RESEARCH REPORT SERIES
(Statistics #2020-01)

A Post-randomization Method for Rigorous Identification Risk Control in Releasing Microdata

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

[1]Facebook AI Applied Research, Menlo Park, CA 94025;
[2]Center for Statistical Research and Methodology, U.S. Census Bureau
and Department of Statistics, George Washington University

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

Report Issued: April 17, 2020

A Post-randomization Method for Rigorous Identification Risk Control in Releasing Microdata *

Xiaoyu Zhai† and Tapan K. Nayak‡

Abstract

One significant concern in releasing survey microdata is the possibility of identifying the records of some survey units by matching the values of some of the variables, called key or pseudo-identifying variables, whose values can be obtained easily from other sources. For categorical key variables, Nayak, Zhang and You [*Int. Stat. Rev*, 86(2), 2018, 300-321] developed a novel approach for measuring and controlling identification risks. For any $\xi > 1/3$, it can guarantee that any unit's probability of correct identification would not exceed ξ. We present another post-randomization method for giving that guarantee more stringently, even for $\xi \leq 1/3$. We use data partitioning and unbiased post-randomization as two effective tools for preserving data utility. We illustrate and assess the procedure by applying it to a U.S. Census Bureau's publicly released data set.

Key words and Phrases: Identity disclosure; key variable; data partitioning; post-randomization block; data utility.

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

†Facebook AI Applied Research, Menlo Park, CA 94025. Her work was completed while she was a doctoral student at The George Washington University.

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

1. Introduction

One basic goal of most statistical agencies is to collect and release data to assist research and inform the public and policy makers. However, the original data may reveal private information about some of the survey participants or units, even if name, social security number and other direct identifiers are removed. In a microdata set that contains each unit's values for many variables, one might be able to correctly identify the records of a target unit by matching the values of some of the variables, such as gender, race and occupation, which can be learned easily from other sources. Then, one can learn the identified unit's values for all other variables. This is called identity disclosure and it is regarded as one of the most severe forms of exposing a respondent's private information.

In this paper, we focus on identity disclosure in microdata release and controlling identification risks. However, data confidentiality breaches may occur in many other forms and even when only data summaries are released. For discussions about other types of disclosure and various disclosure control methods, such as grouping, data swapping, cell suppression, noise addition, synthetic data and post-randomization, we refer interested readers to the books: Willenborg and de Waal (2001), Duncan et al. (2011) and Hundepool et al. (2012).

Protecting data confidentiality, which is required for legal reasons and upholding public trust, usually requires some perturbation of the true data before publishing. That also dilutes and distorts statistical information. So, one should examine the trade-offs between confidentiality protection and data utility to choose a suitable perturbation method. But, formalizing this idea mathematically is difficult largely because neither disclosure risk nor data utility can be defined precisely and also satisfactorily to diverse data subjects and

data users; see e.g., Lambert (1993), Cox et al. (2011) and Skinner (2012). For assessing data utility, we shall compare the original and perturbed data distributions.

Consider a microdata set containing values of multiple variables for each sampled unit. The variables that an intruder may use for matching and identifying the records of a target unit are called key (or pseudo-identifying) variables, whose values should be easily available from other sources. Customarily, it is assumed that the key variables are stated by the agency and all are categorical. Many papers e.g., Paass (1988), Bethlehem et al. (1990), Greenberg and Zayatz (1992), Chen and Keller-McNulty (1998), Skinner and Elliot (2002), Shlomo and De Wall (2008) and Shlomo and Skinner (2010), discuss measuring and controlling identification risk. Bethlehem et al. (1990) defined identification risk as the probability that a unit is population unique, given that it is sample unique, with respect to the key variables. That ignores the effects of data perturbation on identity disclosure. Taking data perturbation into account, Shlomo and Skinner (2010) defined a unit's identification risk as the probability that the unit is correctly identified given that it has a unique match in released data. This and previous measures are difficult to use because they vary with the units and also depend on unknown population level frequencies. Nayak, Zhang and You (2018), henceforth NZY, proposed another approach for controlling identification risk that does not involve unknown parameters. We shall use the NZY framework and so review it in more details below.

As in NZY and previous works, we assume that all key variables are categorical and are specified by the agency. Let X denote the cross-classification of all key variables. Suppose X has k cells, denoted $c_1, \ldots, c_k$. While category and cell are synonymous, for clarity we shall use cell for a cross-classified variable and category for a single variable. For identification risk control, the true values (cells) of X are replaced by their post-

randomized values, as reviewed later. Let Z denote the post-randomized version of X. Most of our discussions involve X and Z, as the values of all other variables are not changed. Let T_j and S_j denote the frequencies of c_j in the original and perturbed data, respectively, and let $\mathbf{T} = (T_1, ..., T_k)'$ and $\mathbf{S} = (S_1, ..., S_k)'$.

Consider an intruder A who wants to identify the records of a target unit B in the released data. Let $X_{(B)}$ denote B's value of X, and suppose $X_{(B)} = c_j$. NZY assumed that (i) A knows $X_{(B)}$, (ii) A knows that B is in the sample and (iii) A randomly selects one of the records in the released data that match $X_{(B)}$, as a match for B. Note that B (with $X_{(B)} = c_j$) matches with S_j records in the released data, and if $S_j = 0$, the intruder declares no match. Thus, B can be matched correctly only if B's value of X remain unchanged during data perturbation. As NZY noted, the assumption in (ii) above is overly stringent because agencies do not reveal which population units are in the sample.

Let CM denote the event that B is *correctly matched* in the preceding setup and define $R_j(a) = P(CM|X_{(B)} = c_j, S_j = a)$ taking randomnesses in both sampling and data perturbation into account. NZY proposed that to limit identification risks, the agency should guarantee, with appropriate data perturbation, that

$$R_j(a) \leq \xi \quad \text{for all } j = 1, \ldots, k, \text{ and all integers } a \geq 1, \tag{1.1}$$

where ξ is specified by the data agency. Thus, no unit's correct identification probability would exceed ξ. In $R_j(a)$, conditioning on $S_j = a$ is quite sensible because an intruder's confidence in a declared match may depend on the number of matches for B in the released data.

