PAIS 2015

Transparency and disclosure risk in data privacy

Vicenç Torra[1]

March, 2015

[1] School of Informatics, University of Skövde, Sweden

Outline

Quantitative measures of risk: record linkage

Transparency principle: publication of data processing methods

- a good practice on data privacy
- similar to the one in cryptography

Risk needs to consider the transparency principle

Outline

1. Introduction
   - Masking methods
   - Disclosure risk assessment
2. Transparency
   - Definition
   - Attacking Rank Swapping
   - Attacking Microaggregation
3. Worst-case scenario when measuring disclosure risk
4. Summary

Introduction

Masking methods

Masking methods

Masking methods.

- Perturbative
- Non-perturbative
- Synthetic data generators

Review

- Microaggregation
- Rank swapping

Rank Swapping

Rank swapping

- For ordinal/numerical attributes
- Applied attribute-wise

Data: $(a_1, \ldots, a_n)$: original data; p: percentage of records
Order $(a_1, \ldots, a_n)$ in increasing order (i.e., $a_i \leq a_{i+1}$) ;
Mark a_i as unswapped for all i ;
for $i = 1$ **to** n **do**
  if a_i *is unswapped* **then**
    Select ℓ randomly and uniformly chosen from the limited range $[i + 1, \min(n, i + p * |X|/100)]$;
    Swap a_i with a_ℓ ;
Undo the sorting step ;

Rank Swapping

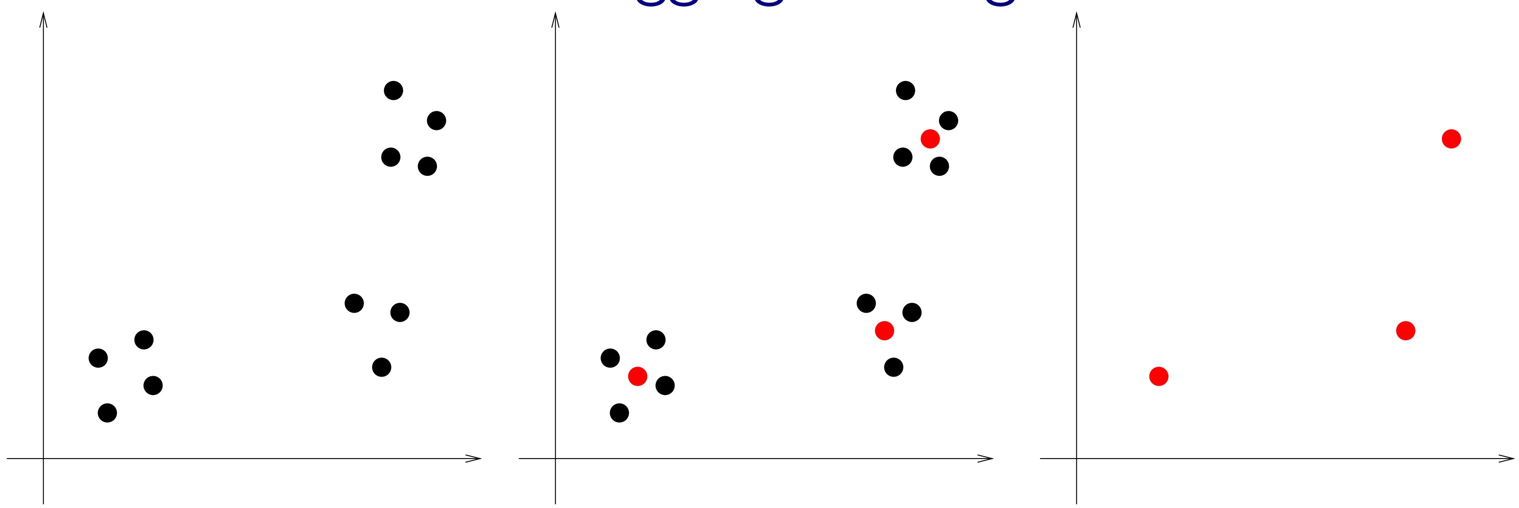
Rank swapping.

- Marginal distributions not modified.
- Correlations between the attributes are modified
- Good trade-off between information loss and disclosure risk

Microaggregation

Microaggregation.

- Case of two attributes microaggregated together



Microaggregation

Microaggregation. Application.

- k: number of records in the cluster
- Partition of the attributes

v_1	v_2	v_3	v_4	v'_1	v'_2	v'_3	v'_4
1	1	1	1	1.66667	2	1.33333	1.66667
2	2	1	2	1.66667	2	1.33333	1.66667
2	3	1	6	1.66667	2	2.33333	5.66667
2	9	1	10	3	7.33333	1.66667	9.66667
3	6	2	2	3	7.33333	1.33333	1.66667
4	1	2	9	4.33333	5	1.66667	9.66667
4	6	2	10	4.33333	5	1.66667	9.66667
4	7	3	2	3	7.33333	2.33333	5.66667
5	8	3	9	4.33333	5	2.33333	5.66667
6	8	4	7	7.66667	8.66667	6	5
8	1	7	2	8.66667	2.66667	6	5
8	9	7	6	7.66667	8.66667	6	5
9	3	8	1	8.66667	2.66667	8.66667	1.33333
9	4	8	2	8.66667	2.66667	8.66667	1.33333
9	9	10	1	7.66667	8.66667	8.66667	1.33333

Introduction

Disclosure risk assesment

Disclosure risk assesment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
  - Attribute disclosure:
    - ⋆ Increase knowledge about an attribute of an individual
  - Identity disclosure:
    - ⋆ Find/identify an individual in a masked file

Disclosure risk assesment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures

Disclosure risk assesment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures
  (minimize information loss vs. multiobjetive optimization)

Disclosure risk assesment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures
  (minimize information loss vs. multiobjetive optimization)

Examples.

