Measuring Identification Risk in Microdata Release and Its Control by Post-randomization

Tapan Nayak, Cheng Zhang

The George Washington University

godeau@gwu.edu

November 6, 2015

What is Statistical Disclosure Control

Disclosure Control

research the issues of privacy and confidentiality that arise in the process collecting data from the public and disclosing the data to a certain group of people.

Statistical Disclosure Control

explores disclosure control issues from a statistical point of view, including (but not limited to) proper measures of privacy and confidentiality, statistical techniques of perturbing the microdata, inference after the perturbation, etc.

Why SDC

- The law requires confidentiality to be preserved, even without publishing the data.
- The need for sharing more microdata with public is becoming stronger than ever.
- Inference issues with the perturbed data.

Previous Work

- Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J. and De Wolf, P.-P. (1998). *"Post randomization for statistical disclosure control: Theory and implementation."* J. Official Statist., 14, 463.
- Shlomo, N. and Skinner, C.J. (2012). *"Privacy protection from sampling and perturbation in survey microdata."* J. Privacy Confidentiality, 4, 155.

Our Contribution

- We focus on identity disclosure based on categorical key variables.
- A new measure for identification risk and a associated disclosure control goal.
- A method that accomplishes the preceding goal, using Post-Randomization (PRAM).
- Effects of our method upon the inference issues.

Definition of Identity Disclosure

Assumptions

- Intruder knows the key variable value of the target.
- Units are non-differentiable with the same key variable values, and the intruder would pick one at random as the record of the target.

Identity Disclosure: Correct Match

A correct match happens to a unit when the intruder correctly matches the unit's record of non-key variable value, among all the units that share the same key variable value with the target.

We measure the risk of identity disclosure by the probability of a unit being correctly matched.

Example

	Key Variable			Non-key variable	
Name	Sex	Race	Residency	VIN	Cross-classification of keys
John	M	White	VA	a	c1
Mike	M	Black	MD	b	c2
Larry	M	Black	VA	c	c3
Susan	F	White	MD	d	c4
Jane	F	Other	MD	e	c5
Rachel	F	White	MD	f	c4

For example,
If the original data is released after the removal of names,

$$P\{\text{John is correctly matched}\} = 1$$

$$P\{\text{Susan is correctly matched}\} = 0.5$$

Disclosure Control Goal

$$P(CM|S_j = a, X_B = c_j) \leq \xi$$

for all $a > 0$ and $j = 1, 2, ..., k$.

- CM stands for the event that the target unit B is correctly matched in the aforementioned scenario and matching scheme.
- $c_1, c_2, ..., c_k$ are all the cells (values of the cross-classified variable formed by all key variables).
- S_j is the count of c_j in the perturbed released data.
- The intruder knows the target's key variable value, $X_B = c_j$

Our Approach: Post-Randomization(PRAM)

What is PRAM

In a nutshell, PRAM is the randomization mechanism of a categorical variable using a transition probability where the transition probability is a function of the data, instead of being predetermined.

EX: A Bernoulli dataset with 10 observations $X_1, X_2, ..., X_{10}$. A PRAM transition matrix could be

$$P = \begin{pmatrix} 1 - \frac{1}{T_0} & \frac{1}{T_0} \\ \frac{1}{T_1} & 1 - \frac{1}{T_1} \end{pmatrix}$$

where T_i is the count of i , and $p_{ij} = P\{j \rightarrow i\}$. If $X_1 = 0$, then change X_1 to 1 with probability $\frac{1}{T_0}$.

Our Approach: Post-Randomization(PRAM)

Our choice of PRAM matrix Let a group contain cells $c_1, c_2, ..., c_k$. Then the transition probability matrix is $P = \big((p_{ij})\big)$ where

$$p_{ii} = 1 - \frac{\theta}{T_i}, p_{ji} = \frac{\theta}{(k-1)T_i}$$

for $i, j = 1, 2, ..., k$, $i \neq j$, $0 \leq \theta \leq 1$, and T_i is the count of c_i in the original dataset.

