# Masking Methods

Vicenç Torra

November, 2022

Umeå University, Sweden

# Outline

1. Introduction

2. Protection for microdata (files)

3. Information loss measures

4. Visualization

5. Summary

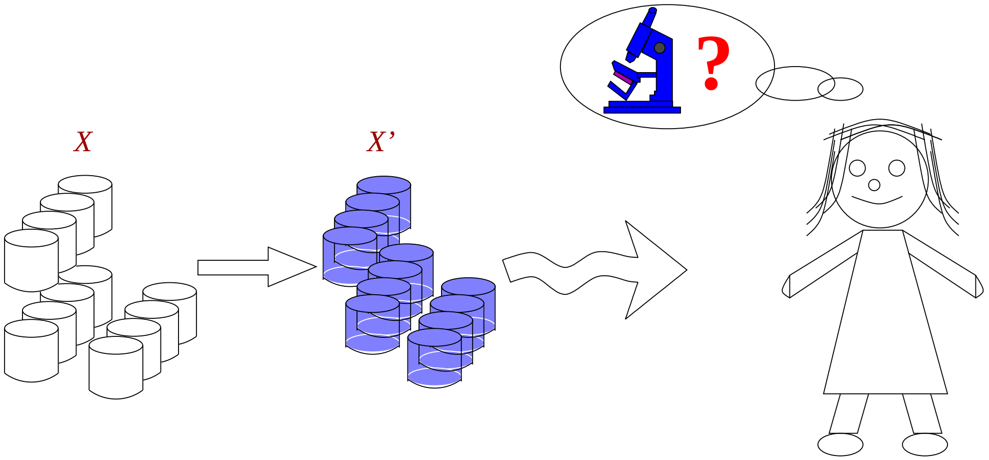
Introduction

Data-driven protection procedures

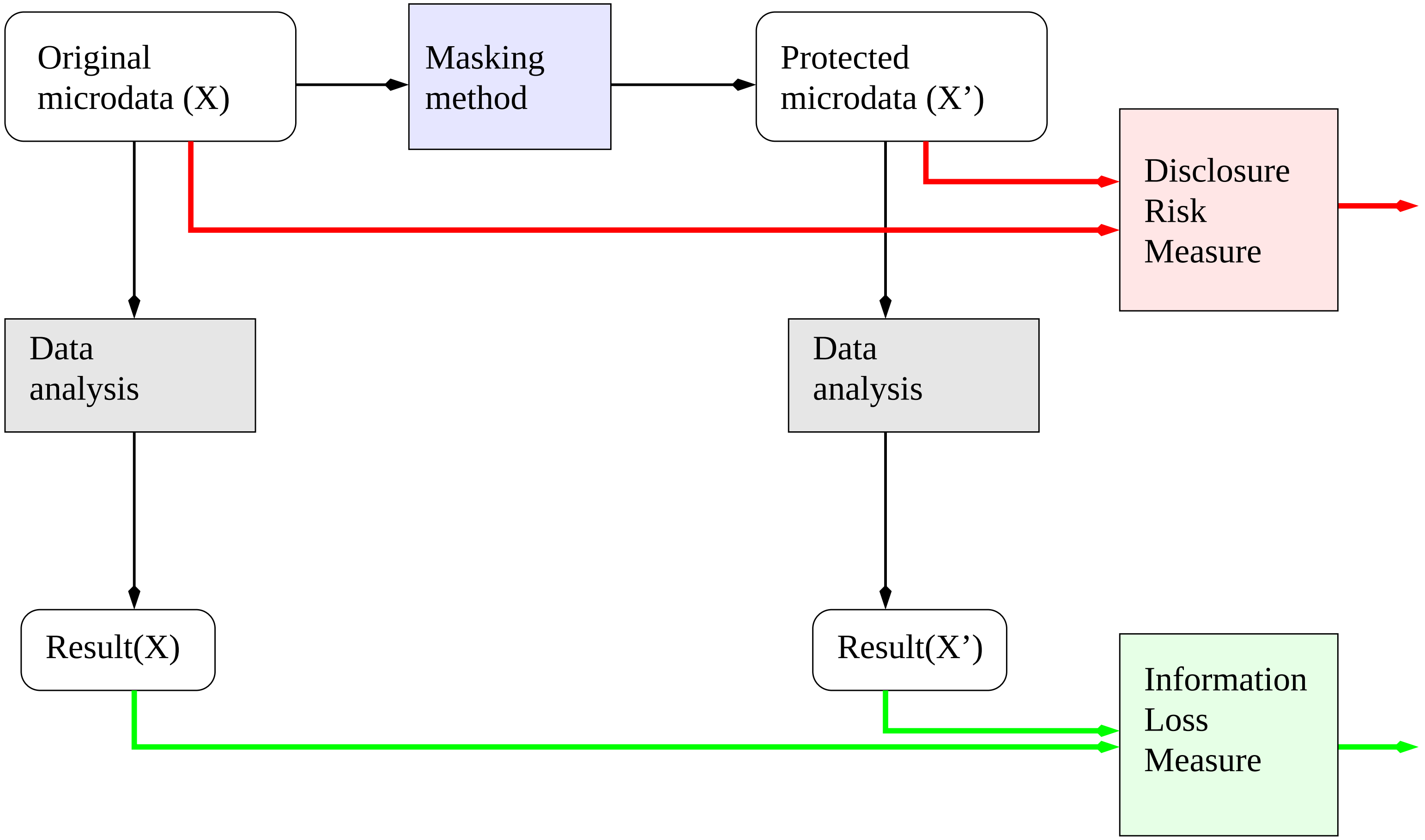
# Data-driven protection methods

Data protection methods

- Data files / Microdata
- Aggregate data / Tabular data
- Other types of data
  - graphs for online social networks
  - search and access logs
  - documents or index of documents



# Data-driven protection methods



# Protection for microdata (files)

# Data-driven protection methods

Data protection methods for datafiles / microdata

- Perturbative methods
- Non-perturbative methods
- Synthetic data generators

# Data-driven protection methods

Data protection methods for datafiles / microdata

- **Perturbative.** The original data set is distorted in some way, and the new data set might contain some erroneous information.

# Data-driven protection methods

Data protection methods for datafiles / microdata

- **Perturbative.** The original data set is distorted in some way, and the new data set might contain some erroneous information.
  For example, noise is added to an attribute following a $N(0, a)$ for a given $a$.
  Some combinations of values disappear, and, new combinations appear in the protected data set.
  At the same time, combinations in the protected data set no longer correspond to the ones in the original data set. This obfuscation makes disclosure difficult for intruders.

# Data-driven protection methods

Data protection methods for datafiles / microdata

- **Non-perturbative.** Protection is achieved through replacing an original value by another one that is not incorrect but less specific.

# Data-driven protection methods

Data protection methods for datafiles / microdata

- **Non-perturbative.** Protection is achieved through replacing an original value by another one that is not incorrect but less specific. For example, we replace a real number by an interval.
  In general, non-perturbative methods reduce the level of detail of the data set. This detail reduction causes different records to have the same combinations of values, which makes disclosure difficult to intruders.

# Data-driven protection methods

Data protection methods for datafiles / microdata

- **Synthetic Data Generators.** In this case, instead of distorting the original data, new artificial data is generated and used to substitute the original values.

  Formally, synthetic data generators build a data model from the original data set and, subsequently, a new (protected) data set is randomly generated constrained by the model computed.

# Data-driven protection methods

Data types

- Numerical data
- Categorical data: ordinal and nominal scale
  - Ordinal: $<$ (elements can be ordered)
  - No order predefined
- Longitudinal data and time series
- Location data
- Graphs and social networks
- Logs
  - search and access logs

# Data-driven protection methods

Data protection methods for datafiles / microdata

- Perturbative methods
  - ○ Noise addition, Microaggregation, Rank Swapping, . . .
- Non-perturbative methods
  - ○ Suppression, top and bottom coding, . . .
- Synthetic data generators
  - ○ IPSO, . . .

