Background

- MSc and PhD in Computer Science (with maths) (U.Polytechn. BCN) 1994
- U. Rovira i Virgili (Tarragona, Catalonia, Spain) - 1999
- Artificial Intelligence Research Institute -
  Spanish National Research Council (IIIA-CSIC, Barcelona) 1999-2014
- Professor (Skövde AI group) at U. of Skövde, 2014-

Research

- Approximate reasoning (since 1994, including non-additive measures, fuzzy sets theory, decision making)
- Data privacy (since 1999/2000)
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. Privacy and graphs

5. Summary
Introduction
Masking methods

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

- Data-driven or general purpose
  → anonymization methods / masking methods
- Computation-driven or specific purpose
  → cryptographic protocols, differential privacy
- Result-driven
Masking methods

Original microdata ($X$)

<table>
<thead>
<tr>
<th>$id$</th>
<th>$X_{nc}$</th>
<th>$X_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifiers</td>
<td>Original non-confidential quasi-identifier attributes</td>
<td>Original confidential attributes</td>
</tr>
</tbody>
</table>

Protected microdata ($X'$)

<table>
<thead>
<tr>
<th>$id$</th>
<th>$X'_{nc}$</th>
<th>$X_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifiers</td>
<td>Protected non-confidential quasi-identifier attributes</td>
<td>Original confidential attributes</td>
</tr>
</tbody>
</table>

anonymization (data masking)
**Approach** valid for different types of data

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

Original microdata (X) → Masking method → Protected microdata (X')

Data analysis

Result(X) → Disclosure Risk Measure

Data analysis

Result(X') → Information Loss Measure
Masking methods

Masking methods. (anonymization methods)
Masking methods. (anonymization methods)

- Perturbative.
  E.g. noise addition/multiplication, microaggregation, rank swapping
Masking methods (anonymization methods)

- Perturbative
  - E.g. noise addition/multiplication, microaggregation, rank swapping
- Non-perturbative
  - E.g. generalization, suppression
Masking methods. (anonymization methods)

- Perturbative.
  E.g. noise addition/multiplication, microaggregation, rank swapping
- Non-perturbative
  E.g. generalization, suppression
- Synthetic data generators
Masking methods

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

$$IL_f(X, X') = \text{divergence}(f(X), f(X'))$$

- $f$: generic vs. specific (data uses)
  - Statistics
  - Machine learning: Clustering and classification
  - ... specific measures for graphs
Masking methods

Disclosure risk. … coming soon
Disclosure risk assessment
Disclosure risk assessment

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 assessment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures
Disclosure risk assessment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures
  (minimize information loss vs. multiobjective optimization)
Disclosure risk.

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

**Examples. Privacy models / disclosure risk measures**

<table>
<thead>
<tr>
<th>Boolean</th>
<th>Attribute disclosure</th>
<th>Identity disclosure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differential privacy</td>
<td></td>
<td>k--Anonymity</td>
</tr>
<tr>
<td>Interval disclosure</td>
<td></td>
<td>Re--identification (record linkage)</td>
</tr>
<tr>
<td>Uniqueness</td>
<td></td>
<td>Uniqueness</td>
</tr>
</tbody>
</table>
Disclosure risk assessment

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 \( \rho \). \( X'_{nc} = \rho(X_{nc}) \).
Disclosure risk assessment

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 \( X \))

---

**Re-identification**

**Record linkage**

- **confidential**
- **quasi-identifiers**
- **identifiers**
Disclosure risk assessment

A scenario for identity disclosure. Reidentification

- Reidentification using the common attributes (quasi-identifiers):
Disclosure risk assessment

A scenario for identity disclosure. Reidentification

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

A scenario for identity disclosure. Reidentification

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

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 assessment

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 assessment

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.
    → intruder with information on only some individuals
Disclosure risk assessment

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.
    → intruder with information on only some individuals
    → intruder with information on only some characteristics
Disclosure risk assessment

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.
    - intruder with information on only some individuals
    - intruder with information on only some characteristics
  - But also,
    - $B$ with a schema different to the one of $A$ (different attributes)
    - Other scenarios. E.g., synthetic data
Worst-case scenario when measuring disclosure risk
Worst-case scenario

Worst-case scenario for disclosure risk assessment (upper bound of disclosure risk)
Worst-case scenario

Worst-case scenario for disclosure risk assessment (upper bound of disclosure risk)

- Maximum information
Worst-case scenario for disclosure risk assessment
(upper bound of disclosure risk)

- Maximum information
- Most effective reidentification method
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
Worst-case scenario

ML for reidentification
(learning distances)
Worst-case scenario for disclosure risk assessment