- Boolean definitions of risk
  - k-Anonymity (Boolean definition / identity disclosure)
  - differential privacy (Boolean definition / attribute disclosure)
- Quantitative measures of risk
  - Re-identification / Record linkage (for identity disclosure)
  - Uniqueness (for identity disclosure)
  - Interval disclosure (for attribute disclosure)

Disclosure risk assesment

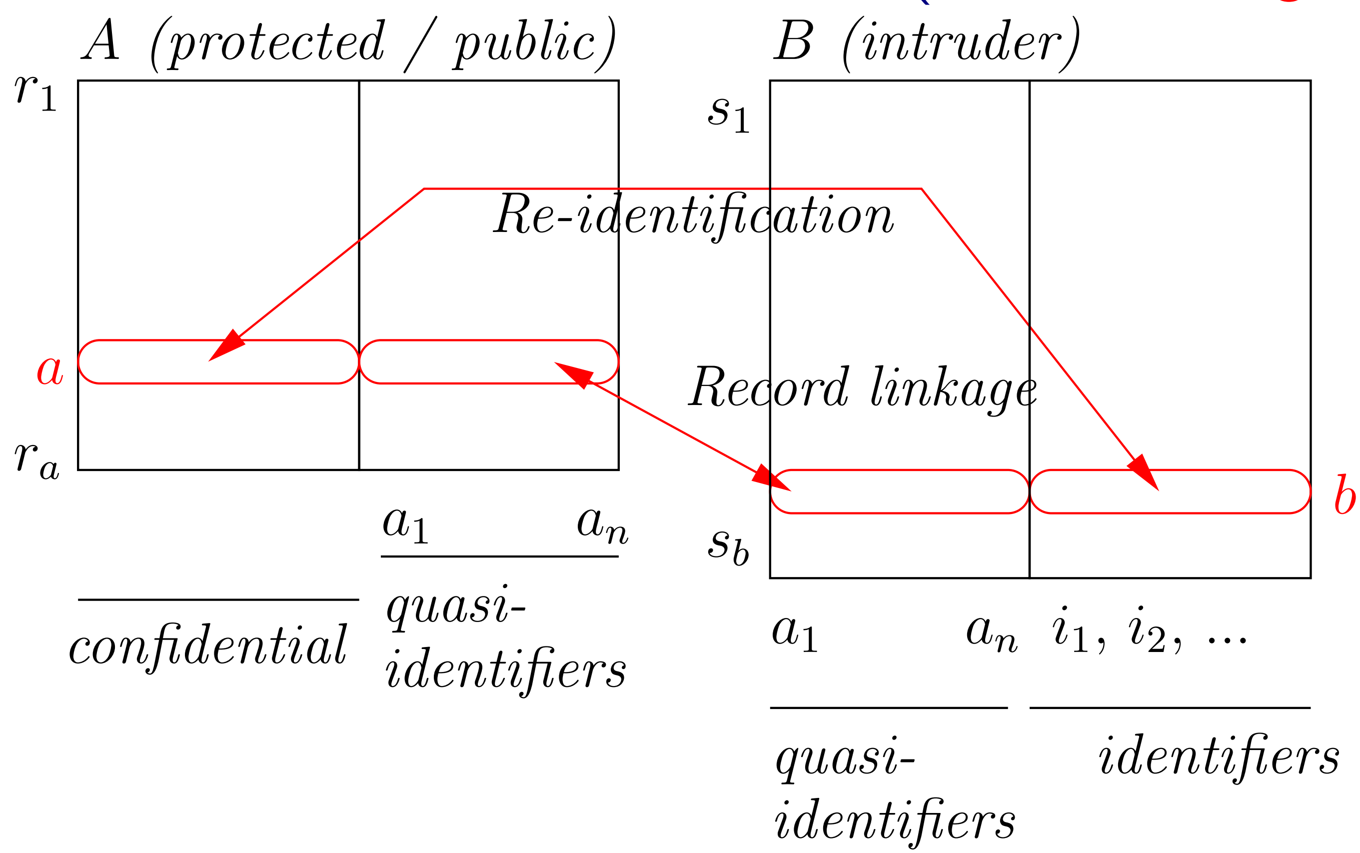
Quantitative measures for identity disclosure

- An scenario for identity disclosure: $X = id||X_{nc}||X_c$
  - Protection of the attributes
    - ⋆ **Identifiers.** Usually removed or encrypted.
    - ⋆ **Confidential.** X_c are usually not modified. $X'_c = X_c$.
    - ⋆ **Quasi-identifiers.** Apply masking method ρ to these attributes. $X'_{nc} = \rho(X_{nc})$.

Disclosure risk assesment

Quantitative measures for identity disclosure

- An scenario for identity disclosure: $X = id||X_{nc}||X_c$
  - A: File with the protected data set
  - B: File with the data from the intruder (subset of original X)



Disclosure risk assesment

Quantitative measures for identity disclosure

- An scenario for identity disclosure
  - Reidentification using the common attributes (quasi-identifiers):

Disclosure risk assesment

Quantitative measures for identity disclosure

- An scenario for identity disclosure
  - Reidentification using the common attributes (quasi-identifiers): identity disclosure

Disclosure risk assesment

Quantitative measures for identity disclosure

- An scenario for identity disclosure
  - Reidentification using the common attributes (quasi-identifiers): identity disclosure
  - Attribute disclosure may be possible

Disclosure risk assesment

Quantitative measures for identity disclosure

- An scenario for identity disclosure
  - Reidentification using the common attributes (quasi-identifiers): identity disclosure
  - Attribute disclosure may be possible
    when reidentification permits to link confidential values to identifiers
    (in this case: identity disclosure implies attribute disclosure)

Disclosure risk assesment

Quantitative measures for identity disclosure

- Flexible scenario for identity disclosure
  - A protected file using a masking method
  - B (intruder's) is a subset of the original file.

Disclosure risk assesment

Quantitative measures for identity disclosure

- Flexible scenario for identity disclosure
  - A protected file using a masking method
  - B (intruder's) is a subset of the original file.

    $\rightarrow$ intruder with information on only some individuals

Disclosure risk assesment

Quantitative measures for identity disclosure

- Flexible scenario for identity disclosure
  - A protected file using a masking method
  - B (intruder's) is a subset of the original file.
    $\rightarrow$ intruder with information on only some individuals
    $\rightarrow$ intruder with information on only some characteristics

Disclosure risk assesment

Quantitative measures for identity disclosure

- Flexible scenario for identity disclosure
  - A protected file using a masking method
  - B (intruder's) is a subset of the original file.

    $\rightarrow$ intruder with information on only some individuals

    $\rightarrow$ intruder with information on only some characteristics
  - But also,
    - $\star$ B with a schema different to the one of A (different attributes)

Disclosure risk assesment

Quantitative measures for identity disclosure

- **Re-identification.** Risk as number of re-identifications that might be obtained by an intruder (estimation).

Disclosure risk assesment

Quantitative measures for identity disclosure

- **Re-identification.** Risk as number of re-identifications that might be obtained by an intruder (estimation).
  - When both files have the same schema: record linkage algorithms.

