Multiple and Probabilistic Swapping of Categorical Keys for Statistical Disclosure Limitation in Microdata

Fang Liu
Merck Research Laboratory, Rahway, NJ 07065

ABSTRACT

Statistical disclosure limitation (SDL) methods aim to protect the privacy of individual cases by statistically modifying raw data, and to release useful information for the purposes of research and decision-making. Multiple and probabilistic swapping of categorical keys (MaPS) is a Bayesian model-based, partial synthesis SDL approach. MaPS swaps the values of the key variables between probabilistically paired cases and releases multiple swapped data sets. MaPS offers an alternative to the model-free data swapping techniques with less distortion of raw information and yet better protection of data confidentiality. Simulation is used to demonstrate the implementation of MaPS and its advantages over simple data swapping techniques.

Key Words: statistical disclosure limitation, Bayesian modeling, swapping odds and probability

1 Introduction

With today's increasing collection and dissemination of information, the techniques for protecting data confidentiality are needed more than ever. Various statistical disclosure limitation (SDL) methods have been proposed for public use microdata over the last several decades. From a statistical viewpoint, these methods can be grouped into two categories: model-free and model-based approaches. Techniques in the former category are easy to implement in practice and provide some confidentiality protection for public use microdata. Unfortunately, most of them offer no rigorous control on the loss of raw information, which could result in limited statistical analysis, invalid inferences, or additional analysis burden (such as artificial missing data) for data users.

Better choices for SDL are model-based approaches, where data are synthesized via statistical modeling (non-, semi- or parametric) based on the information in the raw data. The model-based techniques can be further categorized into full synthesis and partial synthesis methods depending on the proportion of the synthetic part in the released data. In full synthesis, all the cases in the released data are pseudo cases in the sense that all the records are model-generated and do not necessarily correspond to any of those in the original sample. Rubin (1993) proposes to use the multiple imputation (MI) (Rubin,1987) technique to impute the census population and release pseudo samples from multiply imputed populations. Raghu *et. al.* (2003) develop the inferential theory for this MI-based synthesis approach. Reiter (2005a) presents an empirical study via this MI-based approach on the US Current Population Survey. Feinberg *et. al.* (1998) list some general steps of pseudo data generation via the statistical simulation process and suggest to build the information on various sources of survey error (such as editing, matching, nonresponse, etc.) into the simulation process, thus developing an integrated approach to the release and analysis of survey data. In partial synthesis, only part of the released data is synthesized while the remaining is the same as the corresponding part in the raw data set. Little (1993) lists three possible scenarios for partial synthesis – synthesizing all variables of some records, synthesizing some variables for all records, and synthesizing some variables on some records. In the second and third scenarios, noises can be added only to variables with sensitive information (such as HIV status and household income) since they are usually of central interest to data intruders; Alternatively, perturbation can be performed on key variables/keys (e.g. gender, race) to put some obstacle in the way to disclosure since the intruders usually use keys to identify records. Reiter (2003, 2005b) develops inferential methods for partial synthetic data including point estimate, variance estimate, and procedures for Wald and likelihood ratio tests. Little & Liu (2002) propose selected multiple imputation of keys (SMIKe) for only imputing values of key variables of selective cases in a Bayesian framework. An application of SMIKe can be found in Little *et. al.* (2004), which also provides a comprehensive discussion on MI-based synthesis methods.

In this article, we propose a Bayesian model-based SDL approach called multiple and probabilistic swapping of categorical keys (MaPS). MaPS is for categorical keys and aimed at preserving the original cell counts from the cross-tabulation of the categorical keys. In the case of continuous keys or a mixture of categorical and continuous keys, SMIKe or other synthesis techniques are preferred. MaPS can be viewed a stochastic version of the model-free data swapping techniques to swap the values of categorical keys between paired cases in the framework of Bayesian modeling. Unlike simple data swapping where relationship between swapped and unswapped variables could get distorted, MaPS utilizes statistical modeling to preserve the relationship as well as other raw information.

