RESEARCH REPORT SERIES
(Disclosure Avoidance #2016-02)

Measuring Identification Risk in Microdata Release and Its Control by Post-randomization

Tapan K. Nayak, Cheng Zhang and Jiashen You

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

Report Issued: May 2, 2016

Measuring Identification Risk in Microdata Release and Its Control by Post-randomization

Tapan K. Nayak* Cheng Zhang† and Jiashen You‡§

Abstract

Statistical agencies often release a masked or perturbed version of survey data to protect respondents' confidentiality. Ideally, a perturbation procedure should protect confidentiality without much loss of data quality, so that released data may practically be treated as original data for making inferences. One major objective is to control the risk of correctly identifying any respondent's records in released data, by matching the values of some identifying or key variables. For categorical key variables, we propose a novel approach to measuring identification risk and setting strict disclosure control goals. The general idea is to ensure that the probability

*Center for Disclosure Avoidance Research, 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.

‡Bureau of Transportation Statistics, U.S. Department of Transportation, Washington, DC 20590

§The views expressed in this article are those of the authors and not necessarily those of the U.S. Census Bureau or U.S. Department of Transportation.

of correctly identifying any respondent or surveyed unit is at most ξ, which is pre-specified. Then, we develop an unbiased post-randomization procedure that achieves this goal for $\xi > 1/3$. The procedure allows substantial control over possible changes to the original data and the variance it induces is of a lower order of magnitude than sampling variance. We apply the procedure to a real data set, where it performs consistently with the theoretical results and quite importantly, shows very little data quality loss.

Key words and Phrases: Correct match probability; data partitioning; data utility; transition probability matrix; unbiased post-randomization.

1. Introduction

Many statistical agencies aim to collect and release informative data to help policy makers and researchers make appropriate inferences and decisions. But, agencies also need to keep individual or unit level information confidential for legal reasons and upholding public trust and support. So, they often release a perturbed or masked version of the original data. Several books (e.g., Willenborg and de Waal, 2001; Duncan et al., 2011; Hundepool et. al., 2012) and many papers discuss disclosure issues and various data masking techniques, such as grouping, data swapping, cell suppression, random noise infusion and post-randomization.

Data masking reduces data quality; it may suppress, dilute, and even distort some features of the original data. Intuitively, one should examine the trade-offs between disclosure risk and data utility (see e.g., Duncan and Stokes, 2004; Cox et al., 2011) to choose suitable data masking procedures. But, it is hard to define and measure risk and utility, as there are many scenarios and forms of disclosure and likewise a data set may be used in numerous ways by diverse users. Thus, assessing risk-utility trade-offs in practice is a difficult task. Broadly speaking, disclosure occurs when the released data give an intruder R much new knowledge about a target unit B. According to Dalenius (1977), rigorous confidentiality protection (or equivalently disclosure avoidance) means that the released data should not enable any intruder to gain much new information about any target unit. Thus, any intruder's prior and posterior probabilities about any property of any target unit should not differ much. However, this goal is not achievable, as shown in Dwork's (2006) pioneering paper. Thus, data agencies need to determine measures of disclosure risk and data utility and their disclosure control goals specialized for each

application. We believe, these are the fundamental challenges in practical data masking.

Dwork's (2006) work also shows that the main reason why Dalenius's criterion is impractical is that it allows unrestricted choices for intruder's target and prior information, which is overly stringent. Thus, for developing practical disclosure control goals, it is essential for the agency to consider intruders with limited prior information about their target units. One common approach is to specify a subset of the survey variables as identifying or key variables, whose values are fairly easily accessible from other sources, and then assume that intruders may know only the values of some or all of the key variables for their target units. The choice of key variables describes a universe of intruders that is of concern to the agency. In this framework, an identity disclosure occurs when an intruder correctly identifies the records of a target unit (B) in the released data, by matching B's values for the key variables (known to the intruder). Usually, this is considered to be the most serious type of disclosure and has been discussed by many researchers including Bethlehem et al. (1990), Greenberg and Zayatz (1992), Willenborg and de Wall (2001), Skinner and Elliot (2002), Reiter (2005) and Shlomo and Skinner (2010).

In this paper, we focus on identity disclosure based on categorical key variables. First, we shall propose a measure for identification risk and some associated disclosure control goals. We suggest to ensure that no intruder's confidence in his match of a target unit's record in the released data, based on the key variables, can justifiably be larger than ξ, where both ξ and the key variables are specified by the agency. We believe that this goal is strong and easy to communicate. Then, we shall present a method that accomplishes the preceding goal for moderate ξ (essentially $\xi \geq 0.35$). Actually, our method applies PRAM (the Post-randomization Method, introduced by Gouweleeuw et al. (1998)) with suitably chosen transition probabilities. In a sense, our work also gives a useful answer to

the general question of how to choose the transition probability matrix of PRAM.

Valid inferences from released data should account for the masking mechanism that was used by the agency. The probability distribution of the random observables whose values constitute released data depends on both the sampling design and the masking method. So, the agency should give full information about the masking procedure to permit data users to derive appropriate likelihood functions and valid inferences, and investigate properties of statistical methods of their interest. However, as Cox et al. (2011) noted, typically statistical agencies do not give much details about their masking procedures, especially the parameter values of the used procedures, primarily due to disclosure risk concerns. This lack of transparency constrains the scope for deriving principled inferences from released data. Rubin (1993) made a related observation: It is imperative that public-use microdata preserve the user's ability to obtain valid inferences using standard statistical methods, that is, without having to develop complex methods and software. Thus, statistical agencies should apply masking procedures for which common inferential methods for original data would remain valid, at least approximately, for masked data.

The preceding arguments suggest to develop and use data masking procedures which (i) enable the data agency to evaluate, assure and communicate confidentiality protection clearly and (ii) do not require much new theoretical derivations and extensive programming for making valid inferences from masked data. We shall see that our procedure accomplishes both goals fairly well. As our method is fully probabilistic, its properties can be investigated analytically. The method is unbiased in the sense that for each cell, the expected change in its frequency, due to data perturbation, is zero. Moreover, the added variance due to our data masking is negligible in comparison to sampling variance.

In the next section, we briefly review some past approaches to assessing identification

risk and then propose a new measure and an approach to setting strong, precise and workable identification risk control goals. In Section 3, we review the Post-randomization Method (PRAM) and some of its properties and introduce the inverse frequency rule post-randomization (IFPR), which is a cornerstone of the general method developed in this paper. In Section 4, we derive and examine identification risks under PRAM and thereby develop methods for achieving certain disclosure control goals. We present a general post-randomization method, consisting of four steps, in Section 5. It allows us to greatly control the nature or magnitude of changes to the original data via data partitioning. In Section 6, we examine the effects of our procedure on data utility. We show that the additional variance due to our method is negligible in comparison to sampling variance. In Section 7, we apply our method to a real data set containing values of many variables for 59,033 persons. As expected, the empirical and theoretical results agree well and for many sets of variables that we examined, the distributions based on the original and perturbed data sets are very close. We make some concluding remarks in Section 8.

2. Identification Risk and Its Control

2.1. A Review of Past Work

Extending the work of Duncan and Lambert (1986, 1989) and Lambert (1993), Reiter (2005) developed an extensive Bayesian approach for measuring identification disclosure risk. It incorporates the sampling design, dependencies among survey variables and intruder's knowledge about the perturbation mechanism and his prior information about the target. However, the method involves substantial modeling, guessing intruder's behavior,

estimation and computation, and as Shlomo and Skinner (2010) noted, the complexity of the approach may limit its application in practice.

Another approach to measuring identity disclosure risk focuses on the units in the sample that are unique by the key variables; see e.g., Bethlehem et al. (1990), Greenberg and Zayatz (1992), Skinner and Elliot (2002) and Shlomo and Skinner (2010). This essentially requires all key variables to be categorical, because for continuous key variables almost all sampled units will be unique. Thus, suppose all key variables are categorical and X is their cross-classification and that X has k cells (or categories), denoted $c_1, \ldots, c_k$. For confidentiality protection, the agency perturbs the original values of some (or all) of the key variables. Let Z denote the perturbed version of X, with the same set of categories, and for $i = 1, \ldots, k$, let T_i and S_i denote the frequencies of category c_i before and after data perturbation, respectively. The main concern is about an intruder trying to identify a target unit B's records in the released data by directly matching B's X-category.

Suppose that B's X-category is c_j, denoted $X_{(B)} = c_j$, which is known to intruder R. If $T_j = 1$ and R also knows that B is in the sample, then he will correctly identify B's record in the original data. If $T_j = 1$ and R does not know if B is in the sample or not, he will have a correct match if B is unique in the population with respect to X. Thus, the probability that a unit is population unique, given that it is sample unique, has received much attention and several researchers discussed estimating it from the original data, under certain sampling designs and models; see e.g., Bethlehem et al. (1990), Greenberg and Zayatz (1992) and Skinner and Elliot (2002). This conditional probability is useful for identifying sampled units that have high disclosure risk, if the original data are released. However, it does not account for data perturbation and hence it is an indirect (and inadequate) indicator of disclosure risk when a perturbed data set is released.

Shlomo and Skinner (2010) took a more direct approach. In released data, R will find a unique match if $S_j = 1$, but that match may or may not be correct because of data perturbation and sampling. Building on the ideas of Bethlehem et al. (1990) and considering unique match as the worst case (presuming that identification risk will be lower if B's key variables match more than one unit in released data), Shlomo and Skinner (2010) defined *identification risk* (IR) as the probability that a unique match for B is a (unique) correct match (UCM), that is,

$$IR(j) = P(UCM|\text{unique match}) = P(UCM|S_j = 1), \tag{2.1}$$

where the probability is with respect to both sampling and data perturbation. Note that the probability in (2.1) depends on the target unit only through its X-category. For any population unit u, let E_u denote the event that unit u is sampled and $X_{(B)}$ uniquely matches the value of Z for unit u, to be denoted $Z_{(u)}$, in released data. Then, (2.1) can also be expressed as

$$IR(j) = P(E_B) \div \sum_{u \in U} P(E_u), \tag{2.2}$$

where U denotes the set of all units in the population.

