ANONYMIZATION METHODS AS TOOLS FOR FAIRNESS

Salvatore Ruggieri

KDD Lab
Department of Computer Science
University of Pisa, Italy

2

Message of the talk

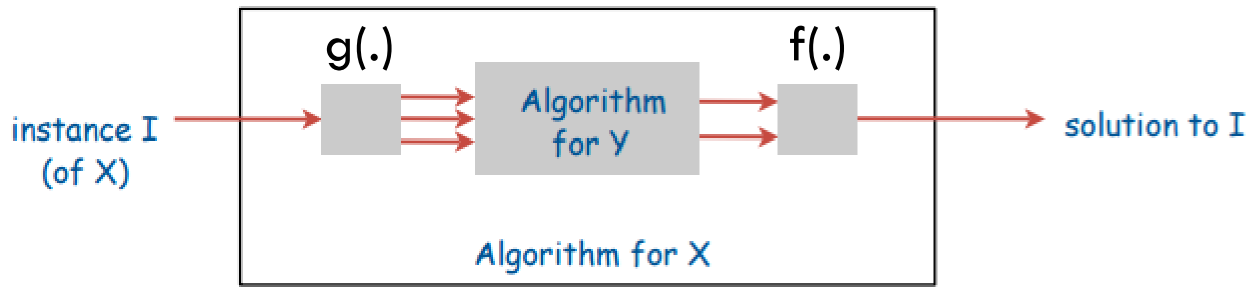
Motivation: risks in data publishing (and in learning from data)

- Privacy risks
  - re-identification or attribute inference
- Discrimination risks?
  - discriminatory decisions
    - An employer may notice from public census data that the race or sex of workers act as proxy of the workers' productivity.
      - The employer may then use those visible traits for hiring decisions.
    - A machine learning model to profile applicants to a bank loan may learn from past application records some patterns of traditional prejudices
      - The model may predict not to loan to a minority group.
- Solutions?
  - dataset sanitization for discrimination prevention

Reductions of problems

Problem X reduces to problem Y if an algorithm that solves Y can be used to solve X

- $sol_X(I) = f(\ sol_Y(\ g(I)\ )\ )$



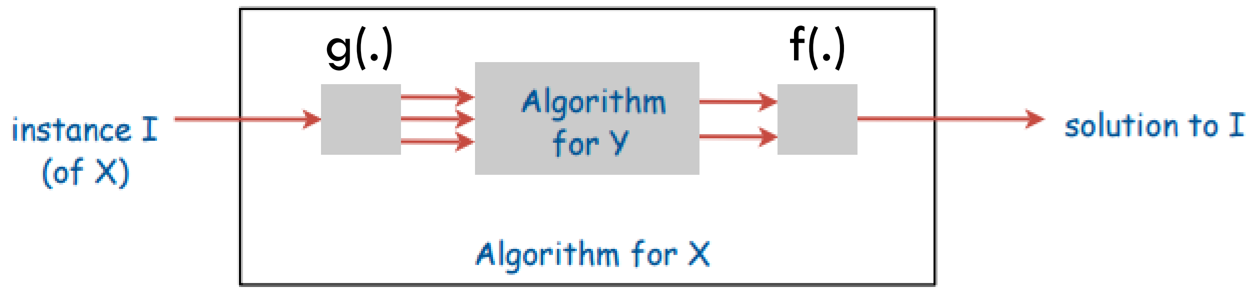
- Widely used concept in
  - Computability
  - Computational complexity
  - Programming
  - …
- Assume that g() and f() are «simple»
  - X is «easier or equal than» than Y
  - If Y reduces to X also

    then X and Y are «equivalent»

Reductions of problems

Problem X reduces to problem Y if an algorithm that solves Y can be used to solve X

- $sol_X(I) = f(\, sol_Y(\, g(I)\,)\,)$



- X = sanitize a dataset for discrimination prevention wrt α-protection
- Y = sanitize a dataset for privacy protection wrt t-closeness

- Message of the talk: X and Y are «equivalent» (in a weak sense)
  - Main reference: *S. Ruggieri. Using t-closeness anonymity to control for non-discrimination. Transactions on Data Privacy 7 (4) : 301-325, 2014.*

Discrimination measures

Discrimination measures

- What is the degree of discrimination suffered?
  - Legal principle of *proportional representation*

city=NYC	benefit denied	benefit granted	total
women	6	4	10
men	1	4	5
total	7	8	15

p_1 = proportion of benefit denied to women = 6/10 = 60%

p_2 = proportion of benefit denied to men = 1/5 = 20%

- Risk difference (RD) is $p_1 - p_2 = 40\%$
- Risk ratio (RR) is $p_1 / p_2 = 3$
- Relative chance (RC) is $(1-p_1)/(1-p_2) = 0.5$
- Odds ratio (OR) is RR / RC = 6

Discrimination measures

- What is the degree of discrimination suffered?
  - Legal principle of *proportional representation*

city=NYC	benefit denied	benefit granted	total
women	6	4 (b)	10
men	1	4	5
total	7	8 (m_2)	15

p_0 = proportion of women in the overall population = 10/15 = 67%

p = proportion of women in the «benefit granted» population = 4/8 = 50%

- Example: jury selection
- Castaneda rule in the U.S. (1977): $p_0 m_2 - b \leq 3\sigma$
  - Binomial distribution $\sigma = \sqrt{m_2 p_0 (1 - p_0)}$

Extensions to account for:

- Lack of comparison term
  - occurs when there are no men (or women) in the context

ZIP=100	benefit denied	benefit granted	total
women	6	4	10
men	0	0	0
total	6	4	10

p_1 = proportion of benefit denied to women = 6/10 = 60%

p_2 = undefined = p_- = proportion of benefit denied to women in the whole dataset

- All discrimination measures extends smoothly

Extensions to account for:

