UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Working Paper M10/14
Methodology

Privacy Protection From Sampling And Perturbation In Survey Microdata

Natalie Shlomo, Chris Skinner

Abstract

We consider the assessment of disclosure risk in the release of microdata from social surveys as public-use files. We consider both identification risk and the notion of differential privacy introduced in the computer science literature. We show that sampling, as a disclosure limitation technique, does not guarantee differential privacy. However, threats to differential privacy, i.e. 'leakage', may have small probability and sampling can provide protection under a broader definition of privacy. Moreover, the occurrence of conditions when such a threat can occur may be unknown to the adversary and require statistical inference. Disclosure limitation techniques that perturb variables in the microdata according to misclassification probabilities guarantee differential privacy provided that there are no zero elements in the misclassification mechanism. Combining sampling and perturbation, especially for rare combinations of identifying variables, will reduce the 'leakage'.

Privacy Protection from Sampling and Perturbation in Survey Microdata

Natalie Shlomo and Chris Skinner

Southampton Statistical Sciences Research Institute, University of Southampton,
Highfield, Southampton SO17 1BJ United Kingdom

Email: N.Shlomo@soton.ac.uk, C.J.Skinner@soton.ac.uk

Abstract: Statistical agencies release microdata arising from social surveys as public-use files after applying statistical disclosure limitation (SDL) techniques. Disclosure risk is assessed in terms of identification risk where small counts on cross-classified indirect identifying key variables, i.e. a key, can be used to make an identification and confidential information may be learnt. In the computer scientist literature, there is no distinction between key variables and sensitive variables, and it is assumed that an adversary, who wishes to learn about a specific target individual, would have complete information about all other units in the population database. Sampling as an SDL technique is examined according to the definition of differential privacy. We show that in special cases when a sample unique in the key is also a population unique, differential privacy is not guaranteed and a small 'leakage' may occur. Indeed, statistical agencies do not have knowledge of the population database since it depends on non-sampled units and statistical inference is used to estimate the population counts. Therefore, the disclosure scenario of differential privacy may be deemed unrealistic. SDL techniques that perturb variables in the microdata according to misclassification probabilities guarantee differential privacy provided that there are no zero elements in the misclassification mechanism. Combining sampling and perturbation, especially for rare combinations of the key, will reduce the 'leakage' and guarantee differential privacy.

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).

In this paper, we focus on 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 must assess the disclosure risk and if required choose appropriate statistical disclosure limitation (SDL) methods to apply to the data. Measuring disclosure risk involves assessing and evaluating numerically the risk of 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.

Agencies usually distinguish 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. 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).

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 sub-sampling (see Willenborg and De Waal, 2001). These methods are the most common for protection 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, Winkler and Creecy, 2002), micro-aggregation (replacing values with their average within groups of records) (Defays and Nanopoulos, 1992), rounding to a pre-selected rounding base, 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 a more general 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, Kooiman, Willenborg, and De Wolf, 1998). For more information on these methods see also: Willenborg and De Waal (2001), Gomatam and Karr (2003), Domingo-Ferrer, Mateo-Sanz, and Torra (2001) and references therein.

In this paper we describe how statistical agencies would define disclosure risk in Section 2. We contrast their approach with the notion of differential privacy as defined in the computer science literature (Dinur and Nissim, 2003; Dwork, McSherry, Nissim and Smith, 2006) in Section 3. We then examine whether the common practice of releasing microdata from social surveys and/or perturbing the microdata prior to release guarantees differential privacy in Sections 4 and 5. 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* 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 a vector $\mathbf{x}_{i_0}$ of *key variabl*es for a target unit i_0 and seeks to use these values to link the unit to a record in the microdata, which we write as an $n \times k$ matrix $\tilde{\mathbf{X}}_s$, with rows corresponding to n units in a sample s and columns corresponding to the values of the key variables, after SDL has been applied. In order that identification risk can 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, Keller and Pannekoek, 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 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. estimating the risk measure from sample data, which may be hard in practice to solve. 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, Winkler, and Creecy, 2002; Domingo-Ferrer and Torra, 2003). It is less easy to assess the impact of sampling, however.