As with all past identification risk measures, $R_j(a)$ also depends on the unknown

population frequencies. To tackle this impediment, NZY cleverly further conditioned on $\mathbf{T}$ and focused on

$$R_j(a, \mathbf{t}) = P(CM|X_{(B)} = c_j, S_j = a, \mathbf{T} = \mathbf{t}), \tag{1.2}$$

which does not involve unknown parameters when the data are perturbed using post-randomization. So, $R_j(a, \mathbf{t})$ can be assessed without estimating any parameter. NZY derived an upper bound for (1.2) and used that to develop a method that can guarantee

$$R_j(a, \mathbf{t}) \leq \xi \quad \text{for } j = 1, \ldots, k, \text{ all } a > 0 \text{ and all } \mathbf{t}, \tag{1.3}$$

for any given $\xi > 1/3$. Obviously, (1.3) implies (1.1).

The NZY procedure works only for $\xi > 1/3$; it cannot ensure (1.3) if $\xi \leq 1/3$. However, a data agency might want to guarantee (1.1) for some $\xi \leq 1/3$. In this paper, we propose another method that can guarantee (1.3) more stringently, even when $\xi \leq 1/3$. In Section 2, we briefly review the main parts of the NZY procedure. In Section 3, we describe our method. Basically, we change the post-randomization part of the NZY method, using a different structure for the randomization probabilities. In Section 4, we present an example to illustrate our method and compare it with the NZY method. In Section 5, we make some concluding remarks.

2. A Review of the NZY Procedure

The probability of correctly identifying any unit in cell c_j (of X) in the original data is $1/t_j$, which exceeds ξ if and only if $t_j < 1/\xi$. So, for given $0 < \xi < 1$, a cell is considered *sensitive* if its frequency is less than $1/\xi$. The NZY procedure satisfies (1.3) by post-

randomizing X for all units in all sensitive cells. To better preserve data utility, it divides the sensitive cells into relatively homogeneous groups, called post-randomization blocks (PRBs), and then post-randomizes X within each PRB. A simple method for creating PRBs is: (i) first partition the data using coarsened versions of the key variables and (ii) then take all sensitive cells in each partition set to form one PRB. For additional clarity and later comparison with our procedure, we next review the following example from NZY.

The example applies the NZY method to the U.S. Census Bureau's 2013 person-level Public Use Microdata Sample (PUMS) for the state of Maryland, available at https://www.census.gov/programs-surveys/acs/data/pums.html. It contains the values of several demographic and economic variables of 59,033 individuals. The example takes gender (2), age (92), race (9), marital status (5) and Public Use Microdata Area (PUMA) (44) as the key variables, where the values in parentheses show the number of categories of the variables. The cross-classification of the five key variables yields 364,320 cells. The example considers $1/3 < \xi < 1/2$, which implies that only singleton and doubleton cells (with frequency 1 and 2, respectively) are sensitive. The data set yields only 25,406 nonempty cells, of which 13,662 are singleton and 4,777 are doubleton. So, the number of sensitive cells is $13,662 + 4,777 = 18,439$, which contain $13,662 + (2 \times 4772) = 23,216$ units, out of 59,033.

In the example, the data are partitioned using gender, 7 age intervals, viz. 0–17, 18–24, 25–34, 35–44, 45–54, 55–64, and 65 and above, and the three race categories: white, black and 'other races.' Their possible combinations divide the data into 42 partition sets. For example, all females of 'other races' with age between 25 and 34 constitute one partition set. All originally singleton and doubleton cells (of the key variables) in a

partition set define one PRB. For the data set, the 42 PRBs contained between 124 and 1480 cells. By applying post-randomization within each PRB, the NZY method keeps the perturbed value of each unit in its PRB. This controls the nature and magnitude of data perturbation. For example, the preceding partitioning preserves gender, age in the broader intervals and race if black or white. But, it permits marital status and PUMA to change freely.

Inspired by randomized response (RR) methods, Gouweleeuw et al. (1998) introduced the post-randomization method (PRAM) as a statistical disclosure control tool for categorical variables. It has been further studied by Van den Hout and Van der Heijden (2002), Van den Hout and Elamir (2006), Cruyff et al. (2008), Shlomo and De Waal (2008), Shlomo and Skinner (2010) and others. PRAM applies randomized response, proposed originally by Warner (1965) and subsequently studied by many others; see Chaudhuri and Mukerjee (1988), Blair et al. (2015) and Nayak et al. (2016) for reviews and additional references. However, as Nayak and Adeshiyan (2016) observed, a salient difference between RR and PRAM is that the transition probabilities may depend on the data in PRAM but not in RR. The choice of the *transition probability matrix* (TPM) is vital to using PRAM. As we describe next, the NZY method uses TPMs with a specific structure and determined by one design parameter.

Suppose a PRB contains l cells, denoted $c_1, \ldots, c_l$, for notational simplicity. Naturally, l is much smaller than the total number of sensitive cells of X. The NZY method

randomizes the true cell of each unit in this PRB with the probabilities

$$p_{ij} = P(Z = c_i | X = c_j) = \begin{cases} 1 - \frac{\theta}{t_j}, & \text{if } i = j; \\ \frac{\theta}{(l-1)t_j}, & \text{if } i \neq j, \end{cases} \tag{2.1}$$

where t_j is the original frequency of c_j and θ is a design parameter, chosen suitably to satisfy (1.3). Specifically, for given $1/3 < \xi < 1$, θ is the solution of $h(\theta) = \xi$, where

$$h(\theta) = \begin{cases} \frac{1-\theta}{1-\theta+\theta^2}, & \text{if } \theta \leq \frac{2}{3}, \\ \frac{2-\theta}{4-2\theta+\theta^2}, & \text{if } \theta > \frac{2}{3}. \end{cases}$$

It was shown that (i) $h(\theta)$ is a strictly decreasing function of θ, with $h(0) = 1$ and $h(1) = 1/3$ and thus for any $1/3 < \xi < 1$, $h(\theta) = \xi$ admits a unique solution for θ in $(0, 1)$ and (ii) this choice of θ ensures (1.3), if $l \geq (1 - \theta)^{-1}$ for all PRBs.

