of IEEE 23rd International Conference on Data Engineering*, pp. 106-115.

[20] Machanavajjhala, A., Gehrke, J., Kifer, D. and Venkitasubramaniam, M. (2007). *l*-Diversity: Privacy beyond k-anonymity. *ACM Trans. Knowledge Discovery from Data*, Vol. 1, No. 1, Article 3.

[21] Nayak, T.K., Zhang, C. and Adeshiyan, S.A. (2015). Emerging applications of randomized response concepts and some related issues. *Model Assisted Statistics and Applications*, 10, 335-344.

[22] Nayak, T.K., Zhang, C., and You, J. (2018). Measuring identification risk in microdata release and its control by post-randomisation. *International Statistical Review*, 86, 300-321.

[23] Samarati, P. (2001). Protecting respondents' identities in microdata release. *IEEE Trans. Knowl. Data Eng.*, 13, 1010-1027.

[24] Shlomo, N., and Skinner, C. (2010). Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *The Annals of Applied Statistics*, 4, 1291-1310.

[25] Skinner, C. J., and Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society, Ser. B*, 64, 855-867.

[26] Sweeney, L. (2002). *k*-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, 10, 557-570.

[27] Torra, V. (2017). *Data Privacy: Foundations, New Developments and the Big Data Challenge*. New York: Springer.

[28] Venkataramanan, N. and Shriram, A. (2017). *Data Privacy: Principles and Practice*. New York, CRC press.

[29] Wang, S., Huang, L., Wang, P., Nie, Y., Xu, H., Yang, W., Li, X-Y. and Qiao, C. (2016). Mutual information optimally local private discrete distribution estimation. *arXiv preprint arXiv:1607.08025*

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

[31] Willenborg, L.C.R.J., De Waal, T., 2001. *Elements of Statistical Disclosure Control*. Springer, New York.

[32] Ye, M. and Barg, A. (2018). Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64, 5662-5676.