Identity disclosure risk control in microdata release via post-randomization

Tapan K. Nayak
Department of Statistics, George Washington University, Washington, DC 20052, USA; e-mail: tapan@gwu.edu

Abstract

Protecting the confidentiality of survey respondents' information in microdata release is a significant concern for most statistical agencies. To address the matter, agencies often publish data that has been perturbed using data swapping, noise addition, multiple imputation and other methods. However, agency reports and research papers in this area rarely state disclosure risk and disclosure protection goals clearly. In this paper, we propose a precise and strict identity disclosure protection goal and then present a post-randomization procedure for achieving that goal, at a minimal loss of data quality. The procedure can also allow agencies to retain selected marginal counts from original data and avoid undesirable changes to the original data.

Keywords: confidentiality protection; sample unique; key variable; correct match probability; data perturbation; data utility.

1. Introduction

The main goal of statistical agencies is to collect and publish data to inform the public, policy makers and researchers, but they also need to protect the confidentiality of unit level information for legal reasons and for upholding public trust. For confidentiality protection, agencies often release a perturbed or masked version of the original data, which also reduces data utility. Various masking methods, such as grouping, cell suppression, data swapping, multiple imputation and random noise infusion have been developed for practical use; see the books by Doyle et al. (2001) and Willenborg and de Waal (2001).

Methods and formulas for analyzing the original data may not be appropriate for analyzing a masked version of it. Data masking may destroy known properties, such as unbiasedness, of standard estimators. The probability distribution of the random observables whose values constitute the data depends on both the sampling design and the masking method. So, agencies should give full information about the masking procedure so that data users can derive suitable inferential methods for applying to released data. It is also desirable that inferential methods for the original data should remain valid, at least approximately, for perturbed data, so that users would not need to develop new methods for data analysis (see Rubin, 1993). Logically, the agency should try to achieve its confidentiality protection goals with minimum loss of data utility. The fundamental challenges in confidentiality protection are defining and measuring disclosure risk and data utility, and determining precise protection goals.

The aforementioned objectives motivated us to seek masking procedures, which (a) enable the agency to evaluate, assure and communicate confidentiality protection clearly and (b) do not require overly complex or burdensome new theoretical derivations and programming for making valid inferences. In the next section, we briefly review some existing approaches to measuring and dealing with disclosure risk and then propose a new disclosure control goal. Specifically, our proposed goal is to ensure, via data perturbation, that no intruder's confidence (probability) in his match of a target unit to a record in the released data, based on a set of categorical key variables, can be larger than ξ, where both ξ and the key variables are specified by the agency. In Section 3, we shall present a procedure that accomplishes this goal.

2. Identification risk and its control

Disclosure is a difficult topic (cf., Lambert, 1993) and it can occur in many forms depending on the disclosure scenario (see Willenborg and de Wall, 2001). Broadly speaking, disclosure occurs when the released data enable an intruder R to predict the values of some confidential variables for a target unit B fairly accurately. To avoid disclosure fully, the released data should not enable any intruder to gain much new information about any target (Dalenius, 1977). However, as Dwork (2006) proved, this goal not achievable, mainly because there is no restriction on the intruder's prior information. Thus, for developing practical disclosure control goals, it seems essential for the agency to draw boundaries on the prior information and thereby consider only a subset of all intruders. One common approach is to specify a subset of the survey variables as identifying or key variables, whose values are easily accessible form other sources, and then assume that intruders may know only the values of some or all of the key variables for their target units. The agency's choice of the key variables describes a universe of intruders that is of concern to the agency. Identity disclosure, which occurs when an intruder R correctly identifies the record of a target unit B in released data by matching the values of the key variables for B, is considered to be the most serious type of disclosure and has received substantial attention; see Bethlehem et al. (1990), Greenberg and Zayatz (1992), Willenborg and de Wall (2001), Skinner and Elliot (2002), Reiter (2005) and Shlomo and Skinner (2010).

