Linköping 2016

On machine learning for data privacy

Vicenç Torra

Dec. 7, 2016

School of Informatics, University of Skövde, Sweden

Outline

Disclosure risk. A quantitative measures: record linkage

- The worst-case scenario
  - Using ML in reidentification
- Transparency principle
  - Transparency attacks

Outline

1. Introduction
2. Disclosure risk assessment
   - Worst-case scenario
   - ML for reidentification
3. Transparency
   - Definition
   - Attacking Rank Swapping
   - Avoiding transparency attack
4. Information loss
5. Summary

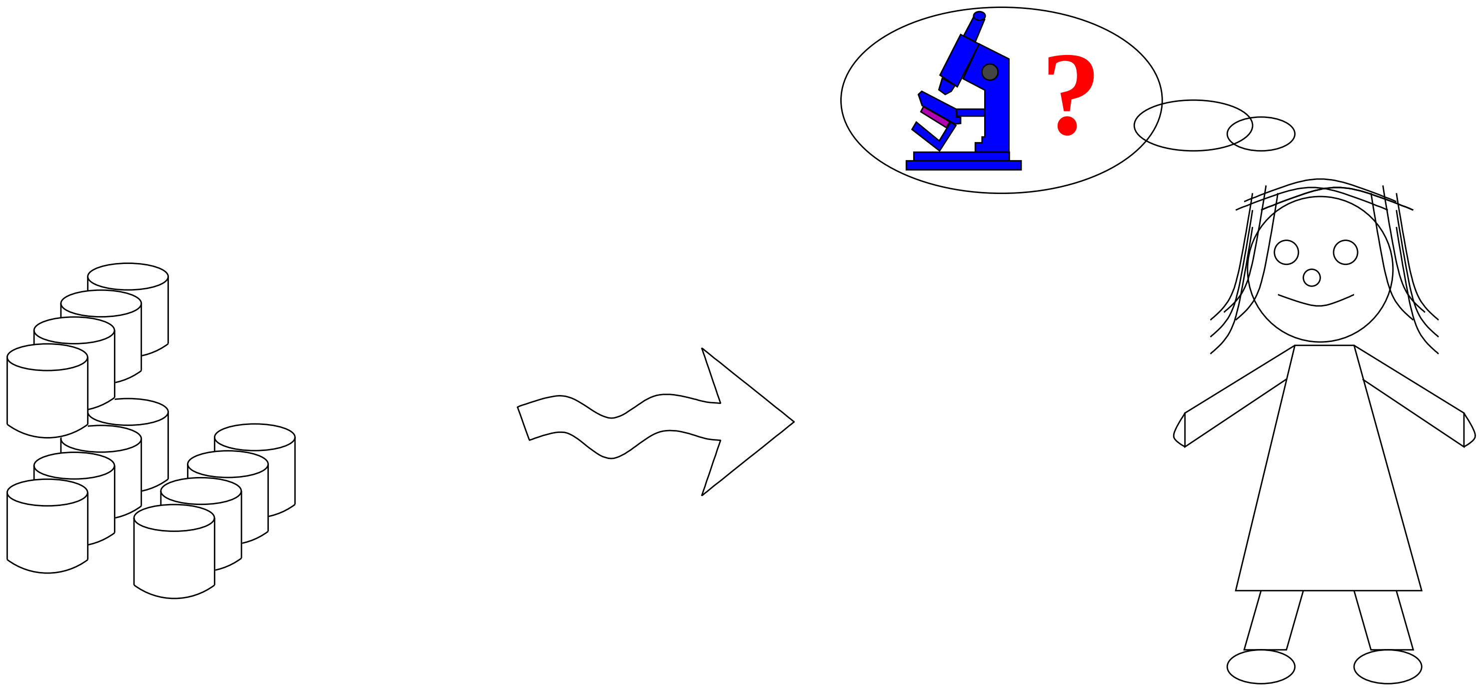
Introduction

Introduction

Masking methods

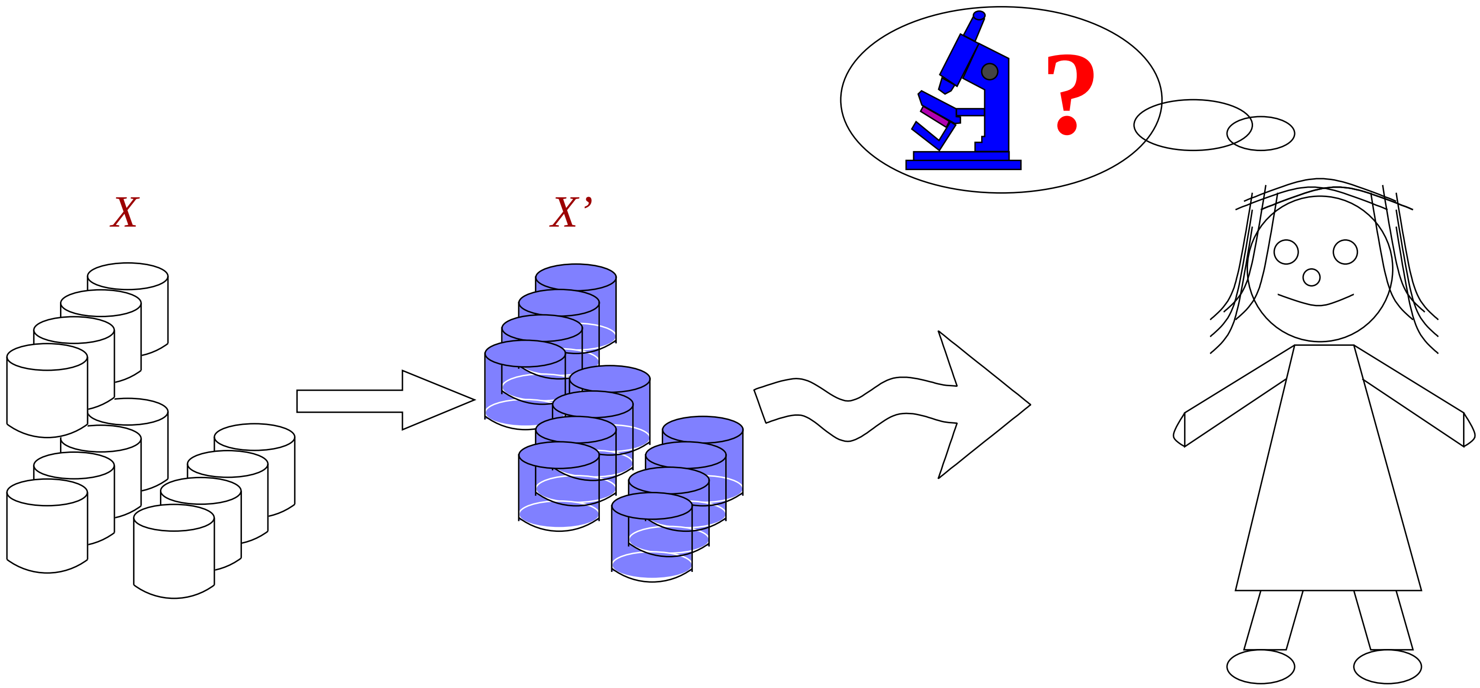
Classification w.r.t. our knowledge on the computation of a third party