The article is organized as follows: the methodology of MaPS and the steps to apply MaPS are described in Section 2. Implementation of MaPS in the case with normally distributed nonkey variables is demonstrated in Section 3. Section 4 presents the inferential methods on MaPS-modified data and discusses disclosure risk issues in MaPS. A simulation study is presented in Section 5. The article ends with some discussion in Section 6.

2 The Methodology of MaPS

Suppose in a data set with n cases, there are K key cells formed by the categorical key variables $\mathbf{X}$, and $\mathbf{y}$ represents p nonkey variables. A prespecified sensitivity threshold s, divides the total sample into $n(\text{sen})$ sensitive cases and $(n - n(\text{sen}))$ nonsensitive ones. A case is defined as sensitive if it belongs to a key cell with size $< s$.

The joint distribution of $\mathbf{X}$ and $\mathbf{y}$ given parameter $\boldsymbol{\theta}$ is denoted by $p(\mathbf{X}, \mathbf{y}|\boldsymbol{\theta})$. By choosing a prior distribution on $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$, we can derive the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})$ of $\boldsymbol{\theta}$ and the posterior predictive distribution $p(\tilde{\mathbf{X}}|\mathbf{X}, \mathbf{y})$ for $\mathbf{X}$. Since $p(\tilde{\mathbf{X}}|\mathbf{X}, \mathbf{y})$ can be rewritten as

$$\int p(\tilde{\mathbf{X}}|\boldsymbol{\theta}, \mathbf{y})\, p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})\, d\boldsymbol{\theta},$$

this indicates that to obtain multiple draws of $\tilde{\mathbf{X}}$ from $p(\tilde{\mathbf{X}}|\mathbf{X}, \mathbf{y})$, we can first draw $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})$, and given the drawn $\boldsymbol{\theta}, \tilde{\mathbf{X}}$ can be drawn from $p(\tilde{\mathbf{X}}|\boldsymbol{\theta}, \mathbf{y})$. $p(\tilde{\mathbf{X}}|\mathbf{X}, \mathbf{y})$ is basic in MaPS. Specifically, suppose cases i and j are two cases with the raw keys $\mathbf{X}_i$ and $\mathbf{X}_j$, respectively. The probabilities of nonswapping and swapping between cases i and j, based on $p(\tilde{\mathbf{X}}|\mathbf{X}, \mathbf{y})$, can be expressed as

$$\begin{aligned} & \Pr(\text{nonswapping}) \\ = \; & p(\tilde{\mathbf{X}}_i = \mathbf{X}_i|\mathbf{X}, \mathbf{y}) \cdot p(\tilde{\mathbf{X}}_j = \mathbf{X}_j|\mathbf{X}, \mathbf{y}) \cdot \delta_{(ij)} \\ & \Pr(\text{swapping}) \\ = \; & p(\tilde{\mathbf{X}}_i = \mathbf{X}_j|\mathbf{X}, \mathbf{y}) \cdot p(\tilde{\mathbf{X}}_j = \mathbf{X}_i|\mathbf{X}, \mathbf{y}) \cdot \delta_{(ij)}, \end{aligned}$$

where $\delta_{(ij)}$ comprises the posterior predictive probabilities associated with other cases than i and j. The swapping odds between cases i and j is given by

$$\begin{aligned} O_{ij} &= \frac{\Pr(\text{swapping})}{\Pr(\text{nonswapping})} \\ &= \frac{p(\tilde{\mathbf{X}}_i = \mathbf{X}_j|\mathbf{X}, \mathbf{y}) \cdot p(\tilde{\mathbf{X}}_j = \mathbf{X}_i|\mathbf{X}, \mathbf{y})}{p(\tilde{\mathbf{X}}_i = \mathbf{X}_i|\mathbf{X}, \mathbf{y}) \cdot p(\tilde{\mathbf{X}}_j = \mathbf{X}_j|\mathbf{X}, \mathbf{y})} \\ &= \left(\frac{p(\tilde{\mathbf{X}}_i = \mathbf{X}_j|\mathbf{X}, \mathbf{y})}{p(\tilde{\mathbf{X}}_i = \mathbf{X}_i|\mathbf{X}, \mathbf{y})}\right) \cdot \left(\frac{p(\tilde{\mathbf{X}}_j = \mathbf{X}_i|\mathbf{X}, \mathbf{y})}{p(\tilde{\mathbf{X}}_j = \mathbf{X}_j|\mathbf{X}, \mathbf{y})}\right) \\ &= \frac{o_i(i,j)}{o_j(i,j)}, \end{aligned}$$

