Data Anonymization: A Tutorial

Josep Domingo-Ferrer

Universitat Rovira i Virgili, Tarragona, Catalonia

UNESCO Chair in Data Privacy

josep.domingo@urv.cat

September 29, 2014

Introduction

- Statistical databases contain statistical information
- They are normally released by:
  - National statistical institutes (NSIs);
  - Healthcare organizations (epidemiology);
  - or private organizations (*e.g.* consumer surveys).

Data formats

- *Tabular data*. Tables with counts or magnitudes (traditional outputs of NSIs).
- *Queryable databases*. On-line databases which accept statistical queries (sums, averages, max, min, etc.).
- *Microdata*. Files where each record contains information on an individual (a physical person or an organization).

Utility vs privacy in statistical databases

- Statistical databases must provide useful statistical information.
- They must also preserve the privacy of respondents, if data are sensitive.

$\Longrightarrow$ statistical disclosure control (SDC) methods are used to protect privacy

$\Longrightarrow$ SDC methods modify data

$\Longrightarrow$ SDC challenge: protect privacy with minimum loss of accuracy.

Disclosure concepts

- Attribute disclosure. It occurs when the value of a confidential attribute of an individual can be determined more accurately with access to the released statistics than without.
- Identity disclosure. It occurs when a record in the anonymised data set can be linked with a respondent's identity.

Note that attribute disclosure does not imply identity disclosure in general, and conversely.

SDC vs other database privacy technologies

- SDC seeks respondent privacy.
- PPDM (privacy-preserving data mining) seeks the data owner's privacy when several owners wish to co-operate in joint analyses across their databases without giving away their data to each other.
- PIR (private information retrieval) seeks user privacy, *i.e.* to allow the user of a database to retrieve some information item without the database knowing which item was recovered.

Brief history of SDC

- Seminal contributions: Dalenius (1974) from NSIs, Schlörer (1975) from the medical community, Denning *et al.* (1979) from the database community.
- Moderate activity in the 1980s, summarized in Adam and Wortmann (1989).
- Renewed interest in the 1990s by NSIs: Eurostat and U.S. Census Bureau promote dedicated conferences and the EU 4th FP funds the SDC project (1996-98).
- Widespread interest since the 2000s: with the advent of WWW, the data mining community enters the field, without much interaction with the NSI's continuing activity.

Outline of this talk

- Tabular data protection.
- Queryable database protection.
- Microdata protection.
- Conclusions.

Tabular data protection

- Goal: Publish *static* aggregate information, *i.e.* tables, in such a way that no confidential information can be inferred on specific individuals to whom the table refers.
- From microdata, tabular data can be generated by crossing one or more categorical attributes.
- Formally, given categorical attributes $X_1, \cdots, X_l$, a table T is a function

$$T : D(X_1) \times D(X_2) \times \cdots \times D(X_l) \Longrightarrow \mathbb{R} \text{ or } \mathbb{N}$$

  where $D(X_i)$ is the domain where attribute X_i takes its values.
- Number of *cells* usually much less than number of respondents.

Types of tables

- Frequency tables: They display the count of respondents (in $\mathbb{N}$) at the crossing of the categorical attributes. *E.g.* number of patients per disease and municipality.
- Magnitude tables: They display information on a numerical attribute (in $\mathbb{R}$) at the crossing of the categorical attributes. *E.g.* Average age of patients per disease and municipality.
- Marginal row and column totals must be preserved.
- Linked tables: Two tables are linked if they share some of the crossed categorical attributes, *e.g.* "Disease" $\times$ "Town" and "Disease" $\times$ "Gender".

Disclosure attacks in tables

Even if tables display aggregate information, disclosure can occur:

- *External attack*. *E.g.*, let a released frequency table "Ethnicity" $\times$ "Town" contain a single respondent for ethnicity E_i and town T_j. Then if a magnitude table is released with the average blood pressure for each ethnicity and each town, the exact blood pressure of the only respondent with ethnicity E_i in town T_j is publicly disclosed.
- *Internal attack*. If there are only two respondents for ethnicity E_i and town T_j, the blood pressure of each of them is disclosed to the other.