Clearly, the TPM $P = ((p_{ij}))$ given by (2.1) is adaptive, viz. it depends on the frequencies of the cells in the PRB. Also, P is determined by a single parameter θ. Under (2.1), a unit's cell is changed with probability θ/t_j, which is inversely proportional to the frequency of the unit's true cell, and when it is changed, the replacement is picked at random from the remaining cells in the PRB. NZY called this IFPR (inverse frequency post-randomization). Also, it is an unbiased procedure in the sense that the expected frequency of any cell after data perturbation is the same as its original frequency.

One limitation of the NZY method is that it can guarantee (1.3) only for $\xi > 1/3$, because if $\xi \leq 1/3$, then $h(\theta) = \xi$ does not admit a solution in $(0, 1)$. Actually, this is a consequence of the structure of the TPM given by (2.1). It changes the cell of a doubleton

unit with probability $\theta/2$, which must be less than 0.5, as $0 < \theta < 1$. This limits the amount of protection it can give to doubleton units. Also, as Table 3 of NZY shows, under IFPR, empirical identification risks of doubleton units are higher than those of singleton units. So, one way to ensure (1.3) when $\xi \leq 1/3$ might be to use TPMs with a different structure, which we do in the next section.

3. The Proposed Method

Our main objective is to develop a method that can satisfy (1.3) more stringently than NZY, i.e., even when $\xi \leq 1/3$. We shall use the idea of data partitioning and form PRBs similarly as in the NZY method. But, we shall use a different post-randomization scheme. Here, it is important to discuss two important points related to the value of ξ. First, as NZY noted, one should not use quite a small value for ξ because (i) an intruder should have strong evidence, viz. a high correct match probability, to declare a match and (ii) the assumption that the intruder knows that the target is in the sample is overly conservative; it ignores the protection given by sampling, which is substantial when the sampling fraction is small, as is the case in most applications. An unduly small value of ξ should not be used to avoid unnecessary data utility loss. Actually, NZY felt that in most applications reasonable values of ξ should be larger than 1/3.

Second, as ξ decreases, the data perturbation needs changes discretely at $\xi = 1/2, 1/3, 1/4, 1/5, \ldots$. Specifically, the "sensitivity" of a cell may change only at these values; recall that a cell is sensitive if its frequency is less than $1/\xi$. For $1/3 \leq \xi < 1/2$, singleton and doubleton cells are sensitive; for $1/4 \leq \xi < 1/3$, tripleton cells are also sensitive and so on. Thus, our method, especially how we form the PRBs, changes discretely at the

above mentioned values of ξ. Considering these two points, we shall present our method for satisfying (1.3) specifically for any $1/4 \leq \xi < 1/3$, i.e., for protecting all singleton, doubleton and tripleton units. However, the basic ideas can be used to guarantee (1.3) even when $\xi < 1/4$. For example, for $1/5 \leq \xi < 1/4$, our method should work if cells with frequency 4 are also included in the PRBs.

3.1. Our Post-randomization Scheme

Suppose a PRB contains l cells. Let $c_1, \ldots, c_l$ denote its cells, T_i and S_i denote the frequency of c_i before and after post-randomization, respectively, and $m = \sum T_i (= \sum S_i)$ denote the total number of units in the PRB. Let $\mathbf{T} = (T_1, \ldots, T_l)'$ and $\mathbf{S} = (S_1, \ldots, T_l)'$. We alert the reader that some of these symbols were used earlier to denote analogous quantities for the whole data set. For notational simplicity, we use them also locally for a PRB. Also, we shall explore effects of our procedure on both identification risk and data utility at PRB level, which should give an insightful understanding of global effects of the procedure.

For given $\mathbf{T} = \mathbf{t}$, a PRAM with TPM P is unbiased if

$$E(\mathbf{S}|\mathbf{t}, P) = \mathbf{t} \quad \text{or} \quad P\mathbf{t} = \mathbf{t} \tag{3.1}$$

where the expectation is with respect to post-randomization. It is seen easily that the solutions of (3.1) for P form a nonempty convex set. Equation (2.1) gives one class of solutions of (3.1), which NZY used. Two specific solutions that are of interest to us are

I (identity matrix) and

$$P_* = \frac{1}{m}\begin{bmatrix} t_1 & t_1 & \dots & t_1 \\ t_2 & t_2 & \dots & t_2 \\ \vdots & \vdots & \ddots & \vdots \\ t_l & t_l & \dots & t_l \end{bmatrix} = \frac{1}{m}\mathbf{t}1'. \tag{3.2}$$

Nayak and Adeshiyan (2016) derived some useful properties of unbiased PRAM. Letting $P_1, \dots, P_l$ denote the columns of a solutions of (3.1), they showed that the conditional variance of $\mathbf{S}$ given $\mathbf{t}$ and P is

$$V(S|\mathbf{t}, P) = \sum_{i=1}^{l} t_i[D_{P_i} - P_i P_i'] = D_{\mathbf{t}} - \sum_{i=1}^{l} t_i P_i P_i', \tag{3.3}$$

where for a vector $a = (a_1, \dots, a_l)'$, D_a denotes the diagonal matrix with diagonal elements $a_1, \dots, a_l$. They also show that (3.3), which may be viewed as the post-randomization induced variance, is the largest when $P = P_*$, as given in (3.2), in the sense that $V(S|\mathbf{t}, P_*) - V(S|\mathbf{t}, P)$ is nonnegative definite for all P satisfying (3.1). Thus, the least and most variance inducing solutions of (3.1) are $P = I$ and $P = P_*$.