Our Approach: Post-Randomization(PRAM)

Physical interpretation of θ:

$$\begin{aligned} & E(\text{number of units moving out of cell i }) \\ = & T_i - E(\text{number of units of does not change in cell i}) \\ = & T_i - T_i \times p_{ii} = \theta \end{aligned}$$

Being independent of θ, this applies to all cells.

Our Approach: Post-Randomization(PRAM)

Example:

	Key Variable			Non-key variable	
Name	Sex	Race	Residency	VIN	Cross-classification of keys
John	M	White	VA	a	c1
Mike	M	Black	MD	b	c2
Larry	M	Black	VA	c	c3
Susan	F	White	MD	d	c4
Jane	F	Other	MD	e	c5
Rachel	F	White	MD	f	c4

$$\begin{pmatrix} c_1 & c_2 & c_3 & c_4 & c_5 \\ 1-\theta & \theta/4 & \theta/4 & \theta/8 & \theta/4 \\ \theta/4 & 1-\theta & \theta/4 & \theta/8 & \theta/4 \\ \theta/4 & \theta/4 & 1-\theta & \theta/8 & \theta/4 \\ \theta/4 & \theta/4 & \theta/4 & 1-\frac{\theta}{2} & \theta/4 \\ \theta/4 & \theta/4 & \theta/4 & \theta/8 & 1-\theta \end{pmatrix}$$

Pros and cons of our approach

Pro:

- Easy to operate: reducing the choosing matrix problem to choosing one parameter for each group
- Unbiased estimators: $E(S_i | T_i) = T_i$
- PRAM matrix, being dependent on the original data, is hard to retrieve;

Con:

- Simple structure costs unnecessary perturbation
- limited to $\xi \geq \frac{1}{3}$

Our perturbation mechanism

- We set $\xi \geq \frac{1}{3}$
- Solve for a common θ for all group.
- Solve for the minimum group size k.
- Subset only the singleton and doubleton cells. Partition the subset into groups of at least k cells.
- PRAM each group independently.

Solution of θ and k

$$\text{ultimate goal: } P(CM|S_j = a, X_B = c_j) \leq \xi$$
$$\Uparrow$$
$$P(CM|S_j = a, X_B = c_j, T = t) \leq \xi,$$
where T is the vector of all cells' counts
$$\Updownarrow$$
$$P(CM|S_j = 1, X_B = c_j, T = t) \leq \xi,$$
$$P(CM|S_j = 2, X_B = c_j, T = t) \leq \xi,$$
$$\xi \geq \frac{1}{3}$$
$$\Uparrow$$
$$P(CM|S_j = 2, X_B = c_j, T = t) \leq P(CM|S_j = 1, X_B = c_j, T = t) \leq \phi(\theta)$$
$$\phi(\theta) = \phi(\theta) = \frac{T_j - \theta}{T_j(T_j - \theta) + \theta^2} \leq \xi$$

Solution

Solve $\phi(\theta) \leq \xi$ for θ. Then plug θ in $P(CM|S_j = 2, X_B = c_j, T = t) \leq P(CM|S_j = 1, X_B = c_j, T = t)$ to solve the smallest possible k.

Data Quality

The exploration on data quality serves mostly as a guide of how to partition all categories into groups, so that the groups are formed in the way that it has a total variation as small as possible.
Numerical findings:

- Total variation from perturbing using PRAM, i.e.

  $\sum var(S_i | T_i)$,
  is ignorable compared to the total variation from sampling.
- Dividing all cells into more groups with smallest possible group size is optimal in terms of lowering the total variation from perturbation.

Future Research

- $\xi < \frac{1}{3}$
- Different form of block transition matrix
- Sampling weights
- Other partitioning criteria
- Variation on the joint distribution between key and non-key variables

Thank You