Disclosure attacks in tables (II)

- *Dominance attack*. If one (or few) respondents dominate in the contribution to a cell in a magnitude table, the dominant respondent(s) can upper-bound the contributions of the rest.
- *E.g.* if the table displays the cumulative earnings for each job type and town, and one individual contributes 90% of a certain cell value, s/he knows her/his colleagues in the town are not doing very well.

SDC methods for tables

- Non-perturbative. They do not modify the values in the cells, but they may suppress or recode them. Best known methods: *cell suppression (CS)*, *recoding of categorical attributes*.
- Perturbative. They modify the values in the cells. Best known methods: *controlled rounding (CR)* and the recent *controlled tabular adjustment (CTA)*.

Cell suppression

1. Identify sensitive cells, using a sensitivity rule.
2. Suppress values in sensitive cells (primary suppressions).
3. Perform additional suppressions (secondary suppressions) to prevent recovery of primary suppressions from row and/or column marginals.

Sensitivity rules

(n, k)-dominance A cell is sensitive if n or fewer respondents contribute more than a fraction k of the cell value.

pq-rule If respondents contributions to the cell can be estimated within q percent before seeing the cell and within p percent after seeing the cell, the cell is sensitive.

$p\%$-rule Special case of the pq-rule with $q = 100$.

Secondary suppression heuristics

- Usually one attempts to minimize either the number of secondary suppressions or their pooled magnitude (complex optimization problems).
- Optimization methods are heuristic, based on mixed linear integer programming or networks flows (the latter for 2-D tables only).
- Implementations in the τ-Argus package.

Controlled rounding and controlled tabular adjustment

- CR rounds values in the table to multiples of a rounding base (marginals may have to be rounded as well).
- CTA modifies the values in the table to prevent inference of sensitive cell values within a prescribed protection interval.
- CTA attempts to find the *closest* table to the original one that protects all sensitive cells.
- CTA optimization is typically based on mixed linear integer programming and entails less information loss than CS.

Queryable database protection

Three main SDC approaches:

- Query perturbation. Perturbation (noise addition) can be applied to the microdata records on which queries are computed (*input perturbation*) or to the query result after computing it on the original data (*output perturbation*).
- Query restriction. The database refuses to answer certain queries.
- Camouflage. Deterministically correct non-exact answers (small interval answers) are returned by the database.

Output perturbation via differential privacy

ε-Differential privacy [Dwork, 2006]

A randomized query function F gives ε-differential privacy if, for all data sets D_1, D_2 such that one can be obtained from the other by modifying a single record, and all $S \subset Range(F)$

$$\Pr(F(D_1) \in S) \leq \exp(\varepsilon) \times \Pr(F(D_2) \in S) \quad (1)$$

- Usually $F(D) = f(D) + Y(D)$, where $f(D)$ is a user query to a database D and $Y(D)$ is a random noise (typically Laplace with zero mean and $\Delta(f)/\varepsilon$, where $\Delta(f)$ is the sensitivity of f and ε is a privacy parameter (the larger, the less privacy)).

Query restriction

- This is the right approach if the user does require deterministically correct answers and these answers have to be exact (*i.e.* a number).
- Exact answers may be very disclosive, so it may be necessary to refuse answering certain queries at some stage.
- A common criterion to decide whether a query can be answered is query set size control: the answer to a query is refused if this query together with the previously answered ones isolates too small a set of records.
- Problems: computational burden to keep track of previous queries, collusion possible.

Camouflage

- Interval answers are returned rather than point answers.
- Unlimited answers can be returned.
- The confidential vector a is camouflaged by making it part of the relative interior of a compact set Π of vectors.
- Each query $q = f(a)$ is answered with an interval $[q^-, q^+]$ containing $[f^-, f^+]$, where f^- and f^+ are, respectively, the minimum and the maximum of f over Π.

Microdata protection

- A *microdata* file $\mathbf{X}$ with s respondents and t attributes is an $s \times t$ matrix where X_{ij} is the value of attribute j for respondent i.
- Attributes can be numerical (*e.g.* age, blood pressure) or categorical (*e.g.* gender, job).

