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

Cheng Zhang

Collaborators: Prof. Tapan K. Nayak and Dr. Jiashen You.

September 26, 2017

Disclosure Control Background

- Statistical agencies aim to collect and release informative data to help policy makers and researchers make appropriate inferences and decisions.
- Agencies need to keep individual or unit level information confidential for legal reasons and upholding public trust and support.
- A perturbed or masked version of the data is usually released instead of the original data.

Localized Problem

We all agree to protect confidentiality with minimal loss of data utility and intuitively data masking procedures should be determined after examining trade-offs between disclosure risk and data utility.
However, on a closer look this is not a well defined objective.

- Impossible to universally define or measure confidentiality, as disclosure takes on different forms and scenarios;
- Impossible to comprehensively examine utility, as released data may be used in many ways by diverse users

The fundamental Challenge of data masking in practice

Agencies need to determine measures of disclosure risk, data utility and their disclosure control goals suitable for each application.

Abstract

- For categorical key variables, we propose a new approach to measuring identity disclosure called **identification risk (IR)** and setting strict disclosure control goals;
- We propose a statistical perturbation method called **Inverse Frequency Post-Randomization (IFPR)** that directly solves the disclosure control goal;
- We show IFPR allows substantial control over possible changes to the original data and retains high level of data utility under multinomial sampling scheme.
- We apply IFPR to 2013 MD PUMS, where it shows very little data quality loss.

PART I: Defining identification risk

Table: 2013 Personal-level Public Use Mircodata for Maryland

Unit	Sex	Age	Race	Marital Status	PUMA	Income, etc.	Match
1	M	60	white	married	1006	.....	Unique
2	F	52	black	married	801	.....	Double
3	F	52	black	married	801	.....	Double
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
59033	...	...	...	...	...	.....	...

Disclosure Scenario

- Intruder knows the Target B is in the sample, and the values of (sex, age, race, marital status and PUMA) for the Target[1];
- Unit 1 can be correctly identified using (Sex,Age, Race, Marital Status, PUMA);
- Unit 2,3 can be correctly identified with probability 0.5[2] ;

[1] (sex (2), age (92), race (9), marital status (5) and PUMA (44)) has 25,406 non-zero cells out of 364,320 possible cells;

[2] Assume intruder chooses one out of two matches at random;

PART I: Defining identification risk

Two types of Variables

Key variables (*identifying variables*), whose values are easily accessible to the public, and **Non-key variables**;

- All key variables are categorical;
- Let **X** denote the cross-classification of all key variables;

e.g., suppose key variables are:
sex $\in \{F, M\}$,
marital status $\in \{$married, widowed, separated, divorced, never married$\}$.

Table: Cross-classification of (Sex, Marital Status)

X	(Sex, Marital Status)	X	(Sex, Marital Status)
c_1	(F, married)	c_6	(M, married)
c_2	(F, widowed)	c_7	(M, widowed)
c_3	(F, separated)	c_8	(M, separated)
c_4	(F, divorced)	c_9	(M, divorced)
c_5	(F, never married)	c_{10}	(M, never married)

PART I: Defining identification risk

• Assume intruder know only X_B where B represents target's unit index;
• A **match** exists if there is a unit with $X = X_B$ in the released data.

Identity Disclosure

Target B's identity is disclosed if intruder successfully declares unit of Target B from the match(es).
When B is identified, we also say a correct match (CM) for B happens, or B is correctly matched.

PART I: Defining identification risk

Identification Risk

$$IR_B(a) = P(\text{CM for B}|S = a)$$

where S is the $\#$ matches in released data.

Disclosure Control Goal

$$IR_B(a) \leq \xi$$

for all $a > 0$, and all $B = 1, 2, ...n$.

- ξ is specified by agency;
- Moderately small ξ would suffice:
  - we consider a very conservative scenario;
  - an intruder needs strong evidence to conclude a disclosure.

PART II: METHOD

- STEP 0: Disclosure Control Goal Specification
- Our method guarantees the disclosure goal for $\xi > \frac{1}{3}$. It can be easily modified to achieve $\xi < \frac{1}{3}$.
- Key variables (X) are perturbed; Non-keys are not changed;
- Application of 2013 personal-level PUMS of Maryland:
  - Choice of key variables: sex (2), age (92), race (9), marital status (5) and PUMA (44);
  - Choice of ξ: $.395$
    $\Rightarrow$ Only singleton (unique match) and doubleton (double matches) needs perturbation:
    13662 + 4777 = 18,439 cells, 72.6% of 25,406 cells;
    13662 singleton units + 9,554 doubleton units = 40% of data.

PART II: METHOD

• STEP 1: Data Partitioning.
Subset all singleton and doubleton units;

To control perturbation magnitude, divide 18,439 cells into homogenous blocks;
when a unit is changed, it changes within its block.
Specifically:

- gender remains unchanged;
- for race, white and black remains unchanged, other races can change within other races;
- age is divided into 7 broader intervals: 0 to 17, 18 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, 65 and above.
- martial status and PUMA are not controlled;

how we partition : create coarsened variables.