- Distance-based record linkage
- Parametric distances with best parameters
  
  E.g.,
  
  o Weighted Euclidean distance
Worst-case scenario for disclosure risk assessment

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

\[ d^2(a, b) = \left\| \frac{1}{n}(a - b) \right\|^2 = \sum_{i=1}^{n} \frac{1}{n} (\text{diff}_i(a, b)) \]

\[ = W M_p(\text{diff}_1(a, b), \ldots, \text{diff}_n(a, b)) \]

with \( p = (1/n, \ldots, 1/n) \) and

\[ \text{diff}_i(a, b) = \left( \frac{(a_i - \bar{a}_i)}{\sigma(a_i)} - \frac{(b_i - \bar{b}_i)}{\sigma(b_i)} \right)^2 \]

- \( p_i = 1/n \) 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)

\[ d^2(a, b) = \text{WM}_p(\text{diff}_1(a, b), \ldots, \text{diff}_n(a, b)) \]

with arbitrary vector \( p = (p_1, \ldots, p_n) \) and

\[ \text{diff}_i(a, b) = ((a_i - \bar{a}_i)/\sigma(a_i) - (b_i - \bar{b}_i)/\sigma(b_i))^2 \]
Worst-case scenario

Worst-case scenario for disclosure risk assessment

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

\[ d^2(a, b) = WM_p(diff_1(a, b), \ldots, diff_n(a, b)) \]

with arbitrary vector \( p = (p_1, \ldots, p_n) \) and

\[ diff_i(a, b) = (((a_i - \bar{a}_i)/\sigma(a_i) - (b_i - \bar{b}_i)/\sigma(b_i))^2 \]

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

• Supervised machine learning approach
• Using an optimization problem
Worst-case scenario for disclosure risk assessment

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

\[
d^2(a, b) = C_p(diff_1(a, b), \ldots, diff_n(a, b))
\]

with parameter \( p \) and

\[
diff_i(a, b) = \left(\frac{(a_i - \bar{a}_i)}{\sigma(a_i)} - \frac{(b_i - \bar{b}_i)}{\sigma(b_i)}\right)^2
\]
Worst-case scenario for disclosure risk assessment

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

$$d^2(a, b) = \mathbb{C}_p(diff_1(a, b), \ldots, diff_n(a, b))$$

with parameter $p$ and

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

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

- Supervised machine learning approach
- Using an optimization problem
Worst-case scenario for distance-based record linkage

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

  \[ A \ (\text{protected} \ / \ \text{public}) \]

  \[ B \ (\text{intruder}) \]

  \[ a_1 \ a_n \]

  confidential   quasi-identifiers

  \[ s_1 \ s_b \]

  quasi-identifiers

  \[ a_1 \ a_n \ i_1, i_2, \ldots \]

  identifiers
Formalization of the problem

Machine Learning for distance-based record linkage

- Generic solution, using
  - an arbitrary combination function \( C \) (aggregation)
  - with parameter \( p \)

\[
d(a_i, b_j) = C_p(d_{1}(a, b), \ldots, d_{n}(a, b))
\]
Formalization of the problem

Machine Learning for distance-based record linkage

• Generic solution, using $C$ 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 $C$ 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,

$$C_p(d_1(a_i, b_j), \ldots, d_n(a_i, b_j)) \geq C_p(d_1(a_i, b_i), \ldots, d_n(a_i, b_i))$$
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
    \[ d(a_i, b_j) \geq d(a_i, b_i) \text{ for all } j \]
  - 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 $C$

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

\[
d(a_i, b_j) + CK_i \geq d(a_i, b_i)
\]

That is,

\[
C_p(\text{diff}_1(a_i, b_j), \ldots, \text{diff}_n(a_i, b_j)) + CK_i \geq C_p(\text{diff}_1(a_i, b_i), \ldots, \text{diff}_n(a_i, b_i))
\]
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
  \[
  d(a_i, b_j) \geq d(a_i, b_i) \text{ for all } j
  \]
  ○ $K_i \in \{0, 1\}$, 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:

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

Subject to:

$$C_p(d_{1}(a_i, b_j), \ldots, d_{n}(a_i, b_j)) - C_p(d_{1}(a_i, b_i), \ldots, d_{n}(a_i, b_i)) + CK_i > 0$$

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

Additional constraints according to $C$
Formalization of the problem

Machine Learning for distance-based record linkage

- Example: the case of the weighted mean $C = WM$
- Formalization:

