Privacy protection from sampling and perturbation in survey microdata

Privacy Protection from Sampling and Perturbation in Survey Microdata

Natalie Shlomo* and Chris Skinner†

Abstract

Statistical agencies release microdata from social surveys as public-use files after applying statistical disclosure limitation (SDL) techniques. Disclosure risk is typically assessed in terms of identification risk, where it is supposed that small counts on cross-classified identifying key variables, i.e. a key, could be used to make an identification and confidential information may be learnt. In this paper we explore the application of definitions of privacy from the computer science literature to the same problem, with a focus on sampling and a form of perturbation which can be represented as misclassification. We consider two privacy definitions: differential privacy and probabilistic differential privacy. Chaudhuri and Mishra (2006) have shown that sampling does not guarantee differential privacy, but that, under certain conditions, it may ensure probabilistic differential privacy. We discuss these definitions and conditions in the context of survey microdata. We then extend this discussion to the case of perturbation. We show that differential privacy can be ensured if and only if the perturbation employs a misclassification matrix with no zero entries. We also show that probabilistic differential privacy is a viable alternative to differential privacy when there are zeros in the misclassification matrix. We discuss some common examples of SDL methods where in some cases zeros may be prevalent in the misclassification matrix.

Keywords: Identification Disclosure, Attribute Disclosure, Differential Privacy, Misclassification.

1 Introduction

Statistical agencies release microdata from social surveys, such as a labour force survey or a survey of incomes, where the units of investigation (households or individuals) have small inclusion probabilities. Provisions for releasing these microdata range from public-use files where the microdata is heavily protected against disclosure risk, microdata-under-contract and special licensed data typically delivered through data archives. In addition, many statistical agencies have facilities for visiting researchers to access unprotected microdata in a safe setting. Microdata from business surveys are generally not released because of their disclosive nature arising from high sampling fractions and skewed distributions. Other types of microdata are also not released in their original form, such as data from a population census. These datasets are typically protected through tabulation and high level aggregation which are released in the form of tables. Alternatively, some statistical agencies have taken the approach of producing synthetically generated multiple datasets of the microdata which retain some of the analytical properties of the original microdata (Rubin, 1993; Reiter, 2005a).

*Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton SO17 1BJ United Kingdom. Email: N.Shlomo@soton.ac.uk

†Department of Statistics, London School of Economics and Political Science, Houghton Street, London WC2A 2AE United Kingdom. Email: C.J.Skinner@lse.ac.uk

We consider here microdata from social surveys released as public-use files. In order to preserve the privacy and confidentiality of individuals responding to social surveys, statistical agencies assess disclosure risk and if required choose appropriate statistical disclosure limitation (SDL) methods to apply to the data. Disclosure risk is usually assessed in terms of the potential for re-identifying statistical units. SDL methods perturb, modify, or summarize the data in order to prevent re-identification by a potential attacker. Higher levels of protection through SDL methods however impact negatively on the utility and quality of the data. The SDL decision problem therefore is based on finding the optimal balance between managing disclosure risk to tolerable thresholds depending on the mode for accessing the data and ensuring high utility in the data.

SDL techniques for microdata include perturbative methods which alter the data and non-perturbative methods which limit the amount of information released in the microdata. Examples of non-perturbative SDL techniques are global recoding, suppression and subsampling (see Willenborg and De Waal, 2001). These methods are the most common for the protection of microdata arising from social surveys. Perturbative methods might be used, either for all records in the microdata or for only those deemed to be at high risk. Perturbative methods for continuous varaibles include adding random noise (Fuller, 1993; Yancey et al., 2002), micro-aggregation (replacing values with their average within groups of records) (Defays and Nanopoulos, 1992), random rounding, and rank swapping (swapping values between pairs of records within small groups) (Dalenius and Reiss, 1982; Fienberg and McIntyre, 2005). Perturbative methods for categorical variables include record swapping (typically swapping geography variables) and post-randomization probability mechanism (PRAM) where categories of variables are changed or not changed according to a prescribed probability matrix and a stochastic selection process (Gouweleeuw et al., 1998). For more information on these methods see also: Willenborg and De Waal (2001), Gomatam and Karr (2003), Domingo-Ferrer et al. (2001), Winkler (2004) and references therein.

In this paper we focus on the assessment of disclosure risk. We explore how definitions of privacy introduced in the computer science literature (Dinur and Nissim, 2003; Dwork et al., 2006) may be applied to the case of survey microdata and discuss their use in comparison to some traditional approaches in the SDL literature.

The traditional SDL literature emphasizes the distinction between an *identifying* or *key variable*, the value of which an adversary is assumed to know (perhaps from public sources) for a target unit, and a sensitive variable, the value of which an adversary wishes to learn for the target unit. It is assumed that in any released microdata, directly identifying variables, such as name, address or identification numbers, are removed. Disclosure risk typically arises when small counts on cross-classified indirect identifying key variables (such as: age, sex, place of residence, marital status, occupation, etc.) can be used to identify an individual and confidential information on a sensitive variable may be learnt. Identifying variables are typically categorical since statistical agencies will often coarsen the data before its release. Therefore, even a variable such as age will often be grouped into categories. Sensitive variables can be continuous (e.g., income) or categorical (e.g., health status).