- Random effects rather than explicit discrimination:
  - Confidence intervals for discrimination measures [Pedreschi et al. 2009]
- Causality in discrimination conclusions:
  - Do women from NYC have the same characteristics of men they are compared with? Or do they differ as per skills or other admissible reasons?
  - Propensity score weighting [Ridgeway2006]

city=NYC	benefit denied	benefit granted	total
women	6	4	10
men	1	4	5
total	7	8	15

Weighted risk difference (wRD) is $p_1 - p_w = 40\%$

$$p_w = \sum_{x \in men \cap denied} w(x) \,/\, \sum_{x \in men} w(x)$$

- $\Pr(x|women) = w(x)\Pr(x|man)$
- $w(x) = \Pr(woman|x) / (1 - \Pr(woman|x))$

Propensity score weights

Discrimination in a dataset

α-protection

city=NYC birth=1965	benefit denied	benefit granted	total
women	1	0	1
men	0	1	1
total	1	1	2

RD = 100% - 0% = 100%

PND attributes (City, Birth date) — PD attribute (Sex)

City	Birth date	Sex	Benefit
NYC	1973	M	No
NYC	1965	F	No
NYC	1965	M	Yes
LA	1973	M	No
…	…	…	…

- A non-empty 4-fold contingency table is α-protective if the discrimination measure is lower or equal than a threshold α
- A dataset if a-protective if all of its non-empty 4-fold contingency tables are α-protective
  - for any **subset** (or conjunction) of PND items

Local approaches [RPT2010@TKDD]

- Extract classification rules:

  gender=women, **B** $\rightarrow$ benefit=denied

  - with **B** providing a context of discrimination
    - E.g., **B** $\equiv$ city = NYC
  - with measure > α

city=NYC	benefit denied	benefit granted	total
women	6	4	10
men	1	4	5
total	7	8	15

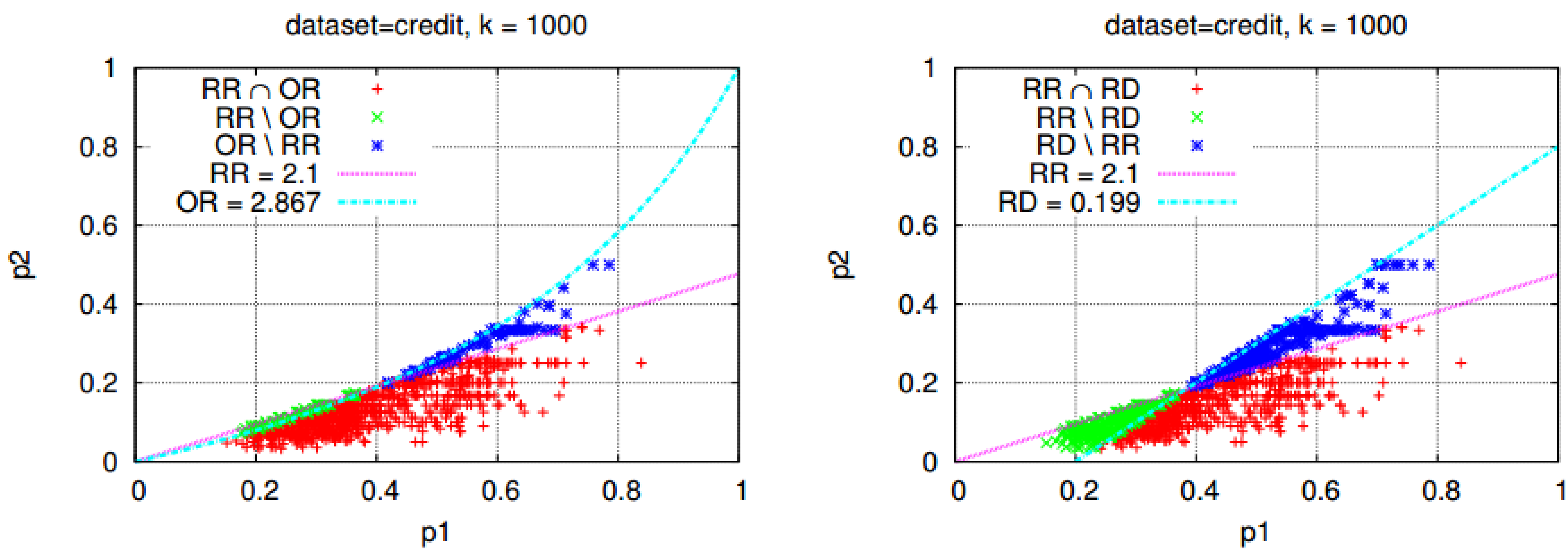
- Notice
  - cover(**B**) is the context of analysis
  - **B** can be a closed itemset (all distinct covers!)

A Java library: dd


```
DDTable tb = new DDTable();
tb.loadFromArff("credit");
tb.setDiscMetaData("personal_status=female_div_or_dep_or_mar,
                    foreign_worker=yes", "class=bad", 20);
tb.closedItemset = true;
tb.extractItemsets();
tb.initScan();
double largest = -1;
ContingencyTable ct = null;
while( (ct = tb.nextCT()) != null) {
        double diff = ct.rd();
        if( diff > largest )
                largest = diff;
}
tb.endScan();
```

- Download it from
  - http://www.di.unipi.it/~ruggieri/software.html

Level curves of top-k tables [PRT@SAC2012]



Privacy in a dataset

k-anonimity

- A partition-based measure of risk in data disclosure

QI attributes (ZIP, Birth date, Sex) — sensitive attribute (Desease)

q-block size = 3

ZIP	Birth date	Sex	Desease
100	1965	F	Yes
100	1965	F	No
100	1965	F	No
101	1973	M	No
…	…	…	…