where $o_i(i,j)$ is the posterior odds of case i falling in cell $\mathbf{X}_j$ vs. in cell $\mathbf{X}_i$; $o_j(i,j)$ is the same odds but for case j. Therefore, the swapping odds O_{ij} between cases i and j can be envisioned as these two cases competing for being in one cell against another. If O_{ij} is too small, or equivalently, $o_i(i,j) \ll o_j(i,j)$, case i is unlikely to be swapped with j and the goal of protection will not be achieved. If $o_i(i,j) \gg o_j(i,j)$, then i has a high probability of swapping with j, which is good for protection, but cases within a cell tend to get homogenized in unswapped variables $\mathbf{y}$ after swapping. That is, the distribution of $\mathbf{y}$ within a cell tends to be over-smoothed during the model-based swapping process, causing inconsistent inferences between the final swapped data and the raw data. If $o_i(i,j) \approx o_j(i,j)$, cases i and j are similar in their odds of falling in cell j against cell i. This equivalent odds (E-odds) rule not only helps to retain the original relationship between $\mathbf{X}$ and $\mathbf{y}$, but also leaves space for introducing noises into the swapping process. To bring this E-odds rule into play in the the construction of swapping probability, we first define a weight function w_{ij} based on O_{ij}:

$$w_{ij} = exp^{-|log(O_{ij})|},$$

which possesses the following properties: (1) $w_{ij} \in (0, 1]$ and reaches the maximum 1 when $O_{ij} = 1$; (2) the further O_{ij} deviates from 1, the smaller w_{ij} is. The swapping probability p_{ij} between cases i and j could be defined as the normalized w_{ij}: $p_{ij} = w_{ij}/(\sum_{j'} w_{ij'})$, where j' is any case available to be swapped with case i. However, there is a potential problem with this formulation: unless the number of the available cases j' is very small or a small proportion of these available cases carry most of the weight, the role of the E-odds rule could be suppressed by the noises introduced into the swapping process. A pre-specified cutpoint w_0 can used to solve this problem by letting $w^*_{ij} = 0$ when $w_{ij} < w_0$ and $w^*_{ij} = w_{ij}$ when $w_{ij} \geq w_0$. Increasing w_0 results in larger number of cases getting null weight, less noises in the swapping process, and consequently less protection of confidentiality but more preservation of raw information. Therefore, w_0 is actually a turning parameter that can be used to adjust the balance between the two ends of SDL. The final swapping probability p_{ij} between cases i and j is presented as

$$p_{ij} = \begin{cases} \frac{w^*_{ij}}{(1+\sum_{k\in\mathcal{A}_i} w^*_{ik})}, & \text{if } i \neq j\,, \\ \frac{1}{(1+\sum_{k\in\mathcal{A}_i} w^*_{ik}}), & \text{if } i = j\,, \end{cases}$$

where $\mathcal{A}_i$ is a set comprising all available cases to be swapped with case i. MaPS swaps paired cases according to p_{ij} sequentially rather than simultaneously. This sequential swapping consists of [$\leq n$(sen)] swapping cycles. In each cycle, a sensitive case is randomly picked, swapping odds and swapping probabilities of this case being swapped with other cases are calculated, and swapping of the key values between two cases is executed given these swapping probabilities. During the whole swapping process, each case is allowed to be swapped only once. Therefore, once two cases are virtually swapped with each other, they will remain in the post-swapping cells and will not be considered for future swapping.