In contrast, there is usually no distinction between key variables and sensitive variables in the computer science literature on privacy. A 'worst case' scenario is allowed for, in which the adversary has complete information about all units in the database other than the target unit. Definitions from both literatures will be discussed in Section 2. The main contributions of the paper are in Sections 3 and 4 where we discuss the application of computer science definitions of privacy in two common settings, one where sampling takes place and one where a perturbative SDL method is employed which can be represented as misclassification. Our discussion of sampling draws on the work of Chaudhuri and Mishra (2006) who obtained some fundamental results on how sampling may or may not preserve

privacy. Our discussion of misclassification refers to similar privacy notions and contrasts cases when there may or may not be zeros in the misclassification matrix. We illustrate these situations with some commonly used SDL methods. We conclude with a discussion in Section 6.

2 Defining Disclosure Risk

In the statistical literature, two broad notions of disclosure risk are used: *identification disclosure*, which refers to the possibility that an adversary can link a microdata record to a known unit in the population, and *attribute* (or *inferential*) *disclosure*, which refers to the possibility that an adversary can learn new information about a target unit in the population (Duncan and Lambert, 1989; Skinner, 1992). The first notion is particularly relevant to survey microdata, since it is often referred to in relevant legislation or professional codes of practice. The fact that identification disclosure does not refer to any particular survey variable also has practical advantages in social surveys where there may be a large number of survey variables. The notion of *differential privacy* in the computer science literature is most closely related to the concept of attribute (inferential) disclosure, by referring to what new information an adversary could learn about a target unit. We now discuss these different notions in more detail.

2.1 Identification Risk

We suppose that an adversary knows the values of some *key variables* for a target unit and seeks to use these values to link the unit to a record in the microdata, which contains values of these variables, after SDL has been applied, for n units in a sample s drawn from a population U. For identification risk to be well-defined, we assume in this section that the records in the released microdata can meaningfully be associated with units in the population. For certain kinds of SDL methods, such as synthetic data or micro-aggregation, this may not be the case.

Identification risk is defined in terms of the probability that such a link is correct (Bethlehem et al., 1990; Reiter, 2005b; Skinner and Shlomo, 2008). If it were the case that (i) no sampling occurs; (ii) the combination of values of the key variables for the target unit is unique in the population and (iii) the key values, as recorded in the microdata, are known by the adversary for the target unit, then the adversary could deduce the correct link and the probability of identification risk might be taken to be unity. The presence of sampling and the use of perturbative methods, leading to departures from (i) and (iii) respectively, are primary ways of reducing the identification risk.

In the presence of sampling, definitions of identification risk will usually depend on population characteristics, which will, in general, be unknown and this creates a problem of statistical inference, i.e. the risk measure must be estimated from sample data. In particular, sample frames that are used to draw the samples for social surveys are typically area frames or address registers and will not include population-wide information on key variables.

One approach to assessing the impact of a perturbative SDL method on identification risk is to start with a record linkage method and a set of key variables, which an adversary is assumed to use, and then to use these to match the protected microdata matched back to the original dataset (Yancey et al., 2002; Domingo-Ferrer and Torra, 2003). It is less easy to assess the impact of sampling, however.

A different approach is through probabilistic models, as first proposed by Bethlehem et al. (1990). Individual per-record risk measures are based on the probability of re-identification. These per-record risk measures are aggregated to obtain global risk measures for the entire file. As mentioned in Section 1, the key variables may be taken to be categorical, defining a contingency table. In this case, let k be the number of cells in the table and write $x_i = j$ if unit i is in cell j, $j \in \{1, \cdots, k\}$. The observed count in cell j in the sample microdata is denoted $f_j = \sum_{i \in s} I(x_i = j)$, where $I(.)$ is the indicator function. The corresponding count in the population is denoted $F_j = \sum_{i \in U} I(x_i = j)$. The identification risk will depend on these population counts F_j, $(j = 1, \cdots, k)$ which will generally be unknown. The probabilistic model makes the natural assumption in the contingency table literature that: $F_j \sim Poisson(\lambda_j)$, where λ_j is the expected population count. If the sample is drawn by Poisson or Bernoulli sampling with a known sampling fraction π_j in cell j then $F_j | f_j \sim Poisson(\lambda_j(1 - \pi_j))$ provides a predictive distribution for inference about the unknown F_j assuming conditional independence. Skinner and Holmes (1998) and Elamir and Skinner (2006) propose using a log-linear model to estimate the parameters λ_j. The sample frequencies f_j are independent Poisson distributed with mean $\mu_j = \pi_j \lambda_j$. A log-linear model for the μ_j is expressed as: $log(\mu_j) = \boldsymbol{z}_j^{'} \boldsymbol{\beta}$ where $\boldsymbol{z}_j$ is a design vector which denotes the main effects and interactions of the model for the key variables. The maximum likelihood (MLE) estimator $\hat{\boldsymbol{\beta}}$ may be obtained by solving the score equations: $\sum_j [f_j - \pi_j exp(\boldsymbol{z}_j^{'} \boldsymbol{\beta})] \boldsymbol{z}_j = 0$. Skinner and Shlomo (2008) discuss goodness of fit criteria to ensure unbiased estimation of μ_j.

