RESEARCH REPORT SERIES
(Disclosure Avoidance #2016-01)

Emerging Applications of Randomized Response

Concepts and Some Related Issues

Tapan K. Nayak, Cheng Zhang and Samson A. Adeshiyan

Center for Disclosure Avoidance Research
U.S. Census Bureau
Washington DC 20233

Report Issued: May 2, 2016

Emerging Applications of Randomized Response Concepts and Some Related Issues

Tapan K. Nayak, * Cheng Zhang† and Samson A. Adeshiyan‡§

Abstract

Randomized response (RR) was introduced as a technique for protecting respondent's privacy in survey interviews regarding sensitive characteristics. In recent years, the basic RR ideas have been used and extended in other contexts. We discuss usage and recent advances of RR in confidentiality protection and in privacy preserving data mining. We discuss important differences between RR surveys and RR for confidentiality protection. In particular, for confidentiality protection, the data may be used to choose randomization probabilities suitably, but by doing so also makes well known inferences derived for RR surveys inapplicable. We examine one privacy breach criterion in data mining and propose a new privacy guarantee and a method for achieving that. We also discuss several new challenges and open problems for future research.

Key words and Phrases: Categorical data; confidentiality protection; privacy breach; sampling design; unbiased estimation; variance inflation.

*Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington, DC 20233 and Department of Statistics, George Washington University, Washington, DC 20052, email: tapan@gwu.edu.

†Department of Statistics, George Washington University, Washington, DC 20052, email: godeau@gwu.edu.

‡U.S. Energy Information Administration, Washington, DC 20585, email: samson.adeshiyan@eia.gov.

§The views expressed in this article are those of the authors and not necessarily those of the U.S. Census Bureau. The analysis and conclusions contained in this paper are those of the authors and do not represent the official position of the U.S. Energy Information Administration (EIA) or the U.S. Department of Energy (DOE).

1. Introduction

The primary reason for using randomized response (RR) techniques in surveys is to offer some protection to respondent's privacy and thereby encourage honest participation, which is especially important when the survey asks for information about sensitive characteristics such as drug abuse, tax evasion, unlawful activity and gambling. For a dichotomous population consisting of a sensitive group A and its complement, A^c, Warner [34] introduced the first RR method, where each respondent is instructed to select one of the two questions Q_1: Do you belong to A? and Q_2: Do you belong to A^c? using a prescribed random mechanism (e.g., drawing a card from a shuffled deck), and then truthfully reply "Yes" or "No" to the selected question. The respondent protects his privacy by not disclosing the question to the interviewer. The probability (p) of selecting Q_1 is a design parameter. The survey organization chooses a value for p and embeds it in the question selection mechanism. In Warner's procedure, a true "Yes" (or No) converts to a false "No" (or Yes) with probability $1-p$. Since the publication of the first RR paper [34] by Warner in 1965, numerous papers offering a wide variety of RR schemes have appeared. While most papers deal with a binary variable with one value (or category) being sensitive, as in [34], the literature on RR methods for categorical and quantitative variables is also substantial; see [6, 7, 8] for reviews of various RR methods and further references.

In this paper, we shall focus only on categorical variables. We do not attempt to cover RR procedures for quantitative variables, as they are inherently different from those for categorical variables. Usually, values of quantitative variables are randomized via noise addition or multiplication (see [4, 13, 17, 27]). Let X be a categorical variable with categories labeled $1, \ldots, k$, and also let $\pi_i = P[X = i], i = 1, \ldots, k$, and $\pi = (\pi_1, \ldots, \pi_k)'$. Typically, π is unknown and we collect data on X in order to make inferences about π and functions of it. To protect respondent's privacy, RR procedures ask each respondent to perform a given random experiment and use its outcome, his true category and pre-specified rules to compute an output and report

it. For simplicity, suppose the output space of the randomization process contains k elements, labeled $1, \ldots, k$, without loss of generality. Let Z denote the output variable that is actually reported by respondents and let $p_{ij} = P(Z = i|X = j)$. Then, $P = ((p_{ij}))$ is the transition probability matrix of the RR procedure and $\sum_i p_{ij} = 1$ for $j = 1, \ldots, k$. We shall focus on RR procedures for which P is known and nonsingular, as those are most convenient for deriving statistical inferences. If P is singular, not all components of π can be estimated unbiasedly.

The essence of any RR procedure is its transition probability matrix $P = ((p_{ij}))$. All properties of an RR procedure, e.g., sampling distribution of all estimators and measures of privacy protection, depend on the randomization mechanism only through P. Thus, if two RR procedures have the same transition probability matrix, then they are statistically equivalent (cf. [24, 25, 29]). Two consequences of this are: (i) a general theory of RR procedures can (and should) be developed in terms of P and (ii) comparison of RR procedures reduces to comparing their transition probability matrices. We may also note that any given P can be implemented by many different random experiments and one may choose one of them based on practical convenience and simplicity.

