Disclosure Risk assessment

Vicenç Torra

October, 2013

Institut d'Investigació en Intel·ligència Artificial (IIIA-CSIC), Bellaterra, Catalonia

Outline

Introduction

Disclosure Risk

Disclosure risk

- Identity disclosure
- Attribute disclosure

Disclosure Risk

Disclosure risk

- Attribute disclosure. Increase knowledge about an attribute of an individual
  - An example: Interval disclosure

    Each attribute is independently ranked and a rank interval is defined around the value the attribute takes on each record. The ranks of values within the interval for an attribute around record r should differ less than p percent of the total number of records and the rank in the center of the interval should correspond to the value of the attribute in record r. Then, the proportion of original values that fall into the interval centered around their corresponding protected value is a measure of disclosure risk.

    A 100 percent proportion means that an attacker is completely sure that the original value lies in the interval around the protected value.

Identity disclosure

Disclosure Risk

An scenario for identity disclosure.

- Classification of attributes in three non-disjoint categories.
  - **Identifiers.** These are attributes that *unambiguously* identify the respondent. Examples are passport number, social security number, full name, etc.
  - **Quasi-identifiers.** These are attributes that, in combination, can be linked with external information to re-identify some of the respondents. Examples are age, birth date, gender, job, zipcode, etc. Although a single attribute cannot identify an individual, a subset of them can.
  - **Confidential.** These are attributes which contain sensitive information on the respondent. For example, salary, religion, political affiliation, health condition, etc.

Disclosure Risk

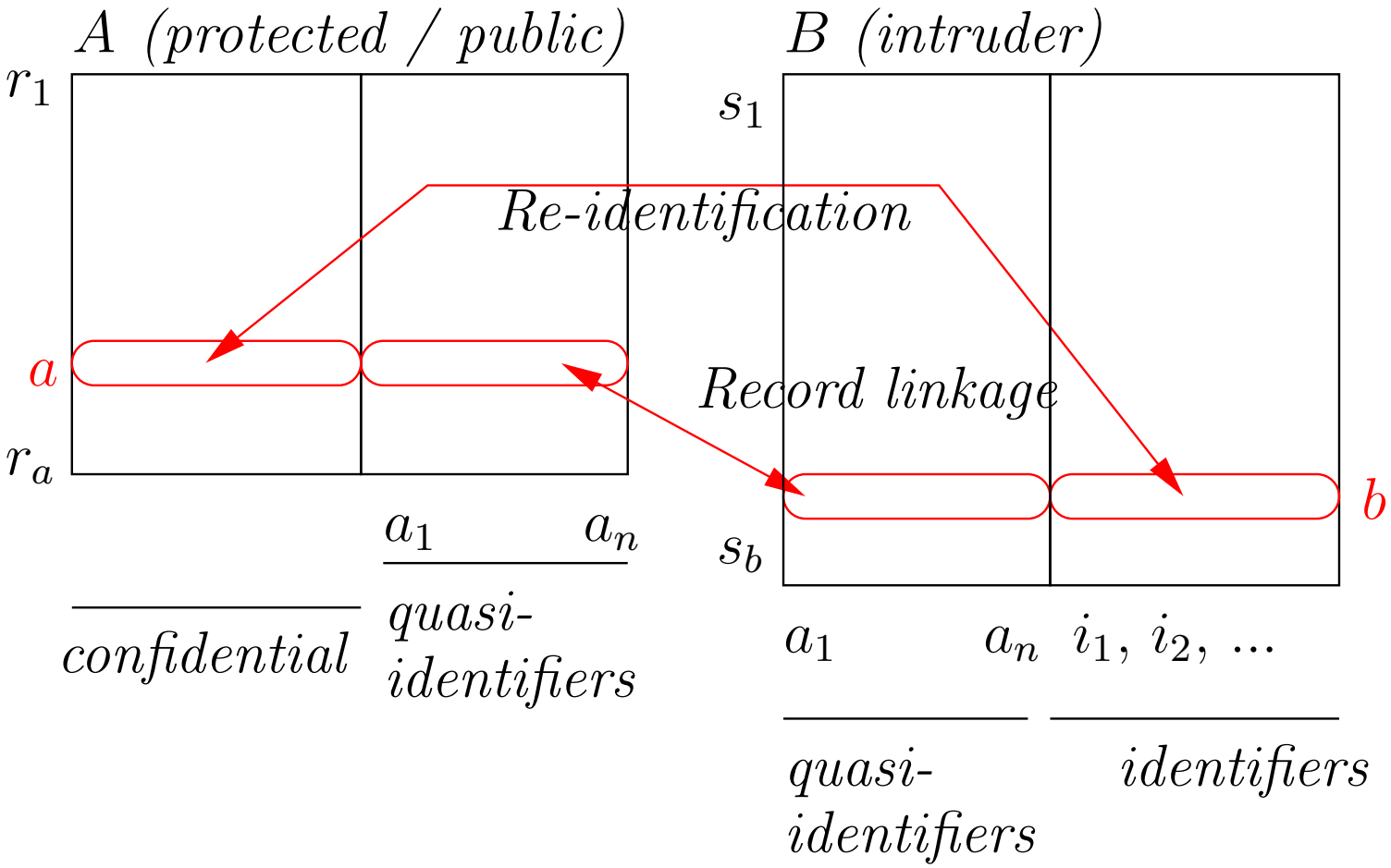
An scenario for identity disclosure: $X = id||X_{nc}||X_c$