In summary, listed below are the steps of MaPS for a microdata set: (1) Specify s and w_0. (2) Choose an appropriate model for the data and a proper prior for the parameters in the model. (3) Obtain the posterior distribution for the parameters and posterior predictive distribution for the key variables. (4) Draw parameters from their posterior distribution and key variables from their posterior predictive distribution. (5) Run the swapping cycle on each sensitive case based on the drawn values in step (4), including calculation of swapping odds and probabilities and execution of swapping. (6) Repeat steps (4) and (5) independently for D times. (7) Steps (2)-(6) can be repeated for various choices of s and w_0 to tune the tradeoff between preservation of raw information and protection of data confidentiality. D swapped data sets can be released once a satisfactory tradeoff is reached.

3 MaPS with Normally Distributed $\mathbf{y}$

This section presents an exemplifying implementation of MaPS with normally distributed $\mathbf{y}$. In the case of categorical $\mathbf{y}$ or a mixture of continuous and categorical $\mathbf{y}$, MaPS can be carried out in a similar manner with an appropriate choice of model for the data.

When $\mathbf{y}$ is normal or approximately so and constant variance of $\mathbf{y}$ is assumed across the K key cells, the general location (GL) model (Olkin & Tate, 1961) is a convenient choice for modeling this kind of data. For notation convenience, we use a one-dimensional x_i to represent the values of $\mathbf{X}_i$ for case i. That is, if a case falls within key cell k $(= 1, \ldots, K)$, then $x_i = k$. The GL model is defined in terms of the marginal distribution of x and the conditional distribution of $\mathbf{y}$ given x:

$$\begin{aligned} p(x_i = k) &= \pi_k, \text{ where } k = 1, \ldots, K; \sum_k \pi_k = 1 \\ p(\mathbf{y}_i|x_i) &\sim N_p(\boldsymbol{\mu}_{x_i}, \Sigma) \text{ for } i = 1, \ldots, n. \end{aligned}$$

The n cases compose an i.i.d. sample from the model. p is the dimension of $\mathbf{y}$ (If $\mathbf{y}$ is nonnormally distributed but still continuous, a transformation on $\mathbf{y}$ or the extended general location model proposed by Liu and Rubin (1998) could be tried). Denote the parameters in the model by $\boldsymbol{\theta} = \{\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma\}$, the log-likelihood for the GL model is

$$\begin{aligned} L(\boldsymbol{\theta}) &= -\frac{1}{2}|\Sigma|^n + \sum_{k=1}^{K} n_k log(\pi_k) - \\ &\quad \frac{1}{2}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(\mathbf{y}_i - \boldsymbol{\mu}_k)^T\Sigma^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k), \end{aligned}$$

where n_k is the size of cell k. If Jeffreys' prior

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \pi_k^{-\frac{1}{2}}|\Sigma|^{-\frac{p+1}{2}}$$

is used, then the posterior distributions of $\boldsymbol{\theta}$ is

$$\begin{aligned} [\boldsymbol{\pi}|x, y] &\sim Dirichlet(n_1 + \frac{1}{2}, \ldots, n_K + \frac{1}{2}) \\ [\Sigma|\boldsymbol{\pi}, x, \mathbf{y}] &\sim Inv - Wishart(S, n - K) \\ [\boldsymbol{\mu}_k|\boldsymbol{\pi}, \Sigma, x, \mathbf{y}] &\sim N_p(\bar{\mathbf{y}}_k, \Sigma/n_k) \text{ for } k = 1, \ldots, K, \end{aligned}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T$, S is the pooled sample variance matrix of n cases, $\boldsymbol{\mu}_k = (\mu_{1k}, \ldots, \mu_{pk})^T$, and $\bar{\mathbf{y}}_k$ is the sample mean of $\mathbf{y}$ in cell k. The full conditional posterior predictive distribution of x_i is given by

$$\begin{aligned} p(\tilde{x}_i = k|\boldsymbol{\theta}, x, \mathbf{y}) &= \frac{\pi_k exp(\phi_{ik})}{\sum_{k'=1}^{K} \pi_{k'} exp(\phi_{ik'})} \\ &\quad \text{for } k = 1, \ldots, K, \end{aligned}$$

where $\phi_{ik} = \mathbf{y}_i^T\Sigma^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\Sigma^{-1}\boldsymbol{\mu}_k$ (similar for $\phi_{ik'}$). Thus, for cases $i = \{i_1, i_2\}$ with raw keys as $k =$