- Q-block = rows with same values for all QIs
- A q-block is k-anonimous if its size is at least k
- A dataset is k-anonimous if every q-block is k-anonimous
  - Any individual cannot is indistinguishable from k-1 others

t-closeness

A partition-based measure of attribute inference risk in data disclosure

QI attributes: ZIP, Birth date, Sex — sensitive attribute: Desease

ZIP	Birth date	Sex	Desease
100	1965	F	Yes
100	1965	F	No
100	1965	F	No
101	1973	M	No
…	…	…	…

q-block (first three rows)
size = 3
p = 33.3%

- A q-block is t-close if it maintains the proportion of sensitive values
  - p = proportion of Yes in the q-block   p* = proportion of Yes in the whole dataset
  - Condition: |p-p*| < t
- A dataset is t-close if every q-block is t-close

Differences between t-close and a-protect

- Distributions
  - t-closeness, **single:** $|p\text{-}p^*| < t$ for every q-block
  - α-protection wrt RD, **joint:** $p_1 - p_2$ for every 4-fold c.t.
- Monotonicity property
  - t-closeness fixes **all** values of QI attributes
  - α-protection fixes **some** values of QI attributes
    - If the rows s.t. city=NYC, birth=X are α-protective , for all X, then the rows s.t. city=NYC may be not α-protective
    - Sympson's paradox
- t-closeness and α-protection are not equivalent models

Sympson's paradox

$$RD = p_1 - p_2$$

dept	*sex*	*admitted*
A	female	no
A	female	no
A	female	no
A	female	no
A	female	yes
A	female	yes
A	female	yes
A	male	no
A	male	yes
A	male	yes

PND itemset *dept*=*A*
$RD = 4/7 - 1/3 = 0.238$

dept	*sex*	*admitted*
B	female	no
B	female	yes
B	male	no
B	male	no
B	male	yes
B	male	yes
B	male	yes
B	male	yes
B	male	yes
B	male	yes

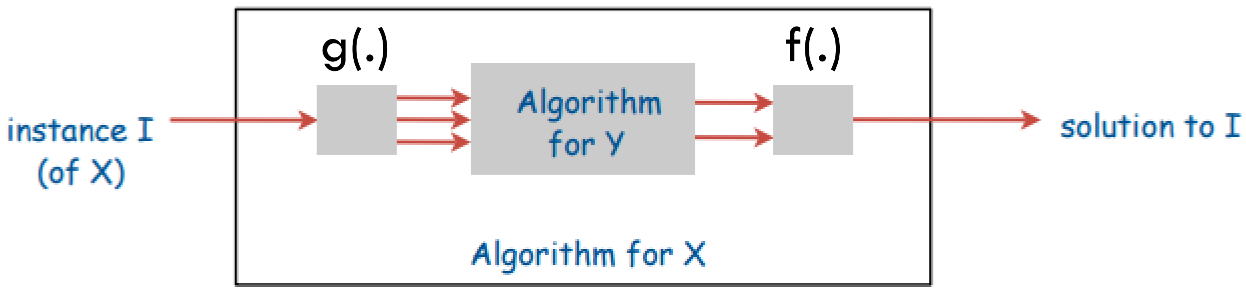
PND itemset *dept*=*B*
$RD = 1/2 - 2/8 = 0.25$

PND itemset empty
(both departments)
$RD = 5/9 - 3/11 = 0.283$

t-closeness reduces to α-protection

Reductions of problems

- X = sanitize a dataset for privacy protection wrt t-closeness
- Y = sanitize a dataset for discrimination prevention wrt α-protection



- g(I) = I+2 new attributes D1,D2 with PND = QI PD = {D1, D2}

QI attributes (ZIP, Birth date, Sex); sensitive attribute (Desease)

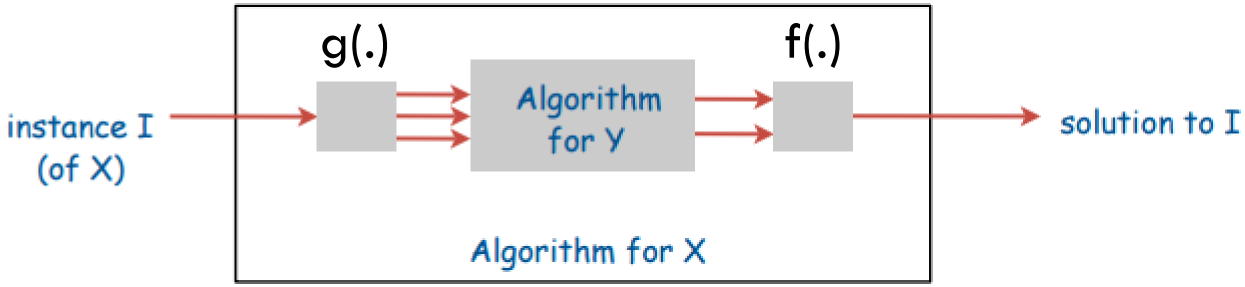
ZIP	Birth date	Sex	Desease
100	1965	F	Yes
100	1965	F	No
100	1965	F	No
101	1973	M	No
…	…	…	…

PND (ZIP, Birth date, Sex); PD (D1, D2); decision attribute (Desease)

ZIP	Birth date	Sex	D1	D2	Desease
100	1965	F	True	False	Yes
100	1965	F	True	False	No
100	1965	F	True	False	No
101	1973	M	True	False	No
…	…	…	…	…	…

Reductions of problems

- X = sanitize a dataset for privacy protection wrt t-closeness
- Y = sanitize a dataset for discrimination prevention wrt α-protection