For an illustration of (2.2), suppose (similar to Shlomo and Skinner (2010)) that if $u \in U$ and $X_{(u)} = c_j$, then u is selected in the sample with probability δ_j and independently of other units. Thus, the inclusion probability of each unit may depend only on its value of X. Also, suppose that the original X-values are perturbed stochastically and independently with known transition probabilities

$$p_{ij} = P(Z = c_i|X = c_j), \quad i, j = 1, \ldots, k.$$

Suppose $X_{(B)} = c_j$ and $X_{(u)} = c_m$ and let F_i denote the population frequency of $X =$

$c_i, i = 1, \ldots, k$. Then,

$$\begin{aligned} P(E_u) &= \delta_m p_{jm}[1 - \delta_m p_{jm}]^{F_m - 1} \prod_{l \neq m} [1 - \delta_l p_{jl}]^{F_l} \\ &= \frac{\delta_m p_{jm} \eta_j}{1 - \delta_m p_{jm}}, \end{aligned} \tag{2.3}$$

where $\eta_j = \prod_{l=1}^{k} [1 - \delta_l p_{jl}]^{F_l}$ (which is the probability of finding no match for $X_{(B)}$ in released data) and for the second equality we assume for simplicity (and realistically) that $\delta_m p_{jm} \neq 1$. Now, from (2.2) and (2.3) it follows that

$$IR(j) = \frac{\delta_j p_{jj}}{1 - \delta_j p_{jj}} \div \sum_{l=1}^{k} \Big(F_l \times \frac{\delta_l p_{jl}}{1 - \delta_l p_{jl}} \Big). \tag{2.4}$$

We may mention that Shlomo and Skinner (2010) assumed that for any unit $u \in U$, its selection probability depends on $Z_{(u)}$ (rather than $X_{(u)}$), in which case (2.4) reduces to

$$IR(j) = \frac{p_{jj}}{1 - \delta_j p_{jj}} \div \sum_{l=1}^{k} \Big(F_l \times \frac{p_{jl}}{1 - \delta_l p_{jl}} \Big). \tag{2.5}$$

Clearly, (2.4) and (2.5) depend on $X_{(B)}$ and also on population frequencies $\{F_i\}$, which are usually unknown. For two data masking procedures, Shlomo and Skinner (2010) estimated (2.5) from original data using a Poisson log-linear model. They also suggested to use the sum of (2.5), over all sample unique units, as an aggregate measure of disclosure risk for a data set. We note some remarks on the preceding approach. First, the measure in (2.1) depends on $\{F_i\}$ and hence cannot be calculated from available information. Evaluating and estimating (2.1) might be difficult for complex sampling designs. Even for simple designs, estimation of (2.1) requires additional work and the results may depend substantially on data modeling. Second, the sum of (2.1) (or (2.5)), over the sample unique units, being small does not imply desirably small disclosure risk for all units. Indeed, some units may carry high disclosure risks even when the aggregate measure is small. Finally,

the search for a suitable masking procedure requires an iterative approach. For example, to apply data swapping, as described in Shlomo and Skinner (2010), one would need to evaluate the aggregate risk measure for various *swap rates* to choose a suitable swap rate for actual application.

2.2. Our Approach

Our framework and perspective are similar to those of Shlomo and Skinner (2010), but we consider more stringent and directly relevant disclosure control goals and eliminate the estimation task. Also, while most approaches require a trial and error process for selecting a suitable masking method, we can find procedures for achieving our disclosure control goals directly. We consider the scenario where an intruder R knows $X_{(B)} = c_j$ for his target unit B and randomly selects one of the units in the released data with $Z = c_j$, if any, and concludes that to be unit B. If $S_j = 0$, R stops searching for B's records in the released data. Then, our general idea for a strong identification disclosure risk control goal is to ensure that no intruder's match, in the preceding scheme, for any unit in the sample, or in important subsets, would be correct with probability larger than ξ, where the value of ξ is chosen by the agency. This will lead to various specific goals when different subgroups of sampled units and probabilities are used.

Considering all units and the conditional probability of correct match given S_j, i.e., the number of matches an intruder finds for his target unit in the released data, the general goal reduces to ensuring that

$$R_j(a) \equiv P(CM|X_{(B)} = c_j, S_j = a) \leq \xi \quad \text{for all } a > 0 \text{ and } j = 1, \ldots, k, \tag{2.6}$$

where CM stands for the event that unit B is correctly matched under the scenario and

matching scheme stated above. Conditioning on S_j is very sensible, perhaps even necessary, because an intruder's confidence in his declared match may depend on the number of matches he finds in the released data for his target unit. We take (2.6) as a fundamental identification disclosure risk control goal. Clearly, (2.6) implies $P(CM|X_{(B)} = c_j) \leq \xi$, i.e., the unconditional correct match probabilities do not exceed ξ.

Regarding the agency's choice of a suitable value of ξ, note that the role of $1-\xi$ is similar to the role of level of significance (α) in hypotheses testing. The intruder should have strong evidence to rationally conclude a match, as in accepting an alternative hypothesis. To declare a match, an intruder should confirm that relevant $R_j(a)$ is substantial, arguably larger than 1/2. We believe that in most practical situations, it is not necessary to use a value for ξ that is less than 1/3. As Lambert (1993) explains, intruders are free to declare matches or draw other conclusions without adequate justification and thereby induce harm, but such actions are beyond any data agency's control.

One difficulty in calculating $\{R_j(a)\}$ and thereby verifying (2.6) is that the probabilities $\{R_j(a)\}$ depend also on the unknown population frequencies. We overcome this challenge (and avoid estimation of $\{R_j(a)\}$) by considering correct match probabilities conditional also on the original frequencies. Letting $\mathbf{T} = (T_1, \ldots, T_k)'$ denote the frequency vector from original data, we can write

$$R_j(a) = \sum_{\mathbf{t}} P(CM|X_{(B)} = c_j, S_j = a, \mathbf{T} = \mathbf{t}) P(\mathbf{T} = \mathbf{t}). \tag{2.7}$$

We shall denote $P(CM|X_{(B)} = c_j, S_j = a, \mathbf{T} = \mathbf{t})$ by $R_j(a, \mathbf{t})$. Then, we can guarantee (2.6) by ensuring that

$$R_j(a, \mathbf{t}) \leq \xi \quad \text{for all } a > 0, j = 1, \ldots, k \text{ and } \mathbf{t}. \tag{2.8}$$

Note that in (2.7), $P(\mathbf{T} = \mathbf{t})$ depends on the population frequencies $\{F_i\}$. But $\{R_j(a, \mathbf{t})\}$

do not depend on $\{F_i\}$ when the data are perturbed using PRAM. In Section 4, we describe methods for achieving (2.8) for moderate to large ξ. Essentially, we take (2.8) as our disclosure control goal, which ensures that any correct match probability would not be larger than ξ, even if the intruder knows the frequencies in the original data.

Remark 2.1. One common concern of data agencies is potential identity disclosure of units falling in low frequency cells. This suggests another disclosure control goal, viz., use a masking procedure under which the correct match probability for any unit in any low frequency cell would be no more than ξ. Formally, this asks us to guarantee that

$$\gamma(j, \mathbf{t}) = P(CM|X_{(B)} = c_j, \mathbf{T} = \mathbf{t}) \leq \xi \tag{2.9}$$

for all j such that $t_j \leq b_0$, where b_0 is a given value (defining 'low frequency'). Note that (2.8) implies (2.9) as

$$\begin{aligned} \gamma(j, \mathbf{t}) &= \sum_a R_j(a, \mathbf{t}) P(S_j = a|X_{(B)} = c_j, \mathbf{T} = \mathbf{t}) \\ &\leq \xi \times P(S_j \geq 1|X_{(B)} = c_j, \mathbf{T} = \mathbf{t}) \leq \xi. \end{aligned}$$

3. The Post-randomization Method (PRAM)

3.1. Basic Mechanism

Motivated by randomized response methods (e.g., Warner, 1965, 1971; Chaudhuri and Mukerjee, 1988; Nayak and Adeshiyan, 2009; Nayak et al., 2016) in survey sampling, Gouweleeuw et al. (1998) introduced the Post-randomization Method (PRAM) for perturbing categorical data for protecting the confidentiality of respondents' information. PRAM can be applied independently to multiple categorical variables, or jointly to their

cross-classification. However, the method can always be viewed as being applied to the cross-classification, and that is also essential for ascertaining the method's effects on confidentiality protection and performing statistical analyses. Interesting properties, variations and applications of PRAM have appeared in many publications including Willenborg and De Waal (2001), Van den Hout and Van der Heijden (2002), Van den Hout and Elamir (2006), Van den Hout and Kooiman (2006), Cruyff et al. (2008), Shlomo and De Waal (2008), Shlomo and Skinner (2010) and Nayak and Adeshiyan (2015).

Let X be a categorical variable (or cross-classification of several variables) with categories (or cells) $c_1, \ldots, c_k$. The basic steps for perturbing data on X using PRAM are: (i) select a matrix of probabilities $P = ((p_{ij}))$ such that $\sum_i p_{ij} = 1$ for $j = 1, \ldots, k$, and then (ii) randomly change any original category c_j to c_i with probability p_{ij} $(i, j = 1, \ldots, k)$. The randomization step is executed independently for all units in the data set. Clearly, $p_{ij} = P(Z = c_i | X = c_j)$, where Z denotes the perturbed variable. In general, proper analyses of perturbed data require P. So, agencies should release P along with the perturbed data to enable data users derive correct inferences.

If P is fixed, post-randomization is mathematically equivalent to a randomized response (RR) survey, and known results in RR theory can be applied to post-randomized data. As before, let $\mathbf{T}$ and $\mathbf{S}$ denote the frequency vectors before and after applying PRAM, respectively. Let $\pi_i = P[X = c_i], i = 1, \ldots, k$, and $\Pi = (\pi_1, \ldots, \pi_k)'$. Let n denote the sample size. Then, under multinomial sampling, and fixed P, $\mathbf{T} \sim Mult(n, \Pi)$ and $\mathbf{S} \sim Mult(n, \lambda)$, where

$$\lambda = P\Pi. \tag{3.1}$$

From this, it follows that $\hat{\Pi} = P^{-1}(\mathbf{S}/n)$ is an unbiased estimator of π and

$$Var(\hat{\Pi}) = \frac{(D_\Pi - \Pi\Pi')}{n} + \frac{[P^{-1}D_\lambda(P^{-1})' - D_\Pi]}{n}, \tag{3.2}$$

where D_Π is a diagonal matrix with diagonal elements being $\pi_1, \ldots, \pi_k$ and D_λ is defined similarly (see Chaudhuri and Mukerjee, 1988, p. 43). The first term on the right side of (3.2) is the variance of $\hat{\Pi}_0 = \mathbf{T}/n$, which is the UMVUE of Π based on the original data, and the last term is the 'variance inflation' due to post-randomization.

We should note that RR theory is of limited help to data users, for two reasons. First, data agencies usually give little information about how they have perturbed the data and rarely release the transition probability matrix. Mostly, data users would not know P and hence would not be able to use results from RR literature. Second, data agencies should select their perturbation procedure after ascertaining particular disclosure control needs for the original data set. For example, if nothing needs protection (i.e., disclosure risk is negligible for all units), the data should not be perturbed at all. Thus, P should not be fixed and should be chosen based on the original data, in which case mathematical results in RR theory (for fixed P) may not apply. One type of data dependent P, which also has some attractive features, is discussed next.

3.2. Unbiased (Invariant) Post-randomization

As we noted earlier, agencies should perturb original data in such a way that standard inferential methods for original data would remain valid, at least approximately, for perturbed data. A common unbiased estimator of Π, based on the original data, is $\hat{\Pi}_0 = \mathbf{T}/n$. So, one might desire $\hat{\Pi}_* \equiv \mathbf{S}/n$ to be a valid (viz. unbiased) estimator of Π. If P depends

on the data only through $\mathbf{T}$, then $E[\mathbf{S}|\mathbf{T}] = P\mathbf{T}$ and it follows that if P also satisfies

$$P\mathbf{T} = \mathbf{T} \quad \text{or equivalently} \quad P\hat{\Pi}_0 = \hat{\Pi}_0, \tag{3.3}$$

then $E[\mathbf{S}|\mathbf{T}] = \mathbf{T}$ and $\hat{\Pi}_* = \mathbf{S}/n$ is an unbiased estimator of Π. So, Π can be estimated without using P or its inverse, essentially by treating released data as original data. Also, $\hat{\Pi}_*$ is always a probability vector, while $P^{-1}(\mathbf{S}/n)$ may not be so. Gouweleeuw et al. (1998) noted these properties and defined a PRAM to be an *invariant* PRAM if P satisfies (3.3).

