Privacy models and disclosure risk

Vicenç Torra

March, 2020

Umeå University, Sweden

Outline

1. Disclosure

2. Privacy models

3. Classification of privacy models and disclosure risk measures
   - Attribute disclosure
   - Identity disclosure
     (record linkage and worst-case scenario)

4. k-Anonymity

Disclosure

Disclosure risk assessment

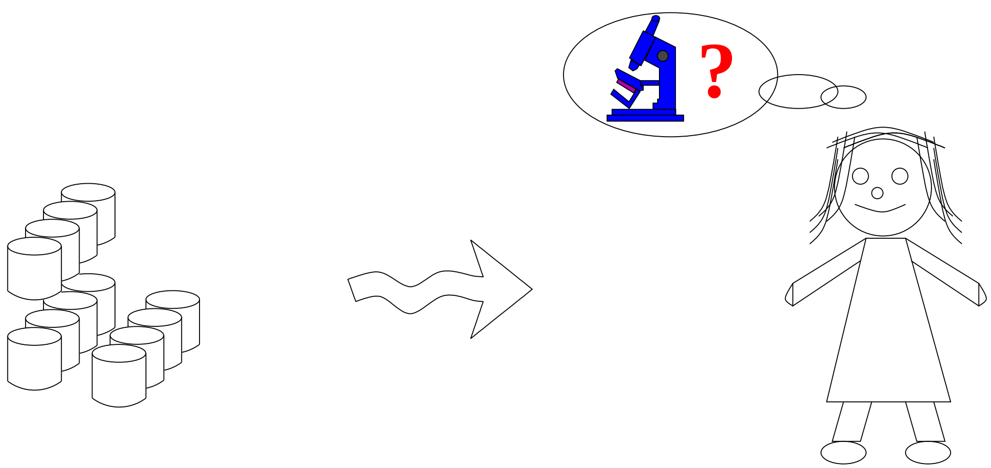
Disclosure risk. Disclosure = leakage of information.

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

Within machine learning, some attribute disclosure is expected.

Privacy models

Privacy models



Privacy models

Privacy models: What is a privacy model ?

- To make a program we need to know what we want to protect

Definition:

- A computational definition for privacy

Privacy models

Privacy models: What is a privacy model ?

- To make a program we need to know what we want to protect

Definition:

- A computational definition for privacy

Quite a large number of *computational definitions*,
they depend on what to protect.

Privacy models

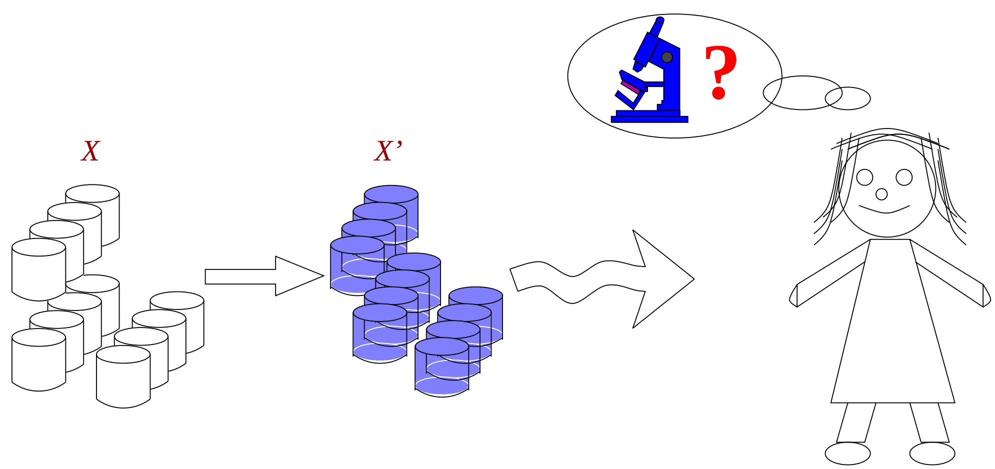
Privacy models. A computational definition for privacy. Examples.

- **Reidentification privacy.** Avoid finding a record in a database.
- **k-Anonymity.** A record indistinguishable with $k - 1$ other records.
- **Secure multiparty computation.** Several parties want to compute a function of their databases, but only sharing the result.
- **Differential privacy.** The output of a query to a database should not depend (much) on whether a record is in the database or not.
- **Result privacy.** We want to avoid some results when an algorithm is applied to a database.
- **Integral privacy.** Inference on the databases. E.g., changes have been applied to a database.
- **Homomorphic encryption.** We want to avoid access to raw data and partial computations.