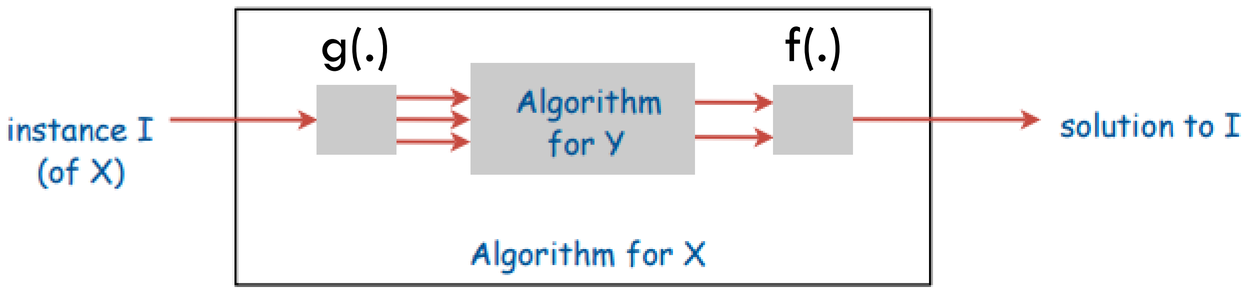
- g(I) = I+2 new attributes D1,D2 with PND = QI PD = {D1, D2}
- contingency tables have no comparison term!
  - Use $p_- =$ proportion of Yes in the whole dataset = p^*
- Fix QIs (eg., ZIP=100, Birth=1965, Sex = F)
  - RD = $p_1 - p_- < \alpha$ for D1=True
  - RD = $p_- - p_1 < \alpha$ for D2=True
  - Thus, $| p_1 - p_- | = | p - p^* | < \alpha$
- α-protection implies t-closeness, for t = α

QI=…	desease =Yes	desease =No	total
D1=True	1	2	3
D1=False	0	0	0
total	1	2	3

α-protection reduces to t-closeness

Reductions of problems

- X = sanitize a dataset for privacy protection wrt t-closeness
- Y = sanitize a dataset for discrimination prevention wrt α-protection



- g(I) = I with QI= PND+PD

Notice that: $p^* = p_{-}$

PND		PD	decision attribute
City	Birth date	Sex	Benefit
NYC	1973	M	No
NYC	1965	F	No
NYC	1965	M	Yes
LA	1973	M	No
…	…	…	…

QI			sensitive attribute
City	Birth date	Sex	Benefit
NYC	1973	M	No
NYC	1965	F	No
NYC	1965	M	Yes
LA	1973	M	No
…	…	…	…

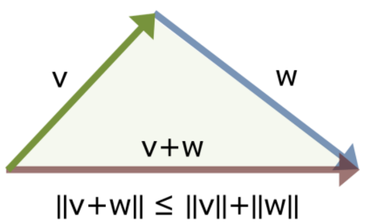
city=NYC	benefit denied	benefit granted	total
women	6	4	10
men	1	4	5
total	7	8	15

p_1 = proportion of benefit denied to women = 6/10 = 60%

p_2 = proportion of benefit denied to men = 1/5 = 20%

Risk difference (RD) is p1 - p2 = 40%

- Triangle inequality
  - RD = p1 - p2 ≤ |p1 – p*| + |p2 – p*|



- If the dataset is t-close …
  - where
    - QI attributes = PND attributes + PD attribute
    - Sensitive attribute = decision attribute
  - RD = p1 - p2 ≤ |p1 – p*| + |p2 – p*| ≤ 2t
- … then it is 2t-protective

Formal results

Theorem 10. *Fix as QIs the set of PND attributes plus the PD attribute, and as sensitive attribute the decision attribute. If the table is t-close then it is $bd_f(t)$-protective w.r.t. $f \in \{ED, RD\}$, where $bd_{RD}(t) = bd_{ED}(t) = min\{2t, t + \hat{p}_-, 1\}$ and $\hat{p}_- = min\{p_-, 1 - p_-\}$.*

A dataset does not contain discrimination (more than $bd_f(t)$) if an attacker cannot be confident (more than a threshold t) on the decision assigned to an individual by exploiting the differences in the fraction of positive and negative decisions between the protected and the unprotected groups.

- The role of an ``attacker" here is played by the anti-discrimination analyst, whose objective is to unveil from data a context where negative decisions are biased against the protected group.

Formal results

Corollary 14. *Fix as QIs the set of PND attributes plus the PD attribute, and as sensitive attribute the decision attribute. If the table is t-close then every PND itemset, possibly with disjunctive items, is $bd_f(t)$-protective w.r.t. $f \in \{ED, RD\}$ and $bf_f()$ as in Thm. 10.*

Disjunctive items $A=v_1 \ \vee \ldots \ A=v_n$

- *Ex., age in [25,30] is a disjunctive item*

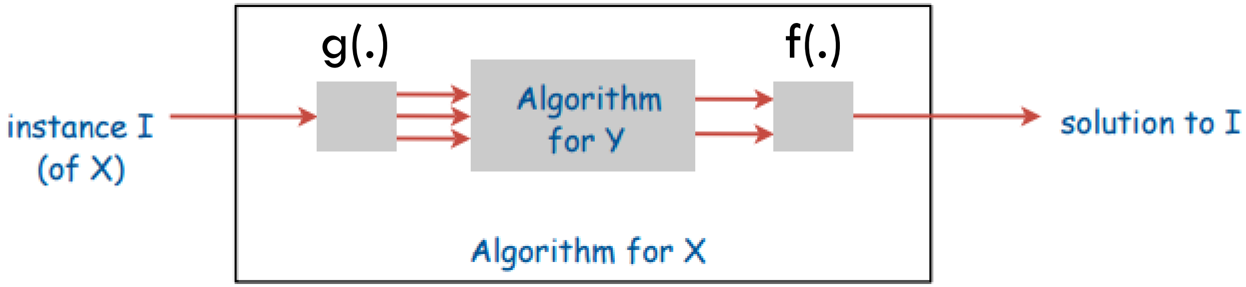
Stronger conclusion than α-protection: context with conjunctions of possibly disjunctive items are covered!!

Patterns used in decision trees, association rule classifiers !!

city=NYC, age in [25,30]	benefit denied	benefit granted	total
women	6	4	10
men	1	4	5
total	7	8	15

Reductions of problems

- X = sanitize a dataset for discrimination prevention wrt α-protection
- Y= sanitize a dataset for privacy protection wrt t-closeness



- The reduction show ONE way to sanitize X using Y. Can all sanitized versions of I be obtained through reduction?