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

Subject to:

$$WM_p(\text{diff}_1(a_i, b_j), \ldots, \text{diff}_n(a_i, b_j)) -$$

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

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

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

$$p_i \geq 0$$
Experiments and distances

Machine Learning for distance-based record linkage

- Distances considered 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 $\mathcal{C}$
  - 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 $C$
  - Choquet integral.
    - Weights: interactions of sets of attributes ($\mu : 2^X \rightarrow [0, 1]$)
    - 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 ($\mu : 2^X \rightarrow [0, 1]$)
    - 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)
  - \textbf{M4-33}: 4 attributes microaggregated in groups of 2 with $k = 3$.
  - \textbf{M4-28}: 4 attributes, 2 attributes with $k = 2$, and 2 with $k = 8$.
  - \textbf{M4-82}: 4 attributes, 2 attributes with $k = 8$, and 2 with $k = 2$.
  - \textbf{M5-38}: 5 attributes, 3 attributes with $k = 3$, and 2 with $k = 8$.
  - \textbf{M6-385}: 6 attributes, 2 attributes with $k = 3$, 2 attributes with $k = 8$, and 2 with $k = 5$.
  - \textbf{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.**

<table>
<thead>
<tr>
<th>Distance</th>
<th>M4-33</th>
<th>M4-28</th>
<th>M4-82</th>
<th>M5-38</th>
<th>M6-385</th>
<th>M6-853</th>
</tr>
</thead>
<tbody>
<tr>
<td>(d^2AM)</td>
<td>84.00</td>
<td>68.50</td>
<td>71.00</td>
<td>39.75</td>
<td>78.00</td>
<td>84.75</td>
</tr>
<tr>
<td>(d^2MD)</td>
<td>94.00</td>
<td>90.00</td>
<td>92.75</td>
<td>88.25</td>
<td>98.50</td>
<td>98.00</td>
</tr>
<tr>
<td>(d^2WM)</td>
<td>95.50</td>
<td>93.00</td>
<td>94.25</td>
<td>90.50</td>
<td>99.25</td>
<td>98.75</td>
</tr>
<tr>
<td>(d^2WM_m)</td>
<td>95.50</td>
<td>93.00</td>
<td>94.25</td>
<td>90.50</td>
<td>99.25</td>
<td>98.75</td>
</tr>
<tr>
<td>(d^2CI)</td>
<td>95.75</td>
<td>93.75</td>
<td>94.25</td>
<td>91.25</td>
<td>99.75</td>
<td>99.25</td>
</tr>
<tr>
<td>(d^2CI_m)</td>
<td>95.75</td>
<td>93.75</td>
<td>94.25</td>
<td>90.50</td>
<td>99.50</td>
<td>98.75</td>
</tr>
<tr>
<td>(d^2SB_{NC})</td>
<td>96.75</td>
<td>94.5</td>
<td>95.25</td>
<td>92.25</td>
<td>99.75</td>
<td>99.50</td>
</tr>
<tr>
<td>(d^2SB)</td>
<td>96.75</td>
<td>94.5</td>
<td>95.25</td>
<td>92.25</td>
<td>99.75</td>
<td>99.50</td>
</tr>
<tr>
<td>(d^2SB_{PD})</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>99.25</td>
</tr>
</tbody>
</table>

\(d_m\): distance; \(d_{NC}\): positive; \(d_{PD}\): positive-definite matrix
Experiments and distances

Machine Learning for distance-based record linkage

- Computation time comparison (in seconds).

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

$1h=3600;\ 1d=86400s$

- Constraints specific to weighted mean and Choquet integral for distances

<table>
<thead>
<tr>
<th></th>
<th>$d^2WM_m$</th>
<th>$d^2CI_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additional Constraints</td>
<td>$\sum_{i=1}^{n} p_i = 1$  $p_i &gt; 0$</td>
<td>$\mu(\emptyset) = 0$  $\mu(V) = 1$  $\mu(A) \leq \mu(B)$ when $A \subseteq B$</td>
</tr>
<tr>
<td>Total Constr.</td>
<td>$N(N - 1) + N + 1 + n$</td>
<td>$N(N - 1) + N + 2 + (\sum_{k=2}^{n} \binom{n}{k} k) + \binom{n}{2}$</td>
</tr>
</tbody>
</table>
## Experiments and distances

### Machine Learning for distance-based record linkage

- **A summary of the experiments**

<table>
<thead>
<tr>
<th></th>
<th>AM</th>
<th>MD</th>
<th>WM</th>
<th>OWA</th>
<th>SB</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Computation</strong></td>
<td>Very fast</td>
<td>Very fast</td>
<td>Fast</td>
<td>regular</td>
<td>Hard</td>
<td>Hard</td>
</tr>
<tr>
<td><strong>Results</strong></td>
<td>Worse</td>
<td>Good</td>
<td>Good</td>
<td>Bad</td>
<td>Very Good</td>
<td>Very Good</td>
</tr>
<tr>
<td><strong>Information</strong></td>
<td>No</td>
<td>No</td>
<td>Few</td>
<td>Few</td>
<td>Large</td>
<td>Large</td>
</tr>
</tbody>
</table>