The fitted values are calculated by: $\hat{\mu}_j = exp(\boldsymbol{z}_j^{'} \hat{\boldsymbol{\beta}})$ and $\hat{\lambda}_j = \hat{\mu}_j / \pi_j$. These are plugged into the expressions: $\hat{\tau}_1 = \sum_j I(f_j = 1) \hat{P}(F_j = 1 | f_j = 1)$ for the number of sample uniques that are population uniques and $\hat{\tau}_2 = \sum_j I(f_j = 1) \hat{E}(1/F_j | f_j = 1)$ the number of correct matches from among the sample uniques. Under the Poisson model: $P(F_j = 1 | f_j = 1) = exp(-\lambda_j(1 - \pi_j))$ and $E(1/F_j | f_j = 1) = [1 - exp(\lambda_j(1 - \pi_j))]/[\lambda_j(1 - \pi_j)]$. Shlomo and Skinner (2010) extended this model to take into account misclassification either arising from errors in the data collection and processing or introduced purposely into the data as an SDL technique, for example PRAM to misclassify categories of categorical variables.

2.2 Attribute Disclosure

Let x denote again the cell value for the key variables, which an adversary is assumed to know for a target unit, and let y denote a sensitive variable, the value of which an adversary wishes to learn for the target unit. A measure of attribute disclosure may then be defined in terms of the predictive probability distribution of y given x and the observable data from the microdata.

2.3 Privacy Measures

In the more recent computer science literature on privacy, there is usually no distinction between key variables and sensitive variables. The starting point is the (original) database of attribute values from which the microdata are generated via the SDL method. It is supposed that an adversary wishes to learn about the attribute values for a specific (target) unit in the database. A 'worst case' scenario is allowed for, in which the adversary has complete information about all other units represented in the database (Dwork et al., 2006). Under this assumption, we again let x denote the cell value, taking possible values $1, \cdots, k$, where the contingency table is now formed by cross-classifying all variables, whether key or sensitive.

In our survey setting, there are two possible definitions of the database: the population 'database' $\mathbf{x}_U = (x_1, \cdots, x_N)$ and the sample 'database' $\mathbf{x}_s = (x_1, \cdots, x_n)$, where N denotes the size of the population $U = \{1, \cdots, N\}$ and, without loss of generality, we write

$s = \{1, \cdots, n\}$. The sample database might be viewed from one perspective as more realistic, since it contains the data collected by the statistical agency, whereas the population database would include values of survey variables for non-sampled units, which are unknown to the agency. A problem with using the sample database in the privacy definition is that it would assume that the adversary knows which units fall in the sample, an assumption referred to as 'response knowledge' by Bethlehem et al. (1990). As they argue, identification can be achieved much more easily by an intruder under this assumption and so the agency must take considerable care to avoid this situation, wherever possible. There may be practical circumstances, when this is infeasible, but we suppose here that it is reasonable to suppose that the adversary does not have response knowledge. We therefore use the population database $\mathbf{x}_U$ to define privacy. We treat the sampling as part of the SDL mechanism and suppose that prior adversary knowledge relates to aspects of $\mathbf{x}_U$.

Let $\tilde{x}_i$ denote the cell value of unit i in the microdata after SDL has been applied and let $\tilde{f}_j = \sum_{i \in s} I(\tilde{x}_i = j)$ denote the corresponding observed count in cell j in the microdata. Supposing that the SDL methods leads to an arbitrary ordering of the records in the microdata, we can view the released data as the vector of counts: $\tilde{\boldsymbol{f}} = (\tilde{f}_1, \tilde{f}_2, \cdots, \tilde{f}_k)$. Let $Pr(\tilde{\boldsymbol{f}}|\mathbf{x}_U)$ denote the probability of $\tilde{\boldsymbol{f}}$ with respect to an SDL mechanism, which includes sampling and/or misclassification, and where $\mathbf{x}_U$ is treated as fixed. There are different definitions of privacy in the computer science literature. As a starting point, we consider the following definition.

Definition (Dwork et al., 2006): ε- differential privacy holds if:

$$max \left| ln \left(\frac{Pr[\tilde{\boldsymbol{f}}|\mathbf{x}_U^{(1)}]}{Pr[\tilde{\boldsymbol{f}}|\mathbf{x}_U^{(2)}]} \right) \right| \leq \varepsilon \tag{1}$$

for some $\varepsilon > 0$, where the maximum is over all pairs $(\mathbf{x}_U^{(1)}, \mathbf{x}_U^{(2)})$, which differ in only one element and across all possible values of $\tilde{\boldsymbol{f}}$.

We also consider an (ε, δ) - probabilistic differential privacy definition, as discussed by Chaudhuri and Mishra (2006), here probabilistic differential privacy holds if (1) applies with probability at least $1 - \delta$ for some $\varepsilon, \delta > 0$. More precisely, this definition holds if the space of possible outcomes $\tilde{\boldsymbol{f}}$ may be partitioned into 'good' and other outcomes, if (1) holds when the outcome is good and if the probability that the outcome is good is at least $1 - \delta$. This definition is essentially the same as the notion of probabilistic differential privacy in Machanavajjhala et al. (2008) where the set of bad outcomes is referred to as the disclosure set.

Based on these definitions, the next two sections consider the question of whether privacy holds in two settings. In Section 3 we consider the effect of sampling but no further SDL method. Section 4 considers whether privacy holds for microdata which have been perturbed by a misclassification-based SDL technique.