As with IFPR, we wanted to consider a class of solutions of (3.1) indexed by a single quantity, for easily assessing identification risks and data utility loss. We decided to

explore convex combinations of the two extreme TPMs, viz.

$$P_\alpha = \alpha T_* + (1-\alpha)I = \begin{bmatrix} 1-\alpha+\alpha\frac{t_1}{m} & \alpha\frac{t_1}{m} & \dots & \alpha\frac{t_1}{m} \\ \alpha\frac{t_2}{m} & 1-\alpha+\alpha\frac{t_2}{m} & \dots & \alpha\frac{t_2}{m} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha\frac{t_l}{m} & \alpha\frac{t_l}{m} & \dots & 1-\alpha+\alpha\frac{t_l}{m} \end{bmatrix} \tag{3.4}$$

for $\alpha \in (0,1)$. Here, the probabilities of changing the true categories increase with α. Next, we discuss how to choose the value of α to satisfy (1.3).

3.2. Assessment and Control of Identification Risk

First, we consider $X_{(B)} = c_1$ and evaluate $R_1(a, \mathbf{t})$, as defined in (1.2), under P_α. One can similarly derive $R_j(a, \mathbf{t})$ for $j = 2, \dots l$. Let U denote the number of units in c_1 excluding B that stay in c_1 after post-randomization and V denote the number of units that move into c_1 from other cells (in the PRB). Under P_α, U and V are independent binomial random variables with $U \sim b(r, \eta_1)$ and $V \sim b(\nu, \eta_1)$, where $r = t_1 - 1, \nu = m - t_1, \eta_1 = 1 - \alpha + \alpha\frac{t_1}{m}$ and $\eta_2 = \alpha\frac{t_1}{m}$. For $a \geq 1$, note that $P(CM|Z_{(B)} = c_1, X_{(B)} = c_1, S_1 = a, \mathbf{T} = \mathbf{t}) = 1/a$ and $P(CM|Z_{(B)} \neq c_1, X_{(B)} = c_1, S_1 = a, \mathbf{T} = \mathbf{t}) = 0$. Using these, as in NZY, and letting

$\beta_i = \eta_i/(1-\eta_i), i = 1, 2$, we obtain

$$\begin{aligned} R_1(a, \mathbf{t}) &= \frac{1}{a} P(Z_{(B)} = c_1 | X_{(B)} = c_1, S_1 = a, \mathbf{T} = \mathbf{t}) \\ &= \frac{1}{a} \frac{P(Z_{(B)} = c_1, S_1 = a | X_{(B)} = c_1, \mathbf{T} = \mathbf{t})}{P(S_1 = a | X_{(B)} = c_1, \mathbf{T} = \mathbf{t})} \\ &= \frac{1}{a} \Big[\frac{\eta_1 P(U+V = a-1)}{\eta_1 P(U+V = a-1) + (1-\eta_1) P(U+V = a)} \Big] \\ &= \frac{1}{a} \Big[1 + \frac{1}{\beta_1} \frac{\Sigma_a}{\Sigma_{a-1}} \Big]^{-1}, \end{aligned} \tag{3.5}$$

where

$$\Sigma_a = \sum_{u=0}^{r} \binom{r}{u} \binom{\nu}{a-u} \beta_1^u \beta_2^{a-u} \tag{3.6}$$

with the understanding that $\binom{b}{c}$ is 0 if $c < 0$ or $c > b$; the actual range of the summation is $\max\{0, a-\nu\} \le u \le \min\{a, r\}$. In particular, $\Sigma_0 = 1, \Sigma_1 = r\beta_1 + \nu\beta_2$ and considering $a = 1$ in (3.5) we get

$$\begin{aligned} R_1(1, \mathbf{t}) &= \Big[1 + \frac{r\beta_1 + \nu\beta_2}{\beta_1} \Big]^{-1} \\ &= \Big[t_1 + \frac{\alpha^2 t_1 (m - t_1)^2}{(m - \alpha t_1)\{m(1-\alpha) + \alpha t_1\}} \Big]^{-1}, \end{aligned} \tag{3.7}$$

as $r = t_1 - 1, \nu = m - t_1$, $\beta_1 = \frac{(1-\alpha)m + \alpha t_1}{\alpha(m - t_1)}$ and $\beta_2 = \frac{\alpha t_1}{m - \alpha t_1}$.

Clearly, $R_1(1, \mathbf{t})$ depends on $\mathbf{t}$ only through t_1. We can verify that $R_1(1, \mathbf{t})$ is a decreasing function of t_1 and thus attains its maximum when $t_1 = 1$. So,

$$R_1(1, \mathbf{t}) \le \Big[1 + \frac{\alpha^2 (m-1)^2}{(m-\alpha)\{m(1-\alpha) + \alpha\}} \Big]^{-1} = \psi(\alpha), \text{ say,} \tag{3.8}$$

for all $\mathbf{t}$. One can verify that $\psi(\alpha)$ in (3.8) is an increasing function of α with $\psi(1) = 1/m$

and $\psi(0) = 1$. So for any $1/m < \xi < 1$, the equation $\psi(\alpha) = \xi$ admits a unique solution for α, say α_ξ. For $1/4 \leq \xi < 1/3$, which is our focus, we need $m > 4$ to find a proper α_ξ. However, this condition is easily satisfied. We use this α_ξ in (3.4) for post-randomizing all units in the PRB. Clearly, this guarantees $R_1(1, \mathbf{t}) \leq \xi$ for all $\mathbf{t}$. We shall show that this also guarantees $R_1(a, \mathbf{t}) \leq \xi$ for all $a \geq 2$ and all $\mathbf{t}$.

