A Methodology to Compare Anonymization Methods Regarding Their Risk-Utility Trade-Off

Sara Ricci

Universitat Rovira i Virgili
Dept. of Computer Engineering and Mathematics
UNESCO Chair in Data Privacy
Av. Països Catalans 26, 43007 Tarragona, Catalonia
e-mail sara.ricci@urv.cat

September 20, 2017

Why anonymization?

Private information is routinely collected and stored.

Entities:

- Companies
- Governments
- Research groups

Purposes:

- Scientific research
- Business management

Problem

Risk-Utility Trade-Off

Statistical disclosure control (SDC)

Statistical disclosure control methods aim at releasing data that preserve their statistical validity while protecting the privacy of each data subject.

Two main approaches exist:

- ***Utility-first:*** Priority is given to preserving certain utility properties. Disclosure risk is assessed *a posteriori* (e.g. global recoding).
- ***Privacy-first:*** A privacy model is adopted to specificy privacy guarantees before anonymization. Utility is assessed *a posteriori* (e.g. k-anonymity).

Note

The greater the amount of masking, the greater are both privacy protection and information loss.

Comparing the risk-utility trade-off in SDC

Note

Comparing the latter regarding the privacy-utility trade-off is not straightforward.

Before

- Select some parameter values for a set of SDC methods.
- Evaluate the disclosure risk and the information loss yielded by the methods for those parameterizations.

Now (Our contribution)

- Set a certain risk level.
- Find which parameter values are needed to attain that risk under different SDC methods.
- Evaluate the utility provided by each method

Note

This permits to rank methods according to their utility preservation, given a certain level of risk and a certain original data set.

Comparing the risk-utility trade-off in SDC (2/2)

If we have two functions

$$\mathcal{U} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$$

$$\mathcal{R} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$$

such that, for any given original data set **X** and anonymized data set **Y**,

- $\mathcal{U}(\mathbf{X}, \mathbf{Y})$ measures the utility of **Y** as a replacement for **X**.
- $\mathcal{R}(\mathbf{X}, \mathbf{Y})$ measures the disclosure risk of **Y** as a replacement for **X**.

Definition

Given two anonymization algorithms M^1 and M^2 of **X**, we say that M^1 is *more utility-preserving* than M^2 at risk level $(\mathcal{R}(\mathbf{X}, M^1_\alpha(\mathbf{X})) = \mathcal{R}(\mathbf{X}, M^2_\beta(\mathbf{X})))$

$$\mathcal{U}(\mathbf{X}, M^1_\alpha(\mathbf{X})) \geq \mathcal{U}(\mathbf{X}, M^2_\beta(\mathbf{X})),$$

And we say that M^1 is *less disclosive* than M^2 at utility level $(\mathcal{U}(\mathbf{X}, M^1_\alpha(\mathbf{X})) = \mathcal{U}(\mathbf{X}, M^2_\beta(\mathbf{X})))$ if

$$\mathcal{R}(\mathbf{X}, M^1_\alpha(\mathbf{X})) \leq \mathcal{R}(\mathbf{X}, M^2_\beta(\mathbf{X})),$$

Comparing the risk-utility trade-off in SDC (2/2)

If we have two functions

$$\mathcal{U} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$$

$$\mathcal{R} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$$

such that, for any given original data set **X** and anonymized data set **Y**,

- $\mathcal{U}(\mathbf{X}, \mathbf{Y})$ measures the utility of **Y** as a replacement for **X**.
- $\mathcal{R}(\mathbf{X}, \mathbf{Y})$ measures the disclosure risk of **Y** as a replacement for **X**.