- Data-driven or general purpose (*analysis not known*)
  → anonymization methods / masking methods
- Computation-driven or specific purpose (*analysis known*)
  → cryptographic protocols, differential privacy
- Result-driven (*analysis known: protection of its results*)



Masking methods

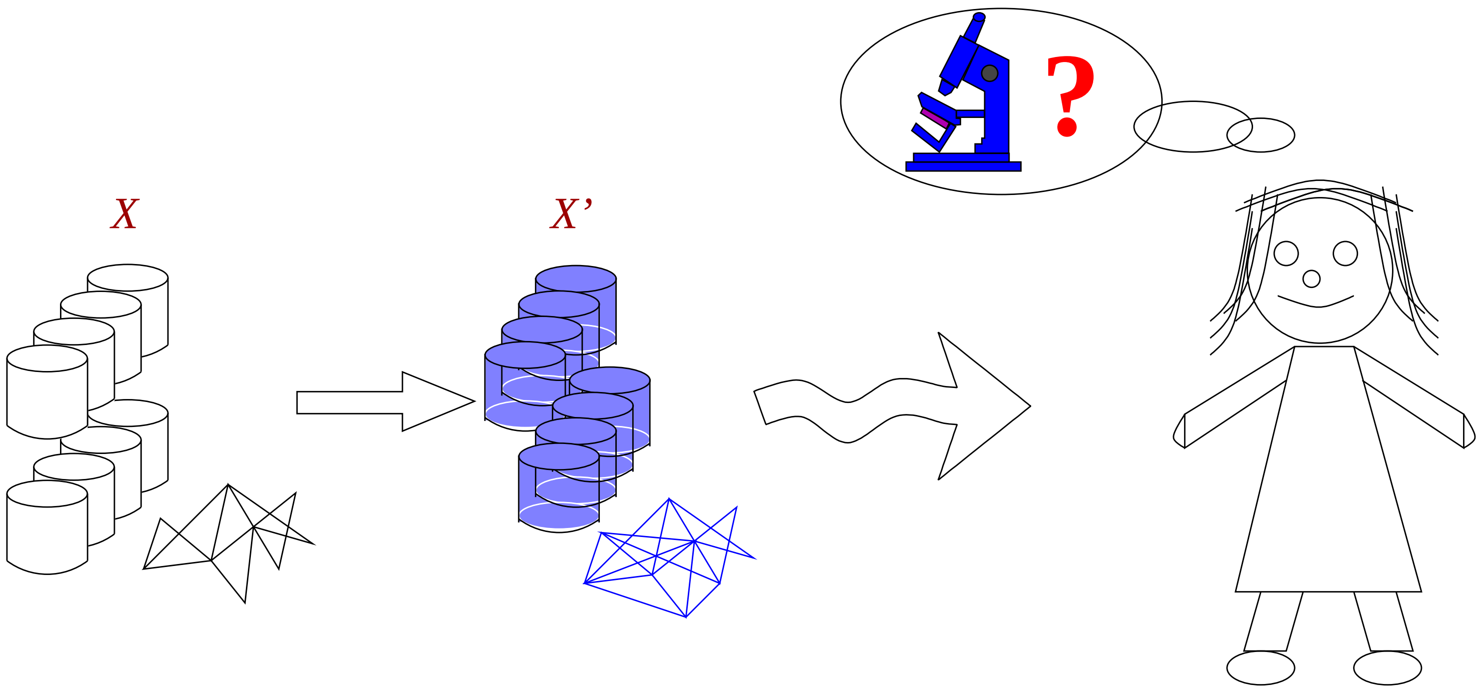
Anonymization/masking method: Given a data file X compute a file X' with data of *less quality*.