- Protection of the attributes
- **Identifiers.** Usually removed or encrypted. So, information cannot be linked to specific respondents.
- **Confidential.** These attributes X_c are usually not modified. So, we have $X'_c = X_c$.
- **Quasi-identifiers.** They cannot be removed as almost all attributes can be quasi-identifiers. Preserve privacy of the individuals applying a protection procedure ρ to these attributes. Therefore, we have $X'_{nc} = \rho(X_{nc})$.

Disclosure Risk

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

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



Disclosure Risk

An scenario for identity disclosure

- Reidentification using the common variables (quasi-identifiers)
- Reidentification permits to link confidential values to identifiers

Disclosure Risk

An scenario for identity disclosure

- Reidentification using the common variables (quasi-identifiers)
- Reidentification permits to link confidential values to identifiers

Identity disclosure implies attribute disclosure

Disclosure Risk

An scenario for identity disclosure: Flexible scenario

- B is a subset of the original file.

Disclosure Risk

An scenario for identity disclosure: Flexible scenario

- B is a subset of the original file.
  $\rightarrow$ the intruder has information on some individuals in the original file

Disclosure Risk

An scenario for identity disclosure: Flexible scenario

- B is a subset of the original file.
  $\rightarrow$ the intruder has information on some individuals in the original file
  $\rightarrow$ the intruder has information on some characteristics of the individuals

Disclosure Risk

An scenario for identity disclosure: Flexible scenario

- B is a subset of the original file.
  $\rightarrow$ the intruder has information on some individuals in the original file
  $\rightarrow$ the intruder has information on some characteristics of the individuals
- A protected file using a masking method
- But also,
  - B with a schema different to the one of A (different attributes)

Disclosure Risk

Measures for identity disclosure

- **Re-identification.** Risk is defined as an estimation on the number of re-identifications that might be obtained by an intruder.

Disclosure Risk

Measures for identity disclosure

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

Disclosure Risk

Measures for identity disclosure

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

Disclosure Risk

Measures for identity disclosure

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

Disclosure Risk

Measures for identity disclosure

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

Uniqueness

Disclosure Risk

Measures for identity disclosure

- **Uniqueness.** Risk is defined as the probability that rare combinations of attribute values in the protected data set are indeed rare in the original population.

Disclosure Risk

Measures for identity disclosure: Uniqueness (categorical data/sampling)

- **File-level uniqueness.** It is defined as the probability that a sample unique (SU) is a population unique (PU). The following expression has been used:

$$P(PU|SU) = \frac{P(PU, SU)}{P(SU)} = \frac{\sum_j I(F_j = 1, f_j = 1)}{\sum_j I(f_j = 1)}$$

  where $j = 1, \ldots, J$ denotes possible values in the sample, F_j is the number of individuals in the population with key value j (frequency of j in the population), f_j is the same frequency for the sample and I stands for the cardinality of the selection.
- **Record-level risk uniqueness.** It is defined as the probability that a particular sample record is re-identified (recognized as corresponding to a particular individual in the population).

Re-identification

Disclosure Risk

Measures for identity disclosure: Re-identification

- Risk is defined as an estimation on the number of re-identifications that might be obtained by an intruder.