Definition

Given two anonymization algorithms M^1 and M^2 of **X**, we say that M^1 is *more utility-preserving* than M^2 at risk level $(\mathcal{R}(\mathbf{X}, M^1_\alpha(\mathbf{X})) = \mathcal{R}(\mathbf{X}, M^2_\beta(\mathbf{X})))$

$$\mathcal{U}(\mathbf{X}, M^1_\alpha(\mathbf{X})) \geq \mathcal{U}(\mathbf{X}, M^2_\beta(\mathbf{X})),$$

And we say that M^1 is *less disclosive* than M^2 at utility level $(\mathcal{U}(\mathbf{X}, M^1_\alpha(\mathbf{X})) = \mathcal{U}(\mathbf{X}, M^2_\beta(\mathbf{X})))$ if

$$\mathcal{R}(\mathbf{X}, M^1_\alpha(\mathbf{X})) \leq \mathcal{R}(\mathbf{X}, M^2_\beta(\mathbf{X})),$$

Empirical measures of disclosure risk

We propose a risk measure based on **record linkage**.

Original data

Age	Height	Income	
55	1.80	2000	rank
44	1.60	1100	(4,2,5)
32	1.83	1500	
67	1.78	900	
36	1.56	750	
72	1.70	1350	
45	1.85	600	
23	1.71	400	

Masked data

Age	Height	Income	
54	1.76	1619	rank
41	1.51	1352	(4,1,6)
32	1.83	1090	
56	1.67	959	
40	1.62	826	
77	1.68	1608	
40	1.80	590	
13	1.72	131	

- We match any original record to the masked record at a minimum distance.
- The mean of these distances gives a measures of the risk of disclosure.

Note

We are not considering the number (or the proportion) of correct re-identifications (linkages).

Empirical measures of disclosure risk

We propose a risk measure based on **record linkage**.

Original data

Age	Height	Income	
55	1.80	2000	
44	1.60	1100	(4,2,5)
32	1.83	1500	
67	1.78	900	
36	1.56	750	
72	1.70	1350	
45	1.85	600	
23	1.71	400	

rank

Masked data

Age	Height	Income	
54	1.76	1619	
41	1.51	1352	(4,1,6)
32	1.83	1090	
56	1.67	959	
40	1.62	826	
77	1.68	1608	
40	1.80	590	
13	1.72	131	

rank

- We match any original record to the masked record at a minimum distance.
- The mean of these distances gives a measures of the risk of disclosure.

Note

We are not considering the number (or the proportion) of correct re-identifications (linkages).

Anonymizations

Age	Height	Income
55	1.80	2000
44	1.60	1100
32	1.83	1500
67	1.78	900
36	1.56	750
72	1.70	1350
45	1.85	600
23	1.71	400

Correlated noise addition

Age	Height	Income
54	1.76	1619
41	1.51	1352
32	1.83	1090
56	1.67	959
40	1.62	826
77	1.68	1608
41	1.80	590
13	1.72	131

Multiplicative noise

Age	Height	Income
63	1.88	2166
39	1.60	1040
33	1.71	1395
63	1.72	891
36	1.74	839
68	1.64	1135
50	1.88	685
22	1.65	500

Multivariate microaggregation

Age	Height	Income
53	1.78	1617
53	1.78	1617
53	1.78	1617
43	1.70	750
43	1.70	750
43	1.70	750
43	1.70	750
43	1.70	750

Rank swapping

Age	Height	Income
55	1.78	1500
36	1.60	900
32	1.85	2000
72	1.80	1100
44	1.56	600
67	1.71	1350
45	1.83	750
23	1.70	400

Anonymizations

Age	Height	Income
55 -1	1.80 -4	2000 -381
44 -3	1.60 -9	1100 +252
32	1.83	1500 -410
56 -11	1.78 -11	500 +59
40 +4	1.56 +6	750 +76
77 +5	1.70 -2	1350 +258
41 -4	1.80 -5	600 -10
13 -10	1.71 +1	400 -269

Correlated noise addition

Age	Height	Income
54	1.76	1619
41	1.51	1352
32	1.83	1090
56	1.67	959
40	1.62	826
77	1.68	1608
41	1.80	590
13	1.72	131

Multiplicative noise

Age	Height	Income
63	1.88	2166
39	1.60	1040
33	1.71	1395
63	1.72	891
36	1.74	839
68	1.64	1135
50	1.88	685
22	1.65	500

Multivariate normally distributed noise is added to the records in the collected data set,

$$Y = X + N(0, \gamma\Sigma),$$

where γ is an input parameter.