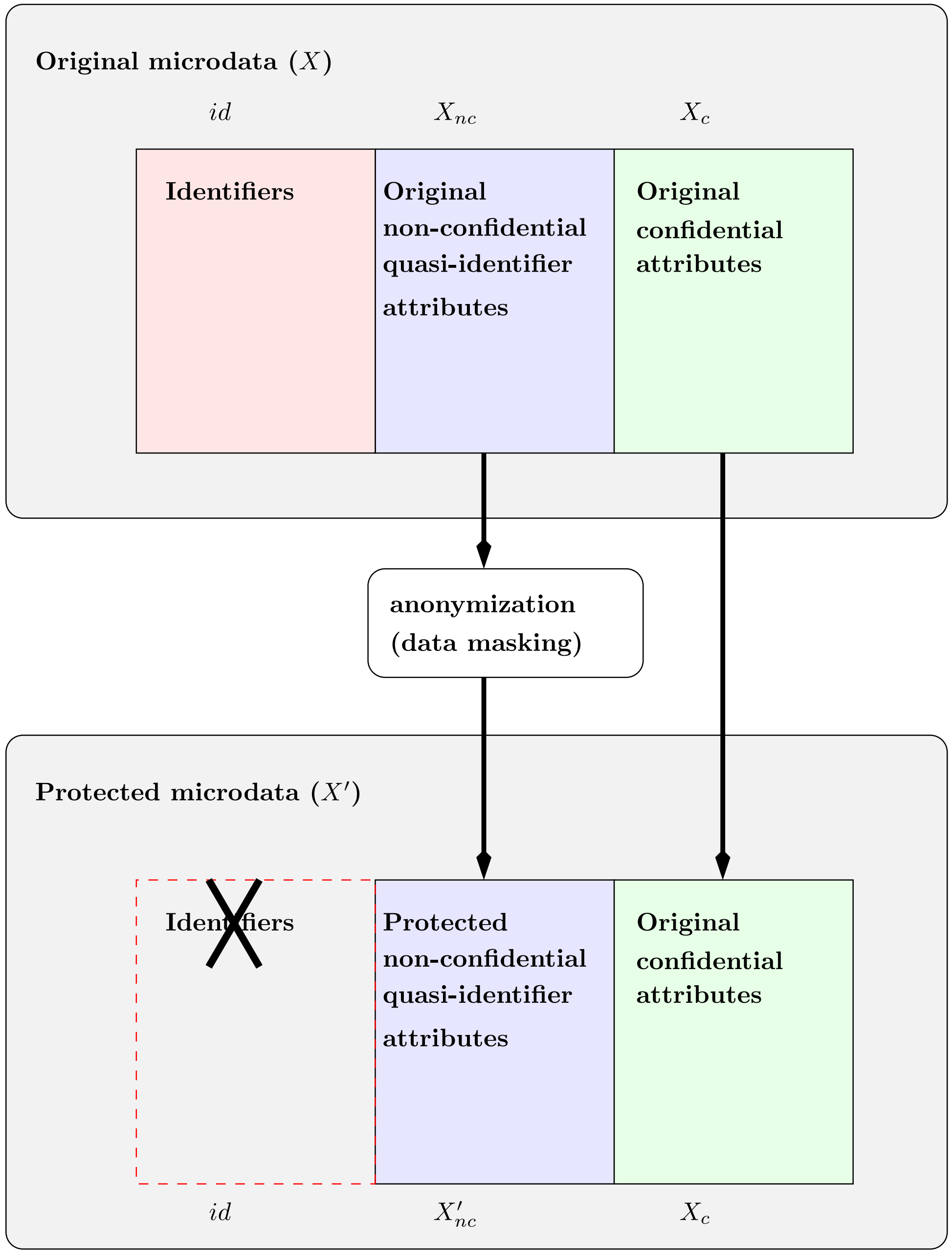
Masking methods

Approach valid for different types of data

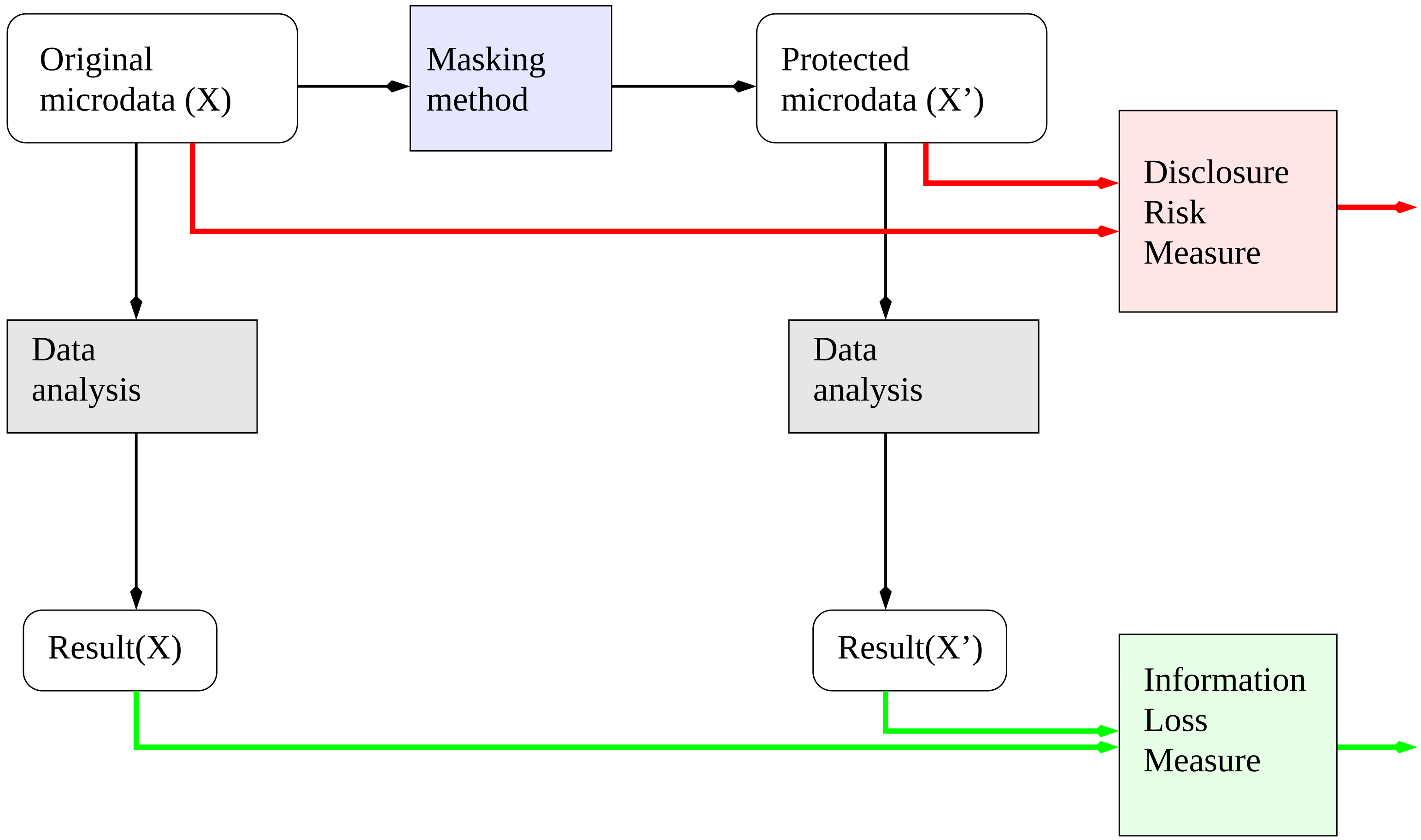
- Databases, documents, search logs, social networks, . . .
  (also masking taking into account semantics: wordnet, ODP)



Masking methods



Research questions



Masking methods

Masking methods. (anonymization methods)

Masking methods

Masking methods. (anonymization methods)

- Perturbative. (less quality=erroneous data)
  E.g. noise addition/multiplication, microaggregation, rank swapping

Masking methods

Masking methods. (anonymization methods)

- Perturbative. (less quality=erroneous data)
  E.g. noise addition/multiplication, microaggregation, rank swapping
- Non-perturbative. (less quality=less detail)
  E.g. generalization, suppression

Masking methods

Masking methods. (anonymization methods)