Let n denote the sample size, T_i and S_i denote the sample frequencies of $X = i$ and $Z = i$, respectively, $\mathbf{T} = (T_1, \ldots, T_k)'$ and $\mathbf{S} = (S_1, \ldots, S_k)'$. Then, under (commonly considered) multinomial sampling, i.e., random sampling from an infinite population or simple random sampling with replacement (SRSWR), if the population is finite, $\mathbf{T} \sim Mult(n, \pi)$ and $\mathbf{S} \sim Mult(n, \lambda)$, where

$$\lambda = P\pi. \tag{1.1}$$

Note that $\mathbf{T}$ is unobservable and hence inferences about π need to be derived based on $\mathbf{S}$. If P is nonsingular and $\tilde{\lambda}$ is an unbiased estimator of λ, then from (1.1) it follows that $\tilde{\pi} = P^{-1}\tilde{\lambda}$ is an unbiased estimator of π. The MLE (and UMVUE) of λ is $\hat{\lambda} = \mathbf{S}/n$ and it yields

$$\hat{\pi} = P^{-1}\hat{\lambda} = P^{-1}(\mathbf{S}/n), \tag{1.2}$$

as an unbiased estimator of π with

$$Var(\hat{\pi}) = \frac{(D_\pi - \pi\pi')}{n} + \frac{[P^{-1}D_\lambda(P^{-1})' - D_\pi]}{n}, \tag{1.3}$$

where D_π is a diagonal matrix with diagonal elements being $\pi_1, \ldots, \pi_k$ and D_λ is defined similarly (see [8], p. 43). The first term on the right side of (1.3) is the sampling variance and the last term is the additional variance due to randomization. Inferences under general sampling designs have been discussed in [5, 25, 26, 28].

While privacy and confidentiality are often used synonymously, we should note a particular difference between the two. Privacy is an individual's right to control access to his information and privacy protection refers to hiding a respondent's true values from everyone, including the interviewer and the survey organization. Protecting confidentiality means not giving identifiable information about any survey participant (or unit) to unauthorized people. Often survey participants give their information to a survey organization trusting that their data will be used by researchers and policy makers only to learn about the population as a whole and not about any individual or survey unit. Privacy arises at data collection stage whereas confidentiality emerges after the data have been collected. Confidentiality protection necessitates physical security of the data. However, often the goal of a survey is to publish data, but the data agency may not release the original data if that would enable others to gain information about any survey participant. Consequently, agencies release a perturbed version of the original data.

Most RR procedures have been developed for using in interview surveys, with particular attention to the randomization process, as the randomization apparatus should be easy to carry for the interviewer and the experiment needs to be easy to understand and perform by respondents. However, recently the basic RR concepts have received considerable attention in on-line surveys, confidentiality protected data dissemination and other contexts, where the actual randomization is carried out more appropriately by computer programs. The main goal of this paper is to discuss such applications and some related issues. Specifically, we shall discuss some

recent advances of RR ideas in confidentiality protection and in privacy preserving data mining (PPDM).

In Section 2, we discuss some recent results on the post-randomization method (PRAM) for confidentiality protection. We shall see that PRAM is similar to RR surveys, but it also raises new issues and questions. While π can be estimated the same way in the two cases, the sampling distribution of the estimator can be different. This has important implications on estimating sampling variance of relevant estimators and constructing confidence intervals. In Section 3, we discuss usage of RR in PPDM, which has become a significant research focus in computer science community. Privacy protection and inferential goals in PPDM are different from those in RR surveys. We review one leading definition of privacy breach and a related result and then propose a simpler and more practical privacy protection goal and characterize all RR procedures that accomplish that goal. We state certain additional special features of PRAM and PPDM as well as some open questions in Section 4.

2. Post-randomization for Statistical Disclosure Control

The main goal of statistical agencies is to collect and publish data to inform the public, policy makers and researchers, but they also need to protect the confidentiality of unit level information for legal reasons and for upholding public trust. Usually, just removing all direct identifiers, such as name, address and social security number, is insufficient because it might be possible to identify a unit in the data set based on easily available values of some variables such as gender, age and occupation, also called key variables. Disclosure is a difficult topic (cf., [22]) and it can occur in different forms depending on the disclosure scenario (see [36]). Broadly speaking, disclosure occurs when the released data enable one to learn much new information about any survey respondent or unit. The most serious is identity disclosure, which happens when an intruder correctly identifies the record of a survey unit by matching externally available values

of some key variables that are included in the survey. For disclosure control, agencies often release a perturbed or masked version of the original data, causing some loss and possibly distortion of information. In general, the goal is to perturb the original data in such a way that disclosure risk is sufficiently low but information loss is small or minimal. However, wide variations in data structures, confidentiality concerns and inferential goals render a uniform treatment of the subject infeasible. Various masking methods, such as grouping, cell suppression, data swapping, multiple imputation and random noise infusion have been developed for practical use; see the books [11, 36] for reviews and additional references.