Extending the work of Duncan and Lambert (1986) and Lambert (1993), Reiter (2005) developed a Bayesian approach for measuring identity disclosure risk. While the procedure can incorporate intruder knowledge and dependencies among survey variables extensively, it also involves guessing intruders' behavior and substantial modeling, estimation and computation, and as Shlomo and Skinner (2010) noted, the complexity of the procedure may limit its application in practice.

Another approach focuses on the units in the sample that are unique by the key variables. Suppose all key variables are categorical and X is their cross-classification and that X has k cells (or categories), denoted $c_1, \ldots, c_k$. For confidentiality protection, the agency perturbs the original values of some (or all) of the key variables. Let Z denote the perturbed version of X, with the same set of categories, and let T_i and S_i denote the frequencies of $X = c_i$ and $Z = c_i$, respectively, for $i = 1, \ldots, k$. Suppose that target unit B's X-category is c_j, denoted $X_B = c_j$, which is known to intruder R. If $T_j = 1$ and R also knows that B is in the sample, then he will correctly identify B's record in the original data. If $T_j = 1$ and R does not know if B is in the sample or not, he will have a correct match if B is unique in the population with respect to X. Thus, the conditional probability that a unit is population unique given that it is sample unique has gained much attention and several researches focused on estimating it from the original data, under certain sampling designs and models; see e.g., Bethlehem et al. (1990), Greenberg and Zayatz (1992), Skinner and Elliot (2002) and Shlomo and Skinner (2010).

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

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

where the probability is with respect to both sampling and data perturbation. However, the IR in (1) depends on the target unit (or rather on the X-category of the target unit) and also on population frequencies of the k cells of X, which are unknown. Thus, (1) cannot be calculated from available information. For two data perturbation procedures, Shlomo and Skinner (2010) estimate (1) from original data using a Poisson log-linear model. The sum (or average) of (1), over all sample unique units, has been used as an aggregate measure of disclosure risk for a data set. However, note that an aggregate measure may be satisfactorily low even when disclosure risk for some units are high.

2a. Our view and proposed goal

Our framework and perspective are similar to those of Shlomo and Skinner (2010), but we consider a more stringent and directly relevant disclosure control goal and eliminate the estimation task. To be more specific, we consider the scenario where an intruder R knows $X_B = c_j$ for his target unit B and randomly selects one of the records in the released data with $Z = c_j$, if any, and takes that as B's record. If $S_j = 0$, R stops searching for B's record. We also assume conservatively that R knows that B is in the sample. The intruder's matching procedure, stated above, seems very reasonable as the intruder is assumed to know the values of only the key variables for his target. As we noted earlier, the agency needs to select the key variables for specifying the intruder set relevant in the particular context. However, we should note that all categorical variables can be allowed to be key variables.

We propose to take the following as the agency's disclosure control goal: select a value ξ and ensure that no intruder's (as described above) match for any target unit in the sample would be correct with probability larger than ξ. Thus, no intruder's confidence (probability) in his match for any target unit can legitimately be larger than ξ. Note that the probability of a correct match for a unit not in the sample is zero. Since intruders would also know the number of units in each category in released data, for measuring *identification disclosure risk* we shall consider $IR(j,a) = P(CM(j)|S_j = a)$ for all $a > 0$, where $CM(j)$ stands for the event that a unit B with $X_B = c_j$ is correctly matched in the aforementioned matching scheme. Thus, our precise disclosure control goal is to ensure that

$$IR(j,a) = P(CM(j)|S_j = a) \leq \xi \quad \text{for all } a > 0 \text{ and } j = 1,\ldots,k. \tag{2}$$

We believe that this goal is precise, objective, quite stringent and compelling. The value of ξ is chosen by the agency. The role of $1-\xi$ is similar to the role of α (level of significance) in hypothesis testing. The intruder should have strong evidence to legitimately conclude a match, as in accepting an alternative hypothesis. In order to conclude a match rationally, an intruder should confirm that the probability in (2) much larger than 1/2. As Lambert (1993) discussed, intruders may also draw conclusions without adequate justification and thereby induce harm, but such actions are beyond agency's control. We believe that in most practical situations, reasonable choices for ξ would be between .75 and .90. In the next section we present a procedure that accomplishes the disclosure control goal in (2) for any $\xi \geq 3/7 = .4286$.