Another approach is through probabilistic models, as first proposed by Bethlehem, Keller, and Pannekoek (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 I, the key variables may be taken to be categorical, defining a contingency table. In this case, redefine the $1\times k$ vector $\mathbf{x}$ so that k is the number of cells in the table and $\mathbf{x}_i=\mathbf{e}_j$ if unit i is in cell j, $j\in\{1,...,k\}$, where $\mathbf{e}_j$ is the $1\times k$ vector with a 1 in the j^{th} position and zeros elsewhere. The vector of counts $\tilde{f}_j$ in the cells j in the microdata (after SDL) may then be expressed as $\tilde{\mathrm{f}}=(\tilde{f}_1,\tilde{f}_2,...,\tilde{f}_k)=\mathbf{1}_n^T\tilde{\mathbf{X}}_s$ where $\mathbf{1}_n$ is the $n\times 1$ vector of ones. In the same way, we define $\mathrm{f}=(f_1,f_2,...,f_k)=\mathbf{1}_n^T\mathbf{X}_s$ where $\mathbf{X}_s$ is the $n\times k$ matrix representing the original unperturbed microdata and $\mathbf{F}=(F_1,F_2,...,F_k)=\mathbf{1}_N^T\mathbf{X}_U$ is the vector of population counts, where $\mathbf{X}_U$ is the $N\times k$ matrix of population values of $\mathbf{x}_i$ and N is the population size. The identification risk will depend on these population counts F_j, $(j=1,...,k)$ 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 sampling fraction π_j in cell j and the sample frequency in cell j is f_j (which is a function of $\mathbf{X}_s$) 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: $\log(\mu_j)=\mathbf{z}_j{}'\beta$ where $\mathbf{z}_j$ is a design vector which denotes the main effects and interactions of the model for the key variables. The maximum likelihood (MLE) estimator $\hat{\beta}$ may be obtained by solving the score equations: $\sum_j[f_j-\pi_j\exp(\mathbf{z}_j{}'\beta)]\mathbf{z}_j=0$. Skinner and Shlomo (2008) discuss goodness of fit criteria to ensure unbiased estimation of μ_j.

The fitted values are calculated by: $\hat{\mu}_j = \exp(\mathbf{z}_j'\hat{\beta})$ and $\hat{\lambda}_j = \hat{\mu}_j / \pi_j$. These are plugged into the expressions: $\hat{\tau}_1 = \sum_j I(f_j = 1)\hat{P}(F_j = 1 \mid f_j = 1)$ for the number of sample uniques that are population uniques and $\hat{\tau}_2 = \sum_j I(f_j = 1)\hat{E}(1/F_j \mid f_j = 1)$ the number of correct matches from among the sample uniques. Under the Poisson model: $P(F_j = 1 \mid f_j = 1) = \exp(-\lambda_j(1-\pi_j))$ and $E(1/F_j \mid f_j = 1) = [1 - \exp(\lambda_j(1-\pi_j))]/[\lambda_j(1-\pi_j)]$. 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 $\mathbf{x}$ denote again the vector of key variables, the value of 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 $\mathbf{x}$ and the observable data from the microdata.

2.3 Differential Privacy

In the computer science literature on differential 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 now let $\mathbf{x}$ denote the full vector of attribute values, not distinguishing between key and sensitive variables.

In our survey setting, there are two possible definitions of the database: the $N \times k$ population ‘database’ $\mathbf{X}_U$ and the $n \times k$ sample 'database' $\mathbf{X}_s$, which is a sub-matrix of $\mathbf{X}_U$. 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 the sample database for differential privacy 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). It is well-known in the statistical literature that making this assumption can increase disclosure risk hugely and that 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 differential privacy. We treat the sampling as part of the SDL mechanism and suppose that prior adversary knowledge relates to aspects of $\mathbf{X}_U$.

Suppose that the SDL methods leads to an arbitrary ordering of the records in the microdata so that we can view the released data as the vector of counts: $\tilde{\mathbf{f}} = (\tilde{f}_1, \tilde{f}_2, ..., \tilde{f}_k)$, as defined in the previous section. Let $\Pr(\tilde{\mathbf{f}} \mid \mathbf{X}_U)$ denote the probability of $\tilde{\mathbf{f}}$ with respect to an SDL mechanism, which includes sampling and/or misclassification, and where $\mathbf{X}_U$ is treated as fixed.