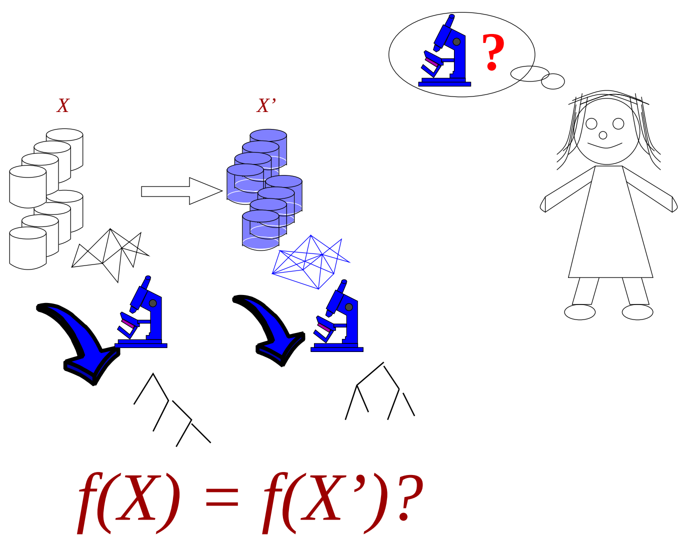
- Perturbative. (less quality=erroneous data)
  E.g. noise addition/multiplication, microaggregation, rank swapping
- Non-perturbative. (less quality=less detail)
  E.g. generalization, suppression
- Synthetic data generators. (less quality=not real data)
  E.g. (i) model from the data; (ii) generate data from model

Masking methods

Information loss measures. Compare X and X' w.r.t. analysis (f)

$$IL_f(X, X') = divergence(f(X), f(X'))$$

- f: generic vs. specific (data uses)
  - Statistics
  - Machine learning: Clustering and classification
    For example, classification using decision trees
  - . . . specific measures for graphs



Masking methods

Dislosure risk. ... coming soon

Introduction

Disclosure risk assesment

Disclosure risk assesment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
  - Attribute disclosure: (e.g. learn about Alice's salary)
    - ⋆ Increase knowledge about an attribute of an individual
  - Identity disclosure: (e.g. find Alice in the database)
    - ⋆ Find/identify an individual in a masked file

Within machine learning, some attribute disclosure is expected.

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)

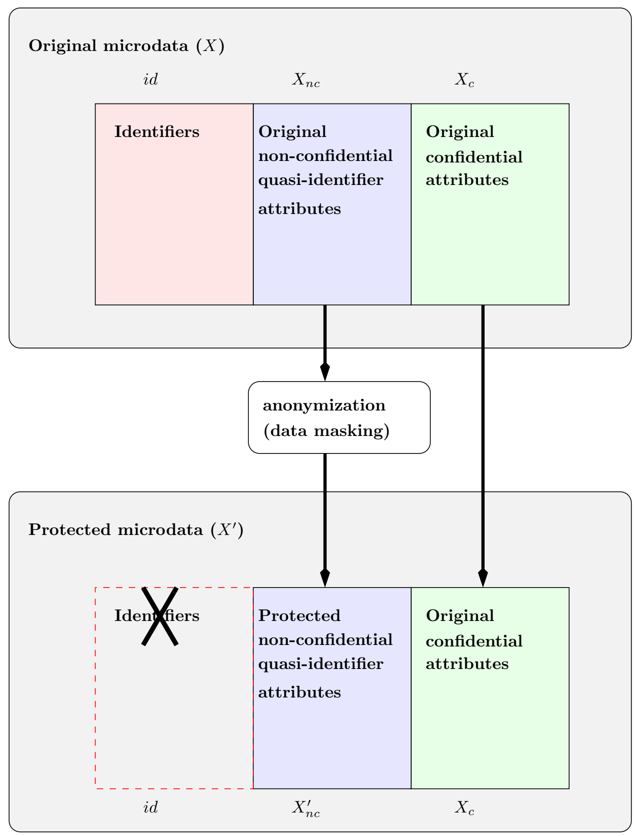
Examples. Privacy models / disclosure risk measures

	Attribute disclosure	Identity disclosure
Boolean	Differential privacy	k−Anonymity
Quantitative	Interval disclosure	Re−identification (record linkage) Uniqueness

Disclosure risk assesment

A scenario for identity disclosure: $X = id||X_{nc}||X_c$

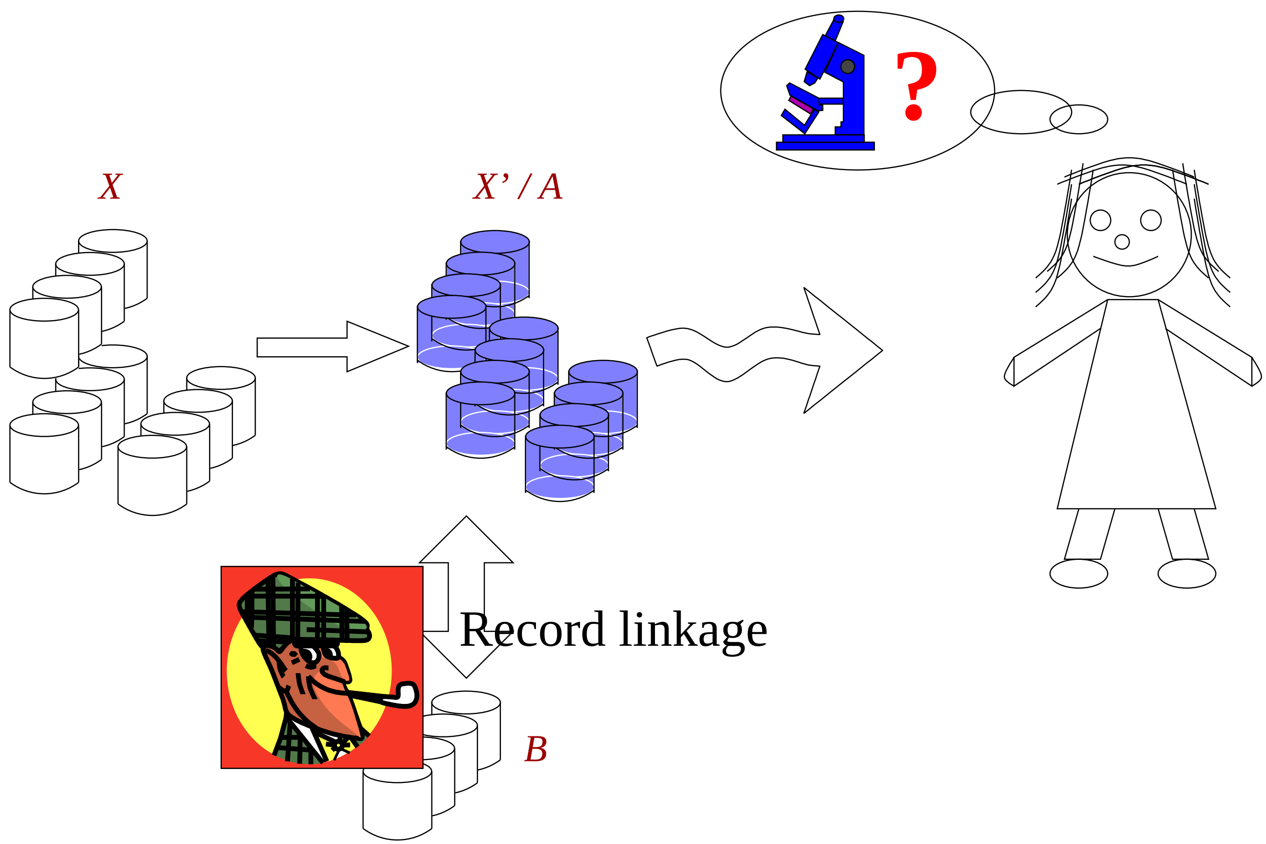
- Protection of the attributes
  - **Identifiers.** Usually removed or encrypted.
  - **Confidential.** X_c are usually not modified. $X'_c = X_c$.
  - **Quasi-identifiers.** Apply masking method ρ. $X'_{nc} = \rho(X_{nc})$.