Clearly, (2) implies that $P(CM(j)) \leq \xi$, i.e., the unconditional probability of correct match for a unit B with $X_B = c_j$ does not exceed ξ. We should also note that the probability in (2) depends also on the unknown population frequencies through sampling. Letting $\mathbf{T} = (T_1,\ldots,T_k)'$ denote the frequency vector from original data, we can write

$$P(CM(j)|S_j = a) = \sum_{\mathbf{t}} P(CM(j)|S_j = a, \mathbf{T} = \mathbf{t})P(\mathbf{T} = \mathbf{t}). \tag{3}$$

In (3), $P(\mathbf{T} = \mathbf{t})$ depends on the population frequencies, but not $P(CM(j)|S_j = a, \mathbf{T} = \mathbf{t})$ and our procedure actually ensures that

$$P(CM(j)|S_j = a, \mathbf{T} = \mathbf{t}) \leq \xi \tag{4}$$

for all $a > 0, j = 1,\ldots,k$ and $\mathbf{t}$. Thus, the correct match probability is no larger than ξ even for an intruder who the frequencies in original data. Obviously, (4) implies (2).

3. Disclosure control by post-randomization

The Post-randomization Method (PRAM) was introduced by Gouweleeuw et al. (1998) as a technique for categorical data perturbation for confidentiality protection. The basic ideas of PRAM are to (i) select a transition probability matrix $P = ((p_{ij}))$, where $\sum_i p_{ij} = 1$ for $j = 1, \dots, k$, and then (ii) randomly change any original category c_j to c_i with probability p_{ij} $(i, j = 1, \dots, k)$. The randomization step is performed for each record in the data set, independently of all other records. Thus, $p_{ij} = P(Z = c_i | X = c_j)$. Importantly, $\{p_{ij}\}$ may depend on the original data, and in particular, for invariant PRAM, P is chosen satisfying

$$P\mathbf{T} = \mathbf{T}, \tag{5}$$

where $\mathbf{T} = (T_1, \dots, T_k)'$ is the frequency vector from original data. Let $\pi_i = P[X = c_i], i = 1, \dots, k$, and $\pi = (\pi_1, \dots, \pi_k)'$. Let n denote the sample size and $\mathbf{S} = (S_1, \dots, S_k)'$. Then, for any invariant PRAM, $\mathbf{S}/n$ is an unbiased estimator of π, like $\mathbf{T}/n$, but the covariance matrix of $\mathbf{S}/n$ is inflated by PRAM.

We shall use a particular invariant PRAM to achieve our disclosure control goal. To explore effects of PRAM on identity disclosure, suppose $X_B = c_1$ (for notational simplicity), which implies that $T_1 \geq 1$. Note that given $\mathbf{T}$ and P, which may be allowed to depend on original data but only through $\mathbf{T}$, we have $S_1 = \sum_{i-1}^{k} V_i$, where $V_i \sim b(T_i, p_{1i}), i = 1, \dots, k$, are independent binomial random variables. We shall only consider the cases where $T_1 \geq 1$ because we shall apply invariant PRAM only to categories with nonzero counts, in which case, $S_1 = 0$ whenever $T_1 = 0$. Then, letting $\alpha_i = p_{1i}$ and $\beta_i = \alpha_i/(1 - \alpha_i), i = 1, \dots, k$ and denoting B's category after applying PRAM by Z_B, it follows that

$$P(S_1 = a, Z_B = c_1 | \mathbf{T}) = \alpha_1 \Big[\prod_{i=1}^{k} (1 - \alpha_i)^{T_i^*} \Big] \sum \prod_{i=1}^{k} \binom{T_i^*}{k} \beta_i^{a_i}, \tag{6}$$

where $T_1^* = T_1 - 1, T_i^* = T_i, i \geq 2$ and the sum is over all integer valued $a_1, \dots, a_k$ such $0 \leq a_i \leq T_i^*$ and $\sum a_i = a - 1$. We shall denote the sum in (6) by Σ_{a-1}. Then, it follows that