Attribute types according to disclosure potential

- *Identifiers*. Attributes that *unambiguously* identify the respondent (*e.g.* passport no., social security no., name-surname, etc.).
- *Quasi-identifiers or key attributes*. They identify the respondent with some ambiguity, but their combination may lead to unambiguous identification (*e.g.* address, gender, age, telephone no., etc.).
- *Confidential outcome attributes*. They contain sensitive respondent information (*e.g.* salary, religion, diagnosis, etc.).
- *Non-confidential outcome attributes*. Other attributes which contain non-sensitive respondent info.

Attribute types according to disclosure potential

- Identifiers are of course suppressed in anonymized data sets.
- Disclosure risk comes from quasi-identifiers (QIs):
  - QIs cannot be suppressed because they often have high analytical value.
  - QIs can be used to link anonymized records to external non-anonymous databases (with identifiers) that contain the same or similar QIs $\Longrightarrow$ re-identification!!!
    Anonymization procedures must deal with QIs.

Approaches to microdata protection

Two main approaches:

- *Masking*. Generate a modified version $\mathbf{X}'$ of the original microdata set $\mathbf{X}$:
  - *Perturbative*. $\mathbf{X}'$ is a perturbed version of $\mathbf{X}$.
  - *Non-perturbative*. $\mathbf{X}'$ is obtained from $\mathbf{X}$ by partial suppressions or reduction of detail (yet the data in $\mathbf{X}'$ are still true).
- *Synthesis*. Generated synthetic data $\mathbf{X}'$ that preserve some preselected properties of the original data $\mathbf{X}$.

Perturbative masking: additive noise

- *Uncorrelated noise addition:* $x'_j = x_j + \epsilon_j$ where $\epsilon_j \sim N(0, \sigma^2_{\varepsilon_j})$, such that $Cov(\varepsilon_t, \varepsilon_l) = 0$ for all $t \neq l$. Neither variances nor correlations are preserved.
- *Correlated noise addition:* As above, but

$$\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n) \sim N(\mathbf{0}, \alpha\Sigma)$$

with Σ being the covariance matrix of the original data. Means and correlations can be preserving by choosing appropriate α
- *Noise addition and linear transformation:* Additional transformations are made to ensure that the sample covariance matrix of the masked attributes is an unbiased estimator for the covariance matrix of the original attributes.

Perturbative masking: additive noise (II)

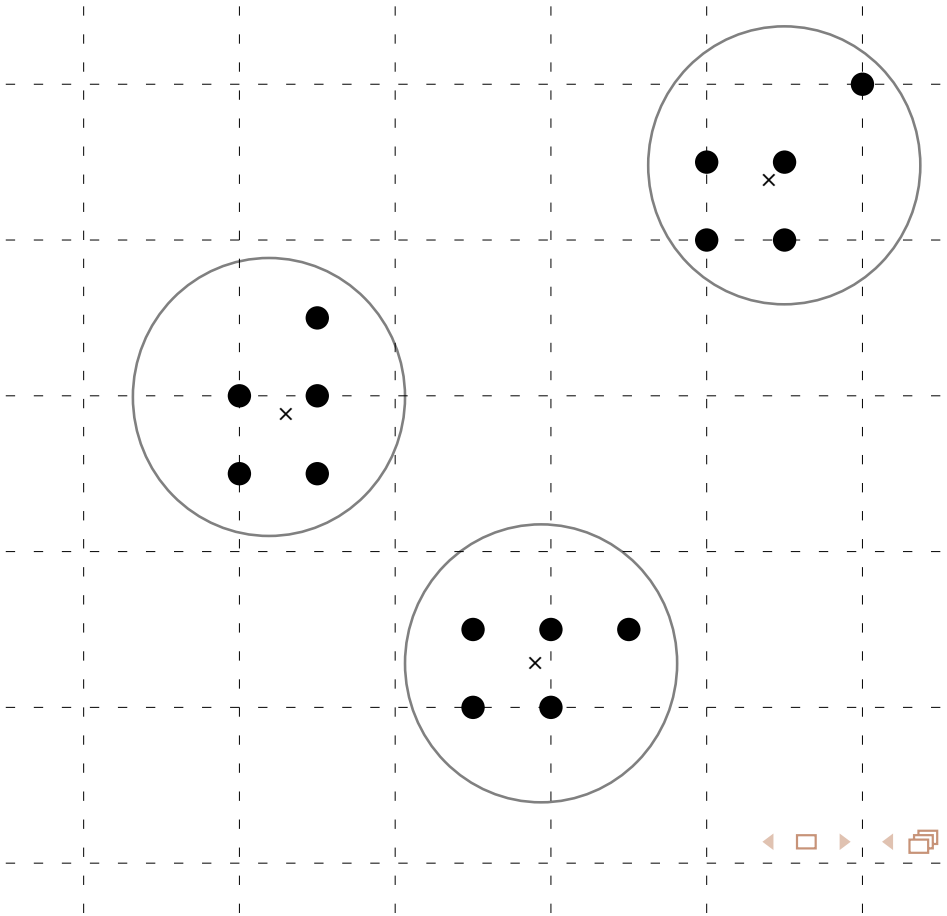
- If using a linear transformation, the protector must decide whether to reveal it to the user to allow for bias adjustment in subpopulations.
- Additive noise is not suited for categorical data.
- It is suited for continuous data.