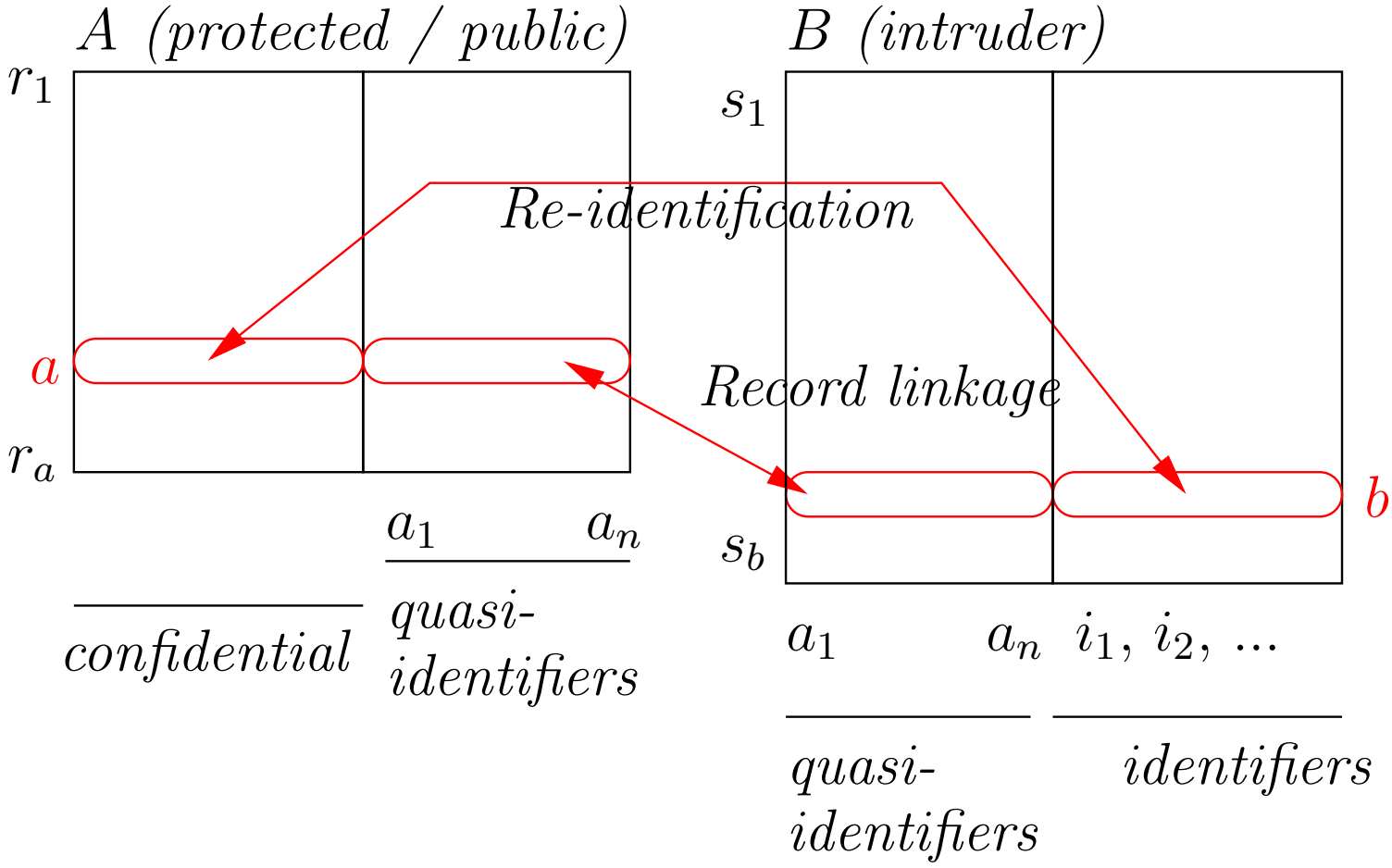
Type of algorithms

- Probabilistic record linkage
- Distance based record linkage

Disclosure Risk

Measures for identity disclosure: Re-identification

- Probabilistic record linkage



- Classification of pairs (a, b) in 3 classes
  Linked pair, non-linked, clerical pair