Disclosure risk assesment

Quantitative measures for identity disclosure

- **Re-identification.** Risk as number of re-identifications that might be obtained by an intruder (estimation).
  - When both files have the same schema: record linkage algorithms.
  - Applicable to different scenarios. E.g., synthetic data

Disclosure risk assesment

Quantitative measures for identity disclosure

- **Re-identification.** Risk as number of re-identifications that might be obtained by an intruder (estimation).
  - When both files have the same schema: record linkage algorithms.
  - Applicable to different scenarios. E.g., synthetic data
- **Uniqueness.** Risk is defined as the probability that rare combinations of attribute values in the protected data set are indeed rare in the original population.

Disclosure risk assesment

Quantitative measures for identity disclosure

- **Re-identification.** Risk as number of re-identifications that might be obtained by an intruder (estimation).
  - When both files have the same schema: record linkage algorithms.
  - Applicable to different scenarios. E.g., synthetic data
- **Uniqueness.** Risk is defined as the probability that rare combinations of attribute values in the protected data set are indeed rare in the original population.
  - Suitable for sampling ($\rho(X)$ is a subset of X).
  - For masked data, the same combination will not appear.

Disclosure risk assesment

Quantitative measures for identity disclosure

- **Re-identification.** Risk as number of re-identifications that might be obtained by an intruder (estimation).
  - Probabilistic and distance-based record linkage

Disclosure risk assesment

Quantitative measures for identity disclosure

- **Re-identification.** Risk as number of re-identifications that might be obtained by an intruder (estimation).
  - Probabilistic and distance-based record linkage

Data: A: masked file; B: intruder's data file (subset of original file)
Result: LP: linked pairs; NP: non-linked pairs
for $a \in \mathbf{A}$ **do**
    b' $=$ arg $\min_{b \in \mathbf{B}} d(a, b)$;
    $\mathbf{LP} = \mathbf{LP} \cup (a, b')$;
    for $b \in \mathbf{B}$ **such that** $b \neq b'$ **do**
        $\mathbf{NP} = \mathbf{NP} \cup (a, b)$;

Transparency

Transparency

Transparency

Transparency: Definition

Transparency

Definition.

- protected/masked data has to be published informing on how the data has been protected

Transparency

Definition.

- protected/masked data has to be published informing on how the data has been protected

Advantage.

- Improve inference/evaluation of some statistics.
  E.g., noise addition with ϵ with $Var(\epsilon) = kVar(X)$,
  - $E(X') = E(X) + E(\epsilon) = E(X)$
  - $Cov(X'_i, X'_j) = Cov(X_i, X_j)$ for $i \neq j$
  - $Var(X') = Var(X) + kVar(X) = (1 + k)Var(X)$
  - $\rho_{X'_i, X'_j} = \frac{Cov(X'_i, X'_j)}{\sqrt{Var(X'_i)Var(X'_j)}} = \frac{Cov(X_i, X_j)}{(1+k)\sqrt{Var(X_i)Var(X_j)}} = \frac{1}{1+k}\rho_{X_i, X_j}$

Transparency

Definition.

- protected/masked data has to be published informing on how the data has been protected

Advantage.

- Improve inference/evaluation of some statistics.
  E.g., noise addition with ϵ with $Var(\epsilon) = kVar(X)$,
  - $E(X') = E(X) + E(\epsilon) = E(X)$
  - $Cov(X'_i, X'_j) = Cov(X_i, X_j)$ for $i \neq j$
  - $Var(X') = Var(X) + kVar(X) = (1+k)Var(X)$
  - $\rho_{X'_i,X'_j} = \frac{Cov(X'_i,X'_j)}{\sqrt{Var(X'_i)Var(X'_j)}} = \frac{Cov(X_i,X_j)}{(1+k)\sqrt{Var(X_i)Var(X_j)}} = \frac{1}{1+k}\rho_{X_i,X_j}$

Inconvenient

- intruders can use this information to attack the data

Transparency

Discussion.

- Cryptography relationship. Encryption method is known.
- Guessing the method. We do not need to worry about the intruder guessing or learning about the method use.
  - Microaggregation find by visual inspection
  - Rank swapping can be guessed if the intruder has a large enough data set.

Transparency

Attacking Rank Swapping

Transparency

Under the transparency principle we publish

- X' (protected data set)

Transparency

Under the transparency principle we publish

- X' (protected data set)
- masking method: rank swapping

Transparency

Under the transparency principle we publish

- X' (protected data set)
- masking method: rank swapping
- parameter of the method: p (proportion of $|X|$)

Then, the intruder can use *(method, parameter)* to attack

Transparency

Under the transparency principle we publish

- X' (protected data set)
- masking method: rank swapping
- parameter of the method: p (proportion of $|X|$)

Then, the intruder can use *(method, parameter)* to attack

$\rightarrow$ *(method, parameter)* $=$ *(rank swapping,* p*)*

Transparency

Intruder perspective.

- All protected values are available.
  I.e.,

Transparency

Intruder perspective.

- All protected values are available.
  I.e.,
  Intruder data are available

Transparency

Intruder perspective.

- All protected values are available.
  I.e.,
  Intruder data are available
  All data in the original data set are also available

Transparency

Intruder perspective.

- All protected values are available.
  I.e.,
  Intruder data are available
  All data in the original data set are also available

Intruder's attack for a single attribute

- Given a value a, we can define the set of possible swaps for a_i
  Proceed as rank swapping does: $a_1, \ldots, a_n$ ordered values If $a_i = a$, it can only be swapped with a_ℓ in the range

$$\ell \in [i + 1, \min(n, i + p * |X|/100)]$$

Transparency

Intruder's attack for a single attribute attribute V_j

- Define $B_j(a)$
  the set of masked records that can be the masked version of a

Transparency

Intruder's attack for a single attribute attribute V_j

- Define $B_j(a)$
  the set of masked records that can be the masked version of a
  No uncertainty on $B_j(a)$

$$x'_\ell \in B_j(a)$$

Intruder's attack for all available attributes

- Define $B_j(a_j)$ for all available V_j
- Intersection attack:

Transparency

Intruder's attack for a single attribute attribute V_j

- Define $B_j(a)$
  the set of masked records that can be the masked version of a
  No uncertainty on $B_j(a)$

$$x'_\ell \in B_j(a)$$

Intruder's attack for all available attributes

- Define $B_j(a_j)$ for all available V_j
- Intersection attack:

$$x'_\ell \in \cap_{1 \leq j \leq c} B_j(x_i).$$

Transparency

Intruder's attack for a single attribute attribute V_j

- Define $B_j(a)$
  the set of masked records that can be the masked version of a
  No uncertainty on $B_j(a)$

$$x'_\ell \in B_j(a)$$

Intruder's attack for all available attributes

- Define $B_j(a_j)$ for all available V_j
- Intersection attack:

$$x'_\ell \in \cap_{1 \leq j \leq c} B_j(x_i).$$

No uncertainty!