Additive noise and differential privacy

- ε-Differential privacy can be also defined on microdata.
- A ε-differentially private data set can be created by pooling the ε-private answers to a query for the content of the i-th data set record, for $i = 1$ to the total number of records.

Perturbative masking: microaggregation

Family of SDC techniques that partition records in groups of at least k (k-partition) and publish the average record of each group.



Perturbative masking: microaggregation (II)

- The optimal k-partition is the one maximizing within-group homogeneity.
- The higher the within-group homogeneity, the lower the information loss when replacing records in a group by the group centroid.
- Usual homogeneity criterion *for numerical data*: minimization of the within-groups sum of squares

$$SSE = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)'(x_{ij} - \bar{x}_i)$$

- In a dataset with several attributes, microaggregation can be performed on all attributes together or independently on disjoint groups of attributes.

Perturbative masking: types of microaggregation

- Fixed group size microaggregation sets the size of all groups of records (except perhaps one) to k, while variable group size allows the size of groups to vary between k and $2k - 1$.
- Exact optimal microaggregation can be computed in polynomial time only for a single attribute; for several attributes, microaggregation is NP-hard and algorithms are heuristic.
- Microaggregation was initially limited to continuous data, but it can also be applied to categorical data, using suitable definitions of distance and average.

Perturbative masking: general fixed-size microaggregation

let X be the original data set
let k be the minimal cluster size

set $i := 0$
while $|X| \geq 2k$ **do**
    $C_i \leftarrow k$ smallest elements from X according to $\leq_i$
    $X := X \setminus C_i$
    $i := i + 1$
end while
$\overline{X} \leftarrow$Replace each record $r \in X$ by the centroid of its cluster

return $\overline{X}$

Perturbative masking: microaggregation and k-anonymity

Domingo-Ferrer and Torra (2005) proposed microaggregation of the projection of records on their quasi-identifiers to achieve k-anonymity:

k-Anonymity [Samarati & Sweeney1998]

A data set is said to satisfy k-anonymity if each combination of values of the quasi-identifier attributes in it is shared by at least k records.

Perturbative masking: swapping

- Data swapping was presented for databases containing only categorical attributes.
- Values of confidential attributes are exchanged among individual records, so that low-order frequency counts or marginals are maintained.
- Rank swapping is a variant of data swapping, also applicable to numerical attributes.
- Values of each attribute are ranked in ascending order and each value is swapped with another ranked value randomly chosen within a restricted range (*e.g.* the ranks of two swapped values cannot differ by more than $p\%$ of the total number of records).

Perturbative masking: PRAM

- The Post-RAndomization Method (PRAM) works on categorical attributes.
- Each value of a categorical attribute is changed to a different value according to a prescribed Markov matrix (PRAM matrix).
- PRAM can be viewed as encompassing noise addition, data suppression and data recoding.
- How to optimally determine the PRAM matrix is not obvious.
- Being probabilistic, PRAM can afford transparency (publishing the PRAM matrix does not allow inverting anonymization).

Non-perturbative masking: sampling

- Instead of publishing the original microdata file, a sample of the original set of records is published.
- Sampling with a low sampling fraction may suffice to anonymize categorical data (probability that a sample unique is also a population unique is low).
- For continuous data it should be combined with other methods: unaltered values of continuous attributes are likely to yield unique matches with external non-anonymous data files (it is unlikely that two different respondents have the same value of a numerical attribute).