$\{k_1, k_2\}$, we can write

$$p(\tilde{x}_{i_1} = k_1|x, \mathbf{y}) = \pi_{k_1}(\mathbf{y}_{i_1}) = \frac{\pi_{k_1} exp(\phi_{i_1 k_1})}{\sum_{k=1}^{K} \pi_k exp(\phi_{i_1 k})}$$
$$p(\tilde{x}_{i_1} = k_2|x, \mathbf{y}) = \pi_{k_2}(\mathbf{y}_{i_1}) = \frac{\pi_{k_2} exp(\phi_{i_1 k_2})}{\sum_{k=1}^{K} \pi_k exp(\phi_{i_1 k})}$$
$$p(\tilde{x}_{i_2} = k_1|x, \mathbf{y}) = \pi_{k_1}(\mathbf{y}_{i_2}) = \frac{\pi_{k_1} exp(\phi_{i_2 k_1})}{\sum_{k=1}^{K} \pi_k exp(\phi_{i_2 k})}$$
$$p(\tilde{x}_{i_2} = k_2|x, \mathbf{y}) = \pi_{k_2}(\mathbf{y}_{i_2}) = \frac{\pi_{k_2} exp(\phi_{i_2 k_2})}{\sum_{k=1}^{K} \pi_k exp(\phi_{i_2 k})},$$

and the swapping odds between cases i_1 and i_2 is

$$\begin{aligned} O(i_1, i_2) &= \frac{\pi_{k_1}(\mathbf{y}_{i_2}) \cdot \pi_{k_2}(\mathbf{y}_{i_1})}{\pi_{k_1}(\mathbf{y}_{i_1}) \cdot \pi_{k_2}(\mathbf{y}_{i_2})} \\ &= exp\{-(\mathbf{y}_{i_1} - \mathbf{y}_{i_2})^T \Sigma^{-1} (\boldsymbol{\mu}_{k_1} - \boldsymbol{\mu}_{k_2})\}. \end{aligned}$$

Suppose i_1 is the sensitive case in a swapping cycle, $O(i_1, i_2)$ will be calculated for each case $i_2 \in \mathcal{A}_{i_1}$, as well as $w_{i_1 i_2}$ and $p_{i_1 i_2}$ given a prespecified w_0. Note the swapping odds with normal $\mathbf{y}$ does not depend on the marginal distribution of $\mathbf{X}$ and its associated parameters $\boldsymbol{\pi}$. This is generally true since $p(\mathbf{X}|\mathbf{y}, \boldsymbol{\theta})$ is proportional to $p(\mathbf{X}|\boldsymbol{\pi}) \cdot p(\mathbf{y}|\mathbf{X}, \boldsymbol{\mu}, \Sigma)$, where the former is cancelled out in the swapping odds calculation.

4 Statistical Inferences and Identification Risk

4.1 Statistical Inferences

For inferences based on D multiple independently swapped data sets via MaPS, we use the inferences methods from Reiter (2003) for partially synthetic data. Suppose γ is a parameter of interest, from each of the swapped data set d, the estimate $\hat{\gamma}_d$ and its variance estimates $\hat{V}_d$ are obtained, then the final estimate for γ is

$$\bar{\gamma} = \sum_{d=1}^{D} \hat{\gamma}_d / D$$

and the variance of $\bar{\gamma}$ is estimated by T

$$\begin{aligned} T &= W + \frac{1}{D} B, \text{ where} \\ W &= \sum_{d=1}^{D} \hat{V}_d / D, \; B = \sum_{d=1}^{D} (\hat{\gamma}_d - \bar{\gamma})^2 / (D-1). \end{aligned}$$

W and B are respectively called the within and between variance of $\bar{\gamma}$. The implementation of the cutpoint parameter w_0 might cause some bias in $\bar{\gamma}$. As shown by our simulation, the bias is mild across a moderate range of w_0. The variance estimate T seems to work well for MaPS in our simulation. More rigorous development on the inference methods for MaPS is in process.