Privacy models

Privacy models. A computational definition for privacy. Publish a DB

- **Reidentification privacy.** Avoid finding a record in a database.
- **k-Anonymity.** A record indistinguishable with $k-1$ other records.
- **k-Anonymity, l-diversity.** l possible categories
- **Interval disclosure.** The value for an attribute is outside an interval computed from the protected value: values different enough.
- **Result privacy.** We want to avoid some results when an algorithm is applied to a database.



Privacy models

Privacy models. A computational definition for privacy. Publish a DB

- Modify DB X to obtain a DB X' compliant with the privacy model.

Original DB X:

Respondent	City	Age	Illness
DRR	Barcelona	30	Heart attack
ABD	Barcelona	32	Cancer
COL	Barcelona	33	Cancer
GHE	Tarragona	62	AIDS
CIO	Tarragona	65	AIDS
HYU	Tarragona	60	Heart attack

Published DB X':

——	City	Age	Illness
—	Barcelona	30	Cancer
—	Barcelona	30	Cancer
—	Barcelona	30	Cancer
—	Tarragona	60	AIDS
—	Tarragona	60	AIDS
—	——	–	——

Privacy models

- Difficulties: naive anonymization does not work
  - (Sweeney, 1997; 2000[1]) on USA population
    - 87.1% (216 /248 million) is likely to be uniquely identified by 5-digit ZIP, gender, date of birth,
    - 3.7% (9.1 /248 million) is likely to be uniquely identified by 5-digit ZIP, gender, Month and year of birth.
- Difficulties: highly identifiable data and high dimensional data
  - Data from mobile devices:
    - two positions can make you unique (home and working place)
  - AOL and Netflix cases (search logs and movie ratings)
  - Similar with credit card payments, shopping carts, search logs, ... (i.e., high dimensional data)

[1] L. Sweeney, Simple Demographics Often Identify People Uniquely, CMU 2000

Privacy models

- Difficulties: Example 1.
  - Q: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)

Privacy models

- Difficulties: Example 1.
  - Q: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)
  - No “personal data”,

    $DB = \{$ (Dublin, CS, No), (Dublin, CS, No),
    (Dublin, CS, Yes), (Maynooth, CS, No), . . . ,
    (Dublin, BA MEDIA STUDIES, No)
    (Dublin, BA MEDIA STUDIES, Yes), . . . $\}$

    is this ok ?

Privacy models

- Difficulties: Example 1.
  - Q: sickness influenced by studies & commuting distance?
  - Records: (where students live, what they study, if they got sick)
  - No "personal data",

    $DB = \{$ (Dublin, CS, No), (Dublin, CS, No),
    (Dublin, CS, Yes), (Maynooth, CS, No), . . . ,
    (Dublin, BA MEDIA STUDIES, No)
    (Dublin, BA MEDIA STUDIES, Yes), . . . $\}$

    is this ok ?

    NO!!:
  - E.g., there is only one student of anthropology living in Enfield.

    (Enfield, Anthropology, Yes)