From (3.5) it follows that for $a \geq 1$, $R_1(a, \mathbf{t}) \geq R_1(a+1, \mathbf{t})$ if and only if

$$\beta_1 \Sigma_a \Sigma_{a-1} + (a+1)\Sigma_{a-1}\Sigma_{a+1} - a\Sigma_a^2 \geq 0. \tag{3.9}$$

For $a = 1$, the left hand side of (3.9) is

$$\begin{aligned} &\beta_1(r\beta_1 + \nu\beta_2) + 2\big[\frac{r(r-1)}{2}\beta_1^2 + r\nu\beta_1\beta_2 + \frac{\nu(\nu-1)}{2}\beta_2^2\big] - (r\beta_1 + \nu\beta_2)^2 \\ &= \nu\beta_1\beta_2 - \nu\beta_2^2 = \nu\beta_2(\beta_1 - \beta_2) > 0, \end{aligned}$$

as $\beta_1 > \beta_2$. This shows that under any P_α, $R_1(1, \mathbf{t}) \geq R_1(2, \mathbf{t})$ for all $\mathbf{t}$ and consequently, P_α with $\alpha = \alpha_\xi$ guarantees that $R_1(2, \mathbf{t}) \leq \xi$ for all $\mathbf{t}$. In the appendix, we prove that under P_α, we also have $R_1(2, \mathbf{t}) \geq R_1(3, \mathbf{t})$ for all $\mathbf{t}$. Note from (3.5), that $R_1(a; t) < 1/4$ for $a \geq 4$. In conclusion, for $1/4 \leq \xi < 1/3$, post-randomization of all units in each PRB using P_α with α as the solution of $\psi(\alpha) = \xi$ guarantees (1.3).

We should note that α_ξ, which is the solution of $\psi(\alpha) = \xi$, depends also on m, the number of units in the PRB. So, α_ξ would vary over the PRBs, which are formed via data partitioning. Table 1 shows α_ξ for some (m, ξ) pairs. As expected, α_ξ increases as ξ decreases. Also, for each ξ, as m increases, α_ξ decreases, but quite slowly over moderate to large m ($m \geq 20$). Thus, when all PRBs are moderate or large ($m \geq 20$),

one may use a common α_ξ in all PRBs for convenience. It can be verified that for large m, $\alpha_\xi \approx [\sqrt{d^2+4d}-d]/2$, where $d = 1/\xi - 1$. In our example in the next section, we calculate α_ξ for each PRB (considering its size) and use it.

	m						
ξ	20	30	40	50	100	500	1000
1/2	0.645	0.636	0.631	0.628	0.623	0.619	0.619
1/3	0.759	0.748	0.743	0.740	0.734	0.729	0.728
1/4	0.827	0.815	0.809	0.805	0.798	0.793	0.792
1/5	0.866	0.853	0.847	0.843	0.836	0.830	0.829
1/6	0.894	0.880	0.874	0.870	0.862	0.856	0.855
1/8	0.930	0.915	0.908	0.904	0.896	0.889	0.888

Table 1: Values of α_ξ for some m and ξ.

3.3. Variance Due to Data Perturbation

Here, we study the effects of our method on the frequencies of the cells of X. Consider a fixed data partitioning, as described in Section 2. Then, each cell falls in a specific partition set, irrespective of the data. Consider a cell with original frequency t_i. Let S_i denote its frequency after data perturbation. Generating S_i depends on the data in two important ways. First, if $t_i = 0$ or $t_i \geq 1/\xi$, then $S_i = t_i$, as the cell is not sensitive and its units are left unperturbed. If $0 < t_i < 1/\xi$, then the cell is included in the PRB for the partition set, and S_i is generated via post-randomization.

For $0 < t_i < 1/\xi$, the value of S_i depends only on the post-randomization in its PRB. However, the composition of each PRB is data dependent, via cell frequencies in the partition set. Let us examine the properties of the perturbed frequencies of the cells of a *given* PRB. Consider a PRB and denote its cells, frequencies etc. as in Section 3.1. Note that here $0 < t_i < 1/\xi$ for $i = 1, \ldots, l$. Since we use an unbiased post-randomization, we

have:

$$E(\mathbf{S}|\mathbf{t}, \text{PRB}) = \mathbf{t}$$

for all $\mathbf{t}$, which implies that the expected perturbed frequency of any cell under our method is the same as its original frequency.

To derive the conditional variance-covarince matrix of $\mathbf{S}$ given $\mathbf{t}$ and the PRB, we use (3.3). Here, $P_i = (1-\alpha)e_i + \alpha\frac{\mathbf{t}}{m}$, where e_i is a vector whose ith component is 1 and the rest are 0. By routine algebra we obtain:

$$V(\mathbf{S}|\mathbf{t}, \text{PRB}) = \alpha(2-\alpha)[D_{\mathbf{t}} - \frac{1}{m}\mathbf{t}\mathbf{t}`]. \tag{3.10}$$

In particular, (3.10) shows that

$$V(S_i|\mathbf{t}, \text{PRB}) = \alpha(2-\alpha)t_i(1-\frac{t_i}{m}) \approx \alpha(2-\alpha)t_i \tag{3.11}$$

when t_i/m is small. Note that (3.11), which is the variance induced by our method, applies only to the cells with $0 < t_i < 1/\xi$. Otherwise, the induced variance is 0.

For comparison, we note from NZY that under IFPR,

$$V(S_i|\mathbf{t}, \text{PRB}) = \theta(2-\frac{\theta}{t_i}) - \theta^2\sum_{j\neq i}\frac{1}{t_j} \leq \theta(2-\frac{\theta}{t_i}) - \frac{\theta^2}{2}, \tag{3.12}$$

where the inequality follows from the fact that each PRB contains at least two cells. However, note that even under a common data partitioning, the PRBs under NZY and our methods are different for $1/3 < \xi \leq 1/4$, where our PRBs also include all tripleton cells. Nonetheless, for $1/3 < \xi < 1/2$, the two methods use the same PRBs and usually

(3.12) is smaller than (3.11). For example, for $\xi = .395, \theta = .8$ and $\alpha_\xi \approx .69$ and for $t_i = 2$, the right hand sides of (3.11) and (3.12) are 1.808 and 0.96, respectively.