Warner [35] had mentioned potential usage of RR techniques for statistical disclosure control, but it has received significant attention only since Gouweleeuw et al. [19] developed the idea as the Post-randomization Method (PRAM) for perturbing data on categorical variables to protect confidentiality while releasing microdata. As before, consider a categorical variable X with possible values $1, \ldots, k$. Just as in RR, to apply PRAM to X, one first selects a transition probability matrix $P = ((p_{ij}))$, also called the PRAM matrix, and then randomly changes any original category $X = j$ to $Z = i$ with probability p_{ij} $(i, j = 1, \ldots, k)$. The randomization step is performed for each record in the data set, independently of all other records.

We should note that PRAM can be applied to more than one categorical variable, independently or jointly, but conceptually, any PRAM can be regarded as being applied to the compound variable created by cross-classifying all variables. This perspective is both convenient and appropriate for logical discussions of PRAM. We refer to [10, 31, 32, 33] for further discussion of PRAM and additional references.

Although RR surveys and PRAM contain some common ideas and mathematical properties, PRAM is critically different in several ways. First, in PRAM, the survey organization randomizes the data and that can be done conveniently and accurately using a computer program (and random number generators). This also makes respondents' perceptions about a randomization

experiment irrelevant. Objective properties of the experiment, captured by the transition probabilities, determine both the degree of protection and information loss. In contrast, RR in a survey should use a randomization experiment that respondents perceive as protective. In practice, a respondent's perception may depend not only on the transition probabilities but also on some ancillary features of the experiment. Second, in surveys, RR is used to randomize values of only sensitive variables. By contrast, for confidentiality protection PRAM can be applied to both sensitive and key variables, to protect against identity disclosure. Another important difference is that in RR surveys the transition probability matrix P is chosen before data collection and hence it cannot depend on the data, but in PRAM, P may depend on the data and hence be random. In particular, P is data dependent in invariant PRAM, to be discussed shortly. This causes certain differences in the variance inflation due to the two procedures, as we discuss below.

Gouweleeuw et al. [19] defined a PRAM to be an invariant PRAM if P satisfies the condition

$$P\mathbf{T} = \mathbf{T} \quad \text{or equivalently} \quad P\hat{\pi}_0 = \hat{\pi}_0, \tag{2.1}$$

where $\hat{\pi}_0 = \mathbf{T}/n$. Under invariant PRAM and multinomial sampling, it follows that $\mathbf{S}/n$ is also an unbiased estimator of π, which one can calculate without knowing P. This is important because data agencies rarely release P or (more generally) much details about how they have perturbed the original data for confidentiality protection. Also, $\mathbf{S}/n$ is always a probability vector, but $\hat{\pi}$ in (1.2) may not be so. The solution space of (2.1) is a non-empty convex set, which also includes the identity matrix. Different methods for solving (2.1) and hence obtaining invariant PRAM matrices have been discussed in [19, 26]. One important point to keep in mind is that if invariant PRAM is applied to several variables independently, it may not be possible to estimate the joint probabilities without knowing the PRAM matrix. So, for invariant PRAM we need to think in terms of the cross classification of all variables.

For deriving statistical inferences, first note that if P depends on the original data, the

distribution of $\mathbf{S}$ is not multinomial and the results of Section 1, developed for RR surveys assuming P is fixed, need not hold. In the following, we review certain results from [26]. Let $P = [P_1 : \ldots : P_k]$ and rewrite (2.1) as

$$\sum_{i=1}^{k} T_i P_i = \mathbf{T}. \tag{2.2}$$

Let F_{ij} denote the number of units whose category changed from i to j due to randomization, and let $\mathbf{F}_i = (F_{i1}, \ldots, F_{ik})'$. Then, $\mathbf{S} = \sum_i^k \mathbf{F}_i$, and moreover, given $\mathbf{T}$ and P, $\{\mathbf{F}_i\}$ are independently distributed with $\mathbf{F}_i \sim Mult(T_i, P_i), i = 1, \ldots, k$. This representation yields certain properties of $\hat{\pi}_* = \mathbf{S}/n$ as an estimator of π.

From the preceding discussion it follows that

$$E(\hat{\pi}_*|\mathbf{T}, P) = \frac{1}{n}\sum_{i=1}^{k} E[\mathbf{F}_i|\mathbf{T}, P] = \frac{1}{n}\sum_{i=1}^{k} T_i P_i = \hat{\pi}_0, \tag{2.3}$$

and hence $E(\hat{\pi}_*) = E(\hat{\pi}_0) = \pi$. Thus, $\hat{\pi}_*$ is an unbiased estimator of π. Standard derivations also yield that

$$V(\hat{\pi}_*|\mathbf{T}, P) = \frac{1}{n^2}\sum_{i=1}^{k} T_i[D_{P_i} - P_i P_i'] = \frac{1}{n}[D_{\hat{\pi}_0} - \sum_{i=1}^{k}(\frac{T_i}{n})P_i P_i'], \tag{2.4}$$

and

$$\begin{aligned} V(\hat{\pi}_*) &= V[E(\hat{\pi}_*|\mathbf{T}, P)] + E[V(\hat{\pi}_*|\mathbf{T}, P)] \\ &= V(\hat{\pi}_0) + \frac{1}{n}[D_\pi - E\{\sum_{i=1}^{k}(\frac{T_i}{n})P_i P_i'\}]. \end{aligned} \tag{2.5}$$