# Perturbative Methods

# Rank swapping

# Masking Methods for Microdata: Perturbative methods

**Rank swapping**:

- Description with parameter $p$
  - Values are ordered in increasing order
    We assume them ordered $x_{ij} \leq x_{lj}$ for all $1 \leq i < l \leq n$
  - Each ranked value $x_{ij}$ is swapped with another ranked value $x_{lj}$ randomly chosen within a restricted range $i < l \leq i + p$
- In applications, each variable is masked independently
- The larger the $p$, the larger the information loss, and the lower the risk

# Masking Methods for Microdata: Perturbative methods

## **Rank swapping**: Example I: Protection

- Four variables with values $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
- rank swapping with $p = 2$.

| Original file | | | | Protected file | | | |
|---|---|---|---|---|---|---|---|
| $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a'_1$ | $a'_2$ | $a'_3$ | $a'_4$ |
| 8 | 9 | 1 | 3 | 10 | 10 | 3 | 5 |
| 6 | 7 | 10 | 2 | 5 | 5 | 8 | 1 |
| 10 | 3 | 4 | 1 | 8 | 4 | 2 | 2 |
| 7 | 1 | 2 | 6 | 9 | 2 | 4 | 4 |
| 9 | 4 | 6 | 4 | 7 | 3 | 5 | 6 |
| 2 | 2 | 8 | 8 | 4 | 1 | 10 | 10 |
| 1 | 10 | 3 | 9 | 3 | 9 | 1 | 7 |
| 4 | 8 | 7 | 10 | 2 | 6 | 9 | 8 |
| 5 | 5 | 5 | 5 | 6 | 7 | 6 | 3 |
| 3 | 6 | 9 | 7 | 1 | 8 | 7 | 9 |

# Masking Methods for Microdata: Perturbative methods

**Rank swapping**: Example II: Information Loss and Disclosure Risk

- Information Loss: IL = 39.22
- Disclosure Risk:
  - DLD = 17.5
  - PLD = 0.0
  - ID = 44.81
  - DR = 0.25DLD + 0.25PLD + 0.5 ID = 26.78
- Score = (IL+DR)/2 = 33
- DR as an average of four scenarios:

| | DBRL | PRL |
|---|---|---|
| $\{V_1\}$ | 0 | 0 |
| $\{V_1, V_2\}$ | 2 | 0 |
| $\{V_1, V_2, V_3\}$ | 4 | 0 |
| $\{V_1, V_2, V_3, V_4\}$ | 1 | 0 |
| Average | 17.5 | 0 |

# Masking Methods for Microdata: Perturbative methods

**Rank swapping**: Specific record linkage for rank swapping

- Rank swapping record linkage (RS-RL).
  - ○ If we know $p$, a given intruder's (original) record can only generate at most $2p$ records

# Masking Methods for Microdata: Perturbative methods

**Rank swapping**: Specific record linkage for rank swapping

- Rank swapping record linkage (RS-RL).
  - ○ If we know $p$, a given intruder's (original) record can only generate at most $2p$ records
  - ○ For each $x_{ij}$ of the intruder,
  - ○ there exists a computable set $B(x_{ij})$ of $2p$ masked records, that can be generated from the original record $x_i$

# Masking Methods for Microdata: Perturbative methods

**Rank swapping**: Example III: Specific record linkage for rank swapping

- Record (intruder) $x_2 = (6, 7, 10, 2)$, $p = 2$ and first variable $x_{21} = 6$
  - $B(x_{21} = 6) = \{(4, 1, 10, 10), (5, 5, 8, 1), (6, 7, 6, 3), (7, 3, 5, 6), (8, 4, 2, 2)\}$

| Original file | | | | Protected 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 | |

# Masking Methods for Microdata: Perturbative methods

**Rank swapping**: Specific record linkage for rank swapping

- Rank swapping record linkage (RS-RL).
  - If we know $p$, a given intruder's (original) record can only generate at most $2p$ records

# Masking Methods for Microdata: Perturbative methods

**Rank swapping**: Specific record linkage for rank swapping

- Rank swapping record linkage (RS-RL).
  - ○ If we know $p$, a given intruder's (original) record can only generate at most $2p$ records
  - ○ For each $x_{ij}$ of the intruder,
  - ○ there exists a computable set $B(x_{ij})$ of $2p$ masked records, that can be generated from the original record $x_i$
- It should happen that the masked record is in all $B(x_{ij})$

$$x'_\ell \in \cap_{1 \leq j \leq c} B(x_{ij}).$$

# Masking Methods for Microdata: Perturbative methods

**Rank swapping**: Example IV: Specific record linkage for rank swapping

- Record (intruder) $x_2 = (6, 7, 10, 2)$, $p = 2$ and 2nd var. $x_{22} = 7$
  - ○ $B(x_{22} = 7) = \{(5, 5, 8, 1), (2, 6, 9, 8), (6, 7, 6, 3), (1, 8, 7, 9), (3, 9, 1, 7)\}$

| Original file | | | | Protected 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 |

# Masking Methods for Microdata: Perturbative methods

**Rank swapping:** Example V: Specific record linkage for rank swapping

- Similarly:
  - $B(x_{21} = 6) = \{(4,1,10,10),(5,5,8,1),(6,7,6,3),(7,3,5,6),(8,4,2,2)\}$
  - $B(x_{22} = 7) = \{(5,5,8,1),(2,6,9,8),(6,7,6,3),(1,8,7,9),(3,9,1,7)\}$
  - $B(x_{23} = 10) = \{(5,5,8,1),(2,6,9,8),(4,1,10,10)\}$
  - $B(x_{24} = 2) = \{(5,5,8,1),(8,4,2,2),(6,7,6,3),(9,2,4,4)\}$
- The intersection of these sets ...
  - is the single record $(5,5,8,1)$.

    $\rightarrow$ this is the correct link

    When several records are present, we apply standard record linkage.

# Masking Methods for Microdata: Perturbative methods

**Rank swapping**: Example VI: Specific record linkage for rank swapping

- Scores:
  - ○ With previous record linkage algorithm:
    - ⋆ DR = 0.25DLD + 0.25PLD + 0.5 ID = 26.78
    - ⋆ Score = (IL+DR)/2 = 33
  - ○ Using only RS-RL:
    - ⋆ DR = 0.5RS-RL + 0.5 ID = 43.655
    - ⋆ Score = (IL+DR)/2 = 41.44

| | DBRL | PRL | RS-RL |
|---|---|---|---|
| $\{V_1\}$ | 0 | 0 | 0 |
| $\{V_1, V_2\}$ | 2 | 0 | 2 |
| $\{V_1, V_2, V_3\}$ | 4 | 0 | 7 |
| $\{V_1, V_2, V_3, V_4\}$ | 1 | 0 | 8 |
| Average | 17.5 | 0 | 42.5 |