Disclosure risk assesment

A scenario for identity disclosure: Reidentification

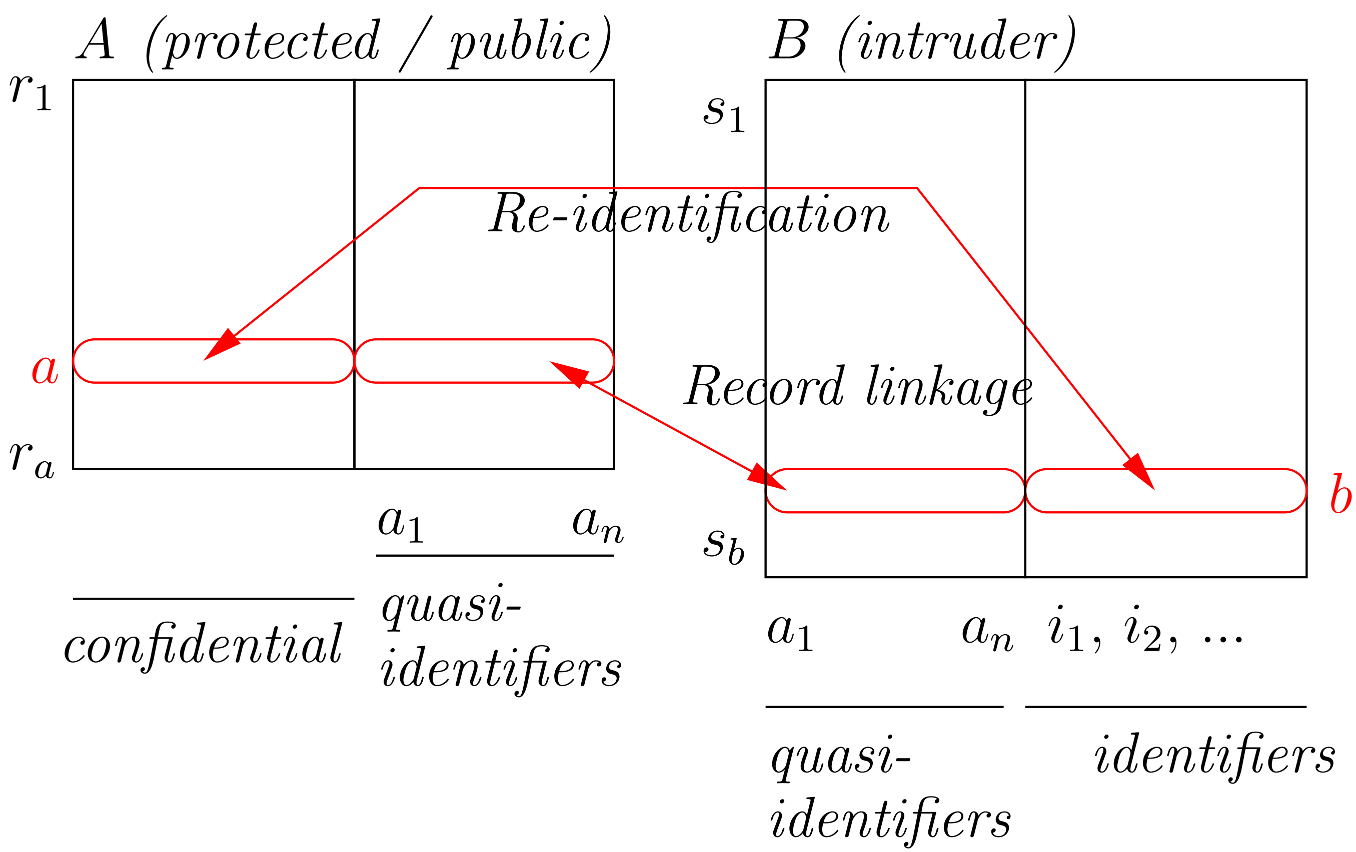
- A: File with the protected data set
- B: File with the data from the intruder (subset of original X)



Disclosure risk assesment

A 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

A scenario for identity disclosure. Reidentification

- Reidentification using the common attributes (quasi-identifiers):

Disclosure risk assesment

A scenario for identity disclosure. Reidentification

- Reidentification using the common attributes (quasi-identifiers): leads to identity disclosure

Disclosure risk assesment

A scenario for identity disclosure. Reidentification

- Reidentification using the common attributes (quasi-identifiers): leads to identity disclosure
- Attribute disclosure may be possible

Disclosure risk assesment

A scenario for identity disclosure. Reidentification

- Reidentification using the common attributes (quasi-identifiers): leads to 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

A scenario for identity disclosure. Reidentification

- 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

A scenario for identity disclosure. Reidentification

- 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

A scenario for identity disclosure. Reidentification

- 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

A scenario for identity disclosure. Reidentification

- 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)
    - $\star$ Other scenarios. E.g., synthetic data