In (2.5), $V(\hat{\pi}_0) = [D_\pi - \pi\pi']/n$ and the last term is variance inflation due to invariant post-randomization, which is markedly different from the last term of (1.3). As (2.1) has many solutions, the expectation in (2.5) also depends on the distribution of P given $\mathbf{T}$, i.e., on the method that the agency used to choose and apply a solution of (2.1). The data agency knows $\mathbf{T}$ and P and hence can assess variance inflation by calculating (2.4). Note that (2.4) can be interpreted as a conditional variance-covariance matrix of $\hat{\pi}_*$ and also as an unbiased estimator

(based on the original data) of the variance inflation term in (2.5). In contrast, if P is not released, a data user shall not be able to estimate $V(\hat{\pi}_*)$ or the variance inflation due to post-randomization.

The true nature and expression of $V(\hat{\pi}_*)$ under invariant PRAM and the difficulty of estimating it from released data had remained unnoticed until the recent work of Nayak and Adeshiyan [26]. This also reveals that invariant PRAM is of limited help, in the sense that if P is not released, a user can still obtain an unbiased estimate of π, using $\hat{\pi}_*$, but not its sampling variance. Thus, a user shall not be able to properly construct confidence intervals or test hypotheses concerning π. However, the following result in [26] can be used to derive conservative large sample confidence intervals and tests.

Theorem 2.1. *An upper bound of* $V(\hat{\pi}_*)$ *is*

$$V_{max}(\hat{\pi}_*) = (2 - \frac{1}{n})[\frac{D_\pi - \pi\pi'}{n}] \tag{2.6}$$

in the sense that $[V_{max}(\hat{\pi}_*) - V(\hat{\pi}_*)]$ *is non-negative definite for any invariant PRAM. Moreover, this upper bound is tight, i.e., there exists an invariant PRAM for which* $V(\hat{\pi}_*) = V_{max}(\hat{\pi}_*)$.

We should remember that the preceding results and discussions of this section are valid for multinomial sampling. Invariant PRAM is defined differently for a general probability sampling. In general, an invariant PRAM matrix P is defined as a solution of

$$P\hat{\pi}_w = \hat{\pi}_w, \tag{2.7}$$

where $\hat{\pi}_w$ is a linear unbiased estimator of π based on the original data and pertinent survey weights; see [19]. Essentially, this replaces $\mathbf{T}$ in (2.1) by a weighted count vector. The main idea is that if we apply PRAM with P satisfying (2.7), then $\hat{\pi}_w$ calculated based on the perturbed data would also be an unbiased estimator of π. Nayak and Adeshiyan [26] derived the variance-covariance matrix of this estimator, which also is difficult to estimate from released data.

3. Randomized Response in Privacy Preserving Data Mining

Privacy and confidentiality have become major issues in e-commerce. Explosive advances in storage, computing and networking and expansion of the Internet have produced large databases containing vast amount of personal and transactional records. Businesses and government agencies have been devoting significant resources to build large databases, using on-line surveys and capturing transactional records, and mining them to develop commercial and policy decisions. Concurrently, consumers' concern that their private information may reach undesired people have been growing. The field of privacy preserving data mining (PPDM) has developed, largely due to contributions from computer scientists, to address relevant privacy and confidentiality concerns. We refer the reader to [1, 9, 16, 18] for review of various privacy concerns, many relevant concepts and methods and further references.

Many papers in PPDM literature do not distinguish between privacy and confidentiality. However, in e-commerce context, privacy and confidentiality are sometimes referred to as B2C (business-to-customer) and B2B (business-to-business) privacy, respectively (cf. [3]). Customers are reluctant to give their information to businesses due to B2C privacy concerns. Businesses need to take appropriate measures to protect B2B privacy, i.e., protect data confidentiality, when they publish their databases or share them with other parties. A challenging task in PPDM is to create a perturbed or masked version of the original data from which valid statistical inferences can be derived but not much respondent specific information can be extracted.

Several papers, e.g., [2, 3, 12, 14, 15, 21, 30], have discussed using RR techniques and ideas in PPDM, especially in association rule mining and building classification trees from transactional data. Suitable choice of randomization probabilities and applications of RR to other types of data have also been discussed. For example, [20] use RR in participatory sensing applications, with wireless sensor network architecture. In the following we examine one concept of privacy breach, introduced by Evfimievski et al. [14] and further discussed in [3, 15] and others.

The PPDM context considered in [14] is similar to that of RR surveys considered in Section 1. In PPDM terminology, clients give randomized versions of their values of survey variables to a server. For categorical variables, we may again think in terms of randomizing one variable (X) that is obtained by cross classifying all categorical variables. We shall use our earlier notations, e.g., denote the randomized output variable by Z, the categories of both X and Z by $1, \ldots, k$, and $P(Z = i|X = j)$ by p_{ij} for $i, j = 1, \ldots, k$. Here, an implicit assumption is that X and Z have the same number of categories. This allows $P = ((p_{ij}))$ to be nonsingular, which is convenient for unbiased estimation of π. Since category labels are arbitrary, there is no loss of generality in labeling categories of X and Z the same way.