We believe that the term 'invariant' is misleading, as it might give the wrong impression that if P satisfies (3.3), then all inferential methods for original data remain valid for post-randomized data. As Nayak and Adeshiyan (2015) noted, (i) under (3.3), $V(\hat{\Pi}_*) - V(\hat{\Pi}_0)$ is positive definite, unless $P = I$, and methods for estimating $V(\hat{\Pi}_0)$ cannot be used to estimate $V(\hat{\Pi}_*)$ and (ii) keeping all inferential methods valid requires a stronger condition, viz. $P\Pi = \Pi$, which cannot be achieved in most applications due to Π being unknown. The main implication of (3.3) is that $\mathbf{S}$ is an unbiasedly perturbed version of $\mathbf{T}$, viz. $E(\mathbf{S}|\mathbf{T}) = \mathbf{T}$. Thus, when P satisfies (3.3), we shall call the procedure *unbiased post-randomization method* (UPRAM).

We should mention that (3.3) always admits a solution for P; in fact the solution space is a convex set containing $P = I$. Thus, UPRAM can always be devised, although not uniquely. One important point to bear in mind is that when there are several variables, applying independent UPRAM to each variable does not amount to applying UPRAM to the cross-classification of all variables, unless they are independently distributed. Under independent UPRAM, the marginal distributions can be estimated unbiasedly from released data (with no adjustment for perturbation), but not joint distributions. Thus, it

is important to apply UPRAM to the cross-classification of all variables.

We now briefly review, largely for later use, some results from Nayak and Adeshiyan (2015) on the distribution of $\mathbf{S}$ under UPRAM. Let $P = [P_1 : \cdots : P_k]$, i.e., denote the ith column of P by P_i $(i = 1, \ldots, k)$, and rewrite (3.3) as

$$\sum_{i=1}^{k} T_i P_i = \mathbf{T}. \tag{3.4}$$

Note that PRAM distributes all units originally falling in category c_i to all categories according to the multinomial distribution with parameters T_i and P_i. Let F_{ij} denote the number of units with original category c_i and perturbed category c_j, and let $\mathbf{F}_i = (F_{i1}, \ldots, F_{ik})'$. Then, $\mathbf{S} = \sum_i^k \mathbf{F}_i$, and given $\mathbf{T}$ and P, $\{\mathbf{F}_i\}$ are independently distributed with $\mathbf{F}_i \sim Mult(T_i, P_i), i = 1, \ldots, k$. These facts and (3.4) yield:

$$E(\mathbf{S}|\mathbf{T}, P) = \sum_{i=1}^{k} E[\mathbf{F}_i|\mathbf{T}, P] = \sum_{i=1}^{k} T_i P_i = \mathbf{T} \tag{3.5}$$

$$V(\mathbf{S}|\mathbf{T}, P) = \sum_{i=1}^{k} T_i[D_{P_i} - P_i P_i'] = D_{\mathbf{T}} - \sum_{i=1}^{k} T_i P_i P_i'. \tag{3.6}$$

From (3.5) and (3.6), it follows that $\hat{\Pi}_*$ is an unbiased estimator of Π,

$$V(\hat{\Pi}_*|\mathbf{T}, P) = \frac{1}{n}[D_{\hat{\Pi}_0} - \sum_{i=1}^{k}(\frac{T_i}{n}) P_i P_i'], \tag{3.7}$$

and

$$\begin{aligned} V(\hat{\Pi}_*) &= V[E(\hat{\Pi}_*|\mathbf{T}, P)] + E[V(\hat{\Pi}_*|\mathbf{T}, P)] \\ &= V(\hat{\Pi}_0) + \frac{1}{n}[D_{\Pi} - E\{\sum_{i=1}^{k}(\frac{T_i}{n}) P_i P_i'\}]. \end{aligned} \tag{3.8}$$

Note that the variance inflations under fixed P and UPRAM, given by the last terms of (3.2) and (3.8), respectively, are quite different. If P is a function of $\mathbf{T}$, which holds true for our choice of P in the sequel, then $E(\hat{\Pi}_*|\mathbf{T}, P) = E(\hat{\Pi}_*|\mathbf{T})$ and $V(\hat{\Pi}_*|\mathbf{T}, P) = V(\hat{\Pi}_*|\mathbf{T})$.

We can estimate $V(\hat{\Pi}_0) = [D_\Pi - \Pi\Pi']/n$ and D_Π by using $\hat{\Pi}_*$ for Π. However, the expectation in (3.8) makes estimation of $V(\hat{\Pi}_*)$ from perturbed data a difficult problem. Nayak and Adeshiyan (2015) discuss this aspect in more details and also prove that

$$V_{max}(\hat{\Pi}_*) = (2 - \frac{1}{n})[\frac{D_\Pi - \Pi\Pi'}{n}], \tag{3.9}$$

is a tight upper bound of $V(\hat{\Pi}_*)$, in the sense that $[V_{max}(\hat{\Pi}_*) - V(\hat{\Pi}_*)]$ is nonnegative definite for all P satisfying (3.3) and $V(\hat{\Pi}_*)$ equals $V_{max}(\hat{\Pi}_*)$ for some P. One can estimate (3.9) from released data by using $\hat{\Pi}_*$ for Π.

3.3. Choice of Transition Probabilities

The choice of the transition probability matrix P (also called the PRAM matrix), by data agency, is crucial and also challenging, especially for large k, as is the case in most applications due to X being the cross-classification of several variables. One practical approach is to divide all categories into several blocks and then choose a PRAM matrix for each block (see Gouweleeuw et al., 1998; Shlomo and De Waal, 2008; Nayak and Adeshiyan, 2015). Thus, instead of choosing a matrix of large dimension, one would choose several matrices of much lower dimensions. This also results in a *block diagonal* P, after suitable rearrangement of the categories. Here, $p_{ij} = 0$ whenever c_i and c_j are in two different blocks, and any perturbation occurs within one block.

Blocking is also very useful for avoiding undesired perturbations. If changing c_i to c_j is unwanted, it can be prevented by putting c_i and c_j in two different blocks. Suppose, for example, in the original data, age is recorded in five year intervals, 0-5, 5-10, etc. Naturally, we would not like an original age of 5-10 to change to 70-75. Here, we may put all categories with age between 0 and 20 in one block, between 20 and 40 in another

block and so on. Other practical ideas can also be used to form blocks. For example, if it is important to retain the original data on some of the variables, we should partition the data (hence form blocks) by categories of those variables. Essentially, if X is the cross-classification of X_1 and X_2, both of which may be compound variables, we may apply UPRAM to X_2 within each cell of X_1, which will perturb X_2 while keeping X_1 unchanged. Finally, a block may also consist of a single cell, in which case, the records of all units falling in that cell will remain unchanged.

Devising a UPRAM boils down to choosing one of many solutions of (3.3). We shall use the following class of solutions in developing our proposed method in later sections. Note that empty cells play no role in (3.3) and they remain empty after UPRAM. (However, a nonempty cell can become empty after UPRAM.) Suppose, without loss of generality, that $T_i > 0$ for $i = 1, \ldots, m \leq k$ and $T_{m+1} = \cdots = T_k = 0$. Let $p_{ii} = 1 - (\theta/T_i), p_{ji} = \theta/[(m-1)T_i]$ for $i, j = 1, \ldots, m$ and $i \neq j$, and $p_{ii} = 1, p_{ji} = 0$ for $i, j = m+1, \ldots, k$ and $i \neq j$. It can be seen that these $\{p_{ij}\}$ satisfy (3.3) for any $0 \leq \theta \leq 1$. This choice of P is closely related to one class of solutions of (3.3) given by Gouweleeuw et al. (1998). For our choice of P, if a unit's original X-category is c_i, then it changes to another category with probability (θ/T_i), which is inversely proportional to the frequency of the category in which the unit falls. Moreover, if the category is changed, the new category is selected at random from one of the remaining nonempty categories. We shall call any PRAM with P having the above structure an *inverse frequency (rule) post-randomization* (IFPR).

We shall use IFPR for its simplicity, easy interpretations and efficacy. The transition probability matrix P of IFPR is determined by a single quantity, or design parameter, θ. So, choosing P reduces to choosing just one number (for θ). Similarly, all effects of P on confidentiality protection and data quality are determined only by θ. Investigating effects

of θ and thereby choosing a suitable value of θ is much easier than similarly choosing a general P. Considering $T_i = 1$, we can interpret θ as the probability of changing the category of any sample unique unit. For any nonempty cell c_i, the number of units that change from $X = c_i$ to $X \neq c_i$, due to IFPR, is a $binomial(T_i, \theta/T_i)$ random variable, whose mean (θ) does not depend on c_i. Thus, another interpretation of θ is that it is the expected number of units that move out of (or move into) *any* nonempty category. In our procedure, described later, we shall actually first form disjoint blocks of nonempty categories and then apply IFPR in each block, with a common θ.

4. Disclosure Control by Post-randomization

4.1. Identification Risk Under PRAM

To devise methods for controlling disclosure risk, we shall evaluate and examine $R_j(a, \mathbf{t})$ first under general PRAM and then under IFPR. Recall that the intruder is assumed to know the target unit B's X-category and that B is in the original sample. For notational simplicity, we shall consider $R_1(a, \mathbf{t})$, supposing that $X_{(B)} = c_1$, which implies that $t_1 \geq 1$. Our arguments and results for $R_1(a, \mathbf{t})$ will hold similarly for $R_j(a, \mathbf{t})$ for any $j \geq 2$. For a PRAM matrix P, which may depend on the original data, but only through $\mathbf{t}$, note that only the nonzero elements in the first row of P affect $R_1(a, \mathbf{t})$. Suppose, for notational simplicity, that $0 < p_{1i} < 1$ for $i = 1, \ldots, k_1$ and $p_{1i} = 0$ for $i > k_1$, for some $k_1 \geq 2$. Note that if $p_{1i} = 0$ for all $i \geq 2$, then for any UPRAM we must have $p_{11} = 1$, which implies that $S_1 = T_1$ with probability 1 and $R_1(1, \mathbf{t}) = 1/t_1$.