Multivariate microaggregation

Age	Height	Income
53	1.78	1617
53	1.78	1617
53	1.78	1617
43	1.70	750
43	1.70	750
43	1.70	750
43	1.70	750
43	1.70	750

Rank swapping

Age	Height	Income
55	1.78	1500
36	1.60	900
32	1.85	2000
72	1.80	1100
44	1.56	600
67	1.71	1350
45	1.83	750
23	1.70	400

Anonymizations

Age	Height	Income
55 * 1.1	1.80 * 1.07	2000 * 1.16
44 * 0.87	1.60 * 0.95	1100 * 0.95
32 * 1.03	1.83 * 0.92	1500 * 0.97
56 * 0.91	1.78 * 0.95	500 * 0.97
40 * 0.99	1.56 * 1.1	750 * 1.09
77 * 0.91	1.70 * 0.92	1350 * 0.85
41 * 1.08	1.80 * 1.04	600 * 1.07
13 * 0.99	1.71 * 0.93	400 * 0.94

Each attribute value $x_j^i \in \mathbf{X}$ is multiplied by $1 \pm N(0, s)$, where s is an input parameter.

Correlated noise addition

Age	Height	Income
54	1.76	1619
41	1.51	1352
32	1.83	1090
56	1.67	959
40	1.62	826
77	1.68	1608
41	1.80	590
13	1.72	131

Multiplicative noise

Age	Height	Income
63	1.88	2166
39	1.60	1040
33	1.71	1395
63	1.72	891
36	1.74	839
68	1.64	1135
50	1.88	685
22	1.65	500

Multivariate microaggregation

Age	Height	Income
53	1.78	1617
53	1.78	1617
53	1.78	1617
43	1.70	750
43	1.70	750
43	1.70	750
43	1.70	750
43	1.70	750

Rank swapping

Age	Height	Income
55	1.78	1500
36	1.60	900
32	1.85	2000
72	1.80	1100
44	1.56	600
67	1.71	1350
45	1.83	750
23	1.70	400

Anonymizations

Age	Height	Income
55	1.80	2000
44	1.60	1100
32	1.83	1500
67	1.78	900
36	1.56	750
72	1.70	1350
45	1.85	600
23	1.71	400

Correlated noise addition

Age	Height	Income
54	1.76	1619
41	1.51	1352
32	1.83	1090
56	1.67	959
40	1.62	826
77	1.68	1608
41	1.80	590
13	1.72	131

Multiplicative noise

Age	Height	Income
63	1.88	2166
39	1.60	1040
33	1.71	1395
63	1.72	891
36	1.74	839
68	1.64	1135
50	1.88	685
22	1.65	500

We partition the records of **X** in groups of k or more records, where records in a group are as similar as possible, and we replace each record by the corresponding centroid.

Multivariate microaggregation

Age	Height	Income
53	1.78	1617
53	1.78	1617
53	1.78	1617
43	1.70	750
43	1.70	750
43	1.70	750
43	1.70	750
43	1.70	750

Rank swapping

Age	Height	Income
55	1.78	1500
36	1.60	900
32	1.85	2000
72	1.80	1100
44	1.56	600
67	1.71	1350
45	1.83	750
23	1.70	400

Anonymizations

Age	Height	Income
55	1.80	2000
44	1.60	1100
32	1.83	1500
67	1.78	900
36	1.56	750
72	1.70	1350
45	1.85	600
23	1.71	400

Correlated noise addition

Age	Height	Income
54	1.76	1619
41	1.51	1352
32	1.83	1090
56	1.67	959
40	1.62	826
77	1.68	1608
41	1.80	590
13	1.72	131

Multiplicative noise

Age	Height	Income
63	1.88	2166
39	1.60	1040
33	1.71	1395
63	1.72	891
36	1.74	839
68	1.64	1135
50	1.88	685
22	1.65	500

Independently for each attribute, this method swaps the attribute's values within a restricted range: the ranks of two swapped values cannot differ by more than p% of the total number of records, where p is an input parameter.