Similar to RR survey contexts, here also the concern is about how much an intruder having access to the server can learn about a client's true value of X from his reported value of Z. However, a significant difference is that X need not have any sensitive categories. It may be just that the clients want to hide their true values from the server. In RR survey context, a respondent wishes the interviewer's (or an intruder's) posterior probability that the respondent belongs to the sensitive group to be small, but in PPDM, a client wants an intruder's opinion or knowledge about the client's true category not to change much from learning the client's reported value of Z. Formally, an intruder's opinion or information should be expressed in terms of probabilities. This means that an intruder's posterior probabilities relating to a client's true value of X should be close to his prior probabilities.

Let α_i denote an intruder's prior probability that $X = i$ and $\vec{\alpha} = (\alpha_1, \ldots, \alpha_k)$. Thus, an intruder is characterized by the vector $\vec{\alpha}$ of his prior probabilities. Obviously, $\vec{\alpha}$ will be different for different intruders. Then, for a given transition probability matrix P, the intruder's posterior probability that $X = i$ given that $Z = j$ is

$$P(X = i|Z = j) = \frac{\alpha_i p_{ji}}{\sum_{i=1}^{k} \alpha_i p_{ji}}. \tag{3.1}$$

Considering general queries or properties about a client's true value, i.e., general events con-

cerning X, Evfimievski et al. [14] introduced the following:

Definition 3.1. *Let Q be a non-null subset of $\{1, \ldots, k\}$ and $0 < \rho_1 < \rho_2 < 1$ be two given numbers. Then, a randomization operator with transition probability matrix P permits an upward ρ_1-to-ρ_2 privacy breach with respect to Q and a prior distribution $\vec{\alpha}$ if for some $1 \leq j \leq k$ with $P(Z = j) > 0$,*

$$P(X \in Q) = \sum_{i \in Q} \alpha_i < \rho_1 \quad \textit{and} \quad P(X \in Q | Z = j) = \sum_{i \in Q} P(X = i | Z = j) > \rho_2. \tag{3.2}$$

The randomization operator is said to allow a downward ρ_2-to-ρ_1 breach if

$$P(X \in Q) = \sum_{i \in Q} \alpha_i > \rho_2 \quad \textit{and} \quad P(X \in Q | Z = j) = \sum_{i \in Q} P(X = i | Z = j) < \rho_1 \tag{3.3}$$

for some $1 \leq j \leq k$ with $P(Z = j) > 0$.

Actually, [14] used $\leq$ and $\geq$ in place of the strict inequalities in (3.2) and (3.3). We made these minor changes for stating some new findings (see Theorem 3.2) conveniently. A reported value $Z = j$ is said to cause an upward ρ_1-to-ρ_2 privacy breach when (3.2) holds. Note that an upward ρ_1-to-ρ_2 privacy breach with respect to Q is equivalent to a downward $(1 - \rho_1)$-to-$(1 - \rho_2)$ privacy breach with respect to Q^c. For privacy protection, we may wish to select and use a randomization operator that does not allow any upward ρ_1-to-ρ_2 or downward ρ_2-to-ρ_1 privacy breach with respect to any Q and any $\vec{\alpha}$. In other words, we want to guarantee that no intruder's opinion (represented by $\vec{\alpha}$) with respect to any property (Q) of a client will change up or downward substantially, as specified by ρ_1 and ρ_2. This is a fairly stringent goal, but it is attainable, as [14] proved using the following concept.

Definition 3.2. *A transition probability matrix P is said to be at most γ-amplifying, for $\gamma \geq 1$, if*

$$\frac{p_{ij}}{p_{il}} \leq \gamma, \quad \textit{or} \quad \frac{P(Z = i | X = j)}{P(Z = i | X = l)} \leq \gamma \quad \textit{for all } i, j, l = 1, \ldots, k. \tag{3.4}$$

Theorem 3.1. *[14] A sufficient condition for a randomization operator with transition probability matrix P to guarantee no upward ρ_1-to-ρ_2 privacy breach with respect to any Q and any $\vec{\alpha}$ is that P is at most γ-amplifying for some*

$$\gamma \leq \frac{\rho_2(1-\rho_1)}{\rho_1(1-\rho_2)}. \tag{3.5}$$

Furthermore, this condition also ensures no downward ρ_2-to-ρ_1 privacy breach.

Definition 3.2 tells how to verify whether a given P is at most γ-amplifying or not for given γ. While a given P can be at most γ-amplifying for many values of γ, (3.5) really refers to the minimum of all such values. We believe it is helpful to define this quantity explicitly.