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

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

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

Based on this definition, the next two sections consider the question of whether ε - differential privacy holds for microdata containing samples from social surveys. Section 3 addresses this question for microdata which have not undergone any SDL techniques. Section 4 considers whether differential privacy holds for microdata from social surveys which have undergone SDL techniques and under what conditions we have differential privacy.

3. Sampling and Differential Privacy

In this section, we suppose that SDL arises solely from sampling and that there is no perturbation, so that $\tilde{\mathbf{X}}_s = \mathbf{X}_s$ and $\tilde{\mathbf{f}}$=$\mathbf{f}$ .et the sample s be drawn by a specified probability sampling scheme with probability $p(s)$ from the population U . For example, under simple random sampling, all possible subsets of specified size *n* have an equal probability of selection from a population of size *N*, i.e. $p(s) = 1/\binom{N}{n}$.

The expression $\Pr(\tilde{\mathbf{f}} \mid \mathbf{X}_U) = \Pr(\mathbf{f} \mid \mathbf{X}_U)$ in (1) may be expressed as $\Pr(\mathbf{f} \mid \mathbf{X}_U) = \sum_{s \in S(\mathbf{f})} p(s)$, where $S(\mathbf{f})$ denotes the set of samples s with $p(s) > 0$ for which $\mathbf{1}_n^T \mathbf{X}_s = \mathbf{f}$. For example, under simple random sampling of size *n* we have

$$\Pr(\mathbf{f} \mid \mathbf{X}_U) = \prod_{j=1}^{k} \binom{F_j}{f_j} \bigg/ \binom{N}{n} . \tag{2}$$

Based on the definition in (1), ε-differential privacy will not hold for some ε iff there is a pair $(\mathbf{X}_U^{(1)}, \mathbf{X}_U^{(2)})$ which differ in only one row and for which $\Pr[\mathbf{f} \mid \mathbf{X}_U^{(1)}] \neq 0$ and $\Pr[\mathrm{f} \mid \mathbf{X}_U^{(2)}] = 0$. Consider an arbitrary value of $\mathbf{f}$. Under many sampling schemes used for drawing samples in social surveys, such as simple random sampling or stratified sampling designs, there will exist $\mathbf{X}_U^{(1)}$ such that $f_j = F_j^{(1)}$ for some j and $\Pr[\mathbf{f} \mid \mathbf{X}_U^{(1)}] \neq 0$, for example, a sample unique in a cell j that is also a population unique. Now if we change the row of $\mathbf{X}_U^{(1)}$ which takes the value e_j and construct a $\mathbf{X}_U^{(2)}$ for which $F_j^{(2)} = F_j^{(1)} - 1 < f_j$ we obtain $\Pr[\mathbf{f} \mid \mathbf{X}_U^{(2)}] = 0$. It is clear that if $F_j^{(1)} = f_j$ then ε- differential privacy will not hold. This result follows for other sample designs based on simple random sampling, for example stratified random sampling or random sampling of clustered data.

There are at least two reasons why a statistical agency might not consider such a breach of ε- differential privacy to be of concern. First, the potential disclosure depends upon an intruder knowing the count $F_j^{(-i_0)}$ for the cell j across the whole of the population excluding the target individual i_0. 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.