3 Sampling and Privacy

In this section, we suppose that microdata are released for a sample obtained using a probability sampling design and that there is no perturbation, so that $\tilde{\mathbf{x}}_s = \mathbf{x}_s$ and $\tilde{\boldsymbol{f}} = \boldsymbol{f}$. Two examples of a probability sampling design are (i) simple random sampling, where all possible subsets of specified size n have an equal probability of selection, and (ii) Bernoulli sampling, as considered by Chaudhuri and Mishra (2006), where each population element is selected independently with probability p.

For these designs $Pr(\tilde{\boldsymbol{f}}|\mathbf{x}_U) = Pr(\boldsymbol{f}|\mathbf{x}_U)$ in (1) may be expressed as:
(i) simple random sampling of size n

$$Pr(\boldsymbol{f}|\mathbf{x}_U) = \prod_{j=1}^{k} \left(\begin{array}{c} F_j \\ f_j \end{array} \right) \Big/ \left(\begin{array}{c} N \\ n \end{array} \right); \tag{2}$$

(ii) Bernoulli sampling with selection probability p

$$Pr(\boldsymbol{f}|\mathbf{x}_U) = \prod_{j=1}^{k} \left(\begin{array}{c} F_j \\ f_j \end{array} \right) p^{f_j}(1-p)^{F_j - f_j},$$

where, in each case, $0 \leq f_j \leq F_j$, $j = 1, \cdots, k$. For each of these and many other sampling schemes, f_j may equal F_j with positive probability. In particular, it is usual, in order to achieve unbiased statistical estimation, for an agency to require of any sampling scheme that all population units have positive inclusion probability and so, if any unit is population unique ($F_j = 1$) there is a positive probability that this unit is sampled in which case $f_j = F_j = 1$. Thus, for given $\boldsymbol{f}$ and any sampling scheme where some element f_j of $\boldsymbol{f}$ may equal F_j with positive probability, there exists a database $\mathbf{x}_U^{(1)}$ such that $f_j = F_j^{(1)} \geq 1$ for some j and $Pr[\boldsymbol{f}|\mathbf{x}_U^{(1)}] \neq 0$. Now if we change an element of $\mathbf{x}_U^{(1)}$ which takes the value j to construct $\mathbf{x}_U^{(2)}$ for which $F_j^{(2)} = F_j^{(1)} - 1 < f_j$ we obtain $Pr[\boldsymbol{f}|\mathbf{x}_U^{(2)}] = 0$. Hence, ε-differential privacy does not hold for a very broad class of sampling schemes, as discussed by Chaudhuri and Mishra (2006) in the case of Bernoulli sampling.

There are at least three reasons why the disclosure implications of this finding might not be considered a cause for concern by a statistical agency.

First, consider the threat that the event $f_j = F_j$ enables an adversary to disclose the cell value of a target individual. Such a disclosure depends upon the adversary knowing the count for the cell j across the whole of the population excluding the target individual. Given this knowledge and the observation that this count equals $f_j - 1$, the intruder could infer that the target individual falls in this cell (and appears in the microdata). For the kinds of large populations of individuals upon which social surveys in most countries are typically based, it may be deemed unrealistic, however, for an intruder to have precise information on all individuals in the population except one. The nearest realistic possibilities are that there exist an external database which either (a) via full population information, enables the population count F_j to be determined together with the identities of these F_j individuals or (b) provides identities of an unknown subset of population individuals in the cell. In neither of these cases would exact disclosure occur. In (a), the key variable value for the target individual would already be known to the intruder. In (b), there would be residual uncertainty.

Second, consider the threat of identification, where an adversary knows both that a target individual belongs to cell j and that the individual is population unique, i.e. $F_j = 1$. In this case, the target individual is sampled (so that $f_j = 1$) then the adversary would be able to identify the individual in the microdata. This possibility is already well-known to agencies as a threat and grounds for ensuring that no microdata are released for which there are combinations of key variables for which an adversary could know that $F_j = 1$ or some other small value, such as in the kind of external database mentioned above.

Third, for any given database, the possible values of $\boldsymbol{f}$ where ε- differential privacy fails may occur only with negligible probability. Therefore, the agency may consider it more appropriate to adopt the (ε, δ)- probabilistic differential privacy definition referred to earlier. Chaudhuri and Mishra (2006) show that, under a number of conditions, Bernoulli sampling can meet this definition. They partition the possible samples s, and by implication the

values $\boldsymbol{f}$, according to whether they are 'good' and show that, ε- differential privacy holds for good samples and that the probability of a good sample may be made greater than $1-\delta$. Bad samples may be referred to as 'leakage'. A basic requirement of a good sample is that it includes no 'rare' cases, that is that $f_j = 0$ for any cells j for which F_j is below $2\varepsilon^{-1}log(2k/\delta)$. Chaudhuri and Mishra (2006) then show that the probability of a good sample may be made greater than $1-\delta$ if the Bernoulli sample selection probability p is sufficiently small.

The conditions required by Chaudhuri and Mishra (2006) to achieve probabilistic differential privacy are, however, severe in common settings considered in statistical disclosure limitation. Consider the following example, for illustration.

Numerical Example: 16 dichotomous key variables generate a key with $k = 2^{16} = 65,536$ cells. In a population of size $N = 1,000,000$ the average cell size F_j in the population is thus 15.3.