# Masking Methods for Microdata: Perturbative methods

**Rank swapping**: Average of 1-7 variables

- Application to Census and EIA files
- DR = 0.1666 RSLD + 0.1666 DLD + 0.166 PLD + 0.5 ID
- DR = 0.25 DLD + 0.25 PLD + 0.5 ID

| | Census | | | | | | EIA | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | IL | RSLD | DLD | PLD | ID | Score | IL | RSLD | DLD | PLD | ID | Score |
| rs 2 | 3.89 | 77.73 | 73.52 | 71.28 | 93.98 | 42.63 | 4.24 | 43.27 | 21.71 | 16.85 | 93.10 | 28.06 |
| rs 4 | 6.54 | 66.65 | 58.40 | 42.92 | 83.09 | 36.67 | 9.67 | 12.54 | 10.61 | 4.79 | 82.09 | 21.89 |
| rs 6 | 10.57 | 54.65 | 43.76 | 22.49 | 72.12 | 31.93 | 14.63 | 7.69 | 7.40 | 2.03 | 72.21 | 21.42 |
| rs 8 | 16.54 | 41.28 | 32.13 | 11.74 | 62.11 | 29.16 | 18.71 | 6.12 | 5.98 | 1.12 | 63.90 | 21.61 |
| rs 10 | 20.18 | 29.21 | 23.64 | 6.03 | 53.28 | 26.31 | 22.87 | 5.60 | 5.19 | 0.69 | 57.09 | 22.37 |
| rs 12 | 23.46 | 19.87 | 18.96 | 3.46 | 47.17 | 24.77 | 26.60 | 5.39 | 4.87 | 0.51 | 51.64 | 23.25 |
| rs 14 | 28.93 | 16.14 | 15.63 | 2.06 | 43.39 | 25.86 | 29.42 | 5.28 | 4.55 | 0.32 | 47.49 | 23.91 |
| rs 16 | 35.16 | 13.81 | 13.59 | 1.29 | 40.78 | 27.97 | 32.38 | 5.19 | 4.54 | 0.23 | 44.19 | 24.82 |
| rs 18 | 32.52 | 12.21 | 11.50 | 0.83 | 38.90 | 25.81 | 34.22 | 5.20 | 4.54 | 0.22 | 41.42 | 25.28 |
| rs 20 | 35.12 | 10.88 | 10.87 | 0.59 | 37.33 | 26.55 | 36.27 | 5.15 | 4.36 | 0.18 | 38.97 | 25.87 |

# Masking Methods for Microdata: Perturbative methods

**Rank swapping**: Rank swapping record linkage (RS-RL)

- Discussion
  - Specific record linkage improve the results of generic record linkage
  - The implementation detects cases where reidentification is achieved and detected (i.e., the intruder knows that reidentification has taken place)
  - Development of masking methods resilient to specific record linkage methods:
    - ⋆ rank swapping p-distribution
    - ⋆ rank swapping p-buckets
  - Microaggregation with individual ranking (univariate) can also be attacked effectively

# Microaggregation

# Data-driven protection methods

**Microaggregation:**

- **Informal definition.** Small clusters are built for the data, and then each record is replaced by a representative.

# Data-driven protection methods

**Microaggregation:**

- **Informal definition.** Small clusters are built for the data, and then each record is replaced by a representative.
- Disclosure risk and information loss
  - Low disclosure is ensured requiring $k$ records in each cluster
  - Low information loss is ensured as clusters are small

# Data-driven protection methods

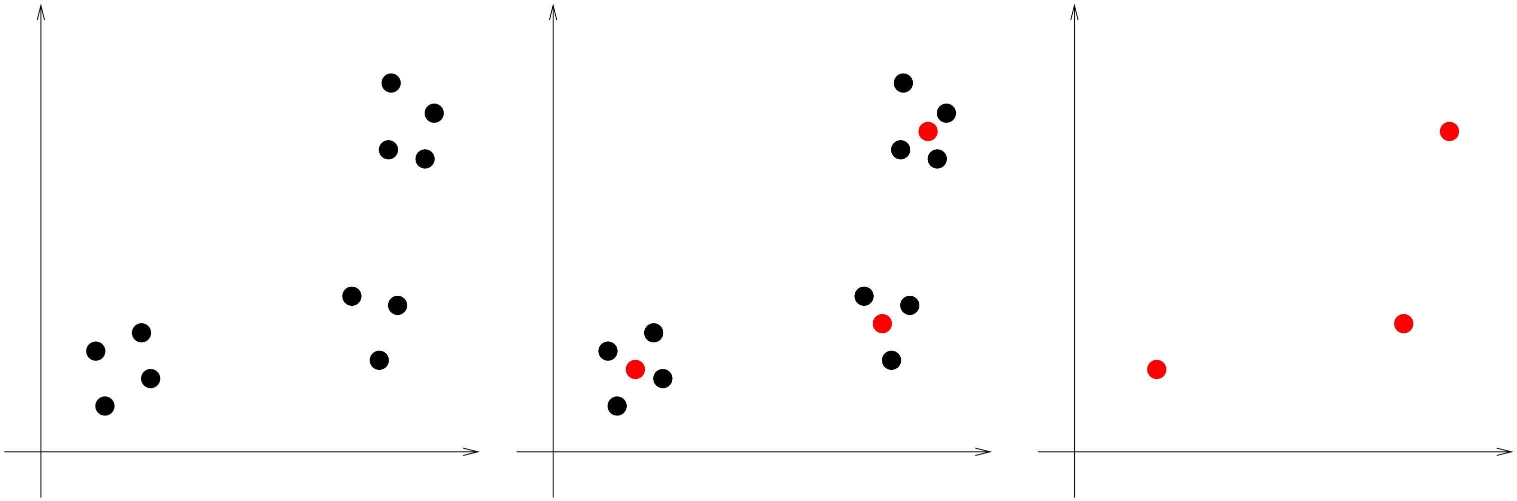
## Microaggregation:

- **Operational definition.** It is defined in terms of
  - **Partition.** Records are partitioned into several clusters, each of them consisting of at least $k$ records.
  - **Aggregation.** For each of the clusters a representative (the centroid) is computed
  - **Replacement.** The original records are replaced by the representative of the cluster to which they belong to.

# Data-driven protection methods

## Microaggregation:

- **Graphical representation of the process.**



# Data-driven protection methods

## Microaggregation:

- **Formalization.** $u_{ij}$ to describe the partition of the records in $X$. That is, $u_{ij} = 1$ if record $j$ is assigned to the $i$th cluster. Let $v_i$ be the representative of the $i$th cluster, then a general formulation of microaggregation with $g$ clusters and a given $k$ is as follows:

  Minimize $SSE = \sum_{i=1}^{g} \sum_{j=1}^{n} u_{ij}(d(x_j, v_i))^2$

  Subject to $\sum_{i=1}^{g} u_{ij} = 1$ for all $j = 1, \ldots, n$

  $2k \geq \sum_{j=1}^{n} u_{ij} \geq k$ for all $i = 1, \ldots, g$

  $u_{ij} \in \{0, 1\}$

# Data-driven protection methods

**Microaggregation:** Optimality

- Polynomial solution when only one variable
- Optimal solution is NP-hard for more than 2 variables
- Heuristic methods have been developed: MDAV, Projected microaggregation

# Data-driven protection methods

**Microaggregation:** Heuristic approaches

- usually follow the operational approach
  - **Build a partition.**
  - **Define an aggregation.** Mean of the records in the cluster
  - **Replacement.**

# Data-driven protection methods

**Microaggregation:** Multivariate

- When a file has several variables
  - Microaggregate all the variables at once
  - Microaggregate sets of variables
  - Microaggregate one variable at a time: individual ranking

# Data-driven protection methods

```
Optimal Univariate Microaggregation one variable/individual ranking begin
    Let $X = (a_1 \dots a_n)$ be a vector of size $n$ containing all the values for the attribute being protected. Sort the values of $X$ in ascending order so that if $i < j$ then $a_i \leq a_j$.
    Given $A$ and $k$, a graph $G_{k,n}$ is defined as follows.
    begin
        Define the nodes of $G$ as the elements $a_i$ in $A$ plus one additional node $g_0$ (this node is later needed to apply the Dijkstra algorithm).
        For each node $g_i$, add to the graph the directed edges $(g_i, g_j)$ for all $j$ such that $i + k \leq j < i + 2k$. The edge $(g_i, g_j)$ means that the values $(a_{i+1}, \dots, a_j)$ might define one of the possible clusters.
        The cost of the edge $(g_i, g_j)$ is defined as the within-group sum of squared error for such cluster. That is, $SSE = \Sigma_{l=i+1}^{j} (a_l - \bar{a})^2$, where $\bar{a}$ is the average record of the cluster.
        The optimal univariate microaggregation is defined by the shortest path algorithm between the nodes $g_0$ and $g_n$. This shortest path can be computed using the Dijkstra algorithm.
```