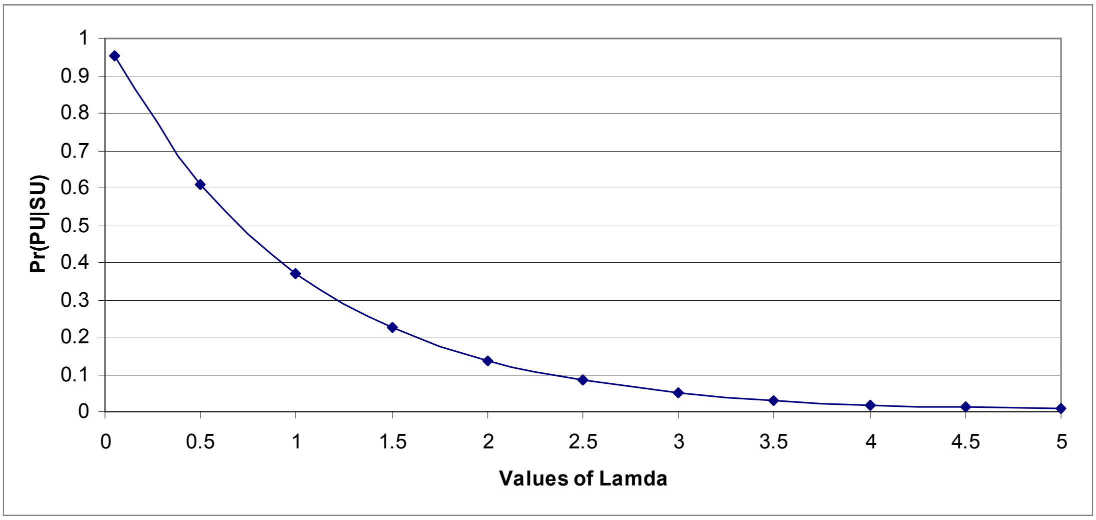
Secondly, the event that the count in the population F_j is exactly equal to the count in the sample f_j might be deemed sufficiently negligible to disregard for survey microdata. By assumption, units in social surveys have small inclusion probabilities and the probability that all population units in 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 following example, this will typically also be unlikely.

Numerical Example: Consider simple random sampling and two samples: Sample 1 with *n*=5,000 and Sample 2 with *n*=10,000, from a population of size *N*=1,000,000, so that inclusion probabilities are $\pi = 0.005$ and $\pi = 0.01$, respectively. This is a realistic sample design at statistical agencies. Let 16 dichotomous key variables be generated independently, each as a 0-1 Bernoulli random variables with the probability of 0.2. This defines a key with 65,536 cells or an average cell size in the population of 15.3. We draw 1000 samples for each of the sample sizes and examine the proportion of cells where $F_j = f_j$ relative to 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.

To reflect such uncertainty the definition of ε - differential privacy might be modified. Machanavajjhala, Kifer, Abowd, Gehrke and Vilhuber (2008) define (ε, δ) probabilistic differential privacy which allows the constraint in (1) to hold with probability at least $1-\delta$. In other words, ε-differential privacy can fail with a small probability, not more than δ. This small probability is known as the leakage. Although sampling is not sufficient to achieve ε- differential privacy, it may achieve (ε, δ) probabilistic differential privacy, with little leakage, as in the numerical example above.

As in (2), theoretical expressions for the probabilities relating to leakage will depend on the population cell counts F_j for the cross-classified key variables and these are generally unknown to statistical agencies since they depend on non-sampled units. In this case, the agency may estimate the proportions according to the probabilistic model described in Section 2.1. For example, the probability that a sample unique is a population unique is $P(F_j = 1 \mid f_j = 1) = \exp[-\lambda_j(1-\pi_j)]$ and the expected population count λ_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. 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.

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



4. **Perturbation and Differential Privacy**

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

$$\Pr(\tilde{\mathbf{x}}_i = \mathbf{e}_{j_1} \,|\, \mathbf{x}_i = \mathbf{e}_{j_2}) = M_{j_1 j_2} \quad , \qquad i = 1, \dots, n, \quad j_1, j_2 = 1, \dots, k \tag{3}$$

where $\tilde{\mathbf{x}}_i$ denotes the i^{th} row 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 **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 $\Pr[\tilde{\mathbf{f}} \,|\, \mathbf{X}_U]$ in 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

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

Suppose that $\mathbf{X}_s^{(1)}$ differs from $\mathbf{X}_s^{(2)}$ only in row i, so that $\mathbf{x}_i^{(1)} \neq \mathbf{x}_i^{(2)}$, then

$$\frac{\Pr[\tilde{\mathbf{X}}_s \,|\, \mathbf{X}_s^{(1)}]}{\Pr[\tilde{\mathbf{X}}_s \,|\, \mathbf{X}_s^{(2)}]} = \frac{\Pr(\tilde{\mathbf{x}}_i \,|\, \mathbf{x}_i^{(1)})}{\Pr(\tilde{\mathbf{x}}_{\mathrm{i}} \,|\, \mathbf{x}_i^{(2)})} = \frac{M_{\tilde{j}j^{(1)}}}{M_{\tilde{j}j^{(2)}}},$$

where $\tilde{j}, j^{(1)}$ and $j^{(2)}$ are the entries of $\tilde{\mathbf{x}}_i$, $\mathbf{x}_i^{(1)}$ or $\mathbf{x}_i^{(2)}$ respectively which take the value 1.