$$P(S_1 = a, Z_B \neq c_1 | \mathbf{T}) = (1 - \alpha_1) \Big[\prod_{i=1}^{k} (1 - \alpha_i)^{T_i^*} \Big] \Sigma_a \tag{7}$$

and hence the correct match (CM) probability for B, conditional on a and $\mathbf{T}$, is

$$P(CM \text{ for } B | S_1 = a, \mathbf{T}) = \frac{1}{a} \Big[\frac{\alpha_1 \Sigma_{a-1}}{\alpha_1 \Sigma_{a-1} + (1 - \alpha_1) \Sigma_a}) \Big] = \frac{1}{a} \Big[1 + \frac{1}{\beta_1} \frac{\Sigma_a}{\Sigma_{a-1}} \Big]^{-1}. \tag{8}$$

Clearly, (8) $\leq 1/a$, which implies that for any $\xi \geq 1/2$, (4) holds for all $a \geq 2$. So, the case of $a = 1$, i.e., when there is a unique match for B, is crucial. For $a = 1$, (8) reduces to

$$P(CM \text{ for } B | S_1 = 1, \mathbf{T}) = \Big[T_1 + \frac{1}{\beta_1} \sum_{i=2}^{k} \beta_i T_i \Big]^{-1}. \tag{9}$$

Clearly, (9) is less than $1/T_1$, which is the probability of correctly identifying B in original data. Thus, PRAM reduces the correct match probability.

Next, we show that for all units with unique match, the correct match probability in (9) can be bounded from above by applying a suitable PRAM to categories with nonzero counts. For notational simplicity, suppose all $T_i \geq 1$ for $i = 1, \ldots, k$; otherwise we would need to change k to some $k' < k$ and re-label the nonempty categories. Now, consider the transition probabilities $p_{ii} = 1 - \theta/T_i, p_{ji} = \theta/[(k-1)T_i]$ for $i, j = 1, \ldots, k$ and $i \neq j$, with $0 \leq \theta \leq 1$. Note that for any $0 \leq \theta \leq 1$, this is invariant PRAM as $\{p_{ij}\}$ satisfy (5). For these $\{p_{ij}\}$, (9) reduces to

$$(9) = (T_1 - \theta)\Big[T_1(T_1 - \theta) + \theta^2 \sum_{i=2}^{k} \frac{T_i}{(k-1)T_i - \theta}\Big]^{-1} \leq \frac{T_1 - \theta}{T_1(T_1 - \theta) + \theta^2} = \psi(T_1, \theta), \text{ say.} \tag{10}$$

The inequality in (10) follows from the fact that the summand is a decreasing function of T_i and hence must be at least $1/(k-1)$.

It can be seen (taking derivative with respect to T_1) that for any $0 < \theta < 1$, $\psi(T_1, \theta)$ is a decreasing function of T_1 over $T_1 \geq 2$. Also, from direct comparison and routine algebra it follows that $\psi(1, \theta) \geq \psi(2, \theta)$ if and only if $\theta \leq 2/3$. Thus, for $\theta \leq 2/3, \psi(T_1, \theta)$ is maximum when $T_1 = 1$ and for $\theta > 2/3, \psi(T_1, \theta)$ is the largest when $T_1 = 2$. Let $g(\theta) = \psi(1, \theta) = (1-\theta)/(1 - \theta + \theta^2)$. It can be seen that $g(\theta)$ is a strictly decreasing function of θ and $g(2/3) = 3/7 = 0.4286$. So, for any $3/7 \leq \xi < 1$, the equation $g(\theta) = \xi$ has a unique solution $\theta_0 \leq 2/3$ and transition probabilities based on this θ_0 ensures that ξ is a uniform upper bound of (9). In summary, we have the following:

Result 1. Let $3/7 \leq \xi < 1$ be a given value and θ be the solution of $g(\theta) = \xi$. Then, applying (invariant) PRAM with $p_{ii} = 1 - \theta/T_i, p_{ji} = \theta/[(k-1)T_i]$ for $i, j = 1, \ldots, k$ and $i \neq j$, ensures that for any unit in the sample with unique match, the probability that the match is correct is at most ξ.