Such a combination of values of k and N seems realistic at statistical agencies. If the agency specifies values $\delta = \varepsilon = 1/3$ (any larger seems unlikely to be deemed much protection) then the threshold $2\varepsilon^{-1}log(2k/\delta) = 77.3$ is well above the average cell size and the requirement for 'good' samples implies suppressing cells with $F_j < 78$ and this may be expected to lead to severe bias in statistical estimation for variables which bear some relation to the key variables. Chaudhuri and Mishra (2006) recognize that they make the 'large assumption' that k is much smaller than N. This assumption is frequently not the case in the kinds of applications often considered by statistical agencies. This questions therefore whether even (ε, δ)- probabilistic differential privacy is achievable in situations commonly considered by statistical agencies.

To explore the probabilistic nature of the threat to privacy further, consider the event that $f_j = F_j$, viewed here as the key threat to ε- differential privacy. By assumption, units in social surveys have small inclusion probabilities and the probability that all population units in a cell j will appear in the sample, i.e. $f_j = F_j = m$, will be very small for $m = 2$ (doubles) and even smaller for $m > 2$. The most realistic outcome is that a sample unique is population unique, i.e. the case $f_j = F_j = 1$ but, as we illustrate in the numerical example, this will typically also be unlikely.

Numerical Example continued: Suppose two simple random samples are drawn from the population of size $N = 1,000,000$. Sample 1 has size $n = 5,000$. Sample 2 has size $n = 10,000$. The inclusion probabilities of population units are thus $\pi = 0.005$ and $\pi = 0.01$, respectively. Let the 16 dichotomous key variables be generated independently, each as a 0-1 Bernoulli random variables with probability 0.2. We draw 1000 samples for each of the sample sizes and examine the proportion of cases where $F_j = f_j$ among cells with a given value of f_j. The average proportion of sample uniques that are population uniques was 0.024 for Sample 1 and 0.035 for Sample 2. The proportions for doubles or triples in the population were minuscule.

Thus, for any observed case in the released sample where f_j is small, and in particular for the case $f_j = 1$, the event that $F_j = f_j$ has a very small probability. Note that, viewed as a probability, this uncertainty relates to F_j conditional on the value of f_j. This conditioning is in the opposite direction to that in the definition of probabilistic differential privacy. There, f_j is treated as the random variable and F_j as fixed. From a statistical disclosure point of view, the former direction captures better the ability of an adversary to learn about unknown characteristics of a target individual given the information released.

Theoretical expressions for conditional probabilities of F_j given f_j may be developed

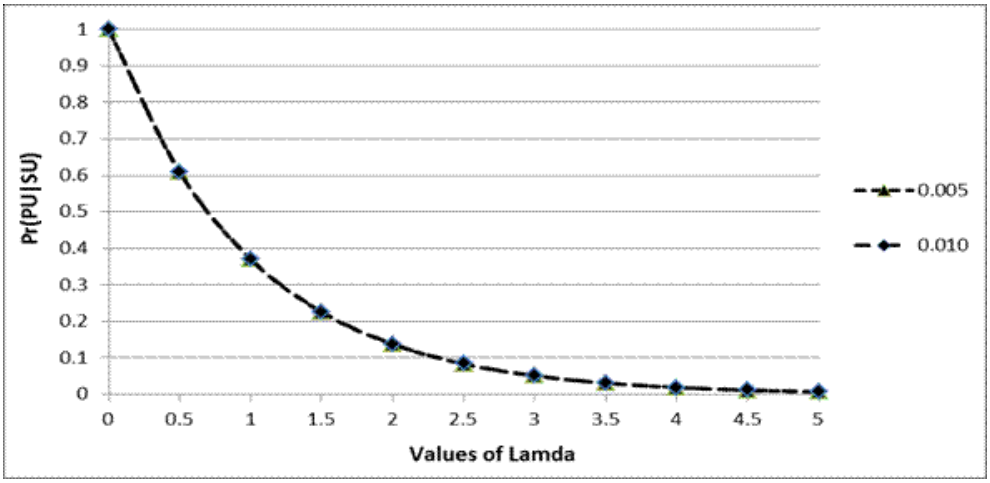

Figure 1: Probability that a sample unique is a population unique, $Pr(PU|SU)$, according to the probabilistic model in Section 2.1 for different values of the expected population count λ_j in a cell j (with $\pi_j = 0.005$ and $\pi_j = 0.010$)

under statistical models such as the Poisson models discussed in Section 2.1. For example, the probability that a sample unique is a population unique is $P(F_j = 1|f_j = 1) = exp[-\lambda_j(1-\pi_j)]$, where π_j is assumed known and λ_j, the expected value of F_j, may be estimated through log-linear modeling (Skinner and Shlomo, 2008). Figure 1 represents the probability of a population unique for different values of λ_j for two known sampling fractions: $\pi_j = 0.005$ and $\pi_j = 0.010$. As can be seen, when the expected population count is less than one, the probability of a population unique may be high, but for larger values of this count, as in the numerical example above where the average cell count is 15, the probability very quickly drops toward zero. There is little difference between the two sampling fractions.