Disclosure Risk

Measures for identity disclosure: Re-identification

- Probabilistic record linkage
  - An index is computed for each pair of records (a, b).
  - The index is computed using the conditional probabilities
    - $P(coincidence|Matching)$: coincidence between both records when there is matching
    - $P(coincidence|Unmatching)$: coincidence between both records when there is no matching
  - Using thresholds, a pair (a, b) is classified as Linked pair, non-linked, clerical pair

Disclosure Risk

Measures for identity disclosure: Re-identification

- Probabilistic record linkage
  - Computation of
    $P(coincidence|Matching)$ and
    $P(coincidence|Unmatching)$:

Disclosure Risk

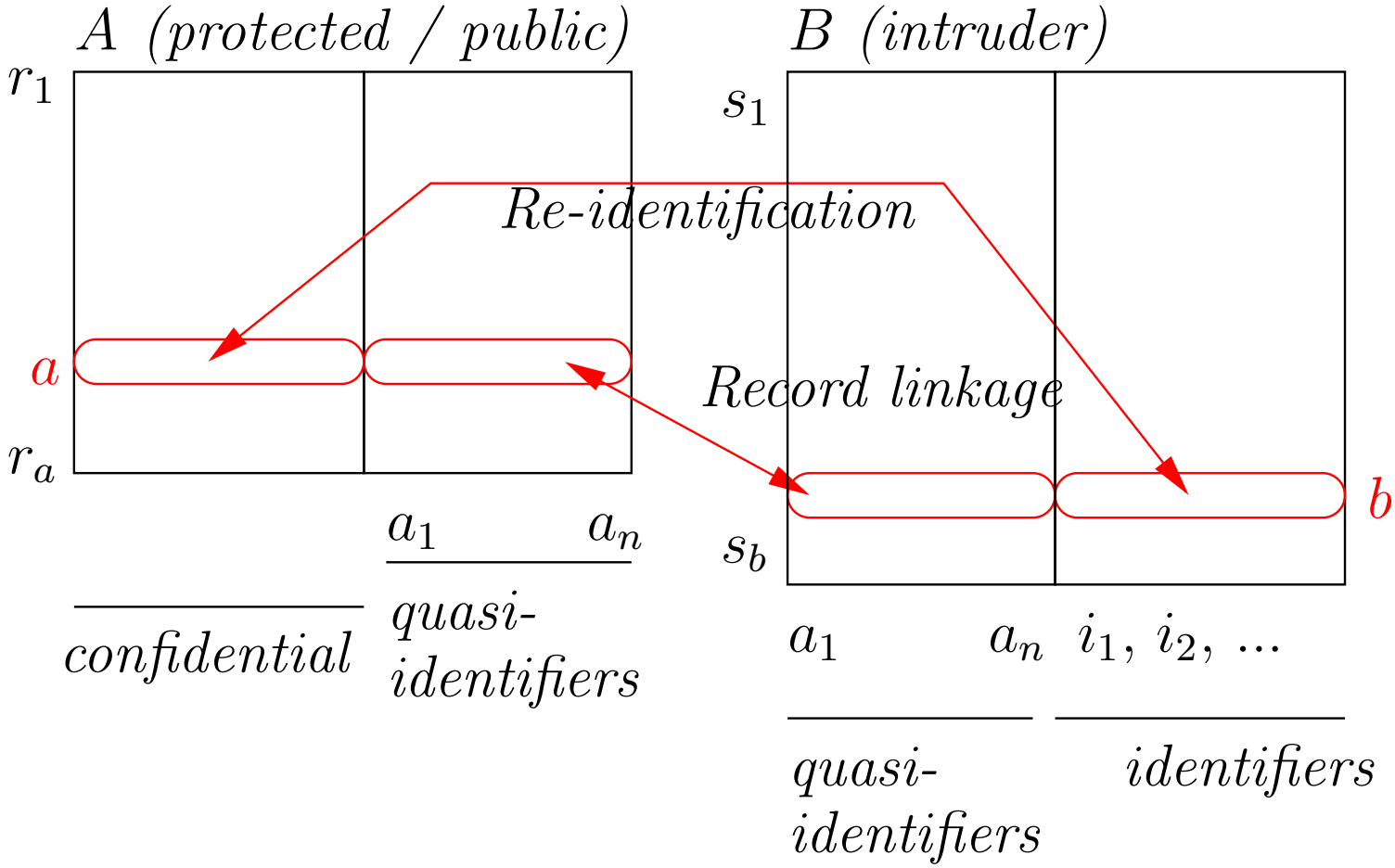
Measures for identity disclosure: Re-identification

- Probabilistic record linkage
  - Computation of
    $P(coincidence|Matching)$ and
    $P(coincidence|Unmatching)$:
    - ⋆ Using EM algorithm
  - Computation of thresholds
    - ⋆ From the probabilities of false positive/negative
      $P(Linkedpair|Unmatching)$
      $P(Nonlinkedpair|Matching)$

Disclosure Risk

Measures for identity disclosure: Re-identification

- Distance based record linkage: assignment to the nearest record



- Classification of a record b in one of the classes in A.