Transparency

Intruder's attack for all available attributes

- Intersection attack:

$$x'_\ell \in \cap_{1 \leq j \leq c} B_j(x_i).$$

- When $|\cap_{1 \leq j \leq c} B_j(x_i)| = 1$, we have a true match
- Otherwise, we can apply record linkage within this set

Data: $Y \subseteq X$: data file of the intruder; X': masked file; p: percentage of records for swapping
Result: linkage between Y and X'
$LP = \emptyset$;
for *each* $x_i \in Y$ **do**
  $B(x_i) = \cap_{1 \leq j \leq c} B_j(x_i)$;
  $x' = \arg\min_{x' \in B(x_i)} d(x', x_i)$;
  $LP = LP \cup (x', x_i)$;
return (LP) ;
Undo the sorting step ;

Transparency

Intruder's attack. Example.

- Intruder's record: $x_2 = (6, 7, 10, 2)$, $p = 2$. First attribute: $x_{21} = 6$
- $B_1(a = 6) = \{(4, 1, 10, 10), (5, 5, 8, 1), (6, 7, 6, 3), (7, 3, 5, 6), (8, 4, 2, 2)\}$

Original file				Masked file				$B(x_{2j})$
a_1	a_2	a_3	a_4	a'_1	a'_2	a'_3	a'_4	$B(x_{21})$
8	9	1	3	10	10	3	5	
6	7	10	2	5	5	8	1	X
10	3	4	1	8	4	2	2	X
7	1	2	6	9	2	4	4	
9	4	6	4	7	3	5	6	X
2	2	8	8	4	1	10	10	X
1	10	3	9	3	9	1	7	
4	8	7	10	2	6	9	8	
5	5	5	5	6	7	6	3	X
3	6	9	7	1	8	7	9	

Transparency

Intruder's attack. Example.

- Intruder's record: $x_2 = (6, 7, 10, 2)$, $p = 2$. Second attribute: $x_{22} = 7$
- $B_2(a = 7) = \{(5, 5, 8, 1), (2, 6, 9, 8), (6, 7, 6, 3), (1, 8, 7, 9), (3, 9, 1, 7)\}$

Original file				Masked file				$B(x_{2j})$	
a_1	a_2	a_3	a_4	a'_1	a'_2	a'_3	a'_4	$B(x_{21})$	$B(x_{22})$
8	9	1	3	10	10	3	5		
6	7	10	2	5	5	8	1	X	X
10	3	4	1	8	4	2	2	X	
7	1	2	6	9	2	4	4		
9	4	6	4	7	3	5	6	X	
2	2	8	8	4	1	10	10	X	
1	10	3	9	3	9	1	7		X
4	8	7	10	2	6	9	8		X
5	5	5	5	6	7	6	3	X	X
3	6	9	7	1	8	7	9		X

Transparency

Intruder's attack. Example.

- Intruder's record: $x_2 = (6, 7, 10, 2)$, $p = 2$.
  - $B_1(x_{21} = 6) = \{(4, 1, 10, 10), (5, 5, 8, 1), (6, 7, 6, 3), (7, 3, 5, 6), (8, 4, 2, 2)\}$
  - $B_2(x_{22} = 7) = \{(5, 5, 8, 1), (2, 6, 9, 8), (6, 7, 6, 3), (1, 8, 7, 9), (3, 9, 1, 7)\}$
  - $B_3(x_{23} = 10) = \{(5, 5, 8, 1), (2, 6, 9, 8), (4, 1, 10, 10)\}$
  - $B_4(x_{24} = 2) = \{(5, 5, 8, 1), (8, 4, 2, 2), (6, 7, 6, 3), (9, 2, 4, 4)\}$
- The intersection is a single record

$$(5, 5, 8, 1)$$

Transparency

Intruder's attack. Application.

- Data:
  - Census (1080 records, 13 attributes)
  - EIA (4092 records, 10 attributes)
- Rank swaping parameter:
  - $p = 2, \ldots, 20$

Transparency

Intruder's attack. Result

	Census			EIA		
	RSLD	DLD	PLD	RSLD	DLD	PLD
rs 2	77.73	73.52	71.28	43.27	21.71	16.85
rs 4	66.65	58.40	42.92	12.54	10.61	4.79
rs 6	54.65	43.76	22.49	7.69	7.40	2.03
rs 8	41.28	32.13	11.74	6.12	5.98	1.12
rs 10	29.21	23.64	6.03	5.60	5.19	0.69
rs 12	19.87	18.96	3.46	5.39	4.87	0.51
rs 14	16.14	15.63	2.06	5.28	4.55	0.32
rs 16	13.81	13.59	1.29	5.19	4.54	0.23
rs 18	12.21	11.50	0.83	5.20	4.54	0.22
rs 20	10.88	10.87	0.59	5.15	4.36	0.18

Transparency

Intruder's attack. Summary

- When $|\cap B_j| = 1$, this is a match.
  25% of reidentifications in this way $\neq$ 25% in distance-based or probabilistic record linkage.
- Approach applicable when the intruder knows a single record
- The more attributes the intruder has, the better is the reidentification.
  Intersection never increases when the number of attributes increases.
- When p is not known, an upper bound can help
  If the upper bound is too high, some $|\cap B_j|$ can be zero

Transparency

Avoiding Transparency Attack in Rank Swapping

Transparency

Avoiding transparency attack in rank swapping.

- Enlarge the B_j set to encompass the whole file.

Transparency

Avoiding **transparency attack** in rank swapping.

- Enlarge the B_j set to encompass the whole file.
- Then,

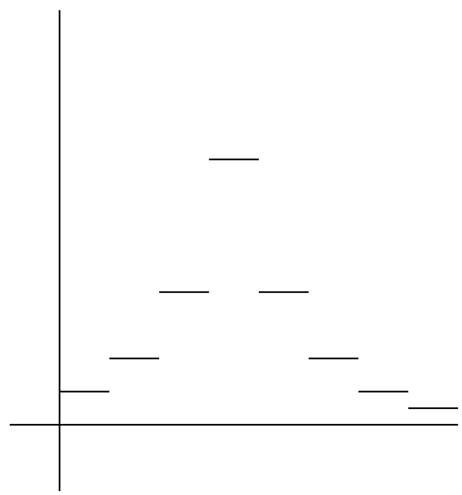
$$\cap B_j = X$$

Transparency

Approaches to avoid transparency attack in rank swapping.