We believe, values of ξ between .75 and .90 would be suitable in most applications, which means our choice of θ should be roughly between .25 and .40, as $g(.40) = .789$ and $g(.25) = .923$. As we noted earlier, for any $\xi \geq 1/2$, (4) holds for all $a \geq 2$, in view of (8). Actually, we can also show that if $\theta \leq (1 - 1/k)$, then $P(CM \text{ for } B|S_1 = 2, \mathbf{T}) \leq P(CM \text{ for } B|S_1 = 1, \mathbf{T})$ for all $\mathbf{T}$. This implies that for $3/7 \leq \xi \leq .5$, the choice of θ in Result 1 also works for all $a \geq 1$, if $k \geq 3$. Thus, we strengthen Result 1 as follows:

Result 2. Suppose $1/2 \leq \xi < 1$ or $3/7 \leq \xi < 1/2$ and $k \geq 3$. Let θ be the solution of $(1-\theta)/(1-\theta+\theta^2) = \xi$. Then, the invariant PRAM with transition probabilities $p_{ii} = 1 - \theta/T_i, p_{ji} = \theta/[(k-1)T_i]$ for $i, j = 1, \ldots, k$ and $i \neq j$, guarantees that (4) holds and hence the probability of correctly matching the record of any survey unit in the released data is at most ξ.

5. Conclusions

In this paper we have proposed a precise and strict identity disclosure protection goal and presented an invariant post-randomization procedure for achieving that goal. The additional variability introduced by our procedure can be evaluated using the results of Nayak and Adeshiyan (2015). Note that since we apply invariant PRAM, the expected number of units moving out of any category is the same as expected number of units moving into that category and hence the expected counts remain the same. Moreover, the expected number of units moving out of any category is θ, which is small (proportionately) for categories with large counts in original data. Thus, our procedure will not affect much the estimates of cell probabilities, unless they are small.

In our procedure, as presented above, the original category of any unit may change to any other category with nonzero probabilities. However, in practice, one may want to avoid switching between certain categories or may wish to preserve certain marginal counts. We can easily enrich our procedure to give the agency more control and flexibility on category switches. One simple idea is to divide all categories into several mutually exclusive and exhaustive groups and then apply invariant PRAM within each group. Finally, we hope that practitioners will find our ideas in this paper appealing and useful.

References

Bethlehem, J.G., Keller, W.J. and Pannekoek, J. (1990). Disclosure control of microdata. *J. Amer. Statist. Assoc.*, 85, 38-45.

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *Statistisk. tidskrift*, 3, 213-225.

Doyle, P., Lane, J., Theeuwes, J. and Zayatz, L. (Ed.) (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies.* Amsterdam: Elsevier.

Duncan, G.T. and Lambert, D. (1986). Disclosure-limited data dissemination. *J. Amer. Statist. Assoc.*, 81, 10-28.

Dwork, C. (2006). Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, ICALP (2), volume 4052 of Lecture Notes in Computer Science, pp. 1-12. Springer.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and De Wolf, P.-P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *J. Official Statist.*, 14, 463–478.

Greenberg, B. and Zayatz, L. (1992). Strategies for measuring risk in public use microdata files. *Statist. Neerland.*, 46, 33-48.

Lambert, D. (1993). Measure of disclosure risk and harm. *J. Official Statist.*, 9, 313-331.

Natak, T.K. and Adeshiyan, S.A. (2015). On invariant post-randomization for statistical disclosure control. *Internat. Statist. Rev.* (to appear).

Reiter, J.P. (2005). Estimating identification risk in microdata. *J. Amer. Statist. Assoc.*, 100, 1101-1113.

Rubin, D.B. (1993). Discussion: Statistical disclosure limitation. *J. Official Statist.*, 9, 462-468.

Shlomo, N. and Skinner, C. (2010). Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *Ann. Appl. Statist.,* 4, 1291-1310.

Skinner, C.J. and Elliot, M.J. (2002). A measure of disclosure risk for microdata. *J. R. Statist. Soc.*, Ser. B, 64, 855-867.

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