4. An Example

We applied our method to the PUMS data set described in Section 2. For fair comparison, we used the same key variables (gender, age, race, marital status and PUMA) and data partitioning as in NZY (and described in Section 2), which yielded 42 partition sets. However, we applied our post-randomization with two values of ξ, viz. $\xi = .395$ (as in NZY) and $\xi = .25$, which the NZY method cannot attain. For $\xi = .395$, only singleton and doubleton cells are sensitive and those within a partition set form a PRB. For $\xi = .25$, tripleton cells are also sensitive and thus included in the PRBs. As stated in Section 2, the data set contains 59,033 observations and shows 13,662 singleton and 4,777 doubleton cells. It also yields 2,360 tripleton cells.

For $\xi = .395$, the number of cells in the PRBs ranged between 124 and 1,480 (as reported in NZY), $\theta = 0.8$ and $\alpha_\xi \approx 0.69$. Thus, our method changes the true cell of each unit in singleton and doubleton cells with probability about 0.69. In contrast, the NZY method (with $\theta = 0.8$) changes the cell of singleton and doubleton units with probabilities 0.8 and 0.4, respectively. As Table 1 shows, for $\xi = .25$, $\alpha_\xi \approx 0.79$. Here, the cell counts in the 42 PRBs varied between 251 and 2,535, which are much larger than the corresponding values for $\xi = .395$. This is because tripleton cells are included in the PRBs when $\xi = .25$ but not for $\xi = .395$. For $\xi = .25$, we could use a finer data partition to improve data utility. Following NZY, we report below some results from two perturbed data sets, generated using our method for $\xi = .395$ and .25, respectively.

4.1. Identification Risk Evaluation

Here, we examine some identification risks associated with our perturbed data sets. To present certain empirical correct match probabilities, we let τ and τ^* denote the number of matches for a unit in the original and perturbed data, respectively. Actually, τ and τ^* depend on the unit, but for simplicity we do not make that explicit in our notation. Then, the correct match probability of a unit in perturbed data is 0 if its category changed by post-randomization, and $1/\tau^*$ otherwise. Suppose that n_{ij} units have $\tau = i$ and $\tau^* = j$, and n^*_{ij} of those units have changed X category in perturbed data. As in NZY, for $i \geq 1, j \geq 1$, we calculate the empirical value of $P(CM|\tau = i, \tau^* = j)$ as $[0 \times n^*_{ij} + \frac{1}{j}(n_{ij} - n^*_{ij})]/n_{ij} = \frac{1}{j}(1 - \frac{n^*_{ij}}{n_{ji}})$.

Table 2 gives some empirical correct match probabilities based on the perturbed data set that we created using our method with $\xi = .25$. For a brief exposition, take $P(CM|\tau = 1, \tau^* = 1)$ as an example. We found 5,066 units that are singleton in the original data and have unique matches in the perturbed data set. Of those, only 969 units are matched correctly. Using these values, we calculate $P(CM|\tau = 1, \tau^* = 1) = 969/5066 = 0.1913$. In Table 2, all empirical conditional correct match probabilities are smaller than the target value 0.25, with the largest being 0.1913 for $\tau = 1, \tau^* = 1$. Also, $P(CM|\tau = i, \tau^* = j)$ decreases with both i and j.

	$\tau = 1$	$\tau = 2$	$\tau = 3$	$1 \leq \tau \leq 3$
$\tau^* = 1$	0.1913	0.1584	0.1416	0.1713
$\tau^* = 2$	0.1598	0.1430	0.1265	0.1458
$\tau^* = 3$	0.1370	0.1187	0.1116	0.1228
$\tau^* \geq 0$	0.1249	0.1131	0.1051	

Table 2: Empirical correct match probabilities for $\xi = 0.25$

Table 3 gives the empirical correct match probabilities for our method with $\xi = 0.395$.

For comparison, we give the corresponding values for the NZY method (as given in their Table 3) in parentheses. All probabilities for our method are smaller than 0.395, as expected, and the largest value is $P(CP|\tau = 1, \tau^* = 1) = 0.3447$. Here also the correct match probabilities decrease as τ and/or τ^* increase. For both $\xi = .395$ and .25, under our method the correct match probabilities for singletons ($\tau = 1$) are higher than those for doubletons ($\tau = 2$). In contrast, under the NZY method, correct match probabilities of doubletons are larger than those of singletons. As intruders would learn τ^* and not τ, the values in the last columns of Tables 2 and 3, which can be regarded as empirical values of $R_j(a)$ (with $a = \tau^*$), are practically most relevant. Note that those values are all much smaller than target ξ (0.25 and 0.395), which indicates that our method is fairly conservative.

	$\tau = 1$	$\tau = 2$	$1 \leq \tau \leq 2$
$\tau^* = 1$	0.3447 (0.2315)	0.2522 (0.3933)	0.3080 (0.2849)
$\tau^* = 2$	0.2592 (0.1961)	0.2178 (0.3477)	0.2405 (0.2827)
$\tau^* \geq 0$	0.2094 (0.1348)	0.1812 (0.3027)	

Table 3: Empirical correct match probabilities for $\xi = 0.395$.