4 Perturbation and Privacy

Assuming now that there is no sampling (so that $s = U$ and $n = N$), we consider misclassification-based SDL techniques which generate the $n \times 1$ vector $\tilde{\mathbf{x}}_s$ from $\mathbf{x}_s$. We define the misclassification matrix as:

$$Pr(\tilde{x}_i = j_1 | x_i = j_2) = M_{j_1 j_2}, \quad i = 1, \cdots, n, \quad j_1, j_2 = 1, \cdots, k \tag{3}$$

where $\tilde{x}_i$ denotes the i^{th} element of $\tilde{\mathbf{x}}_s$. Assuming independent misclassification for different units, we can write the conditional distribution $Pr(\tilde{\mathbf{x}}_s|\mathbf{x}_s)$ in terms of the matrix $\mathbf{M}$.

Suppose first that $\tilde{\mathbf{x}}_s$ can be treated as the released data. Then, using also the fact that $\mathbf{x}_s = \mathbf{x}_U$, we may replace $Pr[\tilde{\boldsymbol{f}}|\mathbf{x}_U]$ by $Pr[\tilde{\mathbf{x}}_s|\mathbf{x}_s]$ in the definition of ε-differential privacy. If we assume independent misclassification for different units then we can write

$$Pr[\tilde{\mathbf{x}}_s|\mathbf{x}_s^{(1)}] = \prod_{i \in s} Pr(\tilde{x}_i|x_i^{(1)}). \tag{4}$$

Suppose that $\mathbf{x}_s^{(1)}$ differs from $\mathbf{x}_s^{(2)}$ only in the i^{th} element, so that $x_i^{(1)} \neq x_i^{(2)}$, then

$$\frac{Pr[\tilde{\mathbf{x}}_s|\mathbf{x}_s^{(1)}]}{Pr[\tilde{\mathbf{x}}_s|\mathbf{x}_s^{(2)}]} = \frac{Pr(\tilde{x}_i|x_i^{(1)})}{Pr(\tilde{x}_i|x_i^{(2)})} = \frac{\mathbf{M}_{\tilde{x}_i x_i^{(1)}}}{\mathbf{M}_{\tilde{x}_i x_i^{(2)}}}.$$

It follows that there exists a finite ε for which ε-differential privacy holds iff all elements of $\mathbf{M}$ are positive (i.e. none are zero).

Note that

$$\begin{aligned}\max\left|ln\left(\frac{Pr[\tilde{\mathbf{x}}_s|\mathbf{x}_s^{(1)}]}{Pr[\tilde{\mathbf{x}}_s|\mathbf{x}_s^{(2)}]}\right)\right| &\leq \max_{\tilde{j},j^{(1)}\neq j^{(2)}}\left|ln\left(\frac{\mathbf{M}_{\tilde{j}j^{(1)}}}{\mathbf{M}_{\tilde{j}j^{(2)}}}\right)\right| \\ &= \max_{\tilde{j}}(\max_j ln(\mathbf{M}_{\tilde{j}j}) - \min_j ln(\mathbf{M}_{\tilde{j}j})) \\ &= \max_{\tilde{j}}(ln[\max_j \mathbf{M}_{\tilde{j}j}] - ln[\min_j \mathbf{M}_{\tilde{j}j}]).\end{aligned}$$

We assumed earlier that the elements of $\tilde{\mathbf{x}}_s$ will be subject to an arbitrary ordering so that it is more appropriate to write $\tilde{\boldsymbol{f}}$ as the released data. Let $\mathbf{a}$ be the $k\times k$ matrix with entries $a_{\tilde{j}j} = \sum_{i\in s} I(\tilde{x}_i = \tilde{j}, x_i = j)$ and note that the vectors of row and column totals of $\mathbf{a}$ are $\boldsymbol{f}$ and $\tilde{\boldsymbol{f}}$, respectively. Then assuming again independent misclassification as in (3) we may write

$$Pr[\tilde{\boldsymbol{f}}|\mathbf{x}_s] = \sum_{\mathbf{a}\in A}\prod_{\tilde{j}}\prod_j \mathbf{M}_{\tilde{j}j}^{a_{\tilde{j}j}}$$

where A is the set of possible values of $\mathbf{a}$ for which the vectors of row and column totals of $\mathbf{a}$ are $\boldsymbol{f}$ and $\tilde{\boldsymbol{f}}$, respectively. Note that, under these assumptions, $Pr[\tilde{\boldsymbol{f}}|\mathbf{x}_s]$ depends on $\mathbf{x}_s$ only via $\boldsymbol{f}$ so that we may write $Pr[\tilde{\boldsymbol{f}}|\mathbf{x}_s] = Pr[\tilde{\boldsymbol{f}}|\boldsymbol{f}]$.

If $\mathbf{x}_s = \mathbf{x}_U$ is changed in just one element then f_j will be increased by 1 for one value of j and decreased by 1 for another value of j. If the values of $\boldsymbol{f}$ before and after the change are denoted $\boldsymbol{f}^{(1)}$ and $\boldsymbol{f}^{(2)}$ respectively we can write $|\boldsymbol{f}^{(1)} - \boldsymbol{f}^{(2)}| = 2$. This is sometimes called the sensitivity of the vector of counts $\boldsymbol{f}$ (Dwork et al., 2006). Abowd and Vilhuber (2008) define ε-differential privacy, with $Pr[\tilde{\boldsymbol{f}}|\boldsymbol{f}]$ replacing $Pr[\tilde{\boldsymbol{f}}|\mathbf{x}_U]$, so that ε-differential privacy holds if (1) holds for all pairs ($\boldsymbol{f}^{(1)}$, $\boldsymbol{f}^{(2)}$) where $|\boldsymbol{f}^{(1)} - \boldsymbol{f}^{(2)}| = 2$.