Sympson's paradox

$RD = p_1 - p_2 \qquad p^* = 0.4$

$p = 0.57$

$p = 0.33$

dept	*sex*	*admitted*
A	female	no
A	female	no
A	female	no
A	female	no
A	female	yes
A	female	yes
A	female	yes
A	male	no
A	male	yes
A	male	yes

PND itemset *dept*=*A*
$RD = 4/7 - 1/3 = 0.238$

$p = 0.5$

$p = 0.25$

dept	*sex*	*admitted*
B	female	no
B	female	yes
B	male	no
B	male	no
B	male	yes
B	male	yes
B	male	yes
B	male	yes
B	male	yes
B	male	yes

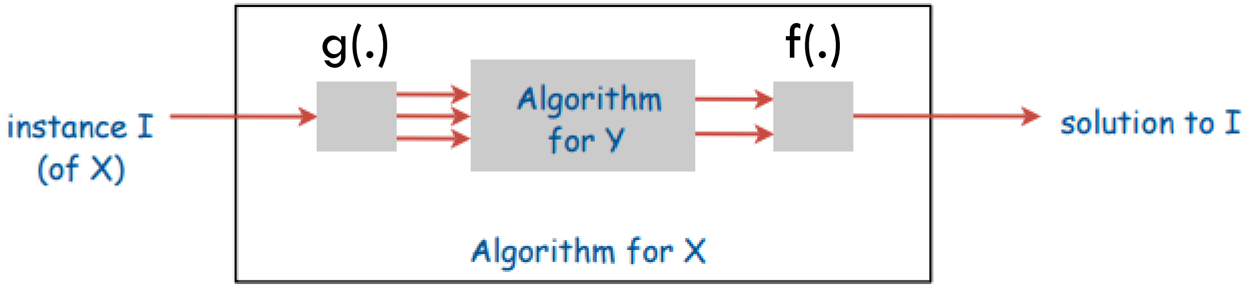
PND itemset *dept*=*B*
$RD = 1/2 - 2/8 = 0.25$

PND itemset empty
(both departments)
$RD = 5/9 - 3/11 = 0.283$

The dataset is 0.17-close

Reductions of problems

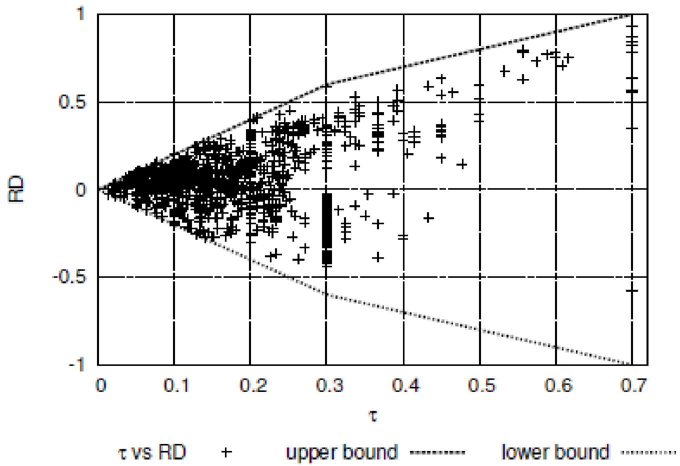
- X = sanitize a dataset for discrimination prevention wrt α-protection
- Y= sanitize a dataset for privacy protection wrt t-closeness

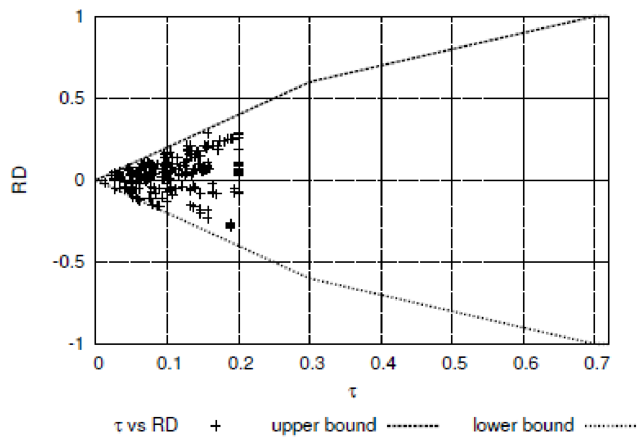


- The reduction show ONE way to sanitize X using Y. Can all sanitized versions of I be obtained through such a reduction?
- Assume the answer is positive:
  - Let I be the Sympon's paradox dataset and α = 0.283 (I is already α-protective)
  - There exists a «empty» sanitization Y of I s.t. 2t $\leq$ 0.283, i.e. t $\leq$ 0.141
  - Impossible because I is only 0.17-close
- Message of the talk: X and Y are «equivalent» (in a weak sense)

Main results and application

- t-closeness implies bd(t)-protection
  - where bd() is a function dependent on the discrimination measure (RD,RR,OR, …)
  - the bound bd(t) can be reached in limit cases



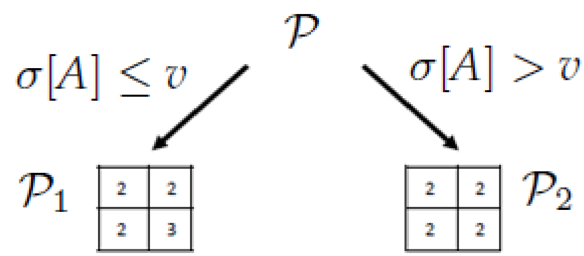


- Application:
  - data anonymization methods can be applied to sanitization wrt non-discrimination

Multidimensional recoding: dMondrian

Algorithm 1 dMondrian.Anonymize($\mathcal{P}$, t)

```
if no d-allowable cut for P then
    return PND_ranges(P)
else
    A ← choose_PND_dimension(P)
    v ← find_median(P, A)
    P1 ← {σ ∈ P | σ[A] ≤ v }
    P2 ← {σ ∈ P | σ[A] > v }
    return Anonymize(P1, t) ∪ Anonymize(P2, t)
end if
```



Definition 5.1: Let p_- be the fraction of the negative decision in a relational table. A cut $V \leq v$ is *d-allowable* if the 4-fold contingency tables of $\mathcal{P}_1$ and $\mathcal{P}_2$ satisfy both $|p_1 - p_-| \leq t$ and $|p_2 - p_-| \leq t$.