Non-perturbative masking: generalization

- Also known as global recoding.
- For a categorical attribute, several categories are combined to form new (less specific) categories.
- For a continuous attribute, it means discretizing (*e.g.* replacing numerical values by intervals).

Example. If there is a record with "Marital status = Widow/er" and "Age = 17", generalization could be applied to "Marital status" to create a broader category "Widow/er or divorced" and decrease the probability of the above record being unique.

Non-perturbative masking: top and bottom coding

- Top and bottom coding apply to attributes that can be ranked (continuous or categorical ordinal).
- Top (resp. bottom) coding lumps values above (resp. below) a certain threshold into a single top (resp. bottom) category.

Non-perturbative masking: local suppression

- Certain values of individual attributes are suppressed in order to increase the set of records agreeing on a combination of quasi-identifier values.
- It can be combined with generalization.
- Local suppression makes more sense for categorical attributes, because any combination of quasi-identifiers involving a continuous attribute is likely to be unique (and hence should be suppressed).

Generalization, suppression and k-anonymity

- The computational approach originally proposed by Samarati and Sweeney to achieve k-anonymity combined generalization and suppression (the latter to reduce the need for the former).
- Most of the k-anonymity literature still relies on generalization, even though:
  - Generalization cannot preserve the numerical semantics of continuous attributes.
  - It uses a domain-level generalization hierarchy, rather than a data-driven one.

Synthetic microdata generation

- Idea: Randomly generate data in such a way that some statistics or relationships of the original data set are preserved.
- Pros: No respondent re-identification seems possible, because data are synthetic.
- Cons:
  - If a synthetic record matches by chance a respondent's attributes, re-identification is likely and the respondent will find little comfort in the data being synthetic.
  - Data utility of synthetic microdata is limited to the statistics and relationships pre-selected at the outset.
  - Analyses on random subdomains are no longer preserved.
  - Partially synthetic or hybrid data are more flexible.

Synthetic data by multiple imputation (Rubin 1993)

1. Let X be microdata set of n records drawn from a much larger population of N individuals, with background attributes A, non-confidential attributes B and confidential attributes C.
2. Attributes A are observed for all N individuals, whereas B and C are only available for the n records in X.
3. For M between 3 and 10, do:
   1. Construct a matrix of (B, C) data for the $N - n$ non-sampled individuals, by drawing from an imputation model predicting (B, C) from A (constructed from the n records in X).
   2. Use simple random sampling to draw a sample Z of n' records from the $N - n$ imputed records with attributes (A, B, C).
   3. Publish synthetic data set Z.

Evaluation of SDC methods

- Evaluation is in terms of two conflicting goals:
  - Minimize the data utility loss caused by the method.
  - Minimize the extant disclosure risk in the anonymized data.
- The best methods are those that optimize the trade-off between both goals.

Utility loss in tabular SDC

- For cell suppression, utility loss is measured as the number of secondary suppressions or their pooled magnitude.
- For controlled tabular adjustment or rounding, it is measured as the sum of distances between true and perturbed cell values.
- The above loss measures may be weighted by cell costs, if not all cells have the same importance.

Disclosure risk in tabular SDC

- Disclosure risk is evaluated by computing the feasibility intervals for sensitive cells (via linear programming constrained by the marginals).
- The table is safe if the feasibility interval for any sensitive cell contains the protection interval previously defined for that cell.
- Tthe protection interval is the narrowest interval interval estimate of the sensitive cell permitted by the data protector.

Utility loss in SDC of queryable databases

- For query perturbation, the difference between the true query response and the perturbed query response is a measure of utility loss $\Longrightarrow$ this can be characterized in terms of the mean and variance of the noise being added (ideally, the mean should be zero and the variance small)
- For query restriction, utility loss can be measured as the number of refused queries.
- For camouflage, utility loss is proportional to the width of the returned intervals.

Disclosure risk in SDC of queryable databases

- If query perturbation is used according to a privacy model like ε-differential privacy, disclosure risk is controlled *a priori* by the ε parameter (the lower, the less risk).
- In query restriction, the query set size below which queries are refused is a measure of disclosure risk (a query set size 1 means total disclosure).
- In camouflage, disclosure risk is inversely proportional to the interval width.