Worst-case scenario

Worst-case scenario when measuring disclosure risk

Worst-case scenario

A scenario for identity disclosure. Reidentification

- Flexible scenario. Different assumptions on what available
  E.g., only partial information on individuals/characteristics
- Worst-case scenario for disclosure risk assessment
  (upper bound of disclosure risk)

Worst-case scenario

A scenario for identity disclosure. Reidentification

- Flexible scenario. Different assumptions on what available
  E.g., only partial information on individuals/characteristics
- Worst-case scenario for disclosure risk assessment
  (upper bound of disclosure risk)
  - Maximum information

Worst-case scenario

A scenario for identity disclosure. Reidentification

- Flexible scenario. Different assumptions on what available
  E.g., only partial information on individuals/characteristics
- Worst-case scenario for disclosure risk assessment
  (upper bound of disclosure risk)
  - Maximum information
  - Most effective reidentification method

Worst-case scenario

A scenario for identity disclosure. Reidentification

- Flexible scenario. Different assumptions on what available
  E.g., only partial information on individuals/characteristics
- Worst-case scenario for disclosure risk assessment
  (upper bound of disclosure risk)
  - Maximum information: Use original file to attack
  - Most effective reidentification method: Use ML
    Use information on the masking method (transparency)

Worst-case scenario

ML for reidentification (learning distances)

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:

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

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

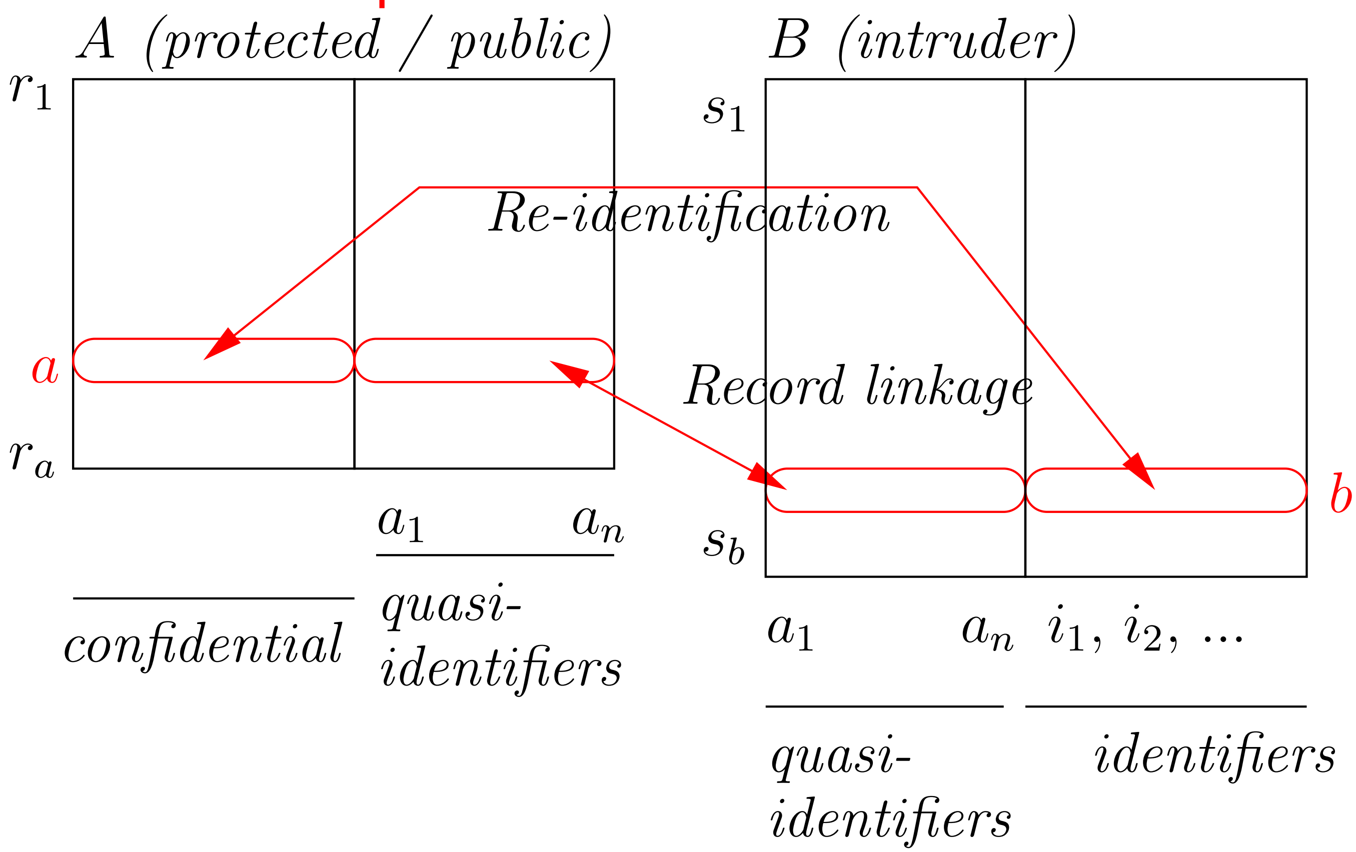
Worst-case: Optimal selection of the parameter p. How??

- 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 $\mathbb{C}$ (aggregation)
  - 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

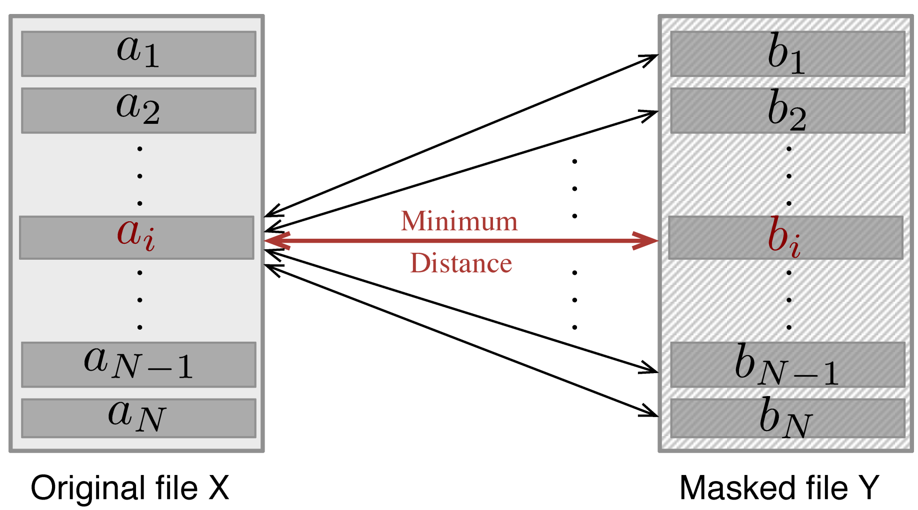
- Generic solution, using $\mathbb{C}$ with parameter p
- Goal (A and B aligned)
  - as much correct reidentifications as possible
  - For record i: $d(a_i, b_j) \geq d(a_i, b_i)$ for all j