# Data-driven protection methods

- **Algorithm** General Multivariate Microaggregation
  - ○ **Input:** X: original data set, k: integer
  - ○ **Output:** X': protected data set
  - ○ $\Pi = \{\pi_1, \ldots, \pi_p\}$ a partition of the variables' set $V = \{V_1, \ldots, V_s\}$
  - ○ **foreach** $\pi \in \Pi$
  - ○     Microaggregate $X$ considering only the variables in $\pi$
  - ○ **end foreach**

# Data-driven protection methods

## Projected Microaggregation

```
begin
    Split the data set $X$ into $r$ sub-data sets $\{X_i\}_{1 \leq i \leq r}$, each one with $a_i$ attributes of the $n$ records, such that $\sum_{i=1}^{r} a_i = A$
    foreach $(X_i \in X)$ do
        Apply a projection algorithm to the attributes in $X_i$, which results in an univariate vector $z_i$ with $n$ components (one for each record)
        Sort the components of $z_i$ in increasing order
        Apply to the sorted vector $z_i$ the following variant of the univariate optimal microaggregation method: use the algorithm defining the cost of the edges $\langle z_{i,s}, z_{i,t} \rangle$, with $s < t$, as the within-group sum of square error for the $a_i$-dimensional cluster in $X_i$ which contains the original attributes of the records whose projected values are in the set $\{z_{i,s}, z_{i,s+1}, \ldots, z_{i,t}\}$
        For each cluster resulting from the previous step, compute the $v_i$-dimensional centroid and replace all the records in the cluster by the centroid
```

# Data-driven protection methods

## MDAV microaggregation

```
begin
    $C = \emptyset$
    while $|X| \geq 3k$ do
        $\bar{x}$ = the average record of all records in $X$
        $x_r$ = the most distant record from $\bar{x}$
        $x_s$ = the most distant record from $x_r$
        $C_r$ = cluster around $x_r$ (with $x_r$ and the $k-1$ closest records to $x_r$)
        $C_s$ = cluster around $x_s$ (with $x_s$ and the $k-1$ closest records to $x_s$)
        Remove records in $C_r$ and $C_s$ from data set $X$
        $C = C \cup \{C_r, C_s\}$

    if $|X| \geq 2k$ then
        $\bar{x}$ = the average record of all records in $X$
        $x_r$ = the most distant record from $\bar{x}$
        $C_r$ = cluster around $x_r$ (with $x_r$ and the $k-1$ closest records to $x_r$)
        $C_s = X \setminus C_r$ (form another cluster with the rest of records)
        $C = C \cup \{C_r, C_s\}$

    else
        $C = C \cup \{X\}$
```

# Data-driven protection methods

## **Microaggregation** Discussion and summary (I)

- The larger the $k$, the lower the risk, the larger the information loss
- Microaggregation is related to $k$-anonymity:
  all variables microaggregated together imply $k$-anonymity
- It is easy to define microaggregation for other types of data
  distance, and aggregation method (plurality rule - most frequent value)

# Data-driven protection methods

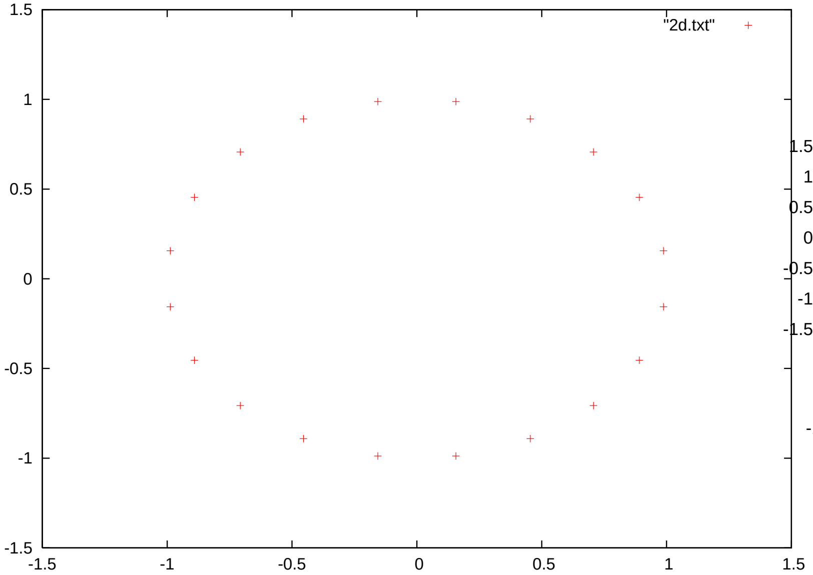
## Microaggregation Discussion and summary (II)

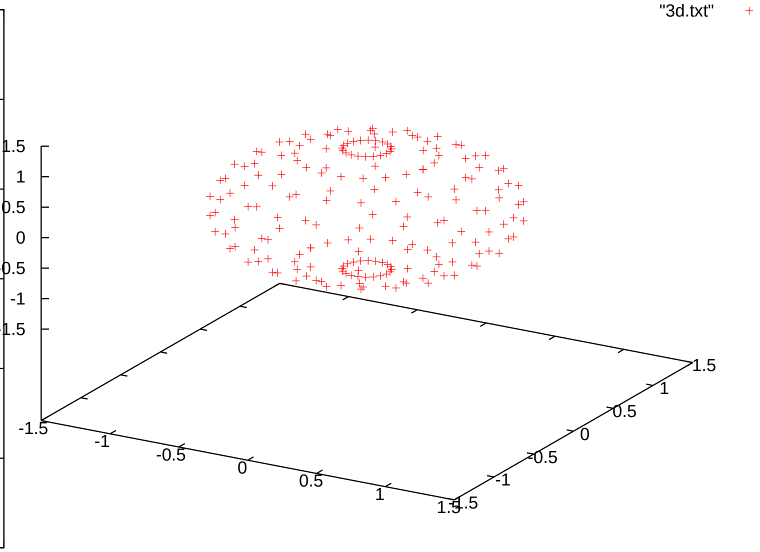
- Correlated variables together or not?

  Most usually correlated variables are microaggregated together to keep correlations in the protected data set.

  Microaggregation of two unrealistic datasets give worse results grouping correlated attributes than not grouping them.

  Not clear conclusion with real data.





# Additive Noise

# Data-driven protection methods

## Additive Noise:

- Description:
  - This method protects data adding noise into the original file. That is,
    $$X' = X + \epsilon,$$
    where $\epsilon$ is the noise.
  - The simplest approach is to require $\epsilon$ to be such that $E(\epsilon) = 0$ and $Var(\epsilon) = kVar(X)$ for a given constant $k$.

# Data-driven protection methods

## Additive Noise:

- Description:
  - This method protects data adding noise into the original file. That is,

    $$X' = X + \epsilon,$$

    where $\epsilon$ is the noise.
  - The simplest approach is to require $\epsilon$ to be such that $E(\epsilon) = 0$ and $Var(\epsilon) = kVar(X)$ for a given constant $k$.
- Properties:
  - It makes no assumptions about the range of possible values for $V_i$ (which may be infinite).
  - The noise added is typically continuous and with mean zero, which suits continuous original data well.
  - No exact matching is possible with external files.