If all elements of $\mathbf{M}$ are positive then $Pr[\tilde{\boldsymbol{f}}|\boldsymbol{f}] > 0$ iff $\tilde{\boldsymbol{f}}^T\mathbf{1}_k = \boldsymbol{f}^T\mathbf{1}_k = n$, where $\mathbf{1}_k$ is a $k\times 1$ vector of 1s.

If all elements of $\mathbf{M}$ are not positive, say $\mathbf{M}_{\tilde{j}j} = 0$ ($\tilde{j}\neq j$), then $\tilde{f}_{\tilde{j}}$ is bounded above by $n - f_j$. Let $\tilde{\boldsymbol{f}}$ be defined by $n - f_j^{(1)}$ in cell $\tilde{j}$, $f_j^{(1)}$ in cell j and 0 in the remaining cells and, assuming independent misclassification and $\mathbf{M}_{jj} > 0$, we have $Pr[\tilde{\boldsymbol{f}}|\boldsymbol{f}^{(1)}] > 0$. Suppose $f_j^{(2)} = f_j^{(1)} + 1$. Then we must have $Pr[\tilde{\boldsymbol{f}}|\boldsymbol{f}^{(2)}] = 0$ since $\tilde{f}_{\tilde{j}}$ is bounded above by $n - f_j^{(1)} - 1$. Hence, ε-differential privacy does not hold.

As before, ε-differential privacy holds iff all elements of $\mathbf{M}$ are positive if we treat the released data as $\tilde{\boldsymbol{f}}$ rather than $\tilde{\mathbf{x}}_s$.

To consider the possible use of probabilistic differential privacy when not all elements of $\mathbf{M}$ are positive, we define an outcome $\tilde{\boldsymbol{f}}$ which could arise from $\tilde{\boldsymbol{f}}^{(1)}$ to be bad if, for some $\tilde{j}$, $\tilde{f}_{\tilde{j}} = n - \sum_{j\in A_j} f_j^{(1)}$ where $A_j = \{j|\mathbf{M}_{\tilde{j}j} = 0\}$. Following a similar argument to that above, ε-differential privacy does not hold for any bad outcome. If the outcome is not bad and $\boldsymbol{f}^{(2)}$ is defined such that $|\boldsymbol{f}^{(1)} - \boldsymbol{f}^{(2)}| = 2$ then $Pr[\tilde{\boldsymbol{f}}|\boldsymbol{f}^{(2)}] \neq 0$. Thus, (ε,δ)- probabilistic differential privacy will hold for some (ε,δ) with $\varepsilon < 1$, $\delta < 1$. For many choices of $\mathbf{M}$ and large values of n, bad samples will be unlikely, although we do not attempt here to quantify this.

We simply illustrate possible choices of the misclassification matrix $\mathbf{M}$ for some common SDL techniques on categorical variables and comment on whether all elements are positive.

Recoding: For the non-perturbative method of recoding, which is the most common SDL technique for microdata arising from social surveys, assume a variable where categories 1 to a are changed to category 1.The misclassification matrix is:

$$\mathbf{M}_{j_1 j_2} = \begin{cases} 1 & j_2 = 1, \cdots, a \text{ and } j_1 = 1, \text{ or } j_2 = a+1, \cdots, k \text{ and } j_1 = j_2 - a + 1 \\ 0 & \text{otherwise} \end{cases}$$

It is clear that with elements equal to zero, ε-differential privacy will not be guaranteed. Categories above a are not changed so increasing a cell count $f_j^{(2)} = f_j^{(1)} + 1$ for one value of j will set the probability in the denominator in (1) equal to zero because of the bounds on the cell counts.

Random Data Swapping: For the perturbative method of random data swapping, the probability of selecting any 2 records for swapping data is $\begin{pmatrix} n \\ 2 \end{pmatrix}^{-1}$. Let f_j be the number of records taking value j, as before, and assume counts f_{j_1} and f_{j_2} are positive, then:

$$\mathbf{M}_{j_1 j_2} = \mathbf{M}_{j_2 j_1} = \frac{f_{j_1} f_{j_2}}{\begin{pmatrix} n \\ 2 \end{pmatrix}}, \quad \mathbf{M}_{jj} = \frac{\begin{pmatrix} f_j \\ 2 \end{pmatrix}}{\begin{pmatrix} n \\ 2 \end{pmatrix}}$$

and, provided there are no zero counts of categories, there are no zero elements in the misclassification matrix. This strategy will guarantee ε-differential privacy as the denominator in (1) will now be positive.

PRAM: The SDL technique of PRAM uses a misclassification (probability) matrix $\mathbf{M}$ to make random changes across categories of a variable. We can also require the property of invariance of the misclassification matrix: $\mathbf{vM} = \mathbf{v}$ where $\mathbf{v}$ is the vector of sample proportions: $\mathbf{v} = \left(\frac{f_1}{n}, \cdots, \frac{f_k}{n}\right)$. This ensures that the perturbed marginal distribution will be similar to the original marginal distribution in the microdata. The misclassification matrix should be defined to have no zero elements in order to ensure differential privacy. Note that in practice, there may be zero elements in the misclassification matrix which represent structural zeros in the population, i.e. impossible combinations of categories such as children having an occupation as a 'doctor'. For these cases, differential privacy is not applicable.