Let $\alpha_i = p_{1i}$ and $\beta_i = \alpha_i/(1 - \alpha_i), i = 1, \ldots, k_1$, and let $Z_{(B)}$ denote B's category after

applying PRAM. Then,

$$\begin{aligned} P(S_1 = a, Z_{(B)} = c_1|\mathbf{T}) &= \alpha_1 \sum \prod_{i=1}^{k_1} \binom{T_i^*}{a_i} \alpha_i^{a_i} (1-\alpha_i)^{T_i^* - a_i} \\ &= \alpha_1 \Big[\prod_{i=1}^{k_1} (1-\alpha_i)^{T_i^*} \Big] \sum \prod_{i=1}^{k_1} \binom{T_i^*}{a_i} \beta_i^{a_i}, \end{aligned} \tag{4.1}$$

where $T_1^* = T_1 - 1, T_i^* = T_i, i \geq 2$ and the sum is over all integer-valued $a_1, \ldots, a_{k_1}$ such that $0 \leq a_i \leq T_i^*$ and $\sum a_i = a - 1$. We shall denote the sum in (4.1) by Σ_{a-1}. Similarly, we obtain

$$P(S_1 = a, Z_{(B)} \neq c_1|\mathbf{T}) = (1-\alpha_1) \Big[\prod_{i=1}^{k_1} (1-\alpha_i)^{T_i^*} \Big] \Sigma_a. \tag{4.2}$$

Now, since $P(CM|S_1 = a, Z_{(B)} \neq c_1, \mathbf{T} = \mathbf{t}) = 0$, we get

$$\begin{aligned} R_1(a, \mathbf{T}) &= P(CM|S_1 = a, Z_{(B)} = c_1, \mathbf{T}) P(Z_{(B)} = c_1|S_1 = a, \mathbf{T}) \\ &= \frac{1}{a} \Big[\frac{\alpha_1 \Sigma_{a-1}}{\alpha_1 \Sigma_{a-1} + (1-\alpha_1)\Sigma_a} \Big] \qquad (4.3) \\ &= \frac{1}{a} \Big[1 + \frac{1}{\beta_1} \frac{\Sigma_a}{\Sigma_{a-1}} \Big]^{-1}. \qquad (4.4) \end{aligned}$$

A common intuitive belief is that disclosure risk is largest when $a = 1$, i.e., $X_{(B)}$ matches exactly one record in the released data (see e.g., Shlomo and Skinner, 2010). To examine this assertion, we observe from (4.3) that in general, $R_1(a, \mathbf{T}) \geq R_1(a+1, \mathbf{T})$ if and only if

$$\alpha_1 \Sigma_{a-1} \Sigma_a + (a+1)(1-\alpha_1) \Sigma_{a-1} \Sigma_{a+1} \geq a(1-\alpha_1)(\Sigma_a)^2, \tag{4.5}$$

and in particular,

$$R_1(1, \mathbf{T}) \geq R_1(2, \mathbf{T}) \Leftrightarrow \alpha_1 \Sigma_1 + (1-\alpha_1)(2\Sigma_2 - \Sigma_1^2) \geq 0, \tag{4.6}$$

as $\Sigma_0 = 1$. Furthermore,

$$\Sigma_1 = \sum_{i=1}^{k_1} T_i^* \beta_i = \sum_{i=1}^{k_1} T_i \beta_i - \beta_1,$$

$$\Sigma_2 = \frac{1}{2}\sum_{j=1}^{k_1} T_j^*(T_j^*-1)\beta_j^2 + \sum_{i<j} T_i^* T_j^* \beta_i \beta_j,$$

and so, $2\Sigma_2 - \Sigma_1^2 = -\sum_{j=1}^{k_1} T_j^* \beta_j^2$ and

$$\begin{aligned} \alpha_1\Sigma_1 + (1-\alpha_1)(2\Sigma_2 - \Sigma_1^2) &= \sum_{j=1}^{k_1} T_j^* \beta_j \{\alpha_1 - (1-\alpha_1)\beta_j\} \\ &= \alpha_1 \sum_{j=2}^{k_1} T_j \beta_j \{1 - \frac{\beta_j}{\beta_1}\}, \end{aligned} \tag{4.7}$$

as $[\alpha_1 - (1-\alpha_1)\beta_1] = 0$ and $T_j^* = T_j$ for $j \geq 2$. Clearly, (4.7) can be negative and in view of (4.6), a unique match in released data is not necessarily the worst case in terms of disclosure risk. In particular, if for a given $\mathbf{t}$, $\beta_1 < \beta_j$ or $\alpha_1 < \alpha_j$ for $j = 2, \ldots, k_1$, then $R_1(2, \mathbf{t}) > R_1(1, \mathbf{t})$. On the other hand, we have:

Proposition 4.1. *For any* $\mathbf{t}$*, a sufficient condition for* $R_1(1, \mathbf{t}) \geq R_1(2, \mathbf{t})$ *to hold is:* $\beta_1 \geq \beta_j$ *or* $\alpha_1 \geq \alpha_j$ *for* $j = 2, \ldots, k_1$.

We may also note that for $a = 1$, (4.4) reduces to

$$\begin{aligned} R_1(1, \mathbf{T}) &= \Big[1 + \frac{1}{\beta_1}\big(\sum_{i=1}^{k_1} \beta_i T_i - \beta_1\big)\Big]^{-1} \\ &= \Big[T_1 + \frac{1}{\beta_1}\sum_{i=2}^{k_1} \beta_i T_i\Big]^{-1}. \end{aligned} \tag{4.8}$$

Clearly, (4.8) is less than $1/T_1$, which is the probability of correctly identifying B in the original data. Thus, PRAM does not increase B's correct match probability when it yields a unique match for B.

Next, we shall investigate effects of IFPR (as described in Section 3.3) on target B's identification risk (assuming as before that $X_{(B)} = c_1$). Actually, our procedure uses the original data to form disjoint blocks of *nonzero* categories of X and then applies IFPR to each block separately but with a common value of θ. Also, not all cells are included

in the IFPR blocks. In particular, the records of all units falling in 'high frequency' cells are not perturbed. Additional guidance on how to create IFPR blocks and choose a value for θ will emerge in subsequent discussions. If the units in c_1 are not perturbed, B's identification risk is $1/T_1$. If c_1 falls in one IFPR block, B's identification risk will depend only on the post-randomization in that block. For notational simplicity, suppose c_1 falls in the block $\{c_1, \ldots, c_{k_1}\}$, with $k_1 \geq 2$. Recall that the transition probabilities for this block are: $p_{ii} = 1-\theta/T_i, p_{ji} = \theta/[(k_1-1)T_i]$ for $i, j = 1, \ldots, k_1$ and $i \neq j$, with $0 \leq \theta \leq 1$.

First, consider the case of $S_1 = 0$, which means B's true category is empty in the released data and hence the intruder terminates his effort to find a match for B. In general, $P(S_1 = 0|\mathbf{T}) = \prod_{i=1}^{k_1}(1-\alpha_i)^{T_i}$ and for the above choice of P, it reduces to

$$P(S_1 = 0|\mathbf{T}) = \Big(\frac{\theta}{T_1}\Big)^{T_1} \prod_{j=2}^{k_1} \Big(1 - \frac{\theta}{(k_1-1)T_j}\Big)^{T_j}. \tag{4.9}$$

Now, using the fact that for $0 < c < 1$ the function $f(t) = (1-c/t)^t$ increases to e^{-c} as t increases from 1 to ∞, we get

$$\Big(\frac{\theta}{T_1}\Big)^{T_1}\Big(1 - \frac{\theta}{k_1-1}\Big)^{k_1-1} \leq P(S_1 = 0|\mathbf{T}) \leq \Big(\frac{\theta}{T_1}\Big)^{T_1} e^{-\theta}. \tag{4.10}$$

The upper bound in (4.10) is independent of k_1 and it increases as θ increases, and decreases as T_1 increases. Actually, the upper bound is fairly small for $T_1 \geq 2$. For example, when $T_1 = 2$, for $\theta = .2, .5, .7, .8$, the upper bounds are 0.0082, 0.0379, 0.0608 and 0.0719, respectively. For $T_1 = 1$ and some values of θ and k_1, the last row of Table 1 gives the upper bounds in (4.10) and the other rows give the lower bounds for corresponding k_1. Table 1 shows that for $k_1 \geq 5$, the upper and lower bounds are fairly close for all θ.

Next, we consider $R_1(1, \mathbf{T})$, the probability that a unique match for B is a correct match. Note that for IFPR, $\beta_1 = (T_1 - \theta)/\theta, \beta_j = \theta/[(k_1-1)T_j - \theta], j = 2, \ldots, k_1$, and

Table 1: Bounds of $P(S_1 = 0|\mathbf{T})$ when $T_1 = 1$

	θ								
k_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2	0.0900	0.1600	0.2100	0.2400	0.2500	0.2400	0.2100	0.1600	0.0900
3	0.0902	0.1620	0.2167	0.2560	0.2812	0.2940	0.2958	0.2880	0.2723
4	0.0903	0.1626	0.2187	0.2604	0.2894	0.3072	0.3154	0.3155	0.3087
5	0.0904	0.1629	0.2196	0.2624	0.2931	0.3132	0.3243	0.3277	0.3247
10	0.0904	0.1634	0.2211	0.2657	0.2989	0.3225	0.3378	0.3461	0.3487
15	0.0905	0.1635	0.2215	0.2666	0.3005	0.3250	0.3414	0.3510	0.3550
UB	0.0905	0.1637	0.2222	0.2681	0.3033	0.3293	0.3476	0.3595	0.3659

(4.8) reduces to

$$\begin{aligned} R_1(1,\mathbf{T}) &= \Big[T_1 + \frac{\theta}{T_1-\theta}\sum_{i=2}^{k_1}\frac{\theta T_i}{(k_1-1)T_i-\theta}\Big]^{-1} \\ &= \big(T_1-\theta\big)\Big[T_1(T_1-\theta)+\theta^2\sum_{i=2}^{k_1}\frac{T_i}{(k_1-1)T_i-\theta}\Big]^{-1} \qquad (4.11) \\ &\le \frac{T_1-\theta}{T_1(T_1-\theta)+\theta^2} = \psi(T_1,\theta), \text{ say.} \qquad (4.12) \end{aligned}$$

The inequality in (4.12) follows from the fact that the summand in (4.11) is a decreasing function of T_i and hence must be at least $1/(k_1-1)$.

Two key features of the upper bound in (4.12) are that (i) it is independent of k_1 and (ii) it depends on $\mathbf{T}$ only through T_1. Taking derivative of $\psi(T_1,\theta)$ with respect to T_1, it can be seen that for any $0<\theta<1$, $\psi(T_1,\theta)$ is a decreasing function of T_1 at all $T_1 \ge 2$. It can also be seen that $\psi(1,\theta) \ge \psi(2,\theta)$ if and only if $\theta \le 2/3$. Thus, if $\theta \le 2/3$, then $\psi(T_1,\theta)$ is maximum when $T_1 = 1$ and for $\theta > 2/3, \psi(T_1,\theta)$ is largest when $T_1 = 2$. Note

that analogous conclusions for any target unit can be reached via appropriate notation change in our derivations.

4.2. Choosing P to Control Identification Risk

Here, we shall use the preceding results and observations to develop methods for accomplishing three increasingly stringent disclosure control goals.

Goal 1. Suppose the agency's primary concern is about correctness of unique matches for sample unique units, and for a chosen value ξ, the agency wants to ensure that

$$R_j(1, \mathbf{T} = \mathbf{t}) \leq \xi \quad \text{for all } j \text{ such that } t_j = 1. \tag{4.13}$$

This goal is of high practical significance. As the need for controlling identification risk of sample unique units is commonly recognized, failure to address that need may harm the agency's reputation substantially. To achieve Goal 1, we shall choose IFPR blocks and θ such that a common upper bound of $\{R_j(1, \mathbf{t})\}$ equals ξ. For any j such that $t_j = 1$, from (4.12) we get

$$R_j(1, \mathbf{t}) \leq \psi(1, \theta) = \frac{1 - \theta}{1 - \theta + \theta^2}.$$

Note that $\psi(1, \theta)$ is a strictly decreasing function of θ with $\psi(1, 0) = 1$ and $\psi(1, 1) = 0$. So, for any $0 < \xi < 1$, the equation $\psi(1, \theta) = \xi$ has a unique solution for θ, say $\theta = \theta_*(\xi)$. Now, Goal 1 can be accomplished as follows. First, form one or more IFPR blocks such that each singleton cell (i.e., with frequency 1) is included in one of the blocks. One option is to form just one block consisting of all singleton cells, if the number of such cells is at least two. If there is only one singleton cell, we would need to combine it with one or more other cells to form one block. Finally, apply IFPR to all units in each block with $\theta = \theta_*(\xi)$. This would guarantee $R_j(1, \mathbf{t}) \leq \psi(1, \theta_*(\xi)) = \xi$ for all j such that $t_j = 1$.