# Data-driven protection methods

**Additive Noise:** Uncorrelated noise

- For variables $V_i$ and $V_j$, noise is such that $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.
  - Uncorrelated additive noise preserves means and covariances.

$$E(X') = E(X) + E(\epsilon) = E(X)$$

$$Cov(X'_i, X'_j) = Cov(X_i, X_j) \text{ for } i \neq j$$

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

$$\begin{aligned} \rho_{X'_i, X'_j} &= \frac{Cov(X'_i, X'_j)}{\sqrt{Var(X'_i)Var(X'_j)}} = \frac{Cov(X_i, X_j)}{(1+k)\sqrt{Var(X_i)Var(X_j)}} \\ &= \frac{1}{1+k}\rho_{X_i, X_j} \end{aligned}$$

# Data-driven protection methods

## Additive Noise: Correlated noise

- $\epsilon$ follows a normal distribution $N(0, k\Sigma)$ where $\Sigma$ is the covariance matrix of $X$.
  - It preserves correlation coefficients and means.

$$E(X') = E(X) + E(\epsilon) = E(X)$$

$$Cov(X'_i, X'_j) = (1+k)Cov(X_i, X_j) \text{ for } i \neq j$$
$$Var(X') = Var(X) + kVar(X) = (1+k)Var(X)$$

$$\begin{aligned}\rho_{X'_i,X'_j} &= \frac{Cov(X'_i, X'_j)}{\sqrt{Var(X'_i)Var(X'_j)}} = \frac{(1+k)Cov(X_i, X_j)}{(1+k)\sqrt{Var(X_i)Var(X_j)}} \\ &= \rho_{X_i,X_j}\end{aligned}$$

# Other methods

# Masking Methods for Microdata: Perturbative methods

**Data Distortion by Probability Distribution:** (synthetic)

- Description:
  1. Identification of the underlying density function and estimation of the parameters.
     - Goodness of fit: Kolmogorov-Smirnov test.
     - Example set of predetermined density functions: Poisson, exponential, normal, gamma, Weibull, log-normal, uniform, triangular, chi-square.
  2. Generation of distorted series for each confidential variable.
  3. Mapping and replacement of the distorted series in place of the confidential series.
     - Only needed if the distorted variables are to be used jointly with other non-distorted variables.

# Masking Methods for Microdata: Perturbative methods

**Resampling:**

- Description:
  1. Take with replacement $t$ independent samples $X_1, \cdots, X_t$ of size $n$ of the values of $V$.
  2. Independently rank each sample (using the same ranking criterion for all samples).
  3. For $j = 1$ to $n$, compute the $j$-th value $v'_j$ of the masked variable $V'$ as the average of the $j$-th ranked values in $X_1, \cdots, X_t$.

# Masking Methods for Microdata: Perturbative methods

## Lossy Compression:

- Description:
  1. The idea is to regard a numerical microdata file as an image
     - records being rows
     - variables being columns
     - values being pixels
  2. Lossy compression (e.g. JPEG) is used on the image
     - Depending on the lossy compression algorithm used, appropriate mappings between variable ranges and color scales will be needed
  3. The compressed image is interpreted as a masked microdata file.

- Example: Description of Lossy Compression using JPEG 80% for a file with 8 records and 8 variables.

# Masking Methods for Microdata: Perturbative methods

**Multiple imputation:** (synthetic)

- Description:
  - Relies on releasing simulated continuous microdata created by multiple imputation techniques.
    - ⋆ A way to perform multiple imputation is on a variable-by-variable basis (using a randomized regression – with normal errors) to impute missing values of each continuous variable

# Masking Methods for Microdata: Perturbative methods

**Camouflage:**

- Description:
  - To give unlimited, correct numerical responses to ad-hoc queries to a database while not compromising confidential numerical data.
    - ⋆ Camouflages the sensitive record (exact answer) providing an interval answer.
- Properties:
  - No probabilistic assumptions are made
  - Optimization techniques are used to camouflage.
  - The information loss is the transformation of a point answer into an interval answer.

# Masking Methods for Microdata: Perturbative methods

**PRAM**: The Post-Randomization Method

- Description:
  - The scores on some categorical variables for certain records in the original file are changed to a different score.
    - ⋆ according to a Markov matrix
- Properties:
  - The Markov approach makes PRAM very general: it encompasses noise addition, data suppression and data recoding.
  - PRAM information loss and disclosure risk largely depend on the choice of the Markov matrix.

# Masking Methods for Microdata: Perturbative methods

**Rounding:**

- Description:
  - Determination of a set of rounding points $p_1, \cdots, p_r$
  - Rounded values are chosen among the set of rounding points.
- Properties:
  - Univariate rounding and multivariate rounding
  - Appropriate only for continuous data

# Masking Methods for Microdata: Perturbative methods

**Rounding:**

- Example (for a continuous variable $V$)
  - Take rounding points as multiples of a base value b:
    - ⋆ $p_i = b \cdot i$ for $i = 1, \cdots, r$
  - Define the set of attraction for each rounding point:
    - ⋆ for $p_i$ for $i = 2, \cdots, r-1$, as the interval $[p_i - b/2, p_i + b/2)$,
    - ⋆ for $p_1$ and $p_r$, respectively, the sets of attraction are $[0, p_1 + b/2)$ and $[p_r - b/2, V_{max}]$, where $V_{max}$ is the largest possible value for variable V.
  - An original value $v$ of $V$ is replaced with the rounding point corresponding to the set of attraction where $v$ lies.

# Nonperturbative Methods

# Masking Methods for Microdata: Nonperturbative methods

**Sampling:**

- Description
  - Instead of publishing $\mathbf{V} : \mathbf{O} \rightarrow D(V_1) \times D(V_2) \times \cdots \times D(V_m)$
  - Publish: $\mathbf{V}' : \mathbf{S} \rightarrow D(V_1) \times D(V_2) \times \cdots \times D(V_m)$
  - where:
    - $\star$ $\mathbf{S} \subset \mathbf{O}$ is a sample of the original set of records
    - $\star$ $\mathbf{V}'$ stands for the original function $\mathbf{V}$ restricted to $\mathbf{S}$.

- Properties:
  - Suitable for categorical microdata
  - Its adequacy for continuous microdata is less clear

- Example: Description of a real-world application of sampling.

# Masking Methods for Microdata: Nonperturbative methods

**Global recoding**:

- Procedure for categorical variables
  - Take a categorical variable $V_i$
  - Several categories are combined to form new categories
  - A new $V_i'$ with $|D(V_i')| < |D(V_i)|$ where $| \cdot |$ is the cardinality operator.
- Procedure for continuous variables
  - Take a continuous variable $V_i$
  - Discretize the $D(V_i')$
  - $V_i'$ which is a discretized version of $V_i$

# Masking Methods for Microdata: Nonperturbative methods

**Global recoding:**

- Properties:
  - More appropriate for categorical microdata
  - High information loss for numerical variables

- Example:
  - consider a record with "Marital status = Widow/er" and "Age = 17"
  - global recoding applied to "Marital status" to create a broader category: "Widow/er or divorced"
  - then, the probability of the above record being unique would diminish

# Masking Methods for Microdata: Nonperturbative methods

**Top and bottom coding:**

- Description
  - A special case of global recoding which can be used on variables that can be ranked
  - Top coding: Top values (above a certain threshold) are lumped together to form a new category
  - Bottom coding: Bottom values (below a certain threshold) are lumped together to form a new category
- Properties:
  - As for global recoding

# Masking Methods for Microdata: Nonperturbative methods

**Local suppression:**

- Description
  - Certain values of individual variables are suppressed
    - ⋆ to increase the set of records agreeing on a combination of values
- Properties:
  - Oriented to categorical variables.
  - Methods to combine local suppression and global recoding implemented in $\mu$-Argus SDC package (Hundepool et al. 1998, De Waal and Willenborg 1995)

# Masking Methods for Microdata: Nonperturbative methods

## Generalization for $k$-anonymity: Mondrian:

```
begin
    if not(partitionable($X$)) then
        return $\{\gamma(x) = \{x \rightarrow summary(\mathsf{X})\} | x \in X\}$

    else
        $V_i$ = select variable from $X$
        $i_0$ = select a value from domain of $V_i$ in $X$
        $lhs = \{x \in X | V_i(x) < i_0\}$
        $rhs = \{x \in X | V_i(x) > i_0\}$
        Distribute records in $\{x \in X | V_i(x) = i_0\}$ between $lhs$ and $rhs$
        return Mondrian($lhs, k$) $\cup$ Mondrian($rhs, k$)
```

# Synthetic data generators

# Synthetic Data Generators

**Synthetic Data Generators:**

→ seldom pay attention to disclosure risk.