Utility and disclosure risk in microdata SDC

- Data utility
  - Data use-specific utility loss measures
  - Generic utility loss measures
- Disclosure risk
  - Fixed *a priori* by a privacy model (ε-differential privacy, k-anonymity)
  - Measured *a posteriori* by record linkage

Microdata use-specific utility loss measures

- If the data protector can anticipate the analyses that the users wish to carry out on the anonymized data, then s/he can choose SDC methods and parameters that, while adequately controlling disclosure risk, minimize the impact on those analyses.
- Unfortunately, the precise user analyses cannot be anticipated when anonymized data are released for general use.
- Releasing different anonymized versions of the same data set optimized for different data uses might result in disclosure $\Longrightarrow$ SDC must often be based on generic utility measures.

Numerical microdata generic utility loss measures

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

All-type microdata generic utility loss measures

- In [Woo *et al.*, 2009] a utility measure applicable to continuous and categorical microdata was proposed.
- Merge the original and anonymized microdata sets and add a binary attribute T with value 1 for the anonymized records and 0 for the original records.
- Regress T on the rest of attributes of the merged data set and call the adjusted attribute $\hat{T}$. Let the propensity score $\hat{p}_i$ of record i of the merged data set be the value of $\hat{T}$ for record i.
- Then utility is high if the propensity scores of the anonymized and original records are similar.
- Hence, if the number of original and anonymized records is the same, a utility measure is

$$U = \frac{1}{N}\sum_{i=1}^{N}[\hat{p}_i - 1/2]^2$$

A priori disclosure risk control in microdata

- Using a privacy model like k-anonymity or differential privacy allows the tolerable disclosure risk to be selected at the outset.
- For k-anonymity the risk of identity disclosure is upper-bounded by $1/k$.
- ε-Differential privacy can ensure a very low identity and disclosure (esp. for small ε), but at the expense of a great utility loss.

k-Anonymity extensions against attribute disclosure

- k-Anonymity does not protect against attribute disclosure in general (*e.g.* if the values of a confidential attribute are very similar in a group of k records sharing quasi-identifier values).
- l-Diversity is an extension requiring that the values of all confidential attributes within a group of k records contain at least l clearly distinct values.
- t-Closeness is another extension requiring that the distribution of the confidential attribute within a group of k records be similar to the distribution of the confidential attribute in the entire data set (at most distance t between both distributions).

A posteriori disclosure risk control in microdata

The uniqueness approach

- Typically used with non-perturbative masking, specifically sampling.
- It measures disclosure risk as the probability that rare combinations of attribute values in the released data are indeed rare in the original population the data come from.
- The probability that a sample unique is a population unique decreases with the sampling fraction (Skinner *et al.*, 1990)
  $\Longrightarrow$ reducing the sampling fraction reduces the disclosure risk.

A posteriori disclosure risk control in microdata (II)

The record linkage approach

- Record linkage software is used to estimate the percentage of valid re-identifications obtainable by an intruder who links via quasi-identifiers the anonymized data with an external non-anonymous data set
  - It can be applied to any type of masking and synthetic data.
  - It can even be applied to measure the actual disclosure risk of ε-differentially private data releases.

Record linkage

- Record linkage (RL) is a procedure to link each record a in file A (*e.g.* anonymized file) to a record b in file B (*e.g.* original file).
- The pair (a, b) is a match is b turns out to be the original record corresponding to a.
- RL was created for data fusion and to increase data quality.
- Two types: distance-based RL and probabilistic RL.

Distance-based record linkage

- Each record a in file A is linked to its nearest neighbor b in file B.
- A record-level distance function is needed to measure nearness.
- The record-level distance can be computed from attribute-level distances[1].
- To combine them, attribute-level distances must be standardized and each attribute must be given a weight.
- Choosing suitable attribute weights and attribute-level distances is not obvious.

[1] *E.g.* Euclidean distance for numerical attributes and a categorical distance for categorical attributes.