Privacy models

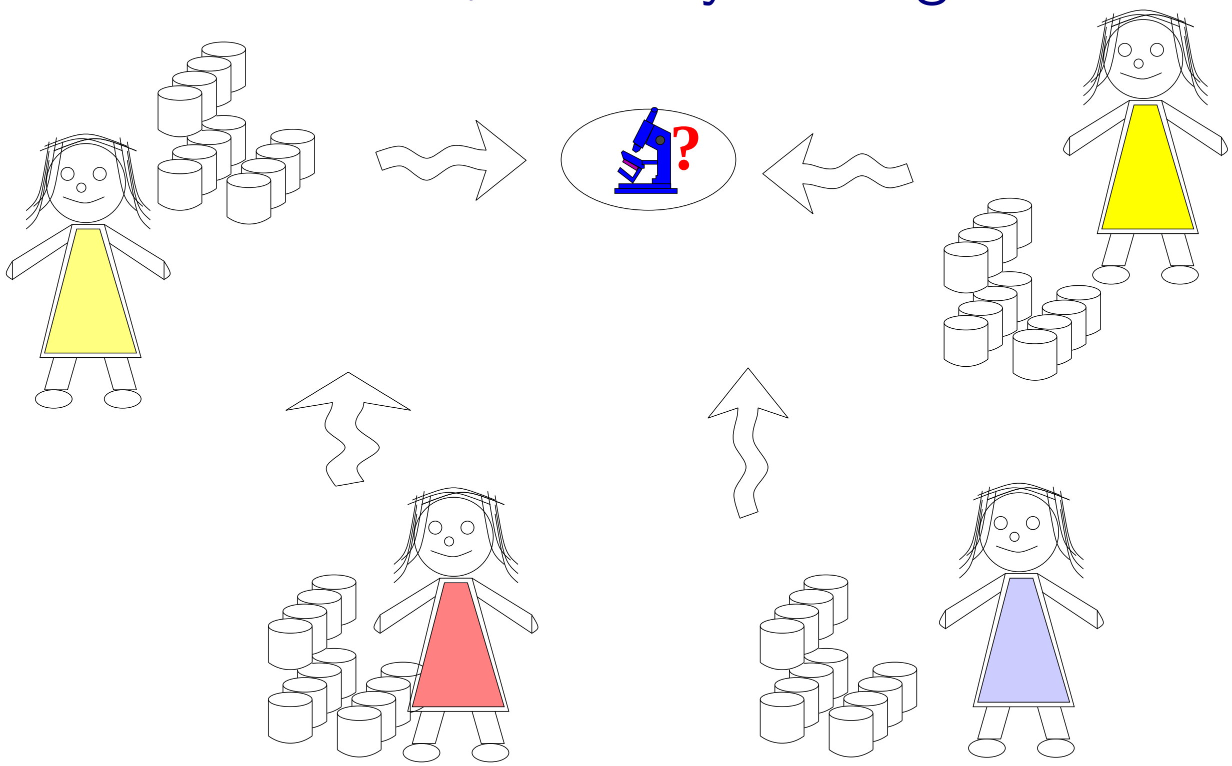
- Difficulties (summary)
  Naive anonymization does not work,
  highly identifiable data, high dimensional data

- Examples of successful reidentification attacks
  Sweeney analysis of USA population, data from mobile data, shopping cards, film ratings

Privacy models

Privacy models. A computational definition for privacy. Share a result

- **Secure multiparty computation.** Several parties want to compute a function of their databases, but only sharing the result.



Privacy models

Privacy models. A computational definition for privacy. Share a result

- Compute

$$f(DB_1, DB_2, DB_3, DB_4)$$

  without sharing DB_1, DB_2, DB_3, DB_4
- Example: national age mean of hospital-acquired infection patients (hospitals do not want to share the age of their infected patients!)

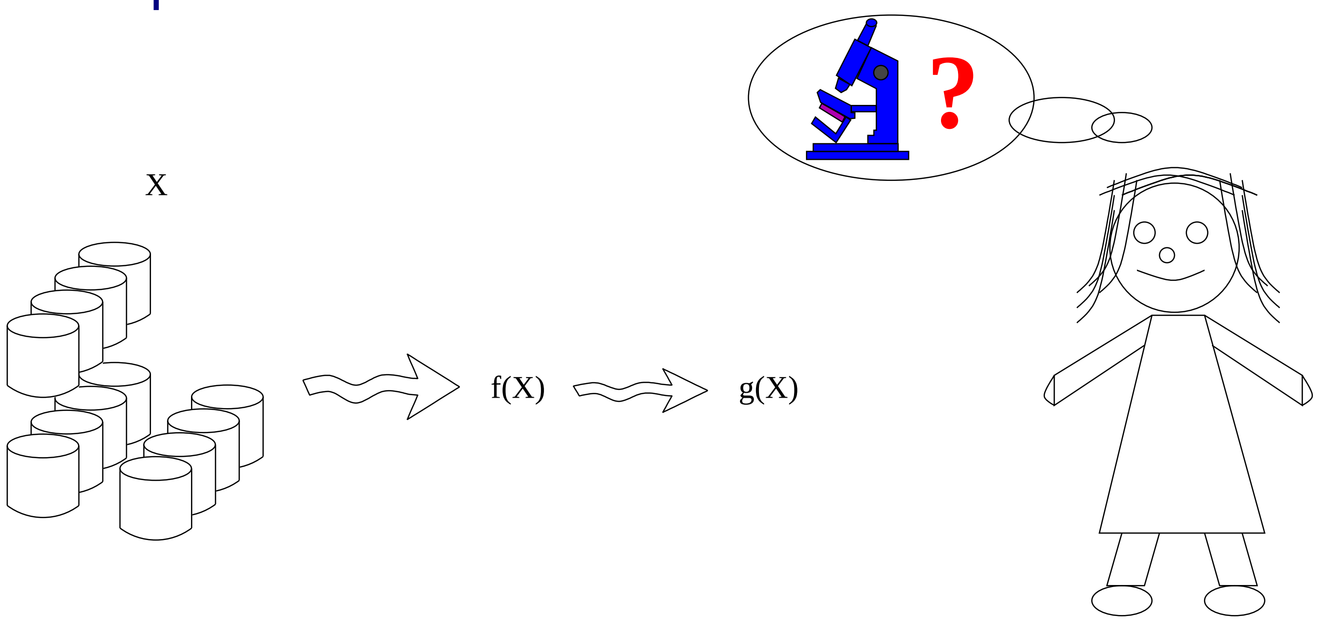
Privacy models

- Difficulties
  - Distributed approach (no trusted-third party) – computational cost of solutions
  - Protocols only valid for a particular function