4.2 Identification Risk

Researchers in the SDL area have been investigating different approaches for a sensible and unified measure on disclosure risk (see Bethlehem *et.al* (1990); Chen *et.al* (1999); Duncan & Lambert (1986); Fienberg et. al. (1997); Fienberg & Markov (1998); Reiter (2005c); Samules (1998); Skinner & Holmes (1998)). Some propose to estimate the number of population unique records or the probability of a case being population unique given the sample data; some use record linkage software to investigate the cases with potentially the highest identification risk; other suggest to model the behavior of intruders to obtain probabilities of identification for sampled units. Without doubt the task of quantifying identification risk is extremely demanding. There are zillions of data intruders who possess different amount of prior information about the released data; and each intruder could have a unique way of stealing information from the released data by using various tools and techniques. Though there is some work on incorporating sources of the uncertainty in a Bayesian framework when formulating identification probabilities, these methods lack practical evidence and remain skeptical due to many strong assumptions made about the behavior of intruders. In the context of MaPS, to measure disclosure is even harder because of the multiplicity feature of the released data sets.

It is not attempted here to provide an explicit measure on identification risk in MaPS-modified data, rather a brief description is given on the aspects where MaPS tries to impose restrictions on record identification. First, MaPS offers two tuning parameters s and w_0 to control the amount of noises that MaPS introduces into the raw data. The larger the s is or the smaller the w_0 is, the more perturbation there will be in the released data. Secondly, MaPS releases a small number D of multiple data sets. The variation of the key information of a case across the multiple data sets may confuse and intimidate intruders more than a single data set does.

5 A Simulation Study

This section presents a simple simulation study where $\mathbf{y}$ has a univariate normal distribution. 1000 data sets

Table 1: Inferences from the Raw Data and the Modified Data via MaPS

				MaPS											
	Raw			$w_0 = 0.9$			$w_0 = 0.8$			$w_0 = 0.7$			$w_0 = 0.6$		
Parameter	Bias	RMSE	CP	Bias	RMSE	CP	Bias	RMSE	CP	Bias	RMSE	CP	Bias	RMSE	CP
β_0	0.001	0.207	95.0	0.014	0.209	94.9	0.041	0.215	94.4	0.068	0.225	94.2	0.098	0.223	94.9
β_1	0.003	0.410	95.4	-0.012	0.415	95.1	-0.046	0.426	95.1	-0.082	0.440	94.3	-0.133	0.459	94.6
β_2	-0.006	0.221	95.2	-0.019	0.223	95.1	-0.045	0.229	94.6	-0.072	0.240	94.1	-0.100	0.238	94.8
β_3	0.001	0.232	94.6	-0.014	0.236	94.8	-0.045	0.244	94.5	-0.076	0.225	94.1	-0.111	0.254	94.7
σ^2	0.003	0.021	95.6	0.003	0.021	95.5	0.004	0.021	95.4	0.006	0.021	95.5	0.010	0.022	95.0

Table 2: Inferences from the RDS/DDS-Modified Data

	Data Swapping					
	RDS			DDS		
Parameter	Bias	RMSE	CP	Bias	RMSE	CP
β_0	0.389	0.453	84.8	1.162	1.662	37.3
β_1	-0.990	1.633	66.8	-3.051	10.048	4.2
β_2	-0.544	0.644	76.4	-1.158	1.709	37.9
β_3	-0.137	0.354	93.4	-1.049	1.448	46.8
σ^2	0.310	0.140	49.6	0.329	0.154	45.9