In general, non-perturbative methods for SDL will not guarantee ε-differential privacy but may under some circumstances uphold (ε, δ)- probabilistic differential privacy whereas perturbative methods which ensure no zero elements in the misclassification matrix will guarantee ε-differential privacy.

5 Conclusion

This paper has focused on microdata from surveys, where sampling is almost always employed. While sampling does not guarantee differential privacy, it does provide probabilistic differential privacy under certain conditions. Through numerical illustration, we have seen that these conditions can be quite severe and so we have also discussed other probabilistic ways of assessing the threat to privacy.

SDL techniques may be used to provide additional protection to sampling. In this paper we have examined perturbation techniques which may be viewed as a form of misclassification. Such perturbation does ensure a form of differential privacy provided the misclassification matrix contains no zeroes. Even if there are zeroes it may provide probabilistic

differential privacy. Further research is needed to establish what conditions on the misclassification mechanism are required to meet specified levels of probabilistic differential privacy. Further work is also needed to assess the combined impact of sampling and perturbation.

References

Abowd, J. and Vilhuber, L. (2008) *How Protective are Synthetic Data? In PSD'2008 Privacy in Statistical Databases, (Eds. J. Domingo-Ferrer and Y. Saygin).* New York: Springer LNCS 5261, 239-246.

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

Chaudhuri, K. and Mishra, N. (2006) When random sampling preserves privacy. *Proceedings of the 26th International Cryptology Conference.*

Dalenius, T. and Reiss, S. (1982) Data swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference, 7, 73-85.*

Defays, D. and Nanopoulos, P. (1992) Panels of enterprises and confidentiality: The small aggregates method. *Proceedings of Statistics Canada Symposium 92, Design and Analysis of Longitudinal Surveys, 195 204.*

Dinur, I. and Nissim, K. (2003) Revealing information while preserving privacy. *PODS 2003, 202-210.*

Domingo-Ferrer, J., Mateo-Sanz, J. and Torra, V. (2001) Comparing sdc methods for microdata on the basis of information loss and disclosure risk. *ETK-NTTS Pre-Proceedings of the Conference, Crete, June 2001.*

Domingo-Ferrer, J. and Torra, V. (2003) Disclosure risk assessment in statistical microdata protection via advanced record linkage. *Statistics and Computing (SAC), Vol. 13(4), 343-354.*

Duncan, G. and Lambert, D. (1989) The risk of disclosure for microdata. *Journal of Business and Economic Statistics 7, 207-217.*

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) *Calibrating Noise to Sensitivity in Private Data Analysis. In Theory of Cryptography TCC (eds. S. Halevi and R. Rabin).* Heidelberg: Springer, LNCS Vol. 3876, 265-284.

Elamir, E. and Skinner, C. (2006) Record-level measures of disclosure risk for survey microdata. *Journal of Official Statistics, 22, 525-539.*

Fienberg, S. and McIntyre, J. (2005) Data swapping: Variations on a theme by dalenius and reiss. *Journal of Official Statistics, 9, 383-406.*

Fuller, W. A. (1993) Masking procedures for micro-data disclosure limitation. *Journal of Official Statistics, 9, 383-406.*

Gomatam, S. and Karr, A. (2003) Distortion measures for categorical data swapping. *Technical Report Number 131, National Institute of Statistical Sciences.*

Gouweleeuw, J., Kooiman, P., Willenborg, L. and De Wolf, P. (1998) Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics, 14, 463-478.*

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J. and Vilhuber, L. (2008) Privacy: Theory meets practice on the map. *In Proceedings of the 24th International Conference on Data Engineering, Cancun, Mexico, 277-286.*

Reiter, J. (2005a) Releasing multiply imputed, synthetic public-use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, A, Vol.168, No.1, 185-205.*

Reiter, J. P. (2005b) Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association 100, 1103-1112.*

Rubin, D. (1993) Satisfying confidentiality constraints through the use of synthetic multiply-imputed microdata. *Journal of Official Statistics, 91, 461-468.*

Shlomo, N. and Skinner, C. (2010) Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *Annals of Applied Statistics, Vol. 4, No. 3, 1291-1310.*

Skinner, C. (1992) On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica, 46, 21-32.*

Skinner, C. and Holmes, D. (1998) Estimating the re-identification risk per record in micro-data. *Journal of Official Statistics 14, 361-372.*

Skinner, C. J. and Shlomo, N. (2008) Assessing identification risk in survey micro-data using log linear models. *Journal of American Statistical Association, Vol. 103, Number 483, 989-1001.*

Willenborg, L. and De Waal, T. (2001) *Elements of Statistical Disclosure Control in Practice. Lecture Notes in Statistics, 155.* New York: Springer-Verlag.

Winkler, W. E. (2004) Masking and re-identification methods for public-use microdata: Overview and research problems. *In: Privacy in Statistical Databases, 2004 (eds. Domingo-Ferrer, J. and Torra, V.), LNCS 3050, 231-246.*

Yancey, W., Winkler, W. and Creecy, R. (2002) *Disclosure Risk Assessment in Perturbative Micro-data Protection.* In: Inference Control in Statistical Databases (ed. J. Domingo-Ferrer), New York: Springer LNCS 2316, 135-151.