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

$$\max\left|\ln\left(\frac{\Pr[\tilde{X}_s \mid X_s^{(1)}]}{\Pr[\tilde{X}_s \mid X_s^{(2)}]}\right)\right| = \max_{\tilde{j}, j^{(1)} \neq j^{(2)}} \left|\ln\left(\frac{M_{\tilde{j}j^{(1)}}}{M_{\tilde{j}j^{(2)}}}\right)\right| = \max_{\tilde{j}}\left(\max_j \ln(M_{\tilde{j}j}) - \min_j \ln(M_{\tilde{j}j})\right)$$

$$= \max_{\tilde{j}}\left(\ln[\max_j M_{\tilde{j}j}] - \ln[\min_j M_{\tilde{j}j}]\right)$$

We assumed earlier that the rows of $\tilde{\mathbf{X}}_s$ will be subject to an arbitrary ordering so that it is more appropriate to write $\tilde{\mathbf{f}}$ as the released data. Let $\mathbf{a}$ be the $k \times k$ matrix with entries $a_{\tilde{j}j} = \sum_{i \in s} I(\tilde{x}_i = e_{\tilde{j}}, x_i = e_j)$ and note that $\mathbf{a}\mathbf{1}_k = \tilde{\mathbf{f}}^T$ and $\mathbf{1}_k^T\mathbf{a} = \mathbf{f}$. Then assuming again independent misclassification as in (3) we may write

$$\Pr[\tilde{\mathbf{f}} \mid \mathbf{X}_s] = \sum_{\mathbf{a} \in \mathbf{A}} \prod_{\tilde{j}} \prod_{j} M_{\tilde{j}j}^{a_{\tilde{j}j}}$$

where $\mathbf{A}$ is the set of possible values of $\mathbf{a}$ for which $\mathbf{a}\mathbf{1}_k = \tilde{\mathbf{f}}^T$ and $\mathbf{1}_k^T\mathbf{a} = \mathbf{f}$. Note that, under these assumptions, $\Pr[\tilde{\mathbf{f}} \mid \mathbf{X}_s]$ depends on $\mathbf{X}_s$ only via $\mathbf{f}$ so that we may write $\Pr[\tilde{\mathbf{f}} \mid \mathbf{X}_s] = \Pr[\tilde{\mathbf{f}} \mid \mathbf{f}]$.

If $\mathbf{X}_s = \mathbf{X}_U$ is changed in just one row 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 $\mathbf{f}$ before and after the change are denoted $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$ respectively we can write $|\mathbf{f}^{(1)} - \mathbf{f}^{(2)}| = 2$. Note that Abowd and Vilhuber (2008) define ε-differential privacy, with $\Pr[\tilde{\mathbf{f}} \mid \mathbf{f}]$ replacing $\Pr[\tilde{\mathbf{f}} \mid \mathbf{X}_U]$, so that ε-differential privacy holds if (1) holds for all pairs $(\mathbf{f}^{(1)}, \mathbf{f}^{(2)})$ where $|\mathbf{f}^{(1)} - \mathbf{f}^{(2)}| = 2$.