Goal 2. Suppose the agency wants to control the probability of any unique match for any target unit being a correct match, irrespective of whether the unit is sample unique in the original data or not. Formally, the agency's goal is to guarantee that

$$R_j(1, \mathbf{t}) \leq \xi \quad \text{for all } j \text{ such that } t_j \geq 1. \tag{4.14}$$

This goal is similar to that of Shlomo and Skinner (2010). Here, we shall describe how Goal 2 can be achieved for any $\xi > 1/3$. Let $h(\theta) = \max_{T_1} \psi(T_1, \theta)$. Then, from earlier discussions we have:

$$h(\theta) = \begin{cases} \psi(1, \theta), & \text{if } \theta \leq \frac{2}{3} \\ \psi(2, \theta), & \text{if } \theta > \frac{2}{3}. \end{cases}$$

It follows that $h(\theta)$ is a strictly decreasing function of θ with $h(0) = 1$ and $h(1) = 1/3$. So, for any $1/3 < \xi \leq 1$, $h(\theta) = \xi$ has a unique solution, say $\theta = \theta_0(\xi)$, in $[0, 1)$. Also note from (4.8) that under any PRAM, $R_j(1, \mathbf{t}) \leq 1/3$ for all j such that $t_j \geq 3$. Thus, to achieve (4.14) with $\xi > 1/3$, it is not necessary to perturb the units falling in categories with frequency 3 or more. We only need to protect all units in singleton and doubleton cells. Thus, for any $\xi > 1/3$, we can achieve Goal 2 by forming IFPR blocks in such a way that each cell with frequency 1 or 2 is included in one of the blocks and then applying IFPR to each block with $\theta = \theta_0(\xi)$.

Goal 3. Suppose the agency wants to guarantee (2.8) to control the probability of any declared match being correct.

Clearly, this is a stronger version of Goal 2, which requires (2.8) to hold only for $a = 1$. As with Goal 2, we shall consider $\xi > 1/3$; actually, our solution works well for $\xi \geq .35$. From (4.4), it follows that (2.8) holds for all $a \geq 3$ if $\xi > 1/3$. So, we need to focus only on $a = 1$ and $a = 2$. For IFPR, i.e., $\alpha_1 = 1 - \theta/t_1$ and $\alpha_i = \theta/[(k_1 - 1)t_j], i = 2, \ldots, k_1$,

Proposition 4.1 shows that $R_1(2,\mathbf{t}) \le R_1(1,\mathbf{t})$ holds true if

$$1-\frac{\theta}{t_1} \ge \frac{\theta}{(k_1-1)t_j} \quad \text{for } 2 \le j \le k_1 \quad \text{or} \quad \theta \le \min_{2\le j\le k_1}\Big[\frac{1}{t_1}+\frac{1}{(k_1-1)t_j}\Big]^{-1}. \tag{4.15}$$

Moreover, the right side of (4.15) $\le (1-1/k_1)$ (attained when $t_1 = t_j = 1$ for some $2 \le j \le k_1$) and so, $R_1(2,\mathbf{t}) \le R_1(1,\mathbf{t})$ will hold if $\theta \le (1-\frac{1}{k_1})$ or $k_1 \ge (1-\theta)^{-1}$. Similarly, we shall have $R_j(2,\mathbf{t}) \le R_j(1,\mathbf{t})$ if category c_j is included in an IFPR block that contains at least $(1-\theta)^{-1}$ cells. This implies that for $\xi > 1/3$, our solution for attaining Goal 2, described above, will also achieve Goal 3 if we put at least $(1-\theta)^{-1}$ cells in each IFPR block. For certain values of θ, Table 2 gives the values of $\psi(1,\theta), \psi(2,\theta), \xi = \max\{\psi(1,\theta), \psi(2,\theta)\}$ and the required block size, calculated as $\lceil (1-\theta)^{-1} \rceil$, the ceiling of $(1-\theta)^{-1}$, i.e., the smallest integer that is not smaller than $(1-\theta)^{-1}$.

Table 2: Minimum Block Size

θ	$\psi(1,\theta)$	$\psi(2,\theta)$	ξ	Required block size
.4	.789	.476	.789	2
.5	.667	.462	.667	2
2/3	.429	.429	.429	3
.75	.308	.408	.408	4
.8	.238	.395	.395	5
.9	.110	.365	.365	10
.95	.052	.350	.350	20
.99	.010	.337	.337	100

Obviously, as θ increases to 1, the required minimum block size $\lceil (1-\theta)^{-1} \rceil$ increases to ∞. However, requiring each block for IFPR to contain too many categories may be

inconvenient and may affect data utility undesirably. Guided by Table 2, we recommend to take $\xi \geq .350$ so that (2.8) can be attained with $\theta \leq .95$ and required minimum block size ≤ 20. In summary, for Goal 3, we suggest to take $\xi \geq .35$ and then achieve it as follows. First, calculate $\theta_0(\xi)$, as the solution of $h(\theta) = \xi$, and $m_0 = \lceil (1-\theta)^{-1} \rceil$. Then, form blocks for IFPR such that they include all cells with frequency 1 or 2 and each block contains at least m_0 cells. Finally, apply IFPR to all units in each block with $\theta = \theta_0(\xi)$.

5. A General Post-randomization Procedure

In this section, we propose a general procedure, by augmenting our approach for Goal 3 in the preceding section with certain suggestions on how to form IFPR blocks before applying post-randomization. This will also further clarify the relevance and usefulness of our previous results and observations. The procedure stated below accomplishes Goal 3, but it can be modified suitably to achieve the other two goals discussed in Section 4. Suppose a data set involves p categorical variables $X_1, \ldots, X_p$ and the data agency has specified a subset of those as key variables and wishes to accomplish Goal 3 with a given value of $\xi > 1/3$. Our procedure does not perturb values of quantitative variables, which may be masked appropriately using other methods, e.g., noise addition or imputation. Also, in many applications, quantitative variables are recorded using interval scales, in which case those may be treated as categorical variables and covered by our procedure. We organize the proposed procedure in 4 steps as follows.

Step 1. Use the given value of ξ to calculate θ_0 by solving $h(\theta) = \xi$, and then compute $m_0 = \lceil (1-\theta)^{-1} \rceil$. As discussed in Section 4, the number of categories in each block for applying IFPR will need to be at least m_0. In practice, we should consider taking ξ to be

0.35 or larger, so that $m_0 \leq 20$.

Step 2. At this step we choose a variable set for post-randomization. Here, our task is to divide the variables $X_1, \ldots, X_p$ into two sets $\mathcal{C}$ and $\bar{\mathcal{C}}$, with the objective of post-randomizing the cross-classification of all variables in $\mathcal{C}$, to be denoted $X_{\mathcal{C}}$, and leaving the values of other variables unchanged. Thus, the frequency of any cell in the cross-classification of the variables in $\bar{\mathcal{C}}$ will remain unchanged. Two main considerations for choosing $\mathcal{C}$ are as follows. First, all key variables should be included in $\mathcal{C}$, for confidentiality protection. Second, we should include certain non-key variables in $\mathcal{C}$ in order to avoid unrealistic combinations of values of $X_1, \ldots, X_p$ in perturbed data (and new edit failures). For example, suppose $\mathcal{C}$ includes gender (say, X_1) but not the number of pregnancies (say, X_2). Then, X_1 will be perturbed and X_2 will remain unchanged, which is likely to generate a perturbed data set that will show a positive number of pregnancies for some males. Here, both X_1 and X_2 should be put either in $\mathcal{C}$ or in $\bar{\mathcal{C}}$. As Nayak and Adeshiyan (2015) noted, UPRAM does not yield any combination of values of the variables in $\mathcal{C}$ that are not in the original data.

To recognize another issue, recall that under UPRAM (which includes IFPR), the perturbed frequency of any category of $X_{\mathcal{C}}$ is an unbiased estimate of the original frequency. However, the same conclusion may not be true for a category of the cross-classification of some variables in $\mathcal{C}$ and some in $\bar{\mathcal{C}}$. Thus, some non-key variables may be included in $\mathcal{C}$ to assure that joint frequencies of those non-key and all key variables are perturbed unbiasedly. Suppose for example, $p = 4$ and X_1, X_2, X_3 and X_4 represent gender, education level, income class and housing tenure, respectively. Suppose also that X_1 and X_2 are key variable and the association between education and income is of substantial inferential interest. Then, we should include X_1, X_2 and X_3 in $\mathcal{C}$ and leave X_4 in $\bar{\mathcal{C}}$.

In principle, all categorical variables $X_1, \ldots X_p$ can be included in $\mathcal{C}$, leaving $\bar{\mathcal{C}}$ empty. However, it is inconvenient to put many variables in $\mathcal{C}$ as $X_{\mathcal{C}}$ would have a very large number of categories, many with low frequencies, viz. 0 or 1 or 2. Also, our next step (data partitioning) offers another substantial opportunity to preserve data quality. Thus, in practice, non-key variables should be added to $\mathcal{C}$ if the need for doing so is strong. Note that our method gives confidentiality protection guarantee with respect to all variables in $\mathcal{C}$. So, it will provide higher level of protection than what the agency requires if $\mathcal{C}$ includes any non-key variable. For notational simplicity, suppose $\mathcal{C} = \{X_1, \ldots, X_r\}$ for some $r \leq p$, and thus $X_{\mathcal{C}}$ represents the cross-classification of $X_1, \ldots, X_r$.

Step 3. Partitioning the data set. As we have seen in Section 4, our approach to attaining Goal 3 only requires that (i) all units in singleton and doubleton cells (i.e., with frequency 1 or 2) be post-randomized and (ii) each block for IFPR contain at least m_0 cells. On the other hand, we shall see in the next section that to minimize data quality loss the number of cells in each IFPR block should be kept as small as possible. Thus, our IFPR blocks must cover all singleton and doubleton cells of $X_{\mathcal{C}}$, and the number of cells in each block should be close to m_0, but not smaller than m_0. These constraints still give us significant freedom for setting up IFPR blocks, which we shall utilize to preserve data quality.

We may not want to perturb the values of certain variables in $\mathcal{C}$ either because of their importance or to comply with regulations. For example, for a person level sample from a state, regulations may require us to preserve the counts of individuals in age intervals (0, 18], (18, 65) and 65 and over, or keep the county of residence of each person unchanged. Similarly, it may be desirable to preserve the original data on employment status and poverty level of each individual. Also, perturbed values of ordinal variables, e.g., income

class and age group, should be reasonably close to the original values. We propose to utilize the flexibility in forming IFPR blocks to control the nature (magnitude in some sense) of data perturbation and prevent undesirable changes to original records, by first partitioning the original data set and then forming IFPR blocks within each partition set.