4.2. Effects on Data Utility

To assess data utility, we shall examine the effects of our method on the distribution of race and some joint distributions. Table 4 gives the frequency distributions of race based on the original data and three perturbed data sets. The first column gives the nine race categories, some of which are abbreviated as follows: Amer Indian = American Indian alone, Alaska Native = Alaska Native alone, Am Ind & AK Native = American Indian and Alaska Native, and Hawaiian & PI = Native Hawaiian and other Pacific Islander. The second column gives the frequencies based on the original data. The frequencies based on

two perturbed data sets created using our method for $\xi = 0.395$ and 0.25, respectively, are given in the columns with heading 'Our.' The column 'NZY' gives the NZY perturbed frequencies for $\xi = 0.395$. The 'diff' columns give the differences between perturbed and true counts. The last column gives the standard error (SE) of the true counts under multinomial sampling. Specifically, for a category with true count t_i, the SE is calculated as $[n(t_i/n)(1 - t_i/n)]^{1/2} = [t_i(1 - t_i/n)]^{1/2}$, where n is the total sample size (59,033).

Race	True Count	$\xi = .395$				$\xi = .25$		Standard Error
		Our	diff	NZY	diff	Our	diff	
White	37201	37201	0	37201	0	37201	0	117.29
Black	15239	15239	0	15239	0	15239	0	106.33
Amer Indian	97	116	19	92	-5	121	24	9.84
Alaska Native	1	1	0	0	-1	4	3	1
Am Ind & AK Native	42	43	1	46	4	39	-3	6.48
Asian	3461	3419	-42	3445	-16	3311	-150	57.08
Hawaiian & PI	20	20	0	21	1	23	3	4.47
Some other race alone	1349	1340	-9	1337	-12	1406	57	36.31
Two or more races	1623	1654	31	1652	29	1689	66	39.73

Table 4: Perturbation effects on the distribution of race.

As we noted in Section 2, the specific data partitioning used in the example preserves race if it is White or Black. Consequently, in Table 4, the perturbed frequencies of those two categories do not differ from the true frequencies. Frequencies of the remaining categories may change due to data perturbation. In Table 4, the differences between perturbed and true frequencies are mostly quite small, especially in comparison to SE. For $\xi = 0.395$, the absolute differences for our method are fairly similar to those for the NZY method. Also, our differences increased in magnitude as ξ is reduced from 0.395 to 0.25, as one would expect intuitively.

Next, we examine the effects of our method on some joint distributions, following the approach and work of NZY. Specifically, we consider the same combinations of variables,

given in Table 5, and use the total variation distance (TVD) to measure the discrepancy between the estimates of a distribution from the original and perturbed data, respectively. The TVD between two discrete distributions $p(x)$ and $q(x)$ over a common sample space $\mathcal{X}$ is defined as

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

It is known that $TVD(p,q) = \sup_{A\subseteq\mathcal{X}} |p(A) - q(A)|$. Thus, $TVD(p,q)$ is a uniform upper bound for $|p(A)-q(A)|$. Consider a set of categorical variables and denote the original and perturbed frequencies of the ith cell by f_i and $\tilde{f}_i$, respectively. Then, the TVD between their estimated joint distributions, based on the original and perturbed data, is

$$TVD = \frac{1}{2}\sum_i |\frac{f_i}{n} - \frac{\tilde{f}_i}{n}| = \frac{1}{2n}\sum_i |f_i - \tilde{f}_i|, \tag{4.1}$$

where n is the sample size. Actually, (4.1) may also be used for a single variable. It can be seen that for race, the TVD values for our method with $\xi = .25$ and $\xi = .395$ and NZY with $\xi = .395$ are 0.00518, 0.00141 and 0.00115, respectively.

For examining joint distributions, we follow NZY and consider some combinations of the five key variables and two non-key variables, viz. class of workers and education level, which have nine and eight categories, respectively. For twelve joint distributions, Table 5 gives the TVD values for our method for $\xi = .395$ and .25 and the NZY method with $\xi = .395$. There, we use the following abbreviated variable names: mar = marital status, edu = level of education, and work = class of workers. As expected, Table 5 shows that our method's TVDs increase as ξ is reduced from .395 to .25; the average increment is about 48%. Table 5 also shows that for $\xi = .395$, our TVDs are larger than the corresponding

NZY values, by about 50% on average. Recall that the NZY method cannot guarantee (1.3) for $\xi \leq 1/3$. Actually, NZY recommended to use their method for $\xi \geq .35$. For such ξ, one should use the NZY method to better preserves data utility. Our proposed method should be used when $\xi < .35$, where the NZY methods fails to give the desired protection against identity disclosure.

	$\xi = .395$		$\xi = .25$		$\xi = .395$		$\xi = .25$
Variables	Our	NZY	Our	Variables	Our	NZY	Our
race, mar	0.0076	0.0028	0.0146	puma, work	0.0348	0.0198	0.0454
race, puma	0.0152	0.0013	0.0233	puma, edu	0.0483	0.0324	0.0649
race, edu	0.0094	0.0088	0.0123	sex, race, mar	0.0088	0.0060	0.0157
race, work	0.0046	0.0035	0.0046	sex, race, edu	0.0107	0.0093	0.0132
mar, edu	0.0135	0.0127	0.0231	mar, race, edu	0.0258	0.0218	0.0397
mar, work	0.0107	0.0070	0.0216	race, sex, work	0.0057	0.0039	0.0058

Table 5: TVD between original and perturbed joint distributions.

5. Discussion

In this paper, we present a variation of the NZY method, via a new post-randomization scheme, that can provide a more stringent identification risk control than the NZY method. The example in Section 4 shows that our method (i) is fairly conservative, as the empirical correct match probabilities in Tables 2 and 3 are much smaller than the nominal values and (ii) it affects the estimates of joint distributions more than the NZY method. Our method can be improved in two ways. First, using smaller PRBs via finer data partitioning is expected to enhance data utility. Second, one may use an α that is smaller than α_ξ to guarantee (1.1). In a given problem, we suggest to calculate the empirical correct match probabilities, as in Tables 2 and 3, for several values of α and thereby find a suitably small (or nearly optimal) value of α for assuring (1.1).