PART II: METHOD

Table: Block distribution

Partition	Sex	Age	Race	# units
1	Male	0 to 17	Others	295
2	Female	0 to 17	Others	314
3	Male	18 to 24	Others	189
4	Female	18 to 24	Others	231
5	Male	25 to 34	Others	656
6	Female	25 to 34	Others	691
7	Male	35 to 44	Others	664
8	Female	35 to 44	Others	726
9	Male	45 to 54	Others	711
10	Female	45 to 54	Others	792
11	Male	55 to 64	Others	739
12	Female	55 to 64	Others	885
13	Male	65 and above	Others	1405
14	Female	65 and above	Others	1926
15	Male	0 to 17	White	438
16	Female	0 to 17	White	476
17	Male	18 to 24	White	168
18	Female	18 to 24	White	196
19	Male	25 to 34	White	446
20	Female	25 to 34	White	523
21	Male	35 to 44	White	525
22	Female	35 to 44	White	616
23	Male	45 to 54	White	631
24	Female	45 to 54	White	716
25	Male	55 to 64	White	562
26	Female	55 to 64	White	732
27	Male	65 and above	White	609
28	Female	65 and above	White	998

PART II: METHOD

Table: Block distribution

Partition	Sex	Age	Race	# units
29	Male	0 to 17	Black/ African American	756
30	Female	0 to 17	Black/ African American	721
31	Male	18 to 24	Black/ African American	278
32	Female	18 to 24	Black/ African American	242
33	Male	25 to 34	Black/ African American	414
34	Female	25 to 34	Black/ African American	423
35	Male	35 to 44	Black/ African American	396
36	Female	35 to 44	Black/ African American	392
37	Male	45 to 54	Black/ African American	324
38	Female	45 to 54	Black/ African American	350
39	Male	55 to 64	Black/ African American	222
40	Female	55 to 64	Black/ African American	317
41	Male	65 and above	Black/ African American	239
42	Female	65 and above	Black/ African American	282

PART II: METHOD

• STEP 2:Post-randomization -IFPR

The **inverse frequency post-randomization** matrix is a block-diagonal matrix where each block features an inverse frequency structure indexed by θ.

Specifically,

i) a singleton unit changes with probability θ;

ii) a doubleton unit changes with probability $\theta/2$;

iii) once a unit is to be changed, it randomly changes to one of the remaining cells from that block.

The value of θ is determined by ξ to meet the disclosure control goal.

Interpretation of θ from theory:

perturbation rate can be interpreted as a linear function of θ ;

PART II: METHOD

Table: θ given ξ

ξ	θ
.789	.4
.667	.5
.429	2/3
.408	.75
.395	.8
.365	.9
.350	.95
.337	.99

PART II: METHOD

Table: Empirical Identification Risks

	$T=1$	$T=2$	
$S=1$	0.2315	0.3933	0.2849
$S=2$	0.1961	0.3477	0.2827
	0.1348	0.3027	

T denote the number of matches in the original dataset;
S denote the number of matches in the released dataset;

PART III: DATA UTILITY

Distribution of X

The variance inflation induced by IFPR is negligible in comparison to the sampling variance, with respect to estimating Π.

PART III: DATA UTILITY

Table: Frequency Distributions of Marital Status

Marital Status	Original Data	Perturbed Data	Difference	SD
Married	24688 (.4182)	24678 (.4180)	10	119.84
Widowed	3156 (.0535)	3180 (.0539)	-24	54.67
Divorced	4742 (.0803)	4704 (.0797)	38	66.03
Seperated	1040 (.0176)	1039 (.0176)	1	31.95
Never married	25407 (.4304)	25432 (.4308)	-25	120.30

PART III: DATA UTILITY

Table: Distribution of Race / Ethnicity

Race or Ethnicity	Original	Perturbed
White	37201 (.6302)	37201 (.6302)
Black	15239 (.2581)	15239 (.2581)
American Indian alone	97 (.0016)	92 (.0015)
Alaska Native alone	1 (.000017)	0 (0)
American Indian & Alaska Native	42 (.0007)	46 (.0008)
Asian	3461 (.0586)	3345 (.0567)
Native Hawaiian & other Pacific Islander	20 (.0004)	21 (.0004)
Some other race alone	1349 (.0228)	1337 (.0227)
Two or more races	1623 (.0275)	1652 (.0280)

PART III: DATA UTILITY

Total Variation Distance (TVD)

$$TVD(p, q) = \sup_{A} |p(A) - q(A)|.$$

TVD measure the divergences of 2 probability measures in terms of how large the 2 probability measures may differ on a given event. It is a mathematical guarantee.

Example: proportion of Asians who are married

- $A = \{race = 6, mar = 1\}$;
- estimate by perturbed data : 0.0336;
- TVD with respect to race and mar $= 0.0028$;
- estimate by original data $\in 0.0336 \pm 0.0028$

PART III: DATA UTILITY

Table: TVD Between Original and Perturbed Distributions

Variables	TVD	Number of cells	Variables	TVD	Number of cells
race, mar	0.0028	45	race, work	0.0035	81
race, puma	0.0013	396	puma, work	0.0198	396
race, edu	0.0088	72	sex, race, mar	0.006	90
puma, edu	0.0324	352	sex, race, edu	0.0093	144
mar, edu	0.0127	40	mar, race, edu	0.0218	360
mar, work	0.007	45	sex, race, work	0.0039	162

THANK YOU