Probabilistic record linkage

- This type of RL (due to Fellegi and Sunter, 1969) computes an index between each pair (a, b) of records in A and B, resp.
- Two thresholds LT and NLT in the index range are used to label a pair a linked ($> LT$), non-linked ($< NLT$) or pair that must be inspected by a human (otherwise).
- If attributes can be assumed independent, the index can be computed from the following probabilities:
  - $P(1|M)$: prob. of coincidence between attribute values in records a and b given that such records are a real match;
  - $P(0|U)$: prob. of non-coincidence given that a and b are a real unmatch.

Probabilistic record linkage (II)

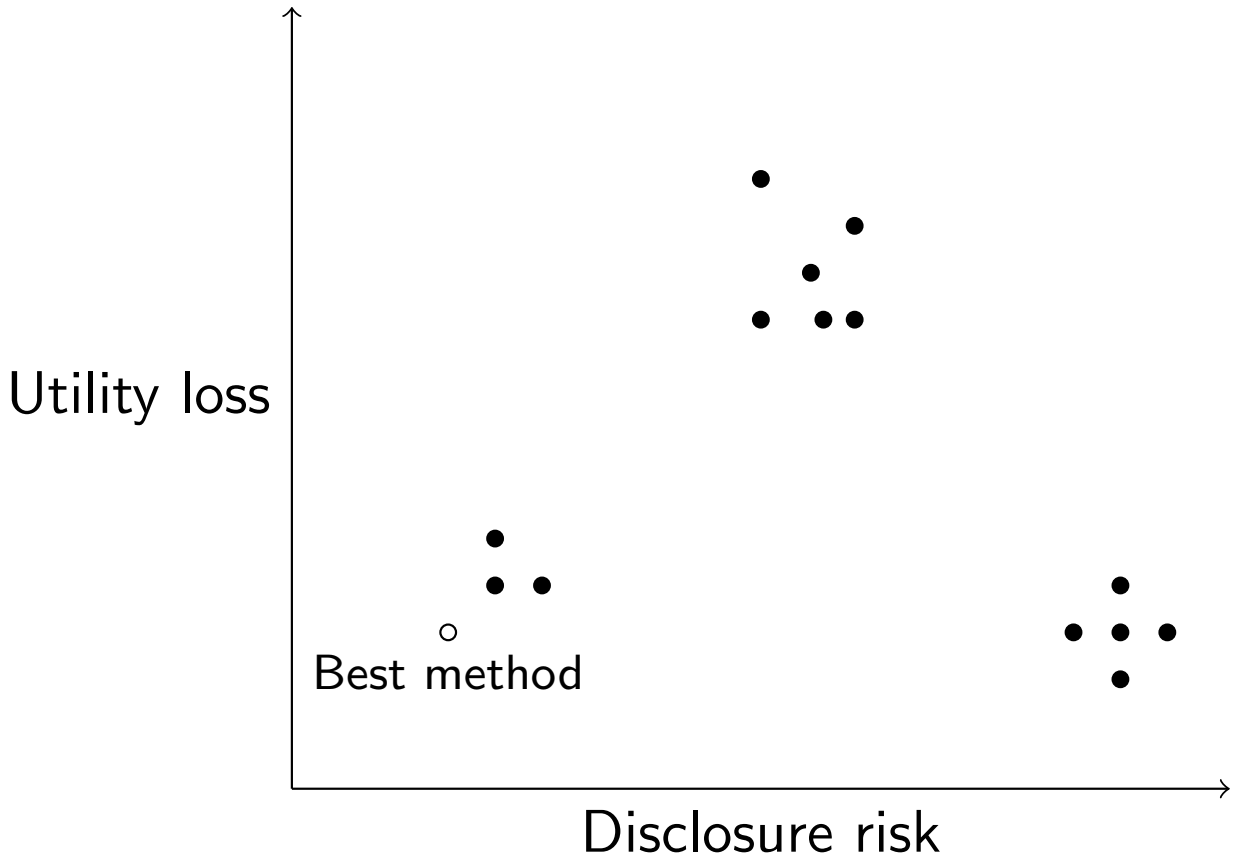
- One needs to set thresholds LT and NLT and estimate $P(1|M)$ and $P(0|U)$.
- Thresholds are computed from:
  - Maximum acceptable prob. $P(LP|U)$ of linking an unmatch (*false positive*).
  - Maximum acceptable prob. $P(NP|M)$ of not linking a match (*false negative*).
- $P(1|M)$ and $P(0|U)$ are estimated using the EM algorithm.