If all elements of $\mathbf{M}$ are positive then $\Pr[\tilde{\mathbf{f}} \mid \mathbf{f}] > 0$ iff $\tilde{\mathbf{f}}^T\mathbf{1}_k = \tilde{\mathbf{f}}^T\mathbf{1}_k = n$

If all elements of $\mathbf{M}$ are not positive, say $M_{\tilde{j}j} = 0$ ($\tilde{j} \neq j$). Then $\tilde{f}_{\tilde{j}}$ is bounded above by $n - f_j$. Let $\tilde{\mathbf{f}}$ 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 $M_{jj} > 0$, we have $\Pr[\tilde{\mathbf{f}} \mid \mathbf{f}^{(1)}] > 0$. Suppose $f_j^{(2)} = f_j^{(1)} + 1$. Then we must have $\Pr[\tilde{\mathbf{f}} \mid \mathbf{f}^{(2)}] = 0$ since $\tilde{f}_{\tilde{j}}$ is bounded above by $n - f_j^{(1)} - 1$. Hence ε-differential privacy does not hold.

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

We next examine misclassification matrices $\mathbf{M}$ for some common SDL techniques on categorical variables and assess 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:

$$M_{jk}=\begin{cases}1 & k=1,...,a \quad and \quad j=1, or \quad j=k-a+1 \quad and \quad j>1\\ 0 & otherwise\end{cases}$$

It is clear that with elements equal to zero, ε -differential privacy will not be guaranteed.

Random Data Swapping: For the perturbative method of random data swapping, the probability of selecting any 2 records for swapping data is $\binom{n}{2}^{-1}$. Let n_j and n_k be the number of records taking values j and k and assume counts n_j and n_k are positive, then:

$$M_{jk}=M_{kj}=\frac{n_j n_k}{\binom{n}{2}},\quad M_{jj}=\frac{\binom{n_j}{2}}{\binom{n}{2}}.$$

and, provided there are no zero counts of categories, there are no zero elements in the misclassification matrix.

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: $\mathbf{v}=\left(\frac{n_1}{n},...,\frac{n_k}{n}\right)$. 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 data, i.e. impossible combinations of categories such as children having an occupation as a 'doctor'.

5. Discussion

We have contrasted alternative approaches to assessing disclosure risk with the release of survey microdata. Sampling alone is not sufficient to guarantee ε -differential privacy. There are, however, at least two reasons why ε -differential privacy might be deemed too strong a condition by statistical agencies. First, the disclosure scenario associated with this definition, that an adversary knows the entire population database except for the target individual, may be deemed unrealistic.

Second, the event that breach of the condition occurs may be very unlikely and a broader definition, such as (ε,δ) probabilistic differential privacy of Machanavajjhala et al. (2008), may better match the disclosure control principles informing the agency's release practice.

The disclosure scenario more usually considered by statistical agencies when considering the release of microdata assumes that an adversary can use key variables to match the data to publicly available external datasets and identify individuals. This leads to the measures of identification risk. The agency will generally not know the entire population database and has to rely on probabilistic models to assess the risk of identification.

Perturbation via misclassification of the identifying categorical key variables does guarantee ε -differential privacy, provided the misclassification matrix do not contain zero elements. It will also generally reduce the risk of identification (Shlomo and Skinner, 2010). The combination of sampling and perturbation will generally lead to greater protection than either method used singly. In particular, the targeting of perturbation at key variable value combinations which are unique in the sample or have low sample counts may be expected to reduce the leakage under (ε,δ) probabilistic differential privacy. Dwork, et al. (2006) define low 'sensitivity' in circumstances like the typical survey set-up described in this paper and conclude that 'only a small perturbation to the proportions should be necessary to achieve ε -differential privacy'.

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.

Dalenius, T. and Reiss, S.P. (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 Micro-Data 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.J. (2006). Record-Level Measures of Disclosure Risk for Survey Microdata. *Journal of Official Statistics, 22, 525-539.*

Fienberg, S.E. 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.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.

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.P. (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.B. (1993). Satisfying Confidentiality Constraints through the Use of Synthetic Multiply-imputed Microdata. *Journal of Official Statistics*, *91*, 461-468.

Shlomo, N. and Skinner, C.J. (2010). Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata. *Annals of Applied Statistics* (to be published).

Skinner, C.J. (1992). On Identification Disclosure and Prediction Disclosure for Microdata. *Statistica Neerlandica*, *46,* 21-32.

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

Skinner, C. J. and Shlomo, N. (2008). Assessing Identification Risk in Survey Microdata Using Log Linear Models. J*ournal 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.

Yancey, W.E., Winkler, W.E., and Creecy, R.H. (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.