Multivariate microaggregation

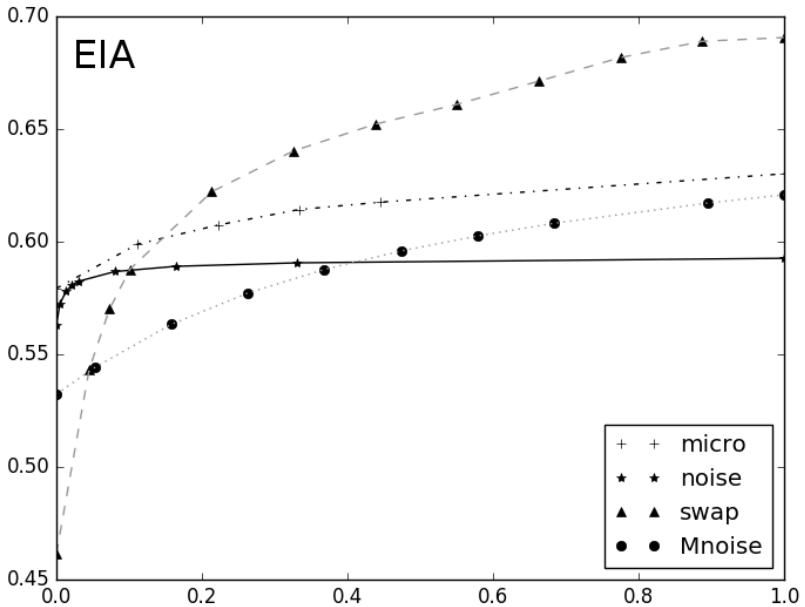
Age	Height	Income
53	1.78	1617
53	1.78	1617
53	1.78	1617
43	1.70	750
43	1.70	750
43	1.70	750
43	1.70	750
43	1.70	750

Rank swapping

Age	Height	Income
55	1.78	1500
36	1.60	900
32	1.85	2000
72	1.80	1100
44	1.56	600
67	1.71	1350
45	1.83	750
23	1.70	400

Disclosure risk assessment

Experiments are conducted by taking as original data the "Census" (1080×13) and "EIA" (4092×11) data sets



For the "EIA" data set, a possible match occurs at $\mathcal{R}(X, Y) = 0.58$:

1. multivariate microaggregation with $k = 5$,
2. correlated noise addition with $\gamma = 0.05$,
3. rank swapping with $p = 0.08$, and
4. multiplicative noise with $s = 0.3$.

Utility loss assessment

Note

The known utility measures considered are Propensity Score and Earth mover's distance.

Methods	CENSUS		EIA	
	Propensity	EMD	Propensity	EMD
Microaggregation	$\mathbf{4.28 \times 10^{-4}}$	**0.16**	$\mathbf{2.17 \times 10^{-5}}$	**0.040**
Correlated noise addition	3.83×10^{-2}	0.38	4.22×10^{-5}	0.065
Rank swapping	3.51×10^{-3}	0.28	9.01×10^{-4}	0.091
Multiplicative noise	6.3×10^{-3}	0.29	9.85×10^{-5}	0.066

Conclusions

- We have described a methodology to compare different anonymizations in terms of the risk-utility trade-off they attain.
- We have proposed a disclosure risk measure based on record linkage.
- The results depends on the data set considered.
- The best strategy seems to be to make several anonymizations at the desired level of disclosure risk and select the one that has the greatest utility.

MANY THANKS

THANKS