Disclosure Risk

Measures for identity disclosure: Re-identification

- Distance-based record linkage

Disclosure Risk

Measures for identity disclosure: Re-identification

- Distance-based record linkage

for $a \in \mathbf{A}$ **do**
  b' = arg $\min_{b \in \mathbf{B}} d(a, b)$
  $\mathbf{LP} = \mathbf{LP} \cup (a, b')$
  for $b \in \mathbf{B}$ **such that** $b \neq b'$ **do**
    $\mathbf{NP} = \mathbf{NP} \cup (a, b)$

Disclosure Risk

Measures for identity disclosure: Re-identification

- Distance-based record linkage
  - Definition of a distance (numerical):
    - ⋆ Euclidean distance
    - ⋆ Kernel distance

$$\begin{aligned} d(a,b)^2 &= ||\Phi(a) - \Phi(b)||^2 = (\Phi(a) - \Phi(b))^2 \\ &= \Phi(a)\cdot\Phi(b) - 2\Phi(a)\cdot\Phi(b) + \Phi(b)\cdot\Phi(b) \\ &= K(a,a) - 2K(a,b) + K(b,b) \end{aligned}$$

    - ⋆ Mahalanobis distance

$$d(a,b)^2 = (a-b)[Var(V^A)+Var(V^B)-2Cov(V^AV^B)]^{-1}(a-b)$$

Disclosure Risk

Measures for identity disclosure: DBRL vs. PRL

- Distance-based record linkage
  - easier to implement
  - easier to include subjective information (in the distance)
  - we can weight different attributes (advantage or incovenient?)
  - difficulty for defining distances for ordinal attributes
- Probabilistic record linkage
  - more difficult to implement
  - parameters are found with EM (only parameters for errors)
  - no possibility for weigths
- DBRL and PRL give similar results
- DBRL with Mahalanobis distance: outperformed once

Disclosure Risk

Extensions for record linkage

- Weighted variables in distance based record linkage
  - Weighted distances
- Non-independent variables
  - Probabilistic record linkage
  - Distance based record linkage:
    Mahalanobis, Choquet integral-based distance

Disclosure Risk

Extensions for record linkage

- Same vs. different schema in the files
  - Tools for schema matching, data cleaning,
- Ontology-based record linkage for textual microdata