We have discussed our method mainly for $1/4 \leq \xi \leq 1/3$. But, we believe that it can guarantee (1.1) also for $\xi < 1/4$. It may be possible to establish this mathematically by proving that $R_1(1, \mathbf{t}) \geq R_1(a, \mathbf{t})$ for all $a \geq 2$ and $\mathbf{t}$. But, in view of the observations from our example and the fact that the set of cells needing protection changes in steps, as discussed in the second para of Section 3, we suggest to take the above mentioned experimental approach in practice to evaluate correct match probabilities and determine a suitable value of α.

Both NZY and we use highly structured TPMs, determined by a single design parameter. This is very convenient for bounding identification risks. But, more general TPMs are likely to better preserve data utility while providing desired risk control. Finding an optimal post-randomization scheme is an important problem for future investigation. We hope that our work will stimulate further research on the theory and applications of identification risk control in releasing microdata.

6. Appendix

Lemma 6.1. *Under any P_α, we have $R_1(2, \mathbf{t}) > R_1(3, \mathbf{t})$ for all $\mathbf{t}$.*

Proof. We shall prove that the inequality in (3.9) holds true for $a = 2$, i.e., $\Sigma_1 \big[\beta_1 \Sigma_2 +$

$3\Sigma_3\big] - 2\Sigma_2^2 > 0$. Using routine algebra and the fact that $\beta_1 > \beta_2$, we obtain:

$$
\begin{aligned}
\Sigma_1\big[\beta_1\Sigma_2 + 3\Sigma_3\big] - 2\Sigma_2^2 &= \big(r\beta_1 + \nu\beta_2\big)\Big[\beta_1\Big\{\binom{r}{2}\beta_1^2 + r\nu\beta_1\beta_2 + \binom{\nu}{2}\beta_2^2\Big\} + 3\Big\{\binom{r}{3}\beta_1^3 \\
&\quad + \binom{r}{2}\nu\beta_1^2\beta_2 + r\binom{\nu}{2}\beta_1\beta_2^2 + \binom{\nu}{3}\beta_2^3\Big\}\Big] - 2\Big[\binom{r}{2}\beta_1^2 + r\nu\beta_1\beta_2 + \binom{\nu}{2}\beta_2^2\Big]^2 \\
&= \big(r\beta_1 + \nu\beta_2\big)\Big[\frac{r(r-1)^2}{2}\beta_1^3 + \frac{r\nu(3r-1)}{2}\beta_1^2\beta_2 + \binom{\nu}{2}(3r+1)\beta_1\beta_2^2 \\
&\quad + 3\binom{\nu}{3}\beta_2^3\Big\}\Big] - 2\Big[\binom{r}{2}\beta_1^2 + r\nu\beta_1\beta_2 + \binom{\nu}{2}\beta_2^2\Big]^2 \\
&= \frac{r\nu(r+1)}{2}\beta_1^3\beta_2 + \frac{r\nu(2\nu-r-3)}{2}\beta_1^2\beta_2^2 + \frac{\nu(\nu-1)(\nu-2r)}{2}\beta_1\beta_2^3 - \frac{\nu^2(\nu-1)}{2}\beta_2^4 \\
&> \frac{r\nu(r+1)}{2}\beta_1^3\beta_2 + \frac{r\nu(2\nu-r-3)}{2}\beta_1^2\beta_2^2 + \Big[\frac{\nu(\nu-1)(\nu-2r)}{2} - \frac{\nu^2(\nu-1)}{2}\Big]\beta_1\beta_2^3 \\
&= \frac{r\nu(r+1)}{2}\beta_1^3\beta_2 + \frac{r\nu(2\nu-r-3)}{2}\beta_1^2\beta_2^2 - r\nu(\nu-1)\beta_1\beta_2^3 \\
&> \Big[\frac{r\nu(r+1)}{2} + \frac{r\nu(2\nu-r-3)}{2} - r\nu(\nu-1)\Big]\beta_1^2\beta_2^2 \\
&= 0,
\end{aligned}
$$

which proves the lemma. □

Acknowledgment. We sincerely thank Eric Slud for giving us some constructive suggestions, which helped to improve the presentation.

References

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

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

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

[4] Chen, G. and Keller-McNulty, S. (1998). Estimation of identification disclosure risk in microdata. *Journal of Official Statistics*, 14, 79-95.

[5] 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 (with discussion and rejoinder). *International Statistical Review*, 79, 160-199.

[6] 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, Ser. B*, 70, 21-30.

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

[8] Greenberg, B. V., and Zayatz, L. V. (1992). Strategies for measuring risk in public use microdata files. *Statistica Neerlandica*, 46, 33-48.

[9] 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. *Journal of Official Statistics*, 14, 463-478.

[10] 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.

[11] Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics*, 9, 313-313.

[12] Nayak, T.K. and Adeshiyan, S. A. (2016). On invariant post-randomization for statistical disclosure control. *International Statistical Review*, 84, 26-42.

[13] Nayak, T.K., Adeshiyan, S.A. and Zhang, C. (2016). A concise theory of randomized response techniques for privacy and confidentiality protection. In *Data Gathering, Analysis and Protection of Privacy through Randomized Response Techniques: Qualitative and Quantitative Human Traits*, Eds. A. Chaudhuri, T.C. Christofides and C.R. Rao, pp. 273-286. New York: Elsevier.

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

[15] Paass, G. (1988). Disclosure risk and disclosure avoidance for microdata. *Journal of Business & Economic Statistics*, 6, 487-500.

[16] Shlomo, N., and De Waal, T. (2008). Protection of micro-data subject to edit constraints against statistical disclosure. *Journal of Official Statistics*, 24, 229-253.

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

[18] Skinner, C. (2012). Statistical disclosure risk: Separating potential and harm (with discussion and rejoinder). *International Statistical Review*, 80, 349-381.

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

[20] 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, 711-731.

[21] 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.

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

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