For simplicity, we suggest to control perturbation at individual variable level, but cross-classifications of subsets of variables can be used similarly. A general framework for implementing this idea is as follows. For each X_i in $\mathcal{C}$, create a new variable X_i^* by merging categories of X_i. Let $X_{\mathcal{C}}^*$ denote the cross-classification of $X_1^*, \ldots, X_r^*$. Then, partition the data set by categories of $X_{\mathcal{C}}^*$. Thus, all units falling in one category of $X_{\mathcal{C}}^*$ constitute one partition set. Note that categories of $X_{\mathcal{C}}^*$ also defines a partition of the original cells of $X_{\mathcal{C}}$. As described in Step 4 below, we apply post-randomization within each partition set. So, our procedure does not change the $X_{\mathcal{C}}^*$-category of any unit. Two extreme types of X_i^* are: (i) $X_i^* = X_i$, in which case, the original data on X_i will remain unchanged and (ii) X_i^* merges all categories of X_i into one category, which means that for X_i, the original category of any unit is permitted to change to any other category.

In Step 4 (below), we require that if a partition set contains at least one singleton or doubleton cell of $X_{\mathcal{C}}$, then it must contain at least m_0 nonempty cells. This condition may not be satisfied if $X_1^*, \ldots, X_r^*$ do not collapse the categories of $X_1, \ldots, X_r$ adequately. If a partition violates the requirement, we would need to merge partition sets or redefine $X_1^*, \ldots, X_r^*$ to coarsen the partition. On the other hand, excessive collapsing would yield a small number of partition sets with large IFPR group sizes, missing further opportunities for perturbation control (and reducing data quality loss).

Step 4. Applying post-randomization. At this step, we apply the followig procedure to each partition set separately. First, we count the number (J) of singleton and doubleton

cells (of $X_{\mathcal{C}}$) in the partition set. If $J = 0$, i.e., there are no singleton or doubleton cells, we leave original categories of all units unchanged. If $J \geq m_0$, we take all singleton and doubleton cells to form one IFPR block and then apply IFPR with $\theta = \theta_0$, as calculated in Step 1, to $X_{\mathcal{C}}$ for all units in this block. If $0 < J < m_0$, we add $(m_0 - J)$ cells with frequency 3 or more to the set of singleton and doubleton cells to form one IFPR block. For minimal impact on data quality, cells with smallest frequencies should be added. Then, we apply IFPR with $\theta = \theta_0$ to all units in the IFPR block. In both cases ($J \geq m_0$ and $J < m_0$) records of all units not included in the IFPR block are left unchanged.

Clearly, the procedure described above may be modified in several ways. As noted earlier, one may use cross-classifications of subsets of $X_1, \ldots, X_r$ for data partitioning. In fact, it is not even necessary to use 'rectangular' data partition. In Step 4, if the number of singleton and doubleton cells in a partition set is $2m_0$ or larger, we may divide those to two (or more) IFPR blocks, each with at least m_0 cells. This can be done by random splitting for convenience or using contextual knowledge and judgment to optimize data quality loss. However, in practice, large data sets need to be perturbed using computer programs, without much human work. We believe, Step 3 provides a simple and practical method for maintaining data quality while controlling identification risk.

6. Effects on Data Quality

In this section, we investigate effects of IFPR on cell frequencies. Taking X as $X_{\mathcal{C}}$, we shall first examine $V(\mathbf{S}|\mathbf{T})$, as $E(\mathbf{S}|\mathbf{T}) = \mathbf{T}$ under IFPR. In our procedure, for any category c_i of $X_{\mathcal{C}}$, any difference between S_i and T_i arises only from the IFPR applied to the block that contains c_i. So, we shall first assess block level effects of IFPR. Thus, consider one

IFPR block and for notational simplicity, suppose it contains the cells $c_1, \ldots, c_m$. Let $\mathbf{S}_1 = (S_1, \ldots, S_m)'$, $\mathbf{T}_1 = (T_1, \ldots, T_m)'$ and let $V = ((v_{ij}))$ denote $V(\mathbf{S}_1|\mathbf{T}_1)$. The ith column of the transition probability matrix (P) for this group is $P_i = (\frac{\theta}{(m-1)T_i}, \ldots, \frac{\theta}{(m-1)T_i}, (1 - \frac{\theta}{T_i}), \frac{\theta}{(m-1)T_i}, \ldots, \frac{\theta}{(m-1)T_i})'$, i.e., the ith element is $(1 - \frac{\theta}{T_i})$ and the rest are $\frac{\theta}{(m-1)T_i}$. Recall that only nonempty cells participate in IFPR and so each $T_i > 0$. Using (3.6) and noting that P is a function of $\mathbf{T}_1$, we obtain:

$$\begin{aligned} v_{ii} &= 2\theta - \theta^2\Big[\frac{1}{(m-1)^2}\sum_{\substack{j=1\\ j\neq i}}^{m}\frac{1}{T_j} + \frac{1}{T_i}\Big] \\ &= \theta\Big(2 - \frac{\theta}{T_i}\Big) - \frac{\theta^2}{(m-1)^2}\sum_{\substack{j=1\\ j\neq i}}^{m}\frac{1}{T_j} \end{aligned} \tag{6.1}$$

for $i = 1, \ldots, m$ and for $i \neq j$,

$$v_{ij} = -\frac{\theta}{m-1}\Big[2 + \theta\Big(\frac{1}{m-1}\sum_{\substack{l=1\\ l\neq i,j}}^{m}\frac{1}{T_l} - \frac{1}{T_i} - \frac{1}{T_j}\Big)\Big]. \tag{6.2}$$

Equation (6.1) shows that $\theta(2 - \theta/T_i) - \theta^2 \leq v_{ii} \leq \theta(2 - \theta/T_i)$ and the lower bound is attained when $m = 2$ and $T_j = 1$ for all $j \neq i$. Thus, if the units in category c_i require confidentiality protection, for minimum variation of the category's count we should put c_i in one IFPR block that contains just one other singleton category. Suppose, for example, $\theta = 0.8, m = 5, T_1 = T_2 = 1, T_3 = T_4 = T_5 = 2$. Then, (6.1) gives $v_{11} = 0.96 - 0.1 = 0.86$. If we add another category with frequency 1 to this block, v_{11} increases to $0.96 - 0.0896 = 0.8704$. Adding 4 more categories with frequencies 2, 3, 3 and 3 increases v_{11} to 0.9205, which is fairly close to 0.96, the upper bound of v_{11}. Thus, (6.1) is fairly flat unless m is very small.

The total variation of $\mathbf{S}_1$, measured by the trace of $V(\mathbf{S}_1|\mathbf{T}_1)$, is

$$V_t = \sum_{i=1}^{m} v_{ii} = 2m\theta - \theta^2\Big(\frac{m}{m-1}\Big)\sum_{i=1}^{m}\frac{1}{T_i}. \tag{6.3}$$

If we add one cell, say c_{m+1}, to the block $\{c_1, \ldots, c_m\}$, (6.3) shows that the total variation for the enlarged block will be

$$V_t^* = 2(m+1)\theta - \theta^2\Big(\frac{m+1}{m}\Big)\sum_{i=1}^{m+1}\frac{1}{T_i}.$$

Then, we obtain:

$$\begin{aligned} V_t^* - V_t &= 2\theta + \frac{\theta^2}{m(m-1)}\sum_{i=1}^{m}\frac{1}{T_i} - \frac{\theta^2(m+1)}{m}\frac{1}{T_{m+1}} \\ &> \theta\Big[2 - \frac{\theta(m+1)}{m}\frac{1}{T_{m+1}}\Big] \\ &\geq \theta\big[2 - (1.5)\theta\big] > 0 \end{aligned}$$

for all $0 < \theta < 1$, as $m \geq 2$ implies $(m+1)/m \leq 1.5$. Thus, as intuition suggests, we should post-randomize the units in only the cells requiring confidentiality protection.

Next, to explore effects of merging (or splitting) IFPR blocks, consider two blocks A_1 and A_2 containing m_1 and m_2 cells, respectively. Let $V_{t|s}$ and $V_{t|c}$ denote the total variation of all $m_1 + m_2$ cells in $A_1 \cup A_2$ under two schemes: (1) apply IFPR to each block separately and (2) combine the two blocks and then apply IFPR. Using (6.3), we get

$$V_{t|s} = 2\theta(m_1+m_2) - \theta^2\Big[\Big(\frac{m_1}{m_1-1}\Big)\sum_{i\in A_1}\frac{1}{T_i} + \Big(\frac{m_2}{m_2-1}\Big)\sum_{i\in A_2}\frac{1}{T_i}\Big],$$

$$V_{t|c} = 2\theta(m_1+m_2) - \theta^2\Big(\frac{m_1+m_2}{m_1+m_2-1}\Big)\Big[\sum_{i\in A_1}\frac{1}{T_i} + \sum_{i\in A_2}\frac{1}{T_i}\Big]$$

and

$$V_{t|c} - V_{t|s} = \frac{\theta^2}{m_1+m_2-1}\Big[\Big(\frac{m_2}{m_1-1}\Big)\sum_{i\in A_1}\frac{1}{T_i} + \Big(\frac{m_1}{m_2-1}\Big)\sum_{i\in A_2}\frac{1}{T_i}\Big] > 0. \qquad (6.4)$$

This shows that to reduce effects of IFPR on data quality, all cells for post-randomization should be divided into disjoint IFPR blocks, each containing the minimally required number of cells for confidentiality protection. We also note that the relative difference between

$V_{t|s}$ and $V_{t|c}$ is significant when m_1, m_2 and $T_i, i \in A_1 \cup A_2$ are very small. For a numerical example, take $\theta = 0.8$. Then, for $m_1 = m_2 = 2$ and $T_i = 1$ for all i, we get $V_{t|c} = 4.6933$ and $V_{t|s} = 1.28$, and so $V_{t|c} = (3.67)V_{t|s}$. In contrast, if $m_1 = m_2 = 5$ and $T_i = 2$ for each of the 10 cells, then $V_{t|c} = 14.8444$, which is only 3.09% larger than $V_{t|s} = 14.4$. Thus, if the required minimum number (m_0) of cells in IFPR blocks is not too small, say $m_0 \geq 5$ (see Table 2), we may not gain much from creating many blocks and satisfying m_0 minimally.

Typically, the number of variables and cross-classified cells are very large and individual cell frequencies are not of much interest. In practice, lower level marginal relative frequencies are of much greater interest. To assess the post-randomization variance of a marginal frequency, we first consider a subset of cells of one IFPR block. Suppose the block consists of the cells $c_1, \ldots c_m$. Let A be a subset of $\{1, \ldots, m\}$ of size b and $S_A = \sum_{i \in A} S_i$ represent the total perturbed count of the cells $c_i, i \in A$. Then, $V(S_A|\mathbf{T}) = \sum_{i,j \in A} \mathrm{cov}(S_i, S_j|\mathbf{T})$, and using (6.1), (6.2) and routine algebra, and letting $A^c = \{1, \ldots, m\} \setminus A$, we find that

$$\begin{aligned} V(S_A|\mathbf{T}) &= \frac{2\theta b(m-b)}{m-1} - \frac{\theta^2}{(m-1)^2}\Big[b^2 \sum_{i \in A^c} \frac{1}{T_i} + (m-b)^2 \sum_{i \in A} \frac{1}{T_i}\Big] \\ &\leq \frac{2\theta b(m-b)}{m-1} \qquad (6.5) \\ &\leq 2\theta b, \qquad (6.6) \end{aligned}$$

as $1 \leq b \leq m$. The right side of (6.5) is the largest when $b = m/2$, and zero when $b = m$.