"Since released microdata are synthetic, no real re-identification is possible".

However, unrealistic assumption, if synthetic data generation is performed on the quasi-identifier attributes. Re-identification can indeed happen if a

snooper is able to link an external identified data source with some record in the released dataset using the quasi-identifier attributes: coming up with a correct pair (identifier, confidential attributes) is indeed a re-identification.

# Synthetic Data Generators: The IPSO family

**IPSO-A:**

- $X$ and $Y$ two sets of attributes
- $X$: confidential outcome attributes
- $Y$: quasi-identifier attributes.
- Then, $X$ are taken as independent and $Y$ as dependent attributes.
- A multiple regression of $Y$ on $X$ is computed and fitted $Y'_A$ attributes are computed. Finally, attributes $X$ and $Y'_A$ are released by IPSO-A in place of $X$ and $Y$.

In the above setting, conditional on the specific confidential attributes $x_i$, the quasi-identifier attributes $Y_i$ are assumed to follow a multivariate normal distribution with covariance matrix $\Sigma = \{\sigma_{jk}\}$ and a mean vector $x_i B$, where $B$ is the matrix of regression coefficients.

# Synthetic Data Generators: The IPSO family

## IPSO B and C:

Let $\hat{B}$ and $\hat{\Sigma}$ be the maximum likelihood estimates of $B$ and $\Sigma$ derived from the complete dataset $(y, x)$. If a user fits a multiple regression model to $(y'_A, x)$, she will get estimates $\hat{B}_A$ and $\hat{\Sigma}_A$ which, in general, are different from the estimates $\hat{B}$ and $\hat{\Sigma}$ obtained when fitting the model to the original data $(y, x)$.

**IPSO-B:** Modifies $y'_A$ into $y'_B$ in such a way that the estimate $\hat{B}_B$ obtained by multiple linear regression from $(y'_B, x)$ satisfies $\hat{B}_B = \hat{B}$.

**IPSO-C:** A more ambitious goal is to come up with a data matrix $y'_C$ such that, when a multivariate multiple regression model is fitted to $(y'_C, x)$, *both* sufficient statistics $\hat{B}$ and $\hat{\Sigma}$ obtained on the original data $(y, x)$ are preserved.

# Synthetic Data Generators: The IPSO family

## Experiments for IPSO-A,B,C:

- EIA dataset (4092 records, 15 attributes); Variables used:

| Quasi-identifier in external $\mathbf{A}$ | Quasi-identifier in released $\mathbf{B}$ |
|---|---|
| $v1$ | $v1_A$ |
| $v1, v7, v8$ | $v1_A, v7_A, v8_A$ |
| $v1, v2, v7, v8, v9$ | $v1_A, v2_A, v7_A, v8_A, v9_A$ |

- Results:

| DBRL1 | DBRL2 | DBRLM-COV0 | DBRLM-COV | KDBRL | PRL |
|---|---|---|---|---|---|
| 14 | 9 | 9 | 9 | 14 | 8 |
| 16 | 15 | 18 | 9 | 16 | 16 |
| 65 | 121 | 3206 | 143 | 63 | 159 |
| 14 | 9 | 9 | 9 | 14 | 8 |
| 17 | 15 | 18 | 8 | 17 | 16 |
| 65 | 120 | 3194 | 135 | 62 | 159 |
| 11 | 11 | 11 | 11 | 11 | 10 |
| 6 | 6 | 14 | 8 | 6 | 5 |
| 53 | 53 | 773 | 46 | 54 | 93 |

# Information Loss Measures

# Information Loss Measures

**Information Loss:** information loss depends on the data uses to be supported by the masked data.

- Let $X$ be the original data set on the domain $D$, and let $X'$ be a protected version of the same data set. Then, for a given data analysis that returns results in a certain domain $D'$ (i.e., $f : D \rightarrow D'$), the information loss of $f$ for data sets $X$ and $X'$ is defined by

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

where $divergence$ is a way to compare two elements of $D'$.

# Information Loss Measures

## Information Loss:

- $X, X', D$ as above, $f : D \to D')$, the information loss of $f$ for data sets $X$ and $X'$ is defined by

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

where $divergence$ is a way to compare two elements of $D'$.

Reasonable to require:

- $divergence(X, X) = 0$ for all $X \in D$
- $divergence(X, Y) \geq 0$ for all $X, Y \in D'$
- $divergence(X, Y) = divergence(Y, X)$

# Information Loss Measures

## Information Loss:

- $X, X', D$ as above, $f : D \to D'$), the information loss of $f$ for data sets $X$ and $X'$ is defined by

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

where $divergence$ is a way to compare two elements of $D'$.

Reasonable to require:

- $divergence(X, X) = 0$ for all $X \in D$
- $divergence(X, Y) \geq 0$ for all $X, Y \in D'$
- $divergence(X, Y) = divergence(Y, X)$
  $\to$ asymetric divergence when e.g. to avoid false positives malfunctioning sensor causes huge damage, undetection no.

# Information Loss Measures

**Information Loss:**

- Generic information loss measures
- Specific information loss measures

# Information Loss Measures

**Information Loss:** Generic information loss measures

- A microdata set is analytically valid (Winkler, 1998):
  1. Means and covariances on a small set of subdomains
  2. Marginal values for a few tabulations of the data
  3. At least one distributional characteristic
- A microdata file is analytically interesting if six variables on important subdomains are provided that can be validly analyzed.

# Information Loss Measures

- Complementary ways to assess the preservation of the structure of the original data set:

1. Compare the data in the original and the masked data sets
    - The more similar the SDC method to the identity function, the less impact
2. Compare some statistics computed on the original and the masked data sets
    - Little information loss should translate to little differences between the statistics
3. Analyze the behavior of the particular SDC method used

# Information Loss Measures

**Generic Information loss measures**:

- Continuous Data
- Categorical Data

# Information Loss Measures: Continuous Data

## Characterization of the information in the dataset

- Assume a microdata set $X$ ($X'$ be the masked microdata set) where:
  - ○ $n$ individuals (records) $I_1, I_2, \cdots, I_n$
  - ○ $p$ continuous variables $Z_1, Z_2, \cdots, Z_p$
- The following tools are useful to characterize the information contained in the data set:
  - ○ Covariance matrices $V$ (on $X$) and $V'$ (on $X'$)
  - ○ Correlation matrices $R$, $R'$
  - ○ Correlation matrices $RF$, $RF'$ between variables and PCA factors $PC_1, \cdots, PC_p$
  - ○ Commonality vectors $C$, $C'$ between variables and the first principal component (Commonality: the percent of each variable that is explained by $PC_1$ (or $PCi$))
  - ○ Factor score coefficient matrices $F$ and $F'$ (factors that should multiply each variable in $X$ to obtain its projection on each principal component)

# Information Loss Measures: Continuous Data

## Matrix divergence

1. Mean square error:

   Sum of squared componentwise differences between pairs of matrices, divided by the number of cells
2. Mean absolute error:

   Sum of absolute componentwise differences between pairs of matrices, divided by the number of cells
3. Mean variation:

   Sum of absolute percent variation of components in the matrix computed on masked data with respect to components in the matrix computed on original data, divided by the number of cells.