Disclosure Risk

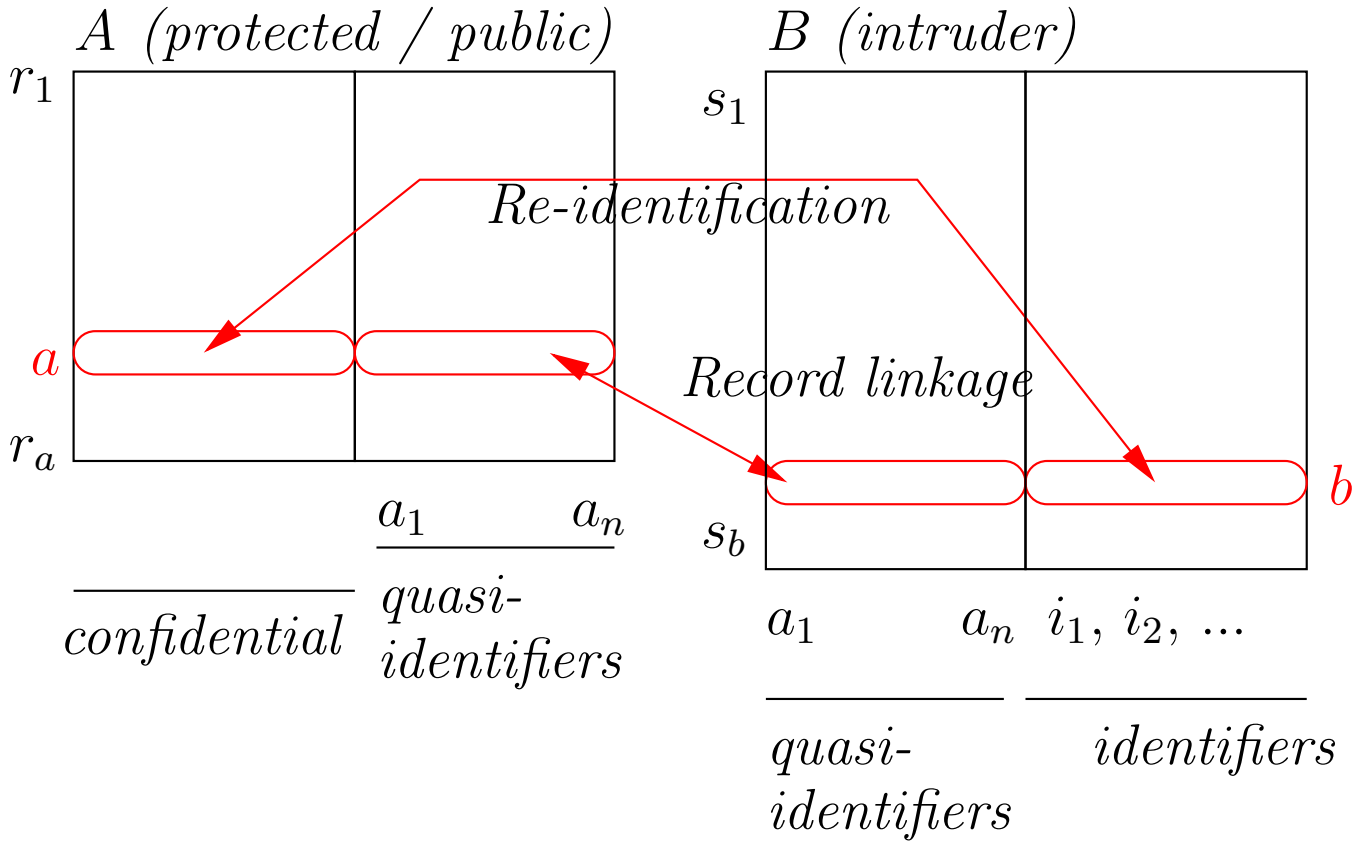
Extensions for record linkage

- Machine learning approaches for distance based record linkage

Disclosure Risk

Extensions for record linkage

- Machine learning approaches for distance based record linkage
- Learn the parameters of the distance
  - Given a distance d_p, supervised approach to learn p.



Disclosure Risk

Machine Learning for DBRL

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

$$Subject\ to:$$

$$\mathbb{C}(d(V_1(a_i), V_1(b_j)), \ldots, d(V_n(a_i), V_n(b_j))) - \\ - \mathbb{C}(d(V_1(a_i), V_1(b_i)), \ldots, d(V_n(a_i), V_n(b_i))) + CK_i > 0 \tag{2}$$

$$K_i \in \{0, 1\} \tag{3}$$

Additional constraints according to $\mathbb{C}$ (4)

Disclosure Risk

Extensions for record linkage

- Generic-disclosure risk measures
- Specific-disclosure risk measures

Definitions of privacy

Disclosure Risk

Definitions of privacy

- k-anonymity
- differential privacy

Disclosure Risk

Definitions of privacy

- k-anonymity
- differential privacy

Forget about disclosure risk, focus on information loss

k-Anonymity

Disclosure Risk

k-Anonymity (Samarati, Sweeney; 1998, 2001, 2002)

- **Definition.** Let $RT(A_1, \ldots, A_n)$ be a table, and QI_{RT} be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in $RT[QI_{RT}]$ appears with at least k occurrences in $RT[QI_{RT}]$.

Disclosure Risk

k-Anonymity (Samarati, Sweeney; 1998, 2001, 2002)

- **Definition.** Let $RT(A_1, \ldots, A_n)$ be a table, and QI_{RT} be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in $RT[QI_{RT}]$ appears with at least k occurrences in $RT[QI_{RT}]$.
- **Example.** k-anonymous table for $k = 2$ when the $QI_{RT} = \{\text{City,age}\}$.