Table 3: A sample from a MaPP-modified Data Set

MaPS-modified x ($w_0 = 0.9$)										y	raw x
3	3	1	3	3	1	4	4	4	3	1.38	1
4	1	1	1	4	4	4	4	4	4	-1.32	1
4	1	1	1	1	1	1	1	1	1	-0.62	1
4	4	1	4	1	4	4	3	4	3	1.60	1
4	2	2	2	3	3	3	2	3	4	1.50	2
3	3	3	3	3	2	3	4	3	4	1.81	2
1	1	1	4	4	1	1	4	4	1	-1.40	4

were simulated from the following model:

$$\begin{aligned} p(x_i = k) &= \pi_k, \text{ for } k = 1, 2, 3, 4; \\ p(y_i|x_i) &\sim N(\mu_{x_i}, \sigma^2) \text{ for } i = 1, \ldots, 100, \end{aligned}$$

where $\mu_{x_i} \in \{\mu_1, \mu_2, \mu_3, \mu_4\}$. Let $\pi_1 = \pi_2 = 0.0625$, and $\pi_3 = 0.5, \pi_4 = 0.375$; $\mu_1 = 0, \mu_2 = 3, \mu_3 = 1.5$, and $\mu_4 = 0.5$; $\sigma^2 = 1$. s was set at 10. The posterior predictive distribution of x_i, the swapping odds and probability can all be obtained from the formulas presented in Section 3. To investigate the effect of w_0, we chose w_0 at $\{0.6, 0.7, 0.8, 0.9\}$ respectively. The model-free data swapping technique was applied to the simulated data sets to provide a comparison to MaPS. Usually data swapping refers to “random data swapping” (RDS) as we call it. In RDS, a certain percentage (p%) of total cases are randomly picked to be swapped with other randomly selected cases. We set p% to be the percentage of sensitive cases in each data set in our simulation. Obviously, RDS leaves some sensitive cases unprotected. On the other hand, RDS is model-free without taking into account the statistical relationship between the swapped and unswapped variables during the swapping process. Ignorance of the relationship consequently leads to possible invalid inferences. Therefore, RDS could perform poorly at both ends of the trade-off. To improve over RDS from the perspective of protection, we also applied “deterministic data swapping” (DDS) where each the sensitive case is bound for swapping with another randomly picked case.

For the analysis model, we fitted a simple linear regression model on y:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2).$$

x_1, x_2, x_3 are the dummy variables created from x. The parameters of interest are $\boldsymbol{\theta} = \{\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2\}$. Table 1 lists a side-by-side comparison between the inferences from the raw data and the MaPS-modified data in terms of bias of the estimate $\hat{\boldsymbol{\theta}}$, root mean square error (RMSE) of $\hat{\boldsymbol{\theta}}$, and the coverage probability (CP) of the nominal 95% confidence interval for each parameter. When $w_0 = 0.9$, the inferences based on MaPS-modified data are comparable with these from the raw data. As w_0 goes down, more noises are introduced into the swapping process, bias and RMSE go up, but CP stays around the nominal level of 95%. As mentioned in Section 4, MaPS is not perfect in inferences. In this simulation, w_0 can be chosen to make the bias small enough to be neglected. The advantages of MaPS over the simple data swapping technique are obvious in Table 2. The inferences based on the RDS- and DDS- modified data are far away from being valid. Especially in DDS, the full protection comes at a high cost in inferences. Table 3 presented some cases from a MaPS-modified simulated data set to show the difficulties that intruders may face to identify records.

6 Discussions

This article has presented a new model-based SDL method – MaPS for public use microdata with categorical key variables. MaPS swaps the values of key variables in probabilistically matched cases via Bayesian modeling. After swapping, the original key cell structure remains unchanged. The well-controlled swapping process in MaPS helps protection and inferences in the swapped data set to reach a satisfactory level, as demonstrated by the simulation study. The increasing popularity of Bayesian methods and the much advanced computation tools (such as MCMC algorithms) make the implementation of MaPS more feasible in practice. Yet more work needs to be done before MaPS can be put into real practice. This includes more rigorous justification of the inferences in MaPS and possible quantification of the identification risk in multiple swapped data sets.