Example

Sample dataset

ID	*purpose*	*emp*	*sex*	*decision*
1	housing	no	female	-
2	housing	no	female	-
3	housing	no	female	+
4	housing	no	male	-
5	housing	no	male	+
6	housing	yes	female	-
7	housing	yes	female	+
8	housing	yes	female	+
9	housing	yes	male	-
10	housing	yes	male	-
11	housing	yes	male	+
12	housing	yes	male	+
13	car	no	female	+
14	car	no	male	-
15	car	no	male	+
16	car	yes	female	-
17	car	yes	male	+

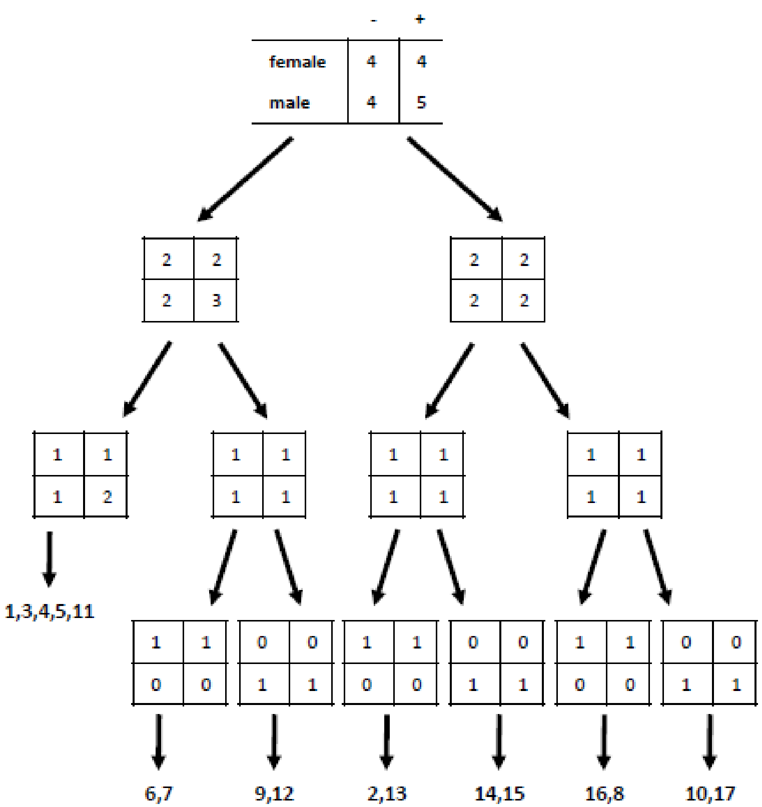
Output of dMondrian

ID	*purpose*	*emp*	*sex*	*decision*
1	housing-car	no	female	-
2	housing-car	no	female	-
3	housing-car	no	female	+
13	housing-car	no	female	+
4	housing-car	no	male	-
14	housing-car	no	male	-
5	housing-car	no	male	+
15	housing-car	no	male	+
6	housing-car	yes	female	-
16	housing-car	yes	female	-
7	housing-car	yes	female	+
8	housing-car	yes	female	+
9	housing-car	yes	male	-
10	housing-car	yes	male	-
11	housing-car	yes	male	+
12	housing-car	yes	male	+
17	housing-car	yes	male	+

emp=no	decision		
sex	-	+	
female	2	2	4
male	2	2	4
	4	4	8

emp=yes	decision		
sex	-	+	
female	2	2	4
male	2	3	5
	4	5	9

Bucketization & redistrib.: dSabre



Output of dSabre

ID	*purpose*	*emp*	*sex*	*decision*
1	housing	no-yes	female	-
3	housing	no-yes	female	+
4	housing	no-yes	male	-
5	housing	no-yes	male	+
11	housing	no-yes	male	+
6	housing	yes	female	-
7	housing	yes	female	+
9	housing	yes	male	-
12	housing	yes	male	+
2	housing-car	no	female	-
13	housing-car	no	female	+
14	car	no	male	-
15	car	no	male	+
16	housing-car	yes	female	-
8	housing-car	yes	female	+
10	housing-car	yes	male	-
17	housing-car	yes	male	+