- Rank swapping p-buckets. Select bucket B_s using

$$Pr[B_s \ is \ choosen \ |B_r] = \frac{1}{K}\frac{1}{2^{s-r+1}}.$$



- Rank swapping p-distribution. Swap a_i with a_ℓ where $\ell = i + r$ and r according to a $N(0.5p, 0.5p)$.

Transparency

Attacking Microaggregation

Microaggregation and transparency

Transparency attack to microaggregation.

- Define $B_j(a)$ as the set of records that can be the masked versio of a for attribute V_j

$$x'_\ell \in B_j(a)$$

  In optimal univariate microaggregation $B_j(a)$ is the union of two clusters $(p_i < a < p_{i+1})$.

- Intersection attack

$$x'_\ell \in \cap_{1 \leq j \leq c} B_j(x_i).$$

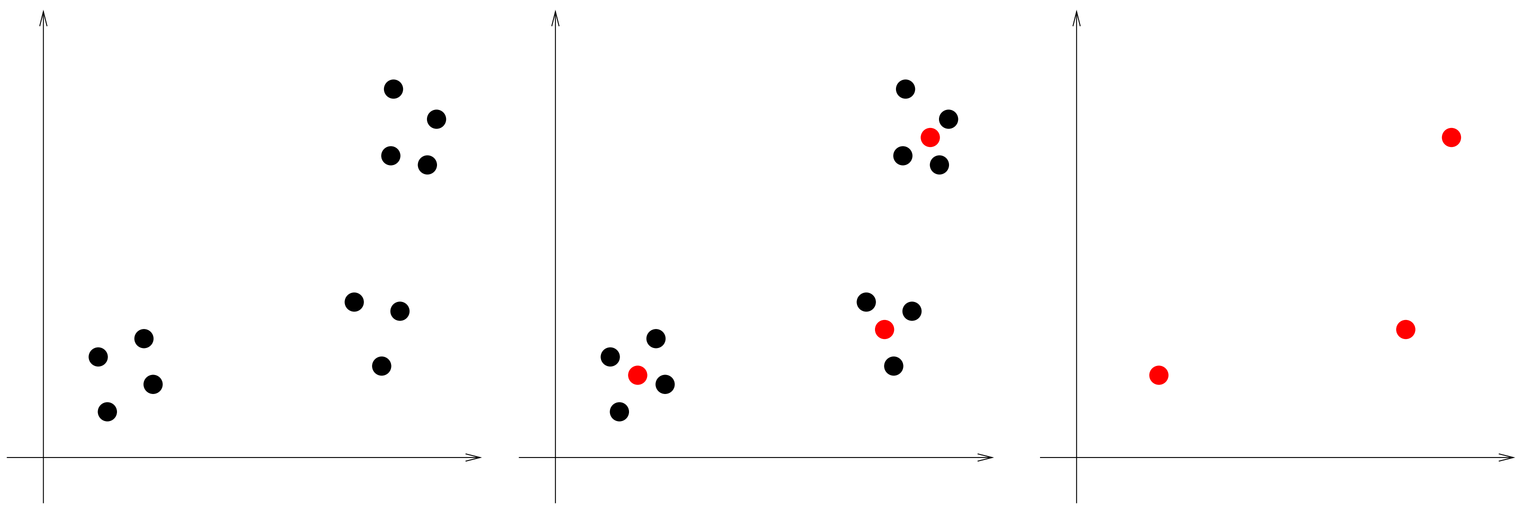
Transparency

Avoiding Transparency Attack in Microaggregation

Microaggregation and transparency

Avoiding transparency attack in microaggregation.

- Fuzzy microaggregation.
  - Construct fuzzy clusters: records belong to several clusters
  - Assign values from cluster centers from a random distribution built from membership functions



Worst-case scenario

Worst-case scenario when measuring disclosure risk

Worst-case scenario

Worst-case scenario for disclosure risk assessment

- Distance-based record linkage
- Parametric distances with best parameters
  E.g.,
  - Weighted Euclidean distance

Worst-case scenario

Worst-case scenario for disclosure risk assessment

- Distance-based record linkage with Euclidean distance equivalent to:

$$d^2(a,b) = \sum_{i=1}^{n} \frac{1}{n} \left(\mathit{diff}_i(a,b) \right)^2$$
$$= WM_p(\mathit{diff}_1(a,b), \ldots, \mathit{diff}_n(a,b))$$

  with $p = (1/n, \ldots, 1/n)$ and
  $\mathit{diff}_i(a,b) = ((a_i - \bar{a}_i)/\sigma(a_i) - (b_i - \bar{b}_i)/\sigma(b_i))^2$
- $p_i = 1/n$ means equal importance to all attributes
- Appropriate for attributes with equal discriminatory power (e.g., same noise, same distribution)

Worst-case scenario

Worst-case scenario for disclosure risk assessment

- Distance-based record linkage with weighted mean distance (weighted Euclidean distance)

$$d^2(a,b) = WM_p(\mathit{diff}_1(a,b), \ldots, \mathit{diff}_n(a,b))$$

with arbitrary vector $p = (p_1, \ldots, p_n)$ and
$\mathit{diff}_i(a,b) = ((a_i - \bar{a}_i)/\sigma(a_i) - (b_i - \bar{b}_i)/\sigma(b_i))^2$

Worst-case scenario

Worst-case scenario for disclosure risk assessment

- Distance-based record linkage with weighted mean distance (weighted Euclidean distance)

$$d^2(a,b) = WM_p(\mathit{diff}_1(a,b), \ldots, \mathit{diff}_n(a,b))$$

with arbitrary vector $p = (p_1, \ldots, p_n)$ and
$\mathit{diff}_i(a,b) = ((a_i - \bar{a}_i)/\sigma(a_i) - (b_i - \bar{b}_i)/\sigma(b_i))^2$

Worst-case: Optimal selection of the weights. How??