Vicenç Torra; Data privacy
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 $\epsilon$ 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, $V = (V_1, \ldots, V_s)$ attributes
  - $B_j(x)$ set of masked records associated to $x$ w.r.t. $j$th variable.
  - Then, for record $x$, the masked record $x_\ell$ corresponding to $x$ is in the intersection of $B_j(x)$.

\[ x_\ell \in \bigcap_j B_j(x). \]

- Worst case scenario in record linkage: upper bound of risk
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 \(\ell\) randomly and uniformly chosen from the limited range \([i + 1, \min(n, i + p \ast |X|/100)]\) ;

Swap \(a_i\) with \(a_\ell\) ;

**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)
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 $|X|$)
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 \textit{(method, parameter)} to attack
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

$\rightarrow (method, parameter) = (rank\ swapping, p)$
Intruder perspective.

- Intruder data are available
Intruder perspective.

- Intruder data are available
- All protected values are available.
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 $a_\ell$ in the range

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

Intruder’s attack for a single attribute \( V_j \)

- Define \( B_j(a) \)
  
  the set of masked records that can be the masked version of \( a \)
Intruder’s attack for a single 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 a single 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:
**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:

\[
x'_{\ell} \in \bigcap_{1 \leq j \leq c} B_j(x_i).
\]
Transparency

Intruder’s attack for a single attribute $V_j$

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

$$\exists \ell \in B_j(a)$$

Intruder’s attack for all available attributes

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

$$\exists \ell \in \bigcap_{1 \leq j \leq c} B_j(x_i).$$

No uncertainty!
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
Intruder’s attack. Example.

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

<table>
<thead>
<tr>
<th>Original file</th>
<th>Masked file</th>
<th>$B(x_{2j})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>$a_2$</td>
<td>$a_3$</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>9</td>
</tr>
</tbody>
</table>
Intruder’s attack. Example.

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

<table>
<thead>
<tr>
<th>Original file</th>
<th>Masked file</th>
<th>$B(x_{21})$</th>
<th>$B(x_{22})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>$a_2$</td>
<td>$a_3$</td>
<td>$a_4$</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>
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

\[(5, 5, 8, 1)\]
Intruder’s attack. Application.

- Data:
  - Census (1080 records, 13 attributes)
  - EIA (4092 records, 10 attributes)
- Rank swapping parameter:
  - $p = 2, \ldots, 20$
### Intruder’s attack. Result

<table>
<thead>
<tr>
<th></th>
<th>Census</th>
<th>EIA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RSLD</td>
<td>DLD</td>
</tr>
<tr>
<td>rs 2</td>
<td>77.73</td>
<td>73.52</td>
</tr>
<tr>
<td>rs 4</td>
<td>66.65</td>
<td>58.40</td>
</tr>
<tr>
<td>rs 6</td>
<td>54.65</td>
<td>43.76</td>
</tr>
<tr>
<td>rs 8</td>
<td>41.28</td>
<td>32.13</td>
</tr>
<tr>
<td>rs 10</td>
<td>29.21</td>
<td>23.64</td>
</tr>
<tr>
<td>rs 12</td>
<td>19.87</td>
<td>18.96</td>
</tr>
<tr>
<td>rs 14</td>
<td>16.14</td>
<td>15.63</td>
</tr>
<tr>
<td>rs 16</td>
<td>13.81</td>
<td>13.59</td>
</tr>
<tr>
<td>rs 18</td>
<td>12.21</td>
<td>11.50</td>
</tr>
<tr>
<td>rs 20</td>
<td>10.88</td>
<td>10.87</td>
</tr>
</tbody>
</table>
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
Avoiding Transparency Attack in Rank Swapping
Avoiding transparency attack in rank swapping.

- Enlarge the $B_j$ set to encompass the whole file.
Avoiding transparency attack in rank swapping.

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

\[ \cap B_j = X \]
Approaches to avoid transparency attack in rank swapping.

- Rank swapping $p$-buckets. Select bucket $B_s$ using

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

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

**Approaches.** As for databases **owner privacy** (vs. user privacy)

- Perturbative. \( X' = X + \epsilon \)
- Nonperturbative. \( X' = \text{generalization}(X) \)
- Synthetic data. \( M = \text{Model}(X). \text{Draw } X' \text{ from } M \)

**Disclosure risk.** Attacks (knowledge)

- degree of a node,
- neighborhood of a node (links and non-links),
- subgraph
Privacy and Graphs

Approaches. Synthetic spatial graphs

- Degree sequence
- Nodes on a map according to a density
- Edges according to nearness

Algorithm.

- Heuristic approach for edge assignment which leads to multigraphs
- Correction of multiple edges
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
Related references.

- V. Torra, Fuzzy microaggregation for the transparency principle, accepted.