MaPS improves over the model-free data swapping approach in both preservation of raw information and protection of data confidentiality, and provides a statistically stricter control over the tradeoff between these two aspects. The simulation presented in this article should ring alarm bell for the SDL practitioners who are attracted to the simple data swapping technique due to its easy implementation. Though it is well understood in the SDL community that the most crucial element in SDL is the trade-off between confidentiality protection and preservation of raw data, yet model-free SDL techniques take a large portion of the SDL market with limited well-justified statistical inference methods or no satisfactory protection. We hope model-based approaches could gain more popularity in the future since they are based on sound statistical theory, offer far more choices than model-free approaches, and help data distributors to better achieve the two-fold objectives of SDL.

References

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

Chen, Guang and Keller-McNulty Sallie (1999), "Estimation of Identification Disclosure Risk in Microdata," *Journal of Official Statistics*, **14**, 79-95.

Duncan, G.T. and Lambert, D. (1986), "Disclosure-limited Data Dissemination," *Journal of the American Statistical Association*, **81**, 10-18.

Fienberg, S.E. and Markov, U.E. (1998), "Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data," *Journal of Official Statistics*, **14**, 385-397.

Fienberg, S.E. Markov, U.E. and Sanil, A.P. (1997), " A Bayesian Approach to Data Disclosure: An optimal Intruder Behavior for Continuous Data," *Journal of Official Statistics*, **13**, 75-89.

Fienberg, S.E. Markov, U.E., and Steele R. J. (1998) "Disclosure Limitation Using Perturbation and Related Methods for Categorical Data", *Journal of Official Statistics*, **14**, 485-502.

Little, R.J.A. (1993), "Statistical Analysis of Masked Data," *Journal of Official Statistics*, **9(2)**, 407-426.

Little, J.A.R., Liu, F. and Raghunathan T. (2004) "Statistical Disclosure Techniques Based on Multiple Imputation," Chapter II.13 of *"Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives : An essential journey with Donald Rubin's statistical family"* edited by Andrew Gelman and Xiao-Li Meng, John Wiley & Sons.

Little, R.J.A. and Liu, F. (2002), "Selective Multiple Imputation for Statistical Disclosure Control in Microdata," *Proceedings of American Statistical Association, Section on Survey Research Methodology, August 2002, New York.*

Liu, C.H. and Rubin, D.B. (1998), "Ellipsoidally Symmetric Extensions of the General Location Model for Mixed Categorical and Continuous Data," *Biometrika*, **85(3)**, 673-688.

Olkin, I. and Tate, R.F. (1961),"Multivariate Correlation Models with Mixture Discrete and Continuous Variables," *Annals of Mathematical Statistics*, **32**, 448-465.

Raghunathan, T.E., Reiter, J.P. and Rubin, D.B. (2002), "Multiple Imputation for Statistical Disclosure Limitation," *Journal of Official Statistics*, **19**, 1-16.

Reiter, J.P. (2003), "Inference for partially synthetic, public use microdata sets," *Survey Methodology*, **29**, 181-188.

Reiter, J.P. (2005a), "Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study," *Journal of Royal Statistical Society, Series A*, **168**, 185-205.

Reiter, J.P. (2005b), "Significance tests for multi-component estimands from multiply-imputed, synthetic microdata," *Journal of Statistical Planning and Inference*, **131**, 365 - 377.

Reiter, J.P. (2005c),"Estimating Probabilities of Identification for Microdata," *Journal of American Statistical Association*

Rubin, D.B.(1987), *Multiple Imputation for Nonresponse in Survey*, New York: John Wiley and Sons.

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

Samuels, S.M.(1998), "A Bayesian, Species-Sampling-Inspired Approach to the Unique Problem in Microdata Disclosure Risk Assessment," *Journal of Official Statistics*, **14(4)**, 373-383.

Skinner, C.J. and Holmes, D.J.(1998), "Estimating the Re-identification Risk per Record in Microdata," *Journal of Official Statistics*, **14(4)**, 361-372.