- Supervised machine learning approach
- Using an optimization problem

Worst-case scenario

Worst-case scenario for disclosure risk assessment

- Distance-based record linkage with parametric distances (distance/metric learning): $\mathbb{C}$ a combination/aggregation function

$$d^2(a,b) = \mathbb{C}_p(\mathit{diff}_1(a,b), \ldots, \mathit{diff}_n(a,b))$$

with parameter p and
$\mathit{diff}_i(a,b) = ((a_i - \bar{a}_i)/\sigma(a_i) - (b_i - \bar{b}_i)/\sigma(b_i))^2$

Worst-case scenario

Worst-case scenario for disclosure risk assessment

- Distance-based record linkage with parametric distances (distance/metric learning): $\mathbb{C}$ a combination/aggregation function

$$d^2(a,b) = \mathbb{C}_p(\mathit{diff}_1(a,b), \ldots, \mathit{diff}_n(a,b))$$

with parameter p and
$\mathit{diff}_i(a,b) = ((a_i - \bar{a}_i)/\sigma(a_i) - (b_i - \bar{b}_i)/\sigma(b_i))^2$

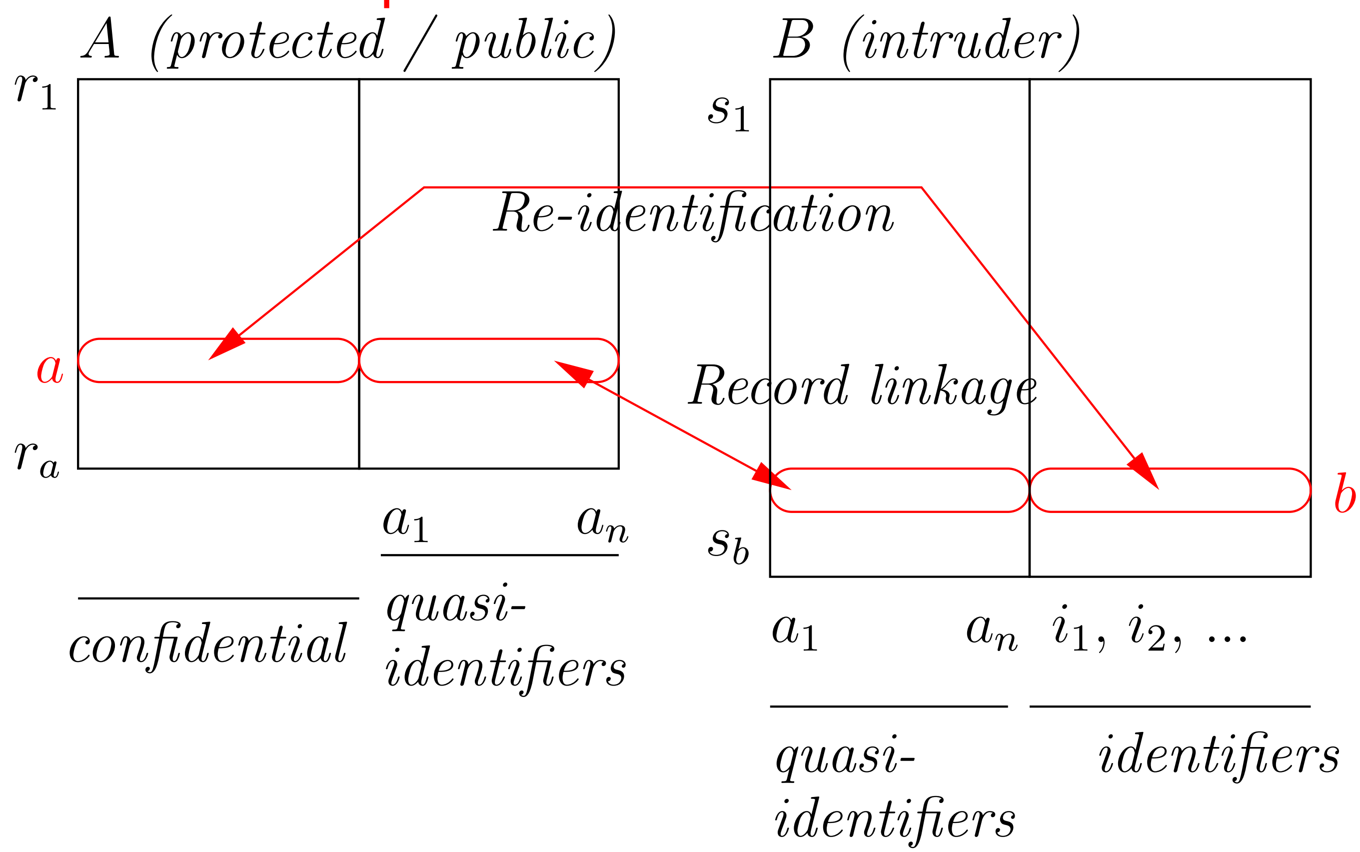
Worst-case: Optimal selection of the parameter p. How??

- Supervised machine learning approach
- Using an optimization problem

Worst-case scenario

Worst-case scenario for distance-based record linkage

- Optimal weights using a supervised machine learning approach
- We need a set of examples from:



Formalization of the problem

Machine Learning for distance-based record linkage

- Generic solution, using
  - an arbitrary combination function $\mathbb{C}$
  - with parameter p

$$d(a_i, b_j) = \mathbb{C}_p(\mathit{diff}_1(a, b), \ldots, \mathit{diff}_n(a, b))$$

Formalization of the problem

Machine Learning for distance-based record linkage

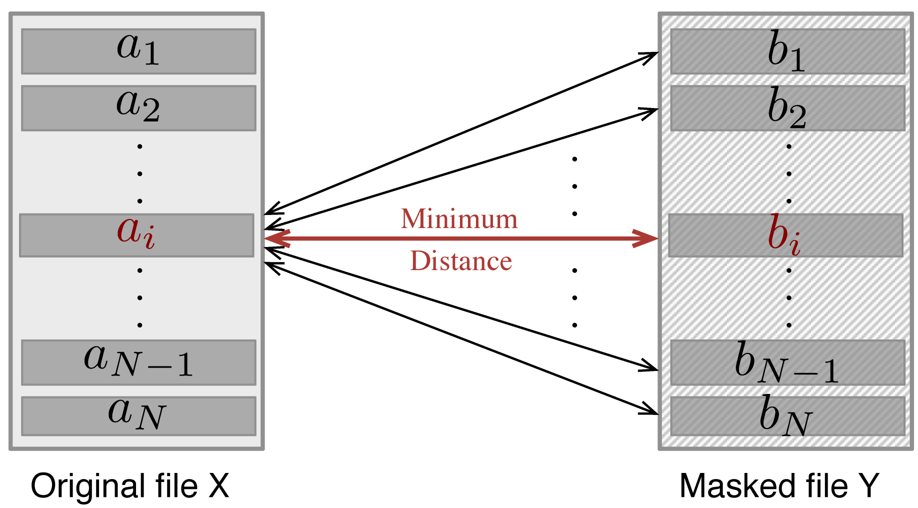
- Generic solution, using $\mathbb{C}$ with parameter p
- Goal
  - as much correct reidentifications as possible
  - For record i: $d(a_i, b_j) \geq d(a_i, b_i)$ for all j