Formalization of the problem

Machine Learning for distance-based record linkage

- Generic solution, using $\mathbb{C}$ with parameter p
- Goal (A and B aligned)
  - 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

$$d(a_i, b_j) \not\geq d(a_i, b_i)$$

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 $\mathbb{C} = WM$
- Formalization:

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

$$Subject\ to:$$

$$\begin{aligned} & WM_p(\mathit{diff}_1(a_i, b_j), \ldots, \mathit{diff}_n(a_i, b_j)) - \\ & \quad - WM_p(\mathit{diff}_1(a_i, b_i), \ldots, \mathit{diff}_n(a_i, b_i)) + CK_i > 0 \\ & K_i \in \{0, 1\} \\ & \sum_{i=1}^{n} p_i = 1 \\ & p_i \geq 0 \end{aligned}$$

Experiments and distances

Machine Learning for distance-based record linkage

- Distances considered through the following $\mathbb{C}$
  - Weighted mean.

    Weights: importance to the attributes

    Parameter: weighting vector n parameters

Experiments and distances

Machine Learning for distance-based record linkage

- Distances considered through the following $\mathbb{C}$
  - Weighted mean.
    Weights: importance to the attributes
    Parameter: weighting vector n parameters
  - OWA - linear combination of order statistics (weighted):
    Weights: to discard lower or larger distances
    Parameter: weighting vector n parameters

Experiments and distances

Machine Learning for distance-based record linkage

- Distances considered through the following $\mathbb{C}$
  - Choquet integral.
    Weights: interactions of sets of attributes ($\mu : 2^X \to [0, 1]$
    Parameter: non-additive measure: $2^n - 2$ parameters

Experiments and distances

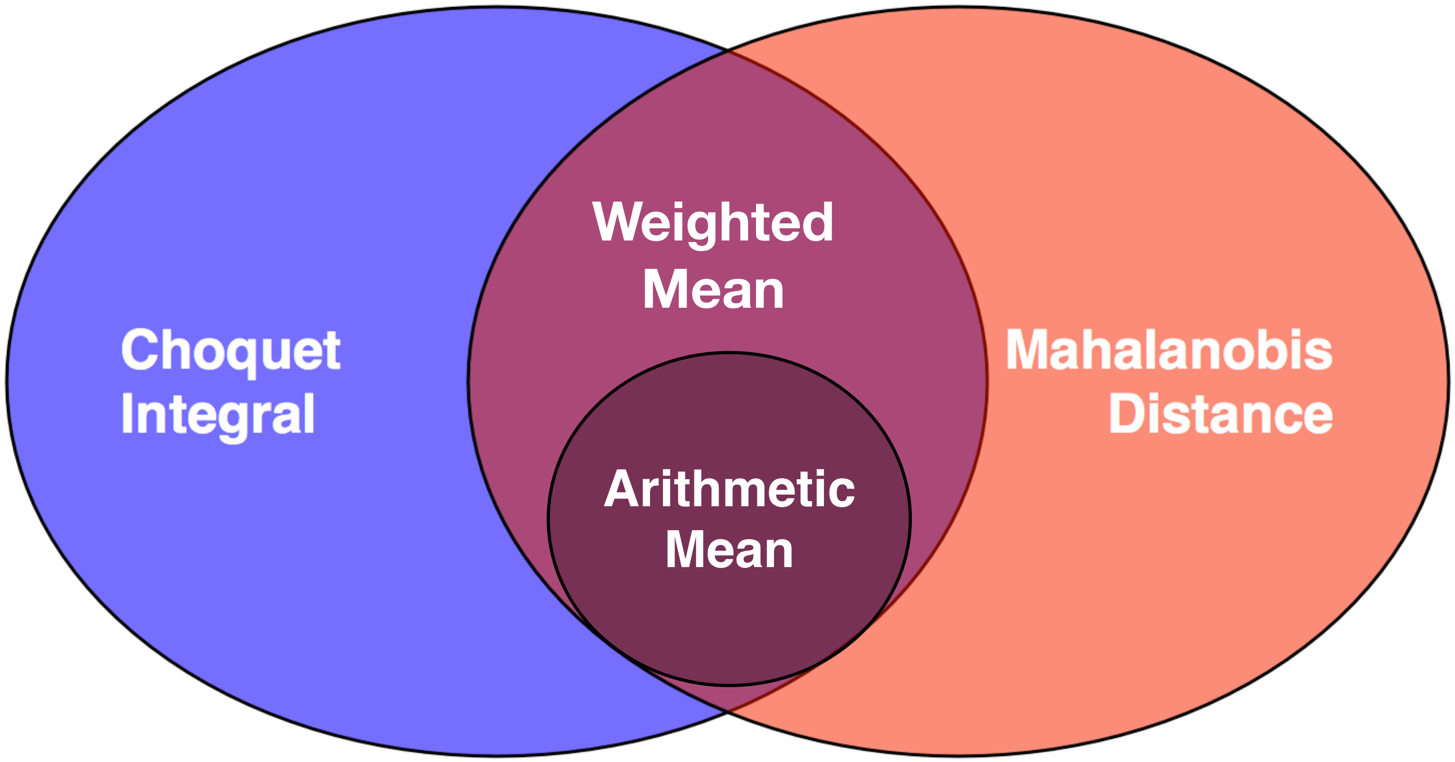
Machine Learning for distance-based record linkage

- Distances considered through the following $\mathbb{C}$
  - Choquet integral.
    Weights: interactions of sets of attributes ($\mu : 2^X \to [0, 1]$
    Parameter: non-additive measure: $2^n - 2$ parameters
  - Bilinear form - generalization of Mahalanobis distance
    Weights: interactions between pairs of attributes
    Parameter: square matrix: $n \times n$ parameters