Trading off utility loss and disclosure risk

If the *a posteriori* approach is chosen, tools to trade off utility loss and disclosure risk include

- *R-U maps*. For each method and parameterization, plot its pair (disclosure risk, utility loss) in a two-dimensional graph having disclosure risk as abscissae and utility loss as ordinates.
- *R-U score*. A score (formula) is constructed that combines one or several utility loss measures and one or several disclosure risk measures. Then the SDC method and parameterization is chosen that minimizes this score.

R-U map



Example R-U score

- The first R-U score was proposed in Domingo-Ferrer *et al.* (2001) and was used to rank parameterized microdata SDC methods in Domingo-Ferrer and Torra (2001).
- For each method M and parameterization P, compute

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

where IL is an information loss measure, DR is a disclosure risk measure, and $\mathbf{X}'$ is the anonymized microdata set obtained from the original $\mathbf{X}$ after applying method M with parameterization P.

Example R-U score (II)

- *E.g.IL* can be computed by averaging the mean variations of $X - X'$, $\bar{X} - barX'$, $V - V'$, $S - S'$ and the mean absolute error of $R - R'$, and multiplying the average by 100.
- *E.g. DR* can be obtained by averaging the percentages of correctly linked pairs via distance-based RL and via probabilistic RL.

Anonymization freeware

- **Argus**, with μ-Argus for microdata and τ-Argus for tables. `http://neon.vb.cbs.nl/casc`
- **sdcMicro**. Statistical Disclosure Control methods for anonymization of microdata and risk estimation. `http://cran.r-project.org/package=sdcMicro`
- **sdcTable**. Methods for statistical disclosure control in tabular data. `http://cran.r-project.org/web/packages/sdcTable/index.html`
- **ARX**, *k*-anonymity, *l*-diversity, *t*-closeness implementation in Java. `http://arx.deidentifier.org`

Bibliography

• J. Domingo-Ferrer (2008) "A critique of *k*-anonymity and some of its enhancements", in *Proc. of ARES/PSAI 2008*, IEEE CS, pp. 990-993.
• J. Domingo-Ferrer (2009) "Statistical databases", in *Wiley Encyclopedia of Computer Science and Engineering*, Wiley, vol. 5, pp. 2810-2820.
• J. Domingo-Ferrer, J. M. Mateo-Sanz and V. Torra (2001) "Comparing SDC methods for microdata on the basis of information loss and disclosure risk", in *Pre-proceedings of ETK-NTTS'2001 (vol. 2)*, Eurostat, pp. 807-826. • J. Domingo-Ferrer and V. Torra (2001) "A quantitative comparison of disclosure control methods for microdata", in *Confidentiality, Disclosure and Data Access*, North Holland, pp. 111-134.
• J. Domingo-Ferrer and V. Torra (2005), "Ordinal, continuous and heterogeneous *k*-anonymity through microaggregation", *Data Mining and Knowledge Discovery*, 11(2):181-193.
• C. Dwork (2006) "Differential privacy", in *ICALP 2006*, LNCS 4052, pp. 1-12.
• I. P. Fellegi and A. B. Sunter (1969) "A theory for record linkage", *Journal of the American Statistical Assoc.*, 64(328):1183-1210.

Bibliography

• A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer and P.-P. de Wolf (2012) *Statistical Disclosure Control*, Wiley.

• P. Samarati and L. Sweeney (1998) "Protecting privacy when disclosing information: *k*-anonymity and its enforcement through generalization and suppression", Technical Report, SRI International.

• M-J. Woo, J.P. Reiter, A. Oganian and A.F. Karr (2009) "Global measures of data utility for microdata masked for disclosure limitation", *J. of Priv. and Conf.*, 1(1):111-124.

• C. Skinner, C. Marsh, S. Openshaw and C. Wymer (1990) "Disclosure avoidance for census microdata in Great Britain", in *Proc. of the 1990 U.S. Bureau of the Census Annual Research Conference*, Arlington VA, Mar. 18-21, 1990, pp. 131-196.

New book on SDC