Note that our IFPR blocks depend on the original data through $\mathbf{T}$, and change in repeated sampling. However, in practice, we are interested in the (total) probabilities of subsets of cells of $X_{\mathcal{C}}$, where the choice of the subsets do not depend on the data. Denote the cells of $X_{\mathcal{C}}$ by $c_1, \ldots, c_k$ and let $\pi_i = P(X_{\mathcal{C}} = c_i), i = 1, \ldots, k$. Then, consider a subset

D of $\{1, \ldots, k\}$ of size b and let $\pi_D = \sum_{i \in D} \pi_i, T_D = \sum_{i \in D} T_i$ and $S_D = \sum_{i \in D} S_i$. For brevity, we shall use D also to denote $\{c_i; i \in D\}$. To investigate $V(S_D|\mathbf{T})$, note that the cells in D fall into different IFPR blocks and hence are perturbed differently. Any given $\mathbf{T} = \mathbf{t}$ yields a specific set of IFPR blocks and a unique distribution of the cells in D into those blocks. Suppose b_U of the b cells in D do not fall in any IFPR block (and hence are not post-randomized), and the remaining $b_R = b - b_U$ cells fall in L different IFPR blocks, denoted $G_1, \ldots, G_L$. Also, let b_i denote the number of cells of D that fall in G_i and let W_i denote the total frequency of those cells after post-randomization. Then,

$$\begin{aligned} V(S_D|\mathbf{T} = \mathbf{t}) &= V\Big(\sum_{i=1}^{L} W_i|\mathbf{t}\Big) \\ &\leq \sum_{i=1}^{L} 2\theta b_i \\ &= 2\theta b_R \leq 2\theta b, \end{aligned} \tag{6.7}$$

by (6.6). Note that L and b_R are functions of $\mathbf{t}$ and the upper bound $(2\theta b)$ in (6.7) is overly conservative when b_U is large.

Next, consider the two estimators $\hat{\pi}_D = T_D/n$ and $\tilde{\pi}_D = S_D/n$, based on original and post-randomized data, respectively, where n is the sample size. Both are unbiased estimators of π_D with $V(\hat{\pi}_D) = [\pi_D(1 - \pi_D)]/n$ and

$$V(\tilde{\pi}_D) = V[E(\tilde{\pi}_D|\mathbf{T})] + E[V(\tilde{\pi}_D|\mathbf{T})] = \frac{\pi_D(1 - \pi_D)}{n} + E[V(\tilde{\pi}_D|\mathbf{T})]. \tag{6.8}$$

Furthermore, (6.7) shows that

$$E[V(\tilde{\pi}_D|\mathbf{T})] \leq \frac{2\theta}{n^2} E(b_R) \leq \frac{2\theta b}{n^2}. \tag{6.9}$$

Note that the first term on the right side of (6.8) is the sampling variance and it is of order $(1/n)$. In contrast, the second term, which is the variance inflation due to post-randomization, is at most of order $(1/n^2)$. Note that $E(b_R)$ is a function also of n and

$E(b_R) \to 0$ and $n \to \infty$. Thus, we establish an important practical conclusion that the randomness and uncertainty induced by our procedure is negligible in comparison to the sampling variance. This feature is also well demonstrated in our example below.

7. An Example

In this section, we use U.S. Census Bureau's 2013 one-year person-level Public Use Microdata Sample (PUMS) for the state of Maryland, to illustrate our procedure (as described in Section 5), examine its empirical performance and affirm our theoretical results. The PUMS data set was extracted from American Community Survey (ACS) data and some values were perturbed by Census Bureau for confidentiality protection, but for our illustration, we shall treat all values as original values. The data set and data description are available at https://www.census.gov/programs-surveys/acs/data/pums.html. The data set contains records of 59,033 persons for many demographic and economic variables.

For illustration, we selected five variables for post-randomization, which are gender (2), age (92), race/ethnicity (9), marital status (5) and Public Use Microdata Area (PUMA) (44), where the values in parentheses are the number of categories of respective variables. Thus, $\mathcal{C}$ comprised these five variables, which we shall denote by $X_1, \ldots, X_5$, respectively. This generated a cross-classified variable $X_{\mathcal{C}}$ with 364,320 cells (or categories). The data set yielded 25,406 nonempty cells, of which 13,662 are singleton and 4,777 are doubleton cells. Thus, if the original data are released, of the 59,033 units (persons) in the data set, 13,662 (or 23.14%) can be identified correctly with certainty and each of another 9,554 (or 16.18%) units (in doubleton cells) can be correctly identified with probability 0.5. Our procedure methodically controls identity disclosure risk of the $13,662 + 9544 = 23,216$

singleton and doubleton units, which constitute 39.33% of all units.

For this illustration, we used $\theta = 0.8$ and $m_0 = 5$, which accomplish Goal 3 with $\xi = .395$; see Table 2. For data partitioning, we created the following five (collapsed) variables: $X_1^* = X_1$ (i.e., no cells are merged); X_2^* collapses X_2 (age) in 7 broader classes, viz. 0 to 17, 18 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, and 65 and above; X_3^* regroups X_3 into the three categories: white, black and other races; $X_4^* \equiv 1$ and $X_5^* \equiv 1$, i.e., X_4^* and X_5^* merge all categories of X_4 and X_5, respectively, into one category. For these choices, our procedure will keep gender unchanged and allow marital status and PUMA (X_4 and X_5) to change freely. All age values will stay within the broader categories of X_2^*. Thus, our procedure will preserve the numbers of voting age (18 or above) and senior (age 65 or above) persons, which may be important in legal and policy studies. Of the nine categories of X_3, white and black are two dominant categories and account for about 89% of all persons in the data set. Our X_3^* preserves the two major race groups and merges the remaining 7 into one category.

The cross-classification of $X_1^*, \ldots, X_5^*$ yielded 42 cells and all units falling in a cell constituted a partition set. The total number of singleton and doubleton cells (of $X_{\mathcal{C}}$) in the 42 partition sets ranged between 124 and 1480, all being much larger than 5. Thus, it was possible to use a finer partition and exercise further control, which we did not pursue. Finally, for each partition set, we took all units falling in singleton and doubleton cells to form one IFPR block and applied IFPR with $\theta = 0.8$ to create a perturbed data set. Thus, our post-randomization changed the original $X_{\mathcal{C}}$-category of each singleton unit with probability 0.8 and each doubleton unit with probability 0.4. When a category was changed, the new category was picked at random from the remaining cells within the IFPR block. The procedure also kept the records of all units falling in $X_{\mathcal{C}}$-cells with frequency 3

or more unchanged. In the following, we report some results from one perturbed data set. However, we had created several perturbed data sets by repeating the post-randomization step (for the same data partitioning) and found similar results in all cases.

7.1. Empirical Identification Risk

Here, we examine several aspects of the perturbed data set for a broad empirical assessment of identification risk. Some basic features of the perturbed data set are as follows. Our procedure changed the $X_{\mathcal{C}}$-category of 10,954 (or 80.18%) of the 13,662 singleton units and 3,807 (or 39.85%) of the 9,554 doubleton units. The procedure also turned 4,983 (or 36.47%) singleton and 360 (or 7.54%) doubleton cells into empty cells. Thus, 36.47% of (originally) singleton and 7.54% of doubleton units had no match in the perturbed data set. These two percentages are close to the corresponding theoretical values of 35.95% and 7.19%, respectively, derived in Section 4.1.

For a given unit in the data set, let τ and τ^* denote the frequency of the unit's true $X_{\mathcal{C}}$-category in original and perturbed data sets, respectively. Alternatively, τ and τ^* are the number of records in original and perturbed data sets, respectively, which match the unit's values of all variables in $\mathcal{C}$. The values of τ and τ^* actually depend on the unit considered, but for brevity we do not make that explicit. For any given unit, the probability of correctly identifying the unit in released data is 0 if its $X_{\mathcal{C}}$-category changed due to post-randomization, and $1/\tau^*$ otherwise. We calculated this probability for all 23,216 originally singleton and doubleton units. Averaging those probabilities over relevant subgroups of units we obtained the empirical conditional correct match probabilities reported in Table 3.

Table 3: Empirical Conditional Probabilities of Correct Match

	$\tau = 1$	$\tau = 2$	
$\tau^* = 1$	0.2315	0.3933	0.2849
$\tau^* = 2$	0.1961	0.3477	0.2827
	0.1348	0.3027	

In our example, 5,066 units satisfied the conditions $\tau = 1$ and $\tau^* = 1$ (i.e., the unit is unique in the original data and has a unique match in released data), of which 1,173 had correct (unique) matches, yielding $P(CM|\tau = 1, \tau^* = 1) = 1173 \div 5066 = 0.2315$, as reported in Table 3. Similarly, 2,570 units satisfied $\tau = 1$ and $\tau^* = 2$, of which 1,008 had their records unchanged. For any of these 2,570 units being the target unit, our intruder would randomly select one of the two units whose released category of $X_{\mathcal{C}}$ equals the target unit's $X_{\mathcal{C}}$-category and take that as the target unit. This would yield a correct match with probability 1/2 for the 1,008 units whose records did not change due to data perturbation, and with probability is 0 for the remaining 1,562 units. Using these, we obtained $P(CM|\tau = 1, \tau^* = 2) = [(1008) \times (0.5) + (1562) \times (0)] \div 2570 = 0.1961$. All other values in Table 3 were calculated similarly.

Here, we note a few points about the empirical probabilities in Table 3. As expected, all values in Table 3 are less than 0.395. Both $P(CM|\tau = 1, \tau^* = 1) = 0.2315$ and $P(CM|\tau = 2, \tau^* = 1) = 0.3933$ are very close to the corresponding upper bounds of 0.238 and 0.395, reported in Table 2. This is not surprising because in our example, the k_1 of (4.11) takes large values (between 124 and 1480 as noted earlier), in which case the sum in (4.11) is nearly a constant with respect to $\mathbf{T}$ and hence the inequality in (4.12)

holds fairly tightly. The overall probability of correct match for an originally $X_{\mathcal{C}}$-unique person is $P(CM|\tau = 1) = 0.1348$, which is fairly low. The corresponding probability for doubleton units, $P(CM|\tau = 2) = 0.3027$, is noticeably larger. This shows that putting too much emphasis on protecting sample unique units or judging a procedure by resulting identification risk for sample unique units may be misleading. In our example, one reason for the difference is that the percentage (36.47%) of singleton units with no match in perturbed data is much larger than the corresponding number (7.54%) for doubleton units.

In practice, an intruder would know the frequency (τ^*) of the $X_{\mathcal{C}}$-category of a target unit in released data, but not the frequency (τ) in the original data. Thus, $P(CM|\tau^*)$ are of much practical interest. As shown in Table 3, $P(CM|\tau^* = 1) = 0.2849$ and $P(CM|\tau^* = 2) = 0.2827$, which can be regarded as empirical values of $R_j(a)$ in (2.6) for $j = 1$ and 2. Interestingly, the above two values are very close, which implies that the correct match probability is almost the same for units showing up as singleton or doubleton units in released data. We note importantly that the perturbed data set has 7,558 singleton and doubleton units, which is 66.44% smaller than the corresponding number (23,216) in the original data.

7.2. Assessment of Data Utility Loss

In this part, we compare several sets of summary statistics from the original and perturbed data, respectively, to assess the procedure's effects on data utility. Our procedure kept the values and hence marginal distributions of all variables, except $X_2, \ldots, X_5$, unchanged. Note that although X_1 (gender) is included in $\mathcal{C}$, it remained unchanged as we used it (with

no collapsing) for data partitioning. On the other extreme, X_4 (marital status) and X_5 (PUMA) were changed freely, without any control. Tables 4 gives frequency distributions of X_4 for the original and perturbed data. In the second and third columns, the values in parentheses are relative frequencies. The 'never married' category includes all persons who are less than 16 years old. The fourth column gives the original count minus perturbed count. The last column gives estimated sampling standard deviations (SD) under simple random sampling with replacement (i.e., under binomial model). For example, the SD for the category 'married' is obtained as $[(59033)(.4182)(1-.4182)]^{1/2} = 119.84$.