Effective to reduce discrimination

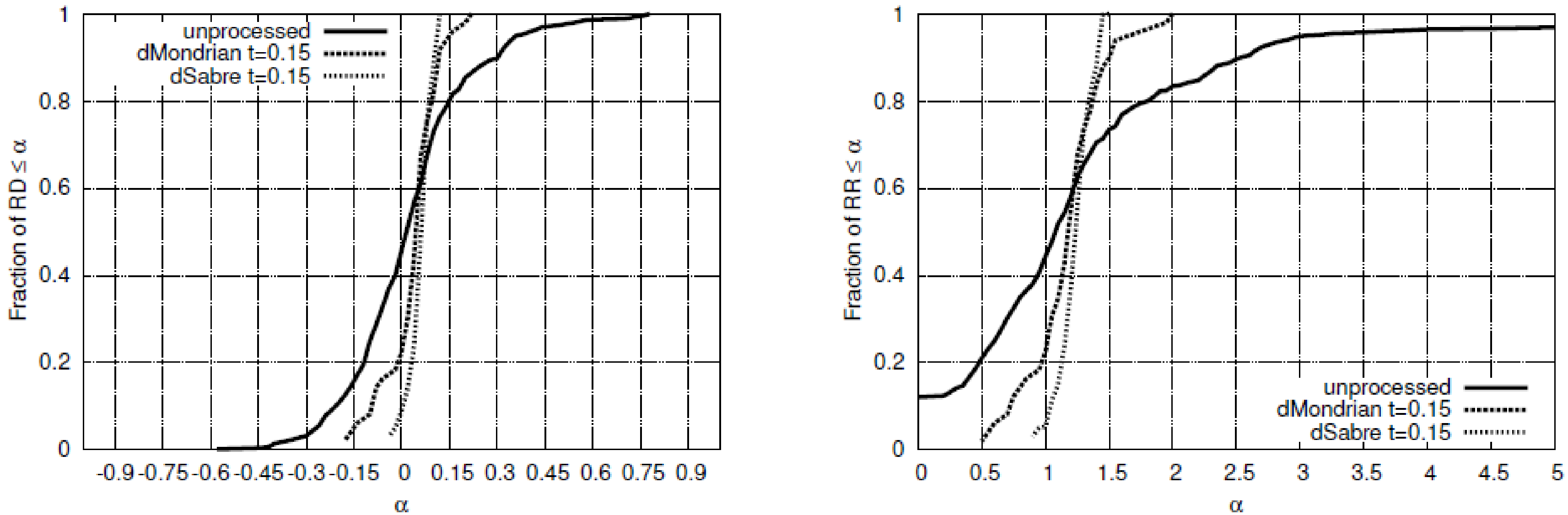
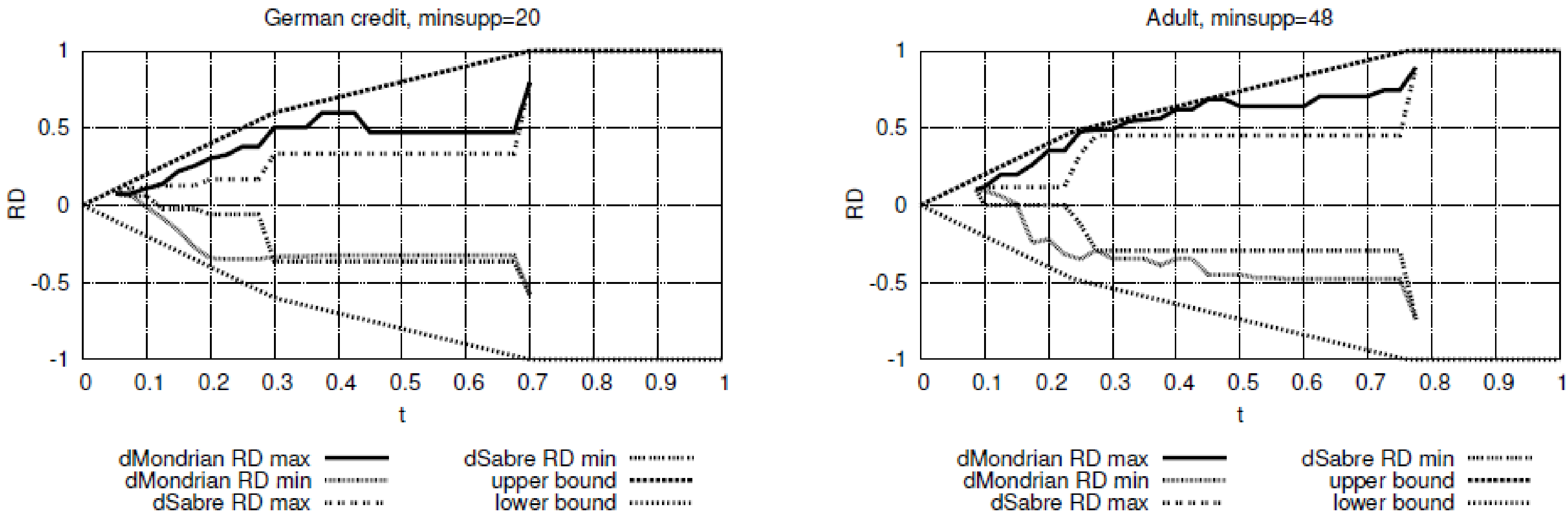
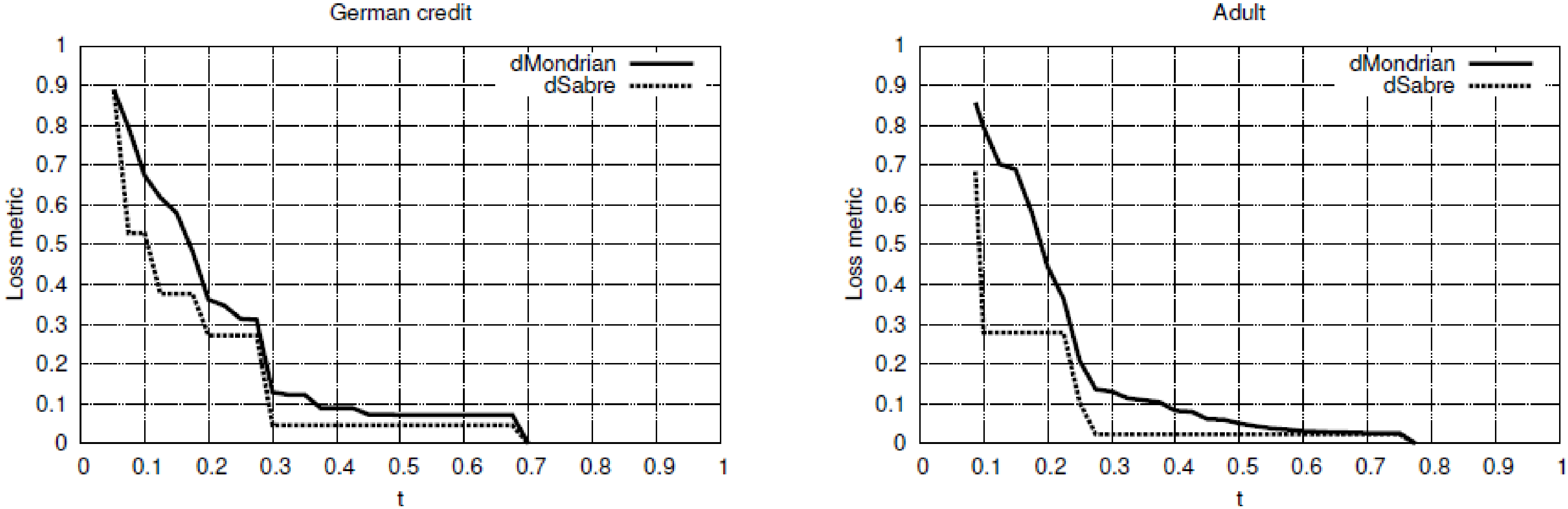