Definition 3.3. *We define the parity of a transition probability matrix P as*

$$\eta(P) = \max\Big\{\frac{p_{ij}}{p_{il}}, i, j, l = 1, \ldots, k\Big\}, \tag{3.6}$$

where we use $0/0 = 1$ and $a/0 = \infty$ for all $a > 0$.

Clearly, $\eta(P) \geq 1$ and it is finite if and only if all elements of P are positive. Also, P is at most γ-amplifying if and only if $\gamma \geq \eta(P)$. The condition in (3.5) can be restated as $\eta(P) \leq \frac{\rho_2(1-\rho_1)}{\rho_1(1-\rho_2)}$. Then, for given ρ_1 and ρ_2, no privacy breaches can be guaranteed by using a P with $\eta(P) = [\rho_2(1-\rho_1)]/[\rho_1(1-\rho_2)]$. The papers [3, 14] give methods for constructing such P matrices. In particular, a P matrix with parity η is obtained by taking $p_{ii} = \eta/[\eta+k-1], i = 1, \ldots, k$, and $p_{ij} = 1/[\eta + k - 1]$ for all $i \neq j$. Obviously, a P matrix chosen to guarantee no privacy breach for specific ρ_1 and ρ_2 may not work for other pairs of values. We may note that Leysieffer and Warner [23] also used the ratios $\{\frac{p_{ij}}{p_{il}}\}$ for comparing RR surveys. Generally, as these ratios get closer to 1, privacy protection increases, as seen also in Theorem 3.1, but accuracy of statistical inferences decreases.

How to choose ρ_1 and ρ_2 in practical applications of Theorem 3.1 is not clear and the question deserves further investigation. For a simpler notion of privacy breach, we shall consider ratios

of posterior to prior probabilities, which are natural indicators of an intruder's information gain about a respondent from the data. Our basic idea is that a privacy breach occurs if the ratio $P(X \in Q|Z = j)/P(X \in Q)$ is too large or too small. Formally, we introduce the following:

Definition 3.4. *We say that a transition probability matrix P admits a β-factor privacy breach, for $\beta \geq 1$, with respect to a subset Q of $\{1, \ldots, k\}$ and a prior distribution $\vec{\alpha}$ if for some $1 \leq j \leq k$ with $P(Z = j) > 0$, either*

$$\frac{P(X \in Q|Z = j)}{P(X \in Q)} > \beta \quad or \quad \frac{P(X \in Q|Z = j)}{P(X \in Q)} < \frac{1}{\beta}. \tag{3.7}$$

We shall say that P guarantees β-factor privacy if it does not allow any breach, with respect to any Q and any $\vec{\alpha}$, as per the preceding criterion. One advantage of Definition 3.4 is that it requires us to specify the value of only one quantity (viz. β). In the following, we establish a connection between the parity of P and its β-factor privacy guarantee. Suppose P has a finite parity $\eta > 1$ and $\vec{\alpha}$ is a prior distribution with $\alpha_i > 0, i = 1, \ldots, k$. Then,

$$\frac{1}{\eta} \leq \frac{p_{ij}}{p_{il}} \leq \eta \quad \text{for all } i, j, l = 1, \ldots, k. \tag{3.8}$$

Using (3.1) and the left inequality in (3.8) we obtain, for all $i, j = 1, \ldots, k$,

$$\begin{aligned} \frac{P(X = i|Z = j)}{P(X = i)} &= \Big[\alpha_i + \sum_{l \neq i} \alpha_l \Big(\frac{p_{jl}}{p_{ji}}\Big)\Big]^{-1} && (3.9) \\ &\leq \Big[\alpha_i + \frac{1}{\eta} \sum_{l \neq i} \alpha_l\Big]^{-1} \\ &= \Big[\alpha_i + \frac{1}{\eta}(1 - \alpha_i)\Big]^{-1} \\ &= \frac{\eta}{1 + (\eta - 1)\alpha_i} \leq \eta, && (3.10) \end{aligned}$$

as $\eta \geq 1$. Thus, $P(X = i|Z = j) \leq \eta\alpha_i$ for all $i, j = 1, \ldots, k$, which implies that for any subset Q of $\{1, \ldots, k\}$, $\sum_{i \in Q} P(X = i|Z = j) \leq \eta \sum_{i \in Q} \alpha_i$, or

$$\frac{P(X \in Q|Z = j)}{P(X \in Q)} \leq \eta. \tag{3.11}$$

As we show next, the inequality in (3.10) is tight in the sense that (3.9) can be made arbitrarily close to η with suitable choice of i, j and $\vec{\alpha}$. Suppose $\eta(P) = \eta > 1$. Then, there exists i, j and l such that $p_{jl}/p_{ji} = 1/\eta$. For notational simplicity, suppose $p_{12}/p_{11} = 1/\eta$. Take $\alpha_2 = a$ and $\alpha_i = (1-a)/(k-1)$ for $i \neq 2$, where $0 < a < 1$. For this $\vec{\alpha}$, we get

$$\frac{P(X=1|Z=1)}{P(X=1)} = \Big[\frac{1-a}{k-1} + a\frac{1}{\eta} + \frac{1-a}{k-1}\sum_{l=3}^{k}\Big(\frac{p_{1l}}{p_{11}}\Big)\Big]^{-1}. \tag{3.12}$$

Clearly, the right side of (3.12) converges to η as $a \to 1$.