Table 4: Frequency Distributions of Marital Status

Marital Status	Original Data	Perturbed Data	Difference	SD
Married	24688 (.4182)	24678 (.4180)	10	119.84
Widowed	3156 (.0535)	3180 (.0539)	-24	54.67
Divorced	4742 (.0803)	4704 (.0797)	38	66.03
Seperated	1040 (.0176)	1039 (.0176)	1	31.95
Never married	25407 (.4304)	25432 (.4308)	-25	120.30

In Table 4, the difference values are quite small, in magnitude, in comparison to SD values. This is consistent with our observation in Sec. 6 that asymptotically, the order of magnitude of perturbation variance is smaller than that of sampling variance. We may also mention that while 39.33% of all units were candidates for randomization, the corresponding percentage was much higher for widowed, divorced and separated categories, which have relatively small frequencies. For example 76.78% of the 4742 units falling in 'divorced' category were subject to post-randomization. Even for these groups, the

original and perturbed relative frequencies are practically the same.

Similarly, the original and perturbed distributions of race (X_3), given in Table 5, are very close. The fact that the original and perturbed counts are exactly the same for white and black is not a chance occurrence. Because of our choice of X_3^* and data partitioning, the race of any white or black person remained unchanged. The original data set contains exactly one person in 'Alaska Native alone' category, who can be identified correctly by matching race alone. This person cannot be identified in our perturbed data set, as the frequency for that category is zero. In any case, the probability of correctly identifying any sample unique person, not just by race but by all five variables in $\mathcal{C}$, in a post-randomized data set (with $\theta = 0.8$) is theoretically less than 0.238 (see Table 2). We had also examined distributions of age based on the original and perturbed data. The plots of the two cumulative distribution functions were almost indistinguishable and hence are not presented here.

We shall next examine certain effects of our procedure on joint distributions of two or more variables. Obviously, our procedure has no effect on the joint distribution of any subset of variables of $\bar{\mathcal{C}}$. Thus, we shall examine combinations of variables from the 5 key variables (sex, age, race, marital status (mar) and puma) and 2 non-key variables, viz. class of workers (work) with 9 categories and education level (edu) with 8 categories: grade 6 or less, grade 7-12 but no high school diploma, high school diploma, some college but not degree, Associate degree, Bachelor's degree, Master's or professional degree, and Doctorate degree.

To compare the original and perturbed joint distributions, we shall use the total variation distance (TVD) between two probability distributions. For a discrete random variable

Table 5: Distribution of Race or Ethnicity

Race or Ethnicity	Original	Perturbed
White	37201 (.6302)	37201 (.6302)
Black	15239 (.2581)	15239 (.2581)
American Indian alone	97 (.0016)	92 (.0015)
Alaska Native alone	1 (.000017)	0 (0)
American Indian & Alaska Native	42 (.0007)	46 (.0008)
Asian	3461 (.0586)	3345 (.0567)
Native Hawaiian & other Pacific Islander	20 (.0004)	21 (.0004)
Some other race alone	1349 (.0228)	1337 (.0227)
Two or more races	1623 (.0275)	1652 (.0280)

X, the TVD between two distributions $p(x)$ and $q(x)$ is

$$TVD(p,q) = \frac{1}{2}\sum_{x} |p(x) - q(x)|.$$

A useful interpretation of TVD comes from the fact that $TVD(p,q) = \sup_A |p(A) - q(A)|$, where the supremum is over all subsets A of the sample space of X. In our applications, p and q will represent relative frequency distributions based on the original and perturbed data. Then, for a given set of variables, letting f_i and $\tilde{f}_i$ denote the frequency of the ith cell in the original and perturbed data, respectively, the TVD measure becomes

$$TVD = \frac{1}{2}\sum_{i} |\frac{f_i}{n} - \frac{\tilde{f}_i}{n}| = \frac{1}{2n}\sum_{i} |f_i - \tilde{f}_i|, \tag{7.1}$$

where n is the sample size.

Gomatam and Karr (2003) used TVD for measuring distortions due to data swapping.

Shlomo and Skinner (2010) used relative absolute average distance ($RAAD$) per cell, which relates to TVD by $RAAD = 100[1 - 2 \times (TVD)]$. For several combinations of variables, Table 6 gives the values of TVD and also the number of cells. All TVD values in Table 6 are quite small and the joint distributions with three largest TVD values have over 350 cells and involve one variable (puma or mar) that was post-randomized without control.

Table 6: Total Variation Distances Between Original and Perturbed Distributions

Variables	TVD	Number of cells	Variables	TVD	Number of cells
race, mar	0.0028	45	puma, work	0.0198	396
race, puma	0.0013	396	puma, edu	0.0324	352
race, edu	0.0088	72	sex, race, mar	0.0060	90
race, work	0.0035	81	sex, race, edu	0.0093	144
mar, edu	0.0127	40	mar, race, edu	0.0218	360
mar, work	0.0070	45	sex, race, work	0.0039	162

8. Discussion

In this paper, we presented a novel approach to measuring identification risk and setting practical identification risk control goals and deductively devising post-randomization procedures for achieving those goals, without having to estimate any unknown parameters. We also exhibited attractive properties, both theoretically and empirically, of our data perturbation procedure. For making statistical inferences, a perturbed data set may be

treated as the original data set without much loss of accuracy.

One limitation of our procedure is that for Goal 3, it works well essentially for $\xi \geq 0.35$. However, our assumptions that the intruder knows the values of all key variables for his target and that the target is in the sample are quite conservative. In practice, an intruder would not know if the target is in the sample or not, especially when an agency releases a subsample as in PUMS, and may know the values of only some (but not all) of the key variables. Thus, the probability of a typical intruder's match being correct would be much smaller than ξ. Moreover, an intruder may not know much about the agency's data perturbation procedure (as noted above), and consequently find it very difficult to assign any probability to his match being correct.

While the general procedure of Section 5 may be improved by modifying our methods for data partitioning and forming IFPR blocks, the limitation mentioned above stems primarily from the simple structure of IFPR involving only one design parameter (θ). Larger classes of PRAM matrices (P) with more complex structures and characterized by multiple design parameters might be more versatile and useful for achieving stricter disclosure control goals, which is a topic for future research. We hope that the ideas and results in this paper will be practically useful and stimulate further research.

Acknowledgment. We thank Professor Bimal Sinha for some helpful discussions and Dr. Martin Klein for some helpful remarks on an earlier draft.

References

[1] Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990), "Disclosure Control of Microdata", *Journal of the American Statistical Association*, 85(409), 38-45.

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

[3] Cox, L. H., Karr, A. F. and Kinney, S. K. (2011), "Risk-utility Paradigms for Statistical Disclosure Limitation: how to think, but not how to act", *International Statistical Review*, 79(2), 160-183.

[4] Cruyff, M. J., Van Den Hout, A., and Van Der Heijden, P. G. (2008), "The Analysis of Randomized Response Sum Score Variables", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 21-30.

[5] Dalenius, T. (1977), "Towards a Methodology for Statistical Disclosure Control", *Statistisk Tidskrift*, 5, 429-444.

[6] Duncan, G.T., Elliot, E. and Juan Jose Salazar, G. (2011), *Statistical Confidentiality: Principles and Practice*, New York: Springer.

[7] Duncan, G. T., and Lambert, D. (1986), "Disclosure-limited Data Dissemination", *Journal of the American statistical association*, 81(393), 10-18.

[8] Duncan, G. T., and Lambert, D. (1989), "The Risk of Disclosure for Microdata", *Journal of Business and Economic Statistics*, 7(2), 207-217.

[9] Duncan, G. T., and Stokes, S. L. (2004), "Disclosure Risk vs Data Utility: The RU confidentiality map as applied to top coding", *Chance*, 17(3), 16-20.

[10] Dwork, C. (2006), "Differential Privacy", in M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, ICALP (2), volume 4052 of *Lecture Notes in Computer Science*, pp. 1-12, Berlin: Springer.

[11] Gomatam, S., and Karr, A. F. (2003), "Distortion measures for categorical data swapping", Technical report 131, National Institute of Statistical Sciences, Research Triangle Park, NC.

[12] Gouweleeuw, J. M., Kooiman, P., and de Wolf, P. P. (1998), "Post Randomisation for Statistical Disclosure Control: Theory and implementation", *Journal of official Statistics*, 14(4), 463.

[13] Greenberg, B. V., and Zayatz, L. V. (1992), "Strategies for Measuring Risk in Public Use Microdata Files", *Statistica Neerlandica*, 46(1), 33-48.

[14] 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: Wiley.

[15] Lambert, D. (1993), "Measures of Disclosure Risk and Harm", *Journal of Official Statistics*, 9(2), 313-313.

[16] Nayak, T. K. and Adeshiyan, S. A. (2009), "A Unified Framework for Analysis and Comparison of Randomized Response Surveys of Binary Characteristics", *Journal of Statistical Planning and Inference*, 139(8), 2757-2766.

[17] Nayak, T. K. and Adeshiyan, S. A. (2015), "On Invariant Post-randomization for Statistical Disclosure Control", *International Statistical Review*, doi: 10.1111/insr.12092.

[18] Nayak, T. K., Adeshiyan, S. A. and Zhang, C. (2016), "A Concise Theory of Randomized Response Techniques for Privacy and Confidentiality Protection ", in A. Chaudhuri, T.C. Christofides and C.R. Rao editors, *Data Gathering, Analysis and Protection*

of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits, (to appear) New York: Elsevier.

[19] Reiter, J. P. (2005), "Estimating Risks of Identification Disclosure in Microdata", *Journal of the American Statistical Association*, 100(472), 1103-1112.

[20] Rubin D. B. (1993), "Discussion: Statistical Disclosure Limitation", *Journal of Official Statistics*, 9(2), 461-468.

[21] Shlomo, N., and De Waal, T. (2008), "Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure", *Journal of Official Statistics*, 24(2), 1-26.

[22] 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(3), 1291-1310.

[23] Skinner, C. J., and Elliot, M. J. (2002), "A Measure of Disclosure Risk for Microdata", *Journal of the Royal Statistical Society: series B (statistical methodology)*, 64(4), 855-867.

[24] Van den Hout, A., and Elamir, E. A. (2006), "Statistical Disclosure Control Using Post Randomisation: Variants and Measures for Disclosure Risk", *Journal of Official Statistics*, 22(4), 711-731.

[25] Van den Hout, A. and Kooiman, P. (2006), "Estimating the Linear Regression Model with Categorical Covariates Subject to Randomized Response", *Computational Statistics and Data Analysis*, 50(11), 3311-3323.

[26] Van den Hout, A. and Van der Heijedn, P. G. (2002), "Randomized Response, Statistical Disclosure Control and Misclassification: a review", *International Statistical Review*, 70, 269-288.

[27] Warner, S. L. (1965), "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias", *Journal of the American Statistical Association*, 60(309), 63-69.

[28] Warner, S. L. (1971), "The Linear Randomized Response Model", *Journal of the American Statistical Association*, 66(336), 884-888.

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