# Information Loss Measures: Continuous Data

| | Mean square error | Mean abs. error | Mean variation |
|---|---|---|---|
| $X-X'$ | $\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$ | $\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{np}$ | $\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$ |
| $V-V'$ | $\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{p(p+1)}{2}}$ | $\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{p(p+1)}{2}}$ | $\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{p(p+1)}{2}}$ |
| $R-R'$ | $\frac{\sum_{j=1}^{p}\sum_{1\le i<j}(r_{ij}-r'_{ij})^2}{\frac{p(p-1)}{2}}$ | $\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{p(p-1)}{2}}$ | $\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{p(p-1)}{2}}$ |
| $RF-RF'$ | $\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}(rf_{ij}-rf'_{ij})^2}{p^2}$ | $\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\lvert rf_{ij}-rf'_{ij}\rvert}{p^2}$ | $\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\frac{\lvert rf_{ij}-rf'_{ij}\rvert}{\lvert rf_{ij}\rvert}}{p^2}$ |
| $C-C'$ | $\frac{\sum_{i=1}^{p}(c_i-c'_i)^2}{p}$ | $\frac{\sum_{i=1}^{p}\lvert c_i-c'_i\rvert}{p}$ | $\frac{\sum_{i=1}^{p}\frac{\lvert c_i-c'_i\rvert}{\lvert c_i\rvert}}{p}$ |
| $F-F'$ | $\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}(f_{ij}-f'_{ij})^2}{p^2}$ | $\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\lvert f_{ij}-f'_{ij}\rvert}{p^2}$ | $\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\frac{\lvert f_{ij}-f'_{ij}\rvert}{\lvert f_{ij}\rvert}}{p^2}$ |

# Information Loss Measures: Categorical Data

**Alternative definitions of information loss measures:**

- Direct comparison of categorical values
- Comparison of contingency tables
- Entropy-based measures

# Information Loss Measures: Categorical Data

## Direct comparison of categorical values:

Comparison of matrices X and X' requires the definition of a distance.

For nominal variables:

$$d_V(c, c') = \begin{cases} 0 & \text{when } c = c' \\ 1 & \text{when } c \neq c' \end{cases}$$

For ordinal variables ($\leq_V$ be the total order):

$$d_V(c, c') = \frac{|\{c'' : min(c, c') \leq_V c'' \leq_V max(c, c')\}|}{|D(V)|}$$

# Information Loss Measures: Categorical Data

## Comparison of contingency tables:

F original data set, G masked data set

$t$-dimensional contingency tables ($t \leq K$)

$x^{file}_{subscripts}$: entry of the contingency table of *file* at position *subscripts*

$$CTBIL(F,G;W,K) = \sum_{\substack{\{V_{j1} \cdots V_{jt}\} \subseteq W \\ |\{V_{j1} \cdots V_{jt}\}| \leq K}} \sum_{i_1 \cdots i_t} |x^F_{i_1 \cdots i_t} - x^G_{i_1 \cdots i_t}|$$

# Information Loss Measures: Categorical Data

## Entropy-based measures:

- Entropy is an information-theoretic measure, but can be used in SDC if the masking process is modeled as the noise that would be added to the original data set in the event of it being transmitted over a noisy channel.
- For PRAM: Assume $P_{V,V'} = \{p(V' = j|V = i)\}$ the PRAM matrix

Procedure:

- The conditional uncertainty of $V$ given that $V' = j$ is:

$$H(V|V' = j) = -\sum_{i=1}^{n} p(V = i|V' = j) log\ p(V = i|V' = j)$$

- Entropy-based information loss measure (EBIL):

$$EBIL(P_{V,V'}, G) = \sum_{r \in G} H(V|V' = j_r)$$

where $j_r$ is the value taken by $V'$ in record $r$. Note $H(V|V' = j_r)$ do not depend on the value in $V$

# Information Loss Measures: Categorical Data

## An alternative Information Loss Measure:

- EBIL a function of the masked data set but does not depend on the original data set.
- Assume that, in a household survey file, variable V contains the town where the household is located. Now consider that V is masked into a new variable V' where the town has been replaced by the state. Locations like "New York City" and "Albany" will be recoded into "NY". Living in Albany is more specific and identifying (in the sense of being less anonymous) than living in New York City. The information loss measure should somehow reflect that there is more information loss when a household in "Albany" becomes a household in "New York State" than when a household in "New York City" becomes a household in "New York State"
- Note that: $P(V = \text{``}Albany''|V' = NY) < P(V = \text{``}NewYorkCity''|V' = NY)$
- According to the USBC American FactFinder, the population of New York State in $2000$ was $17,990,455$, the population of New York City was $7,322,564$ and the population of Albany was $101,082$. Thus, the above probabilities are $P(V = \text{``}Albany''|V' = NY) = 101,082/17,990,455 = 0.05$ and $P(V = \text{``}NewYorkCity''|V' = NY) = 7,322,564/17,990,455 = 0.407$.

# Information Loss Measures: Categorical Data

## An alternative Information Loss Measure:

- The smaller the conditional probability $P(V = i|V' = j)$, the larger the inf. loss.
- Information loss as a function of three elements:
  (i) conditional probability; (ii) original category $i$; (iii) masked category $j$
- Per-record information loss when $V = i$ is masked as $V' = j$ can be defined as:

$$PRIL(PV, V', i, j) = -log P(V = i|V' = j)$$

- The information loss for the entire data sets $F, G$ is

$$IL(PV, V', F, G) = \sum_{r \in G} PRIL(P_{V,V'}, i_r, j_r)$$

where $i_r$ is the value taken by $V$ in record $r$ of $F$ (similarly, $j_r$ in $G$)

# Information Loss Measures

**Specific Information Loss Measure:**

- Types of measures
- Do generic measures approximate specific ones?

# Information Loss Measures

**Specific Information Loss Measure:** An example

- data use: clustering $cl$ with parameter $c$
- divergence: Rand, Jaccard, Adjsted Rand Index, Wallace, Mantaras, ...

$$IL_{Rand,cl}(X, X') = 1 - Rand(cl_c(X), cl_c(X'))$$

$$IL_{Mantaras,cl}(X, X') = Mantaras(cl_c(X), cl_c(X'))$$

# Information Loss Measures

**Specific Information Loss Measure:** Some results:

- Census data set microaggregated (3 vars at a time, different k), k-means with c=15. Cols 2-6: Indices/distance; col 7: averaged probabilistic information loss measure (aPIL); (c) last row is the correlation of the measures and distance with respect to the aPIL.

| | Rand | Jaccard | Adjusted Rand | Wallace | Mantaras | aPIL |
|---|---|---|---|---|---|---|
| Mic3vars.k3 | 0.943 | 0.454 | 0.594 | 0.625 | 0.416 | 15.189 |
| Mic3vars.k4 | 0.943 | 0.464 | 0.602 | 0.633 | 0.425 | 19.325 |
| Mic3vars.k5 | 0.936 | 0.406 | 0.542 | 0.577 | 0.472 | 22.724 |
| Mic3vars.k6 | 0.936 | 0.408 | 0.545 | 0.580 | 0.473 | 25.760 |
| Mic3vars.k7 | 0.929 | 0.367 | 0.499 | 0.537 | 0.500 | 28.750 |
| Mic3vars.k8 | 0.933 | 0.402 | 0.538 | 0.574 | 0.479 | 31.185 |
| Mic3vars.k9 | 0.925 | 0.359 | 0.488 | 0.528 | 0.513 | 33.883 |
| Correlation | -0.930 | -0.882 | -0.887 | -0.882 | 0.931 | 1.000 |

# Information Loss Measures

**Specific Information Loss Measure:** Some results:

- Census data set. Correlations same file.

| Index / Distance | Correlation<br>(a) all<br>(215 files) | Correlation<br>(b) Microaggregation<br>(162 files) |
|---|---|---|
| Rand Index | -0.79281 | -0.86099 |
| Jaccard Index | -0.89094 | -0.94859 |
| Adjusted Rand Index | -0.91609 | -0.96114 |
| Wallace Index | -0.92559 | -0.97593 |
| Mantaras Distance | 0.91617 | 0.97216 |

# Information Loss Measures

**Specific Information Loss Measure:** Some results:

- Census data set. Convergence problems
- Census dataset with additive noise. Fuzzy clustering with $c = 10$. OF for the original file 2851 in the first execution (left), 2829 in the second. $d_1$ distance between cluster centers, $d_2$ distance between membership values.