Disclosure Risk

k-Anonymity (Samarati, Sweeney; 1998, 2001, 2002)

- **Definition.** Let $RT(A_1, \ldots, A_n)$ be a table, and QI_{RT} be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in $RT[QI_{RT}]$ appears with at least k occurrences in $RT[QI_{RT}]$.
- **Example.** k-anonymous table for $k = 2$ when the $QI_{RT} = \{$City,age$\}$.

Respondent	City	age	illness
ABD	Barcelona	30	Cancer
COL	Barcelona	30	Cancer
GHE	Tarragona	60	AIDS
CIO	Tarragona	60	AIDS
HYU	Tarragona	60	Heart attack

Disclosure Risk

k-Anonymity (Samarati, Sweeney; 1998, 2001, 2002)

- **Definition.** A table satisfies k-anonymity when it is partitioned into sets of at least k indistinguishable records.
- **Discussion.**

Disclosure Risk

k-Anonymity (Samarati, Sweeney; 1998, 2001, 2002)

- **Definition.** A table satisfies k-anonymity when it is partitioned into sets of at least k indistinguishable records.
- **Discussion.**
  - k-Anonymity is not a protection procedure in itself but a condition to be satisfied by the protected data set.

Disclosure Risk

k-Anonymity (Samarati, Sweeney; 1998, 2001, 2002)

- **Definition.** A table satisfies k-anonymity when it is partitioned into sets of at least k indistinguishable records.
- **Discussion.**
  - k-Anonymity is not a protection procedure in itself but a condition to be satisfied by the protected data set.
  - Its goal is to avoid disclosure (identity disclosure / attribute disclosure)

Disclosure Risk

k-Anonymity (Samarati, Sweeney; 1998, 2001, 2002)

- **Definition.** A table satisfies k-anonymity when it is partitioned into sets of at least k indistinguishable records.
- **Discussion.**
  - k-Anonymity is not a protection procedure in itself but a condition to be satisfied by the protected data set.
  - Its goal is to avoid disclosure (identity disclosure / attribute disclosure)
  - The definition of k-anonymity makes that algorithm focus on information loss.

Disclosure Risk

k-Anonymity (Samarati, Sweeney; 1998, 2001, 2002)

- **Definition.** A table satisfies k-anonymity when it is partitioned into sets of at least k indistinguishable records.
- **Discussion.**
  - k-Anonymity is not a protection procedure in itself but a condition to be satisfied by the protected data set.
  - Its goal is to avoid disclosure (identity disclosure / attribute disclosure)
  - The definition of k-anonymity makes that algorithm focus on information loss.
  - Different levels of k lead to different protections

Disclosure Risk

k-Anonymity (Samarati, Sweeney; 1998, 2001, 2002)

- **Definition.** A table satisfies k-anonymity when it is partitioned into sets of at least k indistinguishable records.
- **Discussion.**
  - k-Anonymity is not a protection procedure in itself but a condition to be satisfied by the protected data set.
  - Its goal is to avoid disclosure (identity disclosure / attribute disclosure)
  - The definition of k-anonymity makes that algorithm focus on information loss.
  - Different levels of k lead to different protections
  - k-Anonymity through generalization and suppression: NP-Hard problem

Disclosure Risk

k-Anonymity

- Attacks. (I)

Disclosure Risk

k-Anonymity

- Attacks. (I)
  - **Homogeneity attack.** When all indistinguishable records in a cluster are also indistinguishable with respect to a confidential variable, attribute disclosure can take place.

Disclosure Risk

k-Anonymity

- Attacks. (I)
  - **Homogeneity attack.** When all indistinguishable records in a cluster are also indistinguishable with respect to a confidential variable, attribute disclosure can take place.
  - **Example.**

Respondent	City	age	illness
ABD	Barcelona	30	Cancer
COL	Barcelona	30	Cancer
GHE	Tarragona	60	AIDS
CIO	Tarragona	60	AIDS

Disclosure Risk

k-Anonymity

- Attacks. (II)
  - **External knowledge attack.** In this case, some information about an individual is used to deduce information of the same or another individual.

Disclosure Risk

k-Anonymity

- Attacks. (II)
  - **External knowledge attack.** In this case, some information about an individual is used to deduce information of the same or another individual.
  - **Example.** If we are HYU, we can deduce that CIO has AIDS (without reidentification).