Figure 7: *German credit* dataset. Distributions of RD and RR values.

dSabre is better



also regarding information loss



$$\mathrm{LM} = \sum_{\text{QI itemset } \mathbf{Q}} supp(\mathbf{Q})\, L(\mathbf{Q}) \qquad L(\mathbf{Q}) = \sum_{i=1,\ldots,N-1} \frac{range(v_i) - 1}{|dom(A_i)| - 1}$$

Quality: median relative error

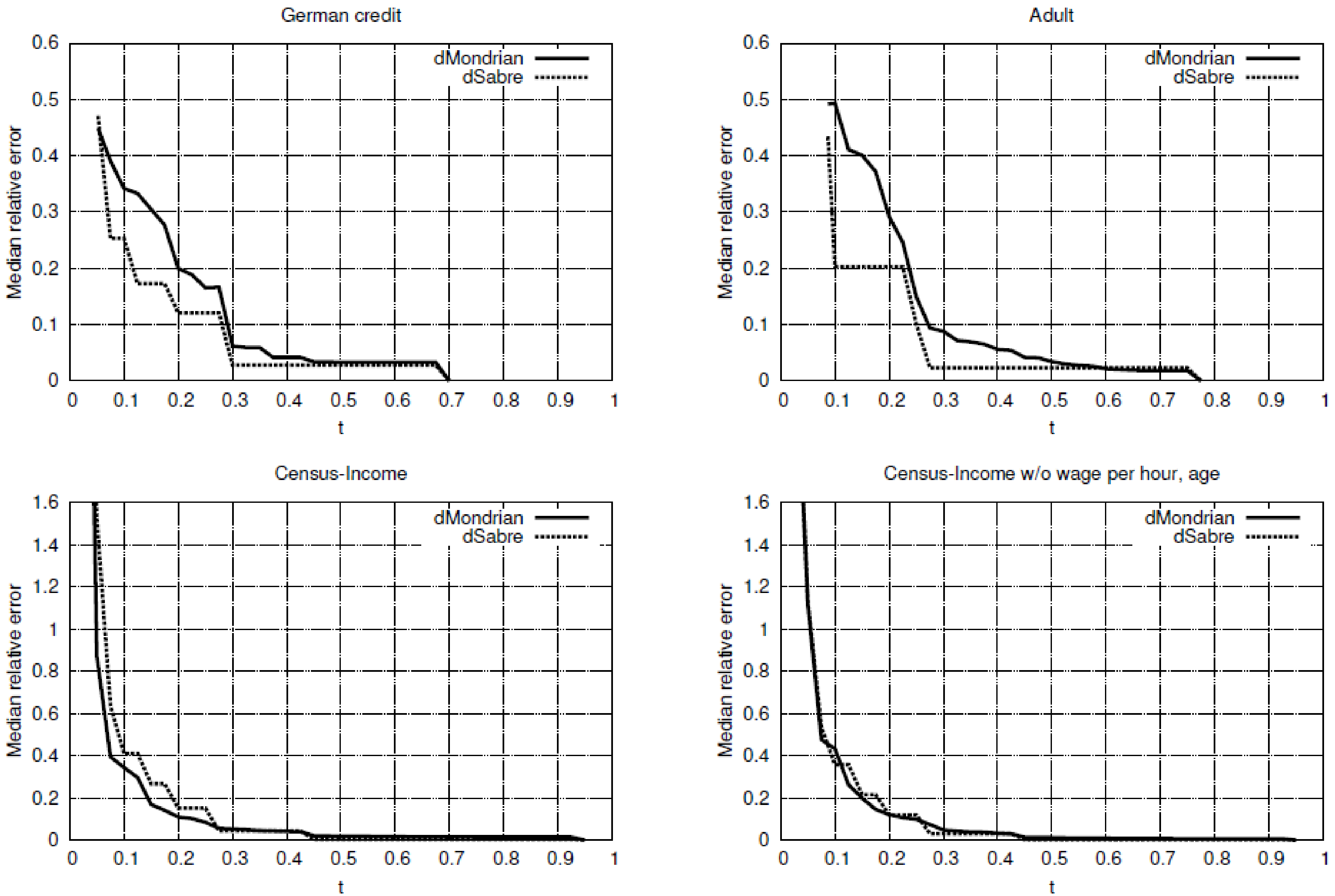
- Count queries are basic elements in classifier construction, e.g., in decision tree or association rule classifiers

Class attribrute

```
SELECT COUNT(*)
FROM dataset
WHERE A_π1 in [v1, w1] AND ... AND A_πn in [vn, wn] AND A_N in [vn+1, wn+1]
```

- Relative error of a count query is $|est - prec|/prec$
  - prec = count over original dataset
  - est = count over sanitized dataset (uniform distribution of values)
- Median relative error is
  - the median error on 10K randomly generated count queries

but more sensitive to high dim/card



Related work

- Incognito-like search for k-anonymous & a-protective sanitization
  - [Haijan & al. @ DAMI 2014]
- Impact of k-anonimity sanitization on a-protection
  - [Haijan & Domingo-Ferrer @ DPADM 2012]
- Techniques for achieving both k-anonimity and a-protection in knowledge disclosure
  - [Haijan et al. @ DPADM 2012]
- Non-discriminatory (fair) classification as a generalization of differential privacy
  - [Dwork et al. @ ITCS 2012], [Zemel et al. @ ICML 2013]

Thanks, questions?