Formalization of the problem

Machine Learning for distance-based record linkage

- Generic solution, using $\mathbb{C}$ with parameter p
- Goal
  - as much correct reidentifications as possible
  - For record i: $d(a_i, b_j) \geq d(a_i, b_i)$ for all j
    That is,

$$\mathbb{C}_p(\mathit{diff}_1(a_i, b_j), \ldots, \mathit{diff}_n(a_i, b_j)) \geq \mathbb{C}_p(\mathit{diff}_1(a_i, b_i), \ldots, \mathit{diff}_n(a_i, b_i))$$



Formalization of the problem

Machine Learning for distance-based record linkage

- Goal
  - as much correct reidentifications as possible
  - Maximize the number of records a_i such that $d(a_i, b_j) \geq d(a_i, b_i)$ for all j
  - If record a_i fails for at least one b_j

$$d(a_i, b_j) \ngeq d(a_i, b_i)$$

Then, let $K_i = 1$ in this case, then for a large enough constant C

$$d(a_i, b_j) + CK_i \geq d(a_i, b_i)$$

Formalization of the problem

Machine Learning for distance-based record linkage

- Goal
  - as much correct reidentifications as possible
  - Maximize the number of records a_i such that $d(a_i, b_j) \geq d(a_i, b_i)$ for all j
  - If record a_i fails for at least one b_j

$$d(a_i, b_j) \ngeq d(a_i, b_i)$$

Then, let $K_i = 1$ in this case, then for a large enough constant C

$$d(a_i, b_j) + CK_i \geq d(a_i, b_i)$$

That is,

$$\mathbb{C}_p(\mathit{diff}_1(a_i, b_j), \ldots, \mathit{diff}_n(a_i, b_j)) + CK_i \geq \mathbb{C}_p(\mathit{diff}_1(a_i, b_i), \ldots, \mathit{diff}_n(a_i, b_i))$$

Formalization of the problem

Machine Learning for distance-based record linkage

- Goal
  - as much correct reidentifications as possible
  - Minimize K_i: minimize the number of records a_i that fail $d(a_i, b_j) \geq d(a_i, b_i)$ for all j
  - $K_i \in \{0, 1\}$, if $K_i = 0$ reidentification is correct

$$d(a_i, b_j) + CK_i \geq d(a_i, b_i)$$

Formalization of the problem

Machine Learning for distance-based record linkage

- Goal
  - as much correct reidentifications as possible
  - Minimize K_i: minimize the number of records a_i that fail
- Formalization:

$$Minimize \sum_{i=1}^{N} K_i$$

$$Subject\ to:$$

$$\mathbb{C}_p(\mathit{diff}_1(a_i, b_j), \ldots, \mathit{diff}_n(a_i, b_j)) -$$
$$- \mathbb{C}_p(\mathit{diff}_1(a_i, b_i), \ldots, \mathit{diff}_n(a_i, b_i)) + CK_i > 0$$

$$K_i \in \{0, 1\}$$

Additional constraints according to $\mathbb{C}$

Formalization of the problem

Machine Learning for distance-based record linkage

- Example: the case of the weighted mean
- Formalization:

$$Minimize \sum_{i=1}^{N} K_i$$

$$Subject\ to:$$

$$WM_p(\mathit{diff}_1(a_i, b_j), \ldots, \mathit{diff}_n(a_i, b_j)) -$$
$$- WM_p(\mathit{diff}_1(a_i, b_i), \ldots, \mathit{diff}_n(a_i, b_i)) + CK_i > 0$$

$$K_i \in \{0, 1\}$$

$$\sum_{i=1}^{n} p_i = 1$$

$$p_i \geq 0$$

Experiments and distances

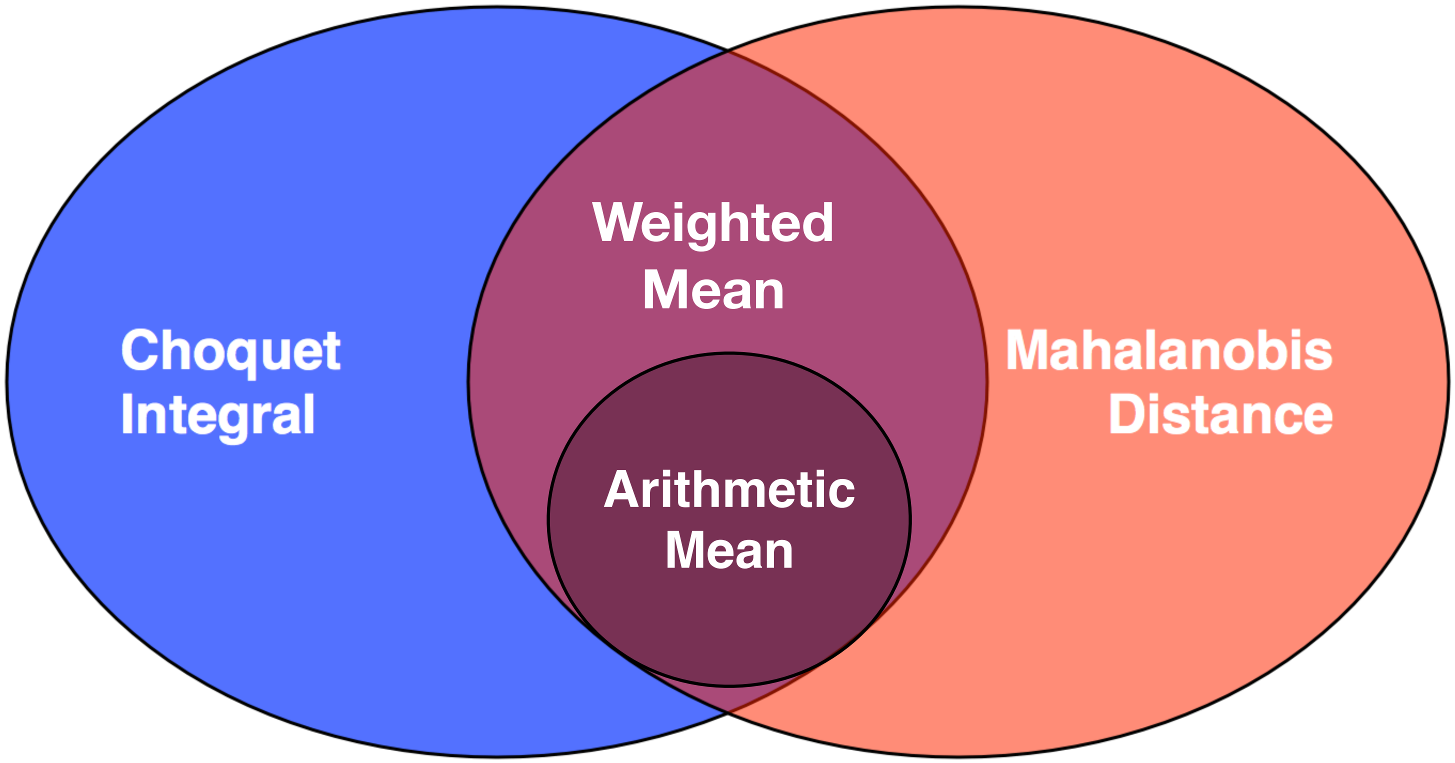
Machine Learning for distance-based record linkage