Using the second inequality of (3.8) in (3.9), we get

$$\begin{aligned}\frac{P(X=i|Z=j)}{P(X=i)} &\geq \Big[\alpha_i + \eta\sum_{l\neq i}\alpha_l\Big]^{-1} \\ &= \Big[\alpha_i + \eta(1-\alpha_i)\Big]^{-1} \\ &= \Big[\eta - (\eta-1)\alpha_i\Big]^{-1} \geq \frac{1}{\eta},\end{aligned} \tag{3.13}$$

as $\eta > 1$. As before, it can be seen that the inequality in (3.13) is tight. Comparing these results with Definition 3.4, we have the following:

Theorem 3.2. *A randomization operator with transition probability matrix P guarantees β-factor privacy with respect to all Q and all $\vec{\alpha}$ if and only if*

$$\eta(P) \leq \beta. \tag{3.14}$$

Thus, we can provide β-factor privacy by constructing and implementing a P matrix with parity β. Theorem 3.2 is well suited for practical applications as it requires us to specify just the value of β. In practice, this value can be determined from a survey of respondents and collecting data on their desired threshold for posterior to prior probability ratio. In practice, eliciting desired values of β would be simple than eliciting values of (ρ_1, ρ_2). We should also note that definitions 3.1 and 3.4 are closely connected and Theorem 3.2 implies, for example, that P guarantees no upward ρ_1-to-ρ_2 privacy breach for all ρ_1 and ρ_2 (and with respect to any Q and any $\vec{\alpha}$) as long as $\rho_2/\rho_1 \geq \eta(P)$.

4. Discussion and Remarks

Fifty years have passed since the publication of the first RR paper [34], due to Warner. Astounding technological innovations during this period have changed the world, including statistical practice and applications, substantively and significantly. The fact that the basic ideas and mathematical results in RR that Warner introduced are being used and further developed for addressing contemporary problems signify the value of his contribution. In this paper, we have discussed adaptations of RR techniques in confidentiality protection and privacy preserving data mining. We have noted some new features of these emerging applications and some related results and issues. However, there are many additional questions that deserve further research, a few of which we mention in the following.

For protecting the confidentiality of a data set, an agency should look at the data set, its features and disclosure issues to choose an appropriate perturbation procedure. Thus, for PRAM, it makes good sense to choose P based on the original data. But, in that case, inferences derived for RR surveys, where P is known and fixed, do not apply. As we have seen, for invariant PRAM, it is difficult to calculate standard errors of estimators and hence confidence intervals. This prompts two immediate questions for further investigation. How can we estimate standard errors of estimators from post-randomized data? Are there suitable choices of P for which valid estimates of π along with their standard errors can be calculated from perturbed data? These questions should be investigated not only for multinomial sampling but also for complex survey designs.

Confidentiality protection is more complex and different from privacy protection in the context of [34]. Disclosure protection goals are difficult to articulate and ascertain and only modest work has been done on effects of the PRAM matrix P on disclosure risk and hence on how to choose P. Much of the work has focused on identity disclosure risk and we refer to [31] for a recent discussion and relevant references. Note that identity disclosure is not an issue in RR

surveys, where the interviewer knows the identity of the respondent and the concern is only about predictive disclosure. Much new research, both theoretical and empirical, on appropriate choices of P for achieving confidentiality protection goals is needed for making PRAM a common data perturbation tool.

Privacy preserving data mining is an important and very active area that offers significant research opportunities to statisticians. As one might expect, there are special data structures and privacy concerns in PPDM that do not arise in RR surveys. One common purpose of data mining is to extract association rules from transactional data. Typically, we have a set of M items and choices (or purchases) of n clients. Essentially, each client selects a subset of the M items, called an itemset. Note that itemsets of different clients may contain different number of items. The data for each client can be recorded as a vector of M binary variables, one for each item, with 1 indicating that a client included the item in his itemset and 0 otherwise. The cross classification of these variables yields one categorical variable with 2^M categories. Here, RR can be used to perturb each client's itemset for privacy protection. However, in many applications, M is moderately large and 2^M is very large, and it is not possible to use (1.2) for estimating π, as it requires inverting a matrix of order 2^M (see [30]). Obviously, here we need to use P with special structures to overcome computational obstacles. There are many other special issues associated with other data mining goals, such as finding classification rules, deserving further research.

References

[1] Aggarwal, C.C. and Yu, P.S. (Eds.) (2008). *Privacy-Preserving Data Mining: Models and Algorithms*, Springer, Berlin.

[2] Agrawal, R. and Srikant, R. (2000). Privacy-preserving data mining. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Dallas, Texas, pp. 439-450.

[3] Agrawal, S., Haritsa, J.R. and Prakash, B.A. (2009). FRAPP: A Framework for high-accuracy privacy-preserving mining. *Data Mining and Knowledge Discovery*, 18, 101-139.