Respondent	City	age	illness
ABD	Barcelona	30	Cancer
COL	Barcelona	30	Cancer
GHE	Tarragona	60	AIDS
CIO	Tarragona	60	AIDS
HYU	Tarragona	60	Heart attack

Disclosure Risk

k-Anonymity: Extensions (I)

- p-**Sensitive** k-**anonymity.** (Truta, Vinay, 2006)
  A data set is said to satisfy p-sensitive k-anonymity for $k > 1$ and $p \leq k$ if it satisfies k-anonymity and, for each group of records with the same combination of values for quasi-identifiers, the number of distinct values for each confidential value is at least p (within the same group).

Disclosure Risk

k-Anonymity: Extensions (II)

- l-**Diversity.** (Machanavajjhala et al. 2006)
  It forces l different categories in each set. However, in this case, categories should have to be *well-represented*. Different meanings have been given to what *well-represented* means.

Disclosure Risk

k-Anonymity: Extensions (III)

- **t-closeness.** (Li, Li, Venkatasubramanian, 2007)
  The distribution of the attribute in any k-anonymous subset of the database is similar to the one of the full database. Similarity is defined in terms of the distance between the two distributions and such distance should be below a given threshold t.
  Low threshold makes the utility of the data doubtful: large information loss.

Differential Privacy

Disclosure Risk

Differential privacy

- **Informal Definition.** Given a database X and a query Q, a data protection method satisfies differential privacy when the outcome of Q for the dataset X does not change significantly for the suppression of any record in X. In other words, the absence or presence of a record is not significant for the output.

Disclosure Risk

Differential privacy

- **Definition.** (Dwork, 2006)
  A randomized function Q gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(Q)$,

$$Pr[Q(D_1) \in S] \leq e^{\epsilon} \cdot Pr[Q(D_2) \in S],$$

  where the probability space in each case is over the coin flips of the mechanism Q.

Disclosure Risk

Differential privacy

- **Definition.** (Dwork, 2006)
  A randomized function Q gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(Q)$,

$$Pr[Q(D_1) \in S] \leq e^{\epsilon} \cdot Pr[Q(D_2) \in S],$$

  where the probability space in each case is over the coin flips of the mechanism Q.
  Note-1. D_1 and D_2 differ in at most one element if one is a proper subset of the other and the larger database contains just one additional row.

Disclosure Risk

Differential privacy

- **Definition.** (Dwork, 2006)
  A randomized function Q gives ϵ-differential privacy if for all data sets D_1 and D_2 differing on at most one element, and all $S \subseteq Range(Q)$,

$$Pr[Q(D_1) \in S] \leq e^{\epsilon} \cdot Pr[Q(D_2) \in S],$$

  where the probability space in each case is over the coin flips of the mechanism Q.
  Note-1. D_1 and D_2 differ in at most one element if one is a proper subset of the other and the larger database contains just one additional row.
  Note-2. Q the function to respond to a particular query; S is the response to the query.

Disclosure Risk

Differential privacy. Discussion

- Focused on a query Q.
- What matters is the difference between the responses of two databases one including an element and the other not.

Disclosure Risk

Differential privacy. Discussion

- Focused on a query Q.
- What matters is the difference between the responses of two databases one including an element and the other not.
- (Dwork, 2008) states that the mechanism Q satisfying the definition "addresses concerns that any participant might have about the leakage of her personal information: even if the participant removed her data from the data set, no outputs (and thus consequences of outputs) would become significantly more or less likely".

Disclosure Risk

Differential privacy. Discussion

- Differential privacy focuses on attribute disclosure

Disclosure Risk

Differential privacy. Discussion

- Differential privacy focuses on attribute disclosure
- Given a level of privacy, algorithms optimize information loss

Disclosure Risk

Differential privacy. Discussion

- Differential privacy focuses on attribute disclosure
- Given a level of privacy, algorithms optimize information loss
- Concerns about differential privacy (Sarathy, Muralidhar, 2011; Bambauer, Muralidhar, Sarathy, 2013)

Disclosure Risk

Differential privacy. Discussion

- Differential privacy focuses on attribute disclosure
- Given a level of privacy, algorithms optimize information loss
- Concerns about differential privacy (Sarathy, Muralidhar, 2011; Bambauer, Muralidhar, Sarathy, 2013)
- The noise needed to ensure differential privacy is too high for data utility

Summary

Disclosure Risk

Disclosure risk measures

Definitions of privacy

- k-Anonymity
- Differential privacy