- Distances considered
  - Weighted mean: importance to the attributes
    Parameter: weighting vector n parameters
  - OWA - linear combination of order statistics (weighted): discard lower or larger distances
    Parameter: weighting vector n parameters
  - Choquet integral: weights to interactions of sets of attributes
    Parameter: non-additive measure: $2^n - 2$ parameters
  - Bilinear form - generalization of the Mahalanobis distance: weights to interactions between pairs of attributes
    Parameter: square matrix: $n \times n$ parameters

Experiments and distances

Machine Learning for distance-based record linkage

- Distances considered



Experiments and distances

Machine Learning for distance-based record linkage

- Data sets considered (from CENSUS dataset)
  - *M4-33*: 4 attributes microaggregated in groups of 2 with $k = 3$.
  - *M4-28*: 4 attributes,2 attributes with $k = 2$, and 2 with $k = 8$.
  - *M4-82*: 4 attributes, 2 attributes with $k = 8$, and 2 with $k = 2$.
  - *M5-38*: 5 attributes, 3 attributes with $k = 3$, and 2 with $k = 8$.
  - *M6-385*: 6 attributes, 2 attributes with $k = 3$, 2 attributes with $k = 8$, and 2 with $k = 5$.
  - *M6-853*: 6 attributes, 2 attributes with $k = 8$, 2 attributes with $k = 5$, and 2 with $k = 3$.

Experiments and distances

Machine Learning for distance-based record linkage

- Percentage of the number of correct re-identifications.

	M4-33	*M4-28*	*M4-82*	*M5-38*	*M6-385*	*M6-853*
d^2AM	84.00	68.50	71.00	39.75	78.00	84.75
d^2MD	94.00	90.00	92.75	88.25	98.50	98.00
d^2WM	95.50	93.00	94.25	90.50	99.25	98.75
d^2WM_m	95.50	93.00	94.25	90.50	99.25	98.75
d^2CI	95.75	93.75	94.25	91.25	**99.75**	99.25
d^2CI_m	95.75	93.75	94.25	90.50	99.50	98.75
d^2SB_{NC}	**96.75**	**94.5**	**95.25**	**92.25**	**99.75**	**99.50**
d^2SB	**96.75**	**94.5**	**95.25**	**92.25**	**99.75**	**99.50**
d^2SB_{PD}	–	–	–	–	–	99.25

Experiments and distances

Machine Learning for distance-based record linkage

- Computation time comparison (in seconds).

	M4-33	*M4-28*	*M4-82*	*M5-38*	*M6-385*	*M6-853*
d^2WM	29.83	41.37	24.33	718.43	11.81	17.77
d^2WM_m	3.43	6.26	2.26	190.75	4.34	6.72
d^2CI	280.24	427.75	242.86	42,731.22	24.17	87.43
d^2CI_m	155.07	441.99	294.98	4,017.16	79.43	829.81
d^2SB_{NC}	32.04	2,793.81	150.66	10,592.99	13.65	14.11
d^2SB	13.67	3,479.06	139.59	169,049.55	13.93	13.70

- Constraints specific to weighted mean and Choquet integral for distances
  N: number of records; n: number of attributes

	d^2WM_m	d^2CI_m
Additional Constraints	$\sum_{i=1}^{n} p_i = 1$ $p_i > 0$	$\mu(\emptyset) = 0$ $\mu(V) = 1$ $\mu(A) \leq \mu(B)$ when $A \subseteq B$ $\mu(A) + \mu(B) \geq \mu(A \cup B) + \mu(A \cap B)$
Total Constr.	$N(N-1) + N + 1 + n$	$N(N-1) + N + 2 + (\sum_{k=2}^{n} \binom{n}{k} k) + \binom{n}{2}$

Experiments and distances

Machine Learning for distance-based record linkage

- A summary of the experiments

	AM	MD	WM	OWA	SBF	CI
Computation	Very fast	Very fast	Fast	regular	Hard	Hard
Results	Worse	Good	Good	Bad	Very Good	Very Good
Information	No	No	Few	Few	Large	Large

Summary

Summary

Experiments and distances

- Quantitative measures of risk
- Transparency and disclosure risk
  - Masking method and parameters published
  - Disclosure risk revisited
  - New masking methods resistant to transparency
- Worst-case scenario for disclosure risk
  - Parametric distances
  - Distance/metric learning

Thank you

* Special thanks to Jordi Nin, Daniel Abril, Guillermo Navarro-Arribas

Experiments and distances

Main references.

- D. Abril, G. Navarro-Arribas, V. Torra, Supervised Learning Using a Symmetric Bilinear Form for Record Linkage, Information Fusion, in press.
- D. Abril, G. Navarro-Arribas, V. Torra, Improving record linkage with supervised learning for disclosure risk assessment, Information Fusion 13:4 (2012) 274-284.
- J. Nin, J. Herranz, V. Torra, On the Disclosure Risk of Multivariate Microaggregation, Data and Knowledge Engineering, 67 (2008) 399-412.
- J. Nin, J. Herranz, V. Torra, Rethinking Rank Swapping to Decrease Disclosure Risk, Data and Knowledge Engineering, 64:1 (2008) 346-364.