[4] Brand, R. (2002). Microdata protection through noise addition. In *Inference Control in Statistical Databases*, J. Domingo-Ferrer (ed.), Springer, Berlin, pp. 97-116.

[5] Chaudhuri, A. (2001). Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. *J. Statist. Plann. Inference* 94, 37-42.

[6] Chaudhuri, A. (2010). *Randomized Response and Indirect Questioning Techniques in Surveys*, CRC Press, Boca Raton.

[7] Chaudhuri, A. and Christofides, T.C.. (2013). *Indirect Questioning in Sample Surveys*. Springer, New York.

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

[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] Cruyff, M.J.L.F., Van den Hout, A. and Van der Heijden, P.G.M. (2008). The analysis of randomized response sum score variables. *J. Royal Statist. Soc.*, Ser. B, 70, 21-30.

[11] Doyle, P., Lane, J., Theeuwes, J., and Zayatz, L. (Ed) (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Elsevier, Amsterdam.

[12] Du, W. and Zhan, Z. (2003). Using randomized response techniques for privacy-preserving data mining. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, Washington DC, pp. 505-510.

[13] Evans, T., Zayatz, L., and Slanta, J. (1998). Using noise for disclosure limitation of establishment tabular data. *J. Official Statist*, 4, 537-551.

[14] Evfimievski, A., Gehrke, J. and Srikant, R. (2003). Limiting privacy breaches in privacy-preserving data mining. *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, San Diego, pp. 211-222.

[15] Evfimievski, A., Srikant, R. Agrawal, R. and Gehrke, J. (2004) Privacy preserving mining of association rules. *Information Systems*, 29, 343-364.

[16] Fienberg, S.E. (2006). Privacy and confidentiality in an ecommerce world: data mining, data warehousing, matching and disclosure limitation. *Statistical Science*, 21, 143-154.

[17] Fuller, W.A. (1993). Masking procedures for microdata disclosure limitation. *J. Official Statist.*, 383-406.

[18] Fung, B.C.M., Wang, K., Chen, R. and Yu, P.S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys*, 42, 1-53.

[19] Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and De Wolf, P.-P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *J. Official Statist.*, 14, 463–478.

[20] Groat, M.M., Edwards, B., Horey, J., He, W. and Forrest, S. (2012). Enhancing privacy in participatory sensing applications with multidimensional data. *Proceedings of 2012 IEEE International Conference on Pervasive Computing and Communications*, Lugano, Switzerland, pp. 144-152.

[21] Guo, L., Ying, X. and Wu, X. (2011). Limiting attribute disclosure in randomization based microdata release. *J. Computing Science and Engineering*, 5, 169-182.

[22] Lambert, D. (1993). Measure of disclosure risk and harm. *J. Official Statist.*, 9, 313-331.

[23] Leysieffer, R.W. and Warner, S.L. (1976). Respondent jeopardy and optimal designs in randomized response models. *J. Amer. Statist. Assoc.*, 71, 649-656.

[24] Nayak, T.K. (1994). On randomized response surveys for estimating a proportion. *Commun. Statist.- Theory Meth.*, 23, 3303-3321.

[25] Nayak, T.K. and Adeshiyan, S.A. (2009). A unified framework for analysis and comparison of randomized response surveys of binary characteristics. *J. Statist. Plann. Inference*, 139, 2757-2766.

[26] Nayak, T.K. and Adeshiyan, S.A. (2015). On invariant post-randomization for statistical disclosure control. *Internat. Statist. Rev.* (to appear). doi:10.1111/insr.12092.

[27] Nayak, T.K., Sinha, B. and Zayatz, L. (2011). Statistical properties of multiplicative noise masking for confidentiality protection. *J. Official Statist.*, 27, 527-544.

[28] Padmawar, V. R., and K. Vijayan (2000). Randomized response revisited. *Journal of statistical planning and inference*, 90.2 : 293-304.

[29] Quatember, A. (2009). A standardization of randomized response strategies. *Survey Methodology*, 35, 143-152.

[30] Rizvi, S.J. and Haritsa, J.R. (2002) Maintaining data privacy in association rule mining. *Proceedings of the 28th International Conference on Very Large Databases (VLDB)*, Hong Kong. pp. 682-693.

[31] Shlomo, N. and Skinner, C. (2010). Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *Ann. Appl. Statist.,* 4, 1291-1310.

[32] Van den Hout, A. and Elamir, E.A.H. (2006). Statistical disclosure control using post randomisation: Variants and measures for disclosure risk. *J. Official Statist.*, 22, 711-731.

[33] Van den Hout, A. and Van der Heijden, P.G.M. (2002). Randomized response, statistical disclosure control and misclassification: A review. *Internat. Statist. Rev.*, 70, 269-288.

[34] Warner, S.L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.*, 60, 63-69.

[35] Warner, S.L. (1971). The linear randomized response model. *J. Amer. Statist. Assoc.*, 66, 884-888.

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