Experiments and distances

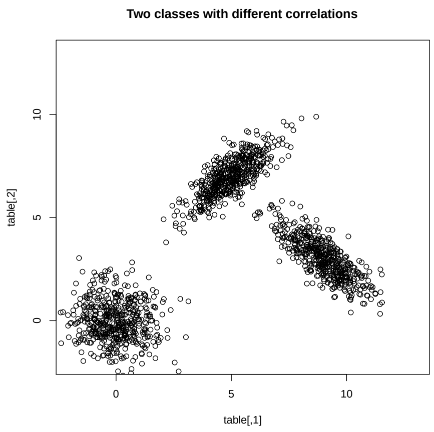
Machine Learning for distance-based record linkage

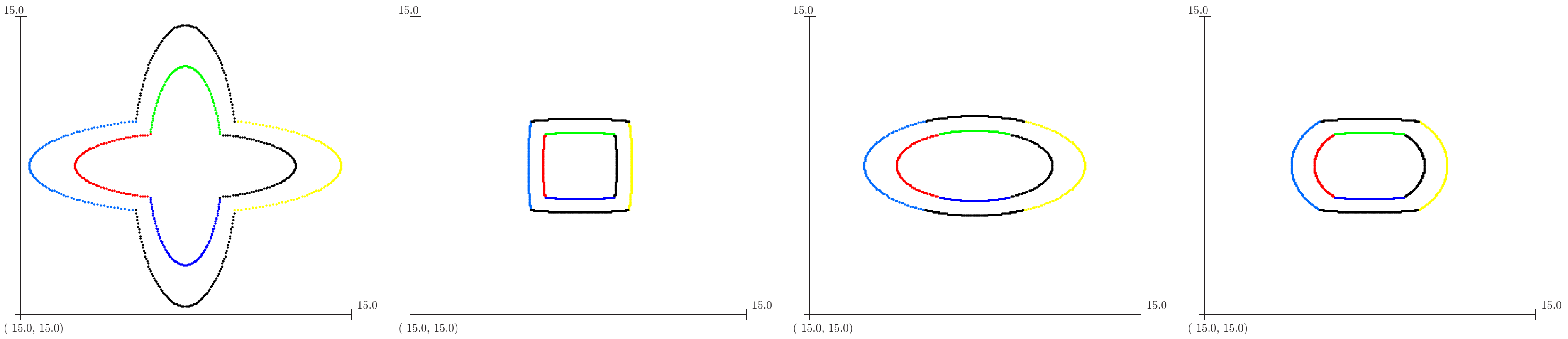
- Distances considered



Choquet integral. A fuzzy integral w.r.t. a fuzzy measure (non-additive measure). CI generalizes Lebesgue integral. Interactions.

Footnote: Mahalanobis / CI





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

d_m: distance; d_{NC}: positive; d_{PD}: positive-definite matrix

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

1h=3600; 1d = 86400s

- 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	SB	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

Transparency

Transparency

Transparency

Transparency: Definition

Transparency

Transparency.

- "the release of information about processes and even parameters used to alter data" (Karr, 2009).

Effect.

- Information Loss. Positive effect, less loss/improve inference
  E.g., noise addition $\rho(X) = X + \epsilon$ where ϵ s.t.
  $E(\epsilon) = 0$ and $Var(\epsilon) = kVar(X)$

$$Var(X') = Var(X) + kVar(X) = (1 + k)Var(X).$$

Transparency

Transparency.

- "the release of information about processes and even parameters used to alter data" (Karr, 2009).

Effect.

- Disclosure Risk. Negative effect, larger risk
  - Attack to single-ranking microaggregation (Winkler, 2002)
  - Formalization of the transparency attack (Nin, Herranz, Torra, 2008)
  - Attacks to microaggregation and rank swapping (Nin, Herranz, Torra, 2008)

Transparency

Transparency.

- "the release of information about processes and even parameters used to alter data" (Karr, 2009).

Effect.

- Disclosure Risk. Formalization
  - X and X' original and masked files, $\mathbf{V} = (V_1, \ldots, V_s)$ attributes
  - $B_j(x)$ set of masked records associated to x w.r.t. jth variable.
  - Then, for record x, the masked record x_ℓ corresponding to x is in the intersection of $B_j(x)$.

$$x_\ell \in \cap_j B_j(x).$$

- Worst case scenario in record linkage: upper bound of risk

Transparency

Attacking Rank Swapping

Transparency

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 ;

Transparency

Rank swapping.

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

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 $|X|$)

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.

- Intruder data are available

Transparency

Intruder perspective.

- Intruder data are available
- All protected values are available.

Transparency

Intruder perspective.

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

Transparency

Intruder perspective.

- Intruder data are available
- All protected values are available.
  I.e.,
  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 $B_j(a)$

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

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

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,

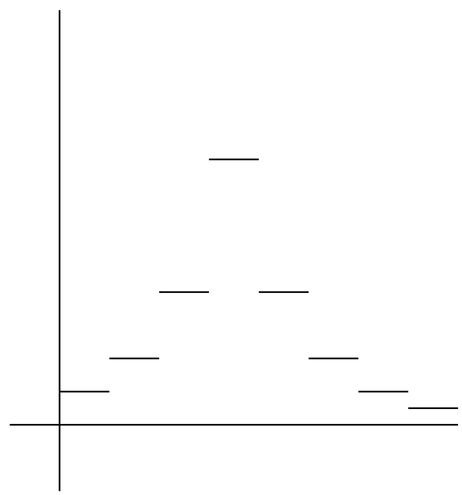
$$\cap B_j = X$$

Transparency

Approaches to avoid transparency attack in rank swapping.

- Rank swapping p-buckets. Select bucket B_s using

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



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

Information Loss

Information Loss

Information Loss

Information Loss. Compare X and X' w.r.t. analysis

$$IL_f(X, X') = divergence(f(X), f(X'))$$

- f: clustering (k-means).
  - Comparison of clusters by means of Rand, Jaccard indices
  - Comparison of clusters by means of F-measure
- f: classification (SVM, Naïve classifiers, k-NN, Decision Trees)
  - Comparison of accuracy

Summary

Summary

Experiments and distances

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

Thank you

References

Related references.

- D. Abril, G. Navarro-Arribas, V. Torra, Supervised Learning Using a Symmetric Bilinear Form for Record Linkage, Information Fusion 26 (2015) 144-153.
- 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.
- V. Torra, Fuzzy microaggregation for the transparency principle, accepted.
- V. Torra, A. Jonsson, G. Navarro-Arribas, J. Salas, Generation of spatial graphs for a given degree sequence, submitted.