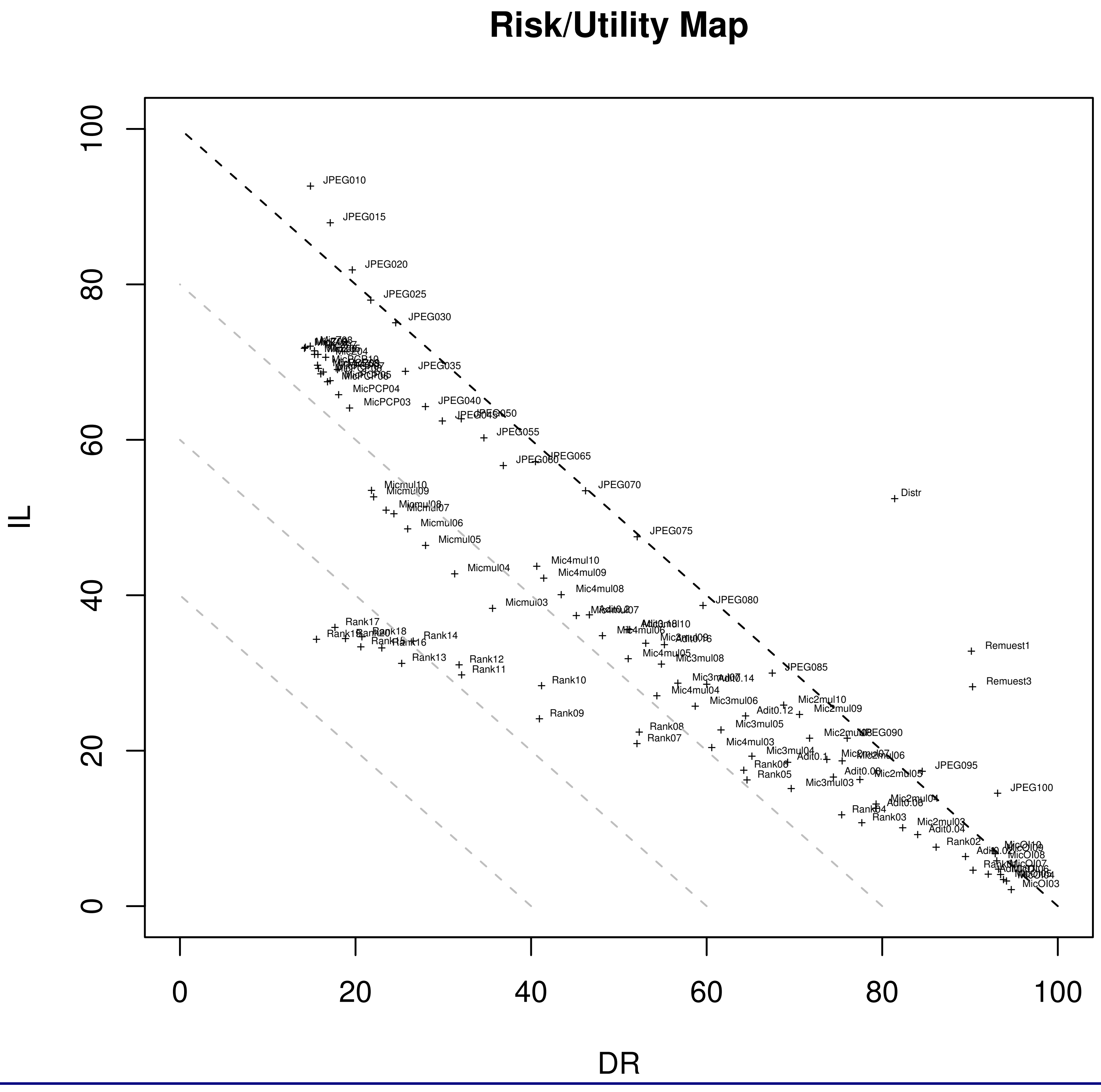
| Noise | $d_1$ | $d_2$ | $O.F.$ | $d_1$ | $d_2$ | $O.F.$ |
|---|---|---|---|---|---|---|
| 0.0 | 3.21 | 40.73 | 2826.0 | 3.93 | 91.3 | 2826 |
| 0.1 | 3.21 | 40.67 | 2827.0 | 3.97 | 91.45 | 2827 |
| 0.2 | 3.17 | 40.86 | 2829.0 | 3.94 | 90.89 | 2829 |
| 0.4 | 0.32 | 0.92 | 2859.0 | 4.07 | 93.05 | 2835 |
| 0.6 | 3.28 | 42.09 | 2844.0 | 6.92 | 113.76 | 2867 |
| 0.8 | 3.48 | 43.48 | 2862.0 | 4.19 | 91.53 | 2862 |
| 1.0 | 3.55 | 48.87 | 2886.0 | 4.37 | 99.33 | 2886 |
| 1.2 | 2.24 | 55.56 | 2908.0 | 2.75 | 68.04 | 2903 |
| 1.4 | 1.44 | 18.35 | 2935.0 | 4.53 | 99.53 | 2918 |
| 1.6 | 2.27 | 36.83 | 2978.0 | 6.98 | 103.84 | 2978 |
| 1.8 | 2.71 | 45.59 | 3006.0 | 4.68 | 99.20 | 2989 |
| 2.0 | 4.24 | 96.87 | 3028.0 | 2.70 | 31.17 | 3013 |

# Visualization

# Visualization

**Trade-off:**

- Information loss and disclosure risk are usually in conflict
- R-U maps
  - Graphical representation
- Score
  - R-U maps


Risk/Utility Map
IL
DR
0
20
40
60
80
100
0
20
40
60
80
100
JPEG010
JPEG015
JPEG020
JPEG025
JPEG030
JPEG035
JPEG040
JPEG045
JPEG050
JPEG055
JPEG060
JPEG065
JPEG070
JPEG075
JPEG080
JPEG085
JPEG090
JPEG095
JPEG100
MicZ08
MicZ09
MicZ07
MicZ05
MicZ04
MicPCP10
MicPCP09
MicPCP07
MicPCP08
MicPCP05
MicPCP06
MicPCP04
MicPCP03
Micmul10
Micmul09
Micmul08
Micmul07
Micmul06
Micmul05
Micmul04
Micmul03
Mic4mul10
Mic4mul09
Mic4mul08
Mic4mul07
Adit0.2
Adit0.18
Mic3mul10
Mic4mul06
Mic3mul09
Adit0.16
Mic4mul05
Mic3mul08
Mic3mul07
Adit0.14
Mic4mul04
Mic3mul06
Adit0.12
Mic2mul10
Mic2mul09
Mic3mul05
Mic2mul08
Mic4mul03
Mic3mul04
Adit0.1
Mic2mul07
Mic2mul06
Adit0.08
Mic2mul05
Mic3mul03
Mic2mul04
Adit0.06
Mic2mul03
Adit0.04
Adit0.02
Rank20
Rank19
Rank18
Rank17
Rank16
Rank15
Rank14
Rank13
Rank12
Rank11
Rank10
Rank09
Rank08
Rank07
Rank06
Rank05
Rank04
Rank03
Rank02
Rank01
MicOI10
MicOI09
MicOI08
MicOI07
MicOI06
MicOI05
MicOI04
MicOI03
Distr
Remuest1
Remuest3

# Visualization

**Trade-off:** Score

$$Score(X, X') = \frac{IL(X, X') + DR(X, X')}{2}$$

# Summary

# Visualization

Data privacy

- Masking methods
- Information loss
- Disclosure risk
- Visualization