Privacy models

Privacy models. A computational definition for privacy. Compute result

- **Differential privacy.** The output of a query to a database should not depend (much) on whether a record is in the database or not.
- **Integral privacy.** Inference on the databases. E.g., changes have been applied to a database.
- **Homomorphic encryption.** We want to avoid access to raw data and partial computations.



Privacy models

- Difficulties. Output of a function can be sensitive. Example 2
  - Mean income of admitted to hospital unit (e.g., psychiatric unit)
  - Mean salary of participants in Alcoholics Anonymous by town

  Is this ok? NO!!

  - disclosure of a rich person in the database

Privacy models: Summary

- Privacy models: quite a few *competing* models
  - differential privacy
  - secure multiparty computation
  - k-anonymity
  - k-Anonymity, l-diversity
  - computational anonymity
  - reidentification (record linkage)
  - uniqueness
  - result privacy
  - interval disclosure
  - integral privacy

Privacy models: Summary

- Privacy models: quite a few *competing* models
  - differential privacy
  - secure multiparty computation
  - k-anonymity
  - k-Anonymity, l-diversity
  - computational anonymity
  - reidentification (record linkage)
  - uniqueness
  - result privacy
  - interval disclosure
  - integral privacy
- ... and combined:
  - secure multiparty computation + differential privacy

Classification of privacy models and disclosure risk measures

Disclosure risk assessment

Disclosure risk.

- Boolean vs. quantitative privacy models
  - Boolean: Disclosure either takes place or not. Check whether the definition holds or not. Includes definitions based on a threshold.
  - Quantitative: Disclosure is a matter of degree that can be quantified. Some risk is permitted.
- Implication when selecting a method
  - minimize information loss (max. utility) vs. multiobjetive optimization

Disclosure risk assessment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures/models

Disclosure risk assessment

Disclosure risk.

- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures/models

Classification of privacy models (and measures)

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

Disclosure risk assessment

Boolean definitions of risk.

- k-Anonymity (Boolean definition / identity disclosure)
- Secure multiparty computation (Boolean / identity and attribute disclosure)
- Result privacy (Boolean definition / attribute disclosure)
- Differential privacy (Boolean definition / attribute disclosure)

Quantitative measures of risk. alternative measures.

- Re-identification (for identity disclosure). Different ways to evaluate re-identification by means of record linkage.
- Uniqueness (for identity disclosure).
- Interval disclosure (for attribute disclosure). Several definitions for different types of attributes.

Disclosure risk assessment

Classification of privacy models (and measures)

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

Other privacy models

- Other models combining features: l-diversity, secure multiparty computation ensuring differential privacy
- Alternative but related models: k-confusion, k-concealment

Attribute disclosure (database protection)

Attribute disclosure

- **Algorithm** Rank-based interval disclosure: $rid(X, V, V', x, p)$
  - **Input:** X: Original file; V: Original attribute; V': Masked attribute; x: record; p: percentage
  - **Output:** Attribute disclosure for attribute V' of record x
  - $R(V) :=$ Rank data for attribute V'
  - $i :=$ position of $V'(x)$ in $R(V)$
  - $w := p \cdot |X|/2/100$ (width of the interval)
  - $I(x) = [R[\min(i - w, 0)], R[\max(i + w, |X| - 1)]]$ (definition of the interval for record x)
  - $rid := V(x) \in I(x)$
  - **Return** rid

Attribute disclosure

- **Algorithm** Standard deviation-based interval disclosure: $sdid(X, V, V', x, p)$
  - **Input** X: Original file; V: Original attribute; V': Masked attribute; x: record; p: percentage
  - **Output** Attribute disclosure for attribute V' of record x
  - $sd(V) :=$ standard deviation of V
  - $sdid := |V(x) - V'(x)| \leq p \cdot sd(V)/100$
  - **Return** $sdid$

Uniqueness (database protection)

Uniqueness

- **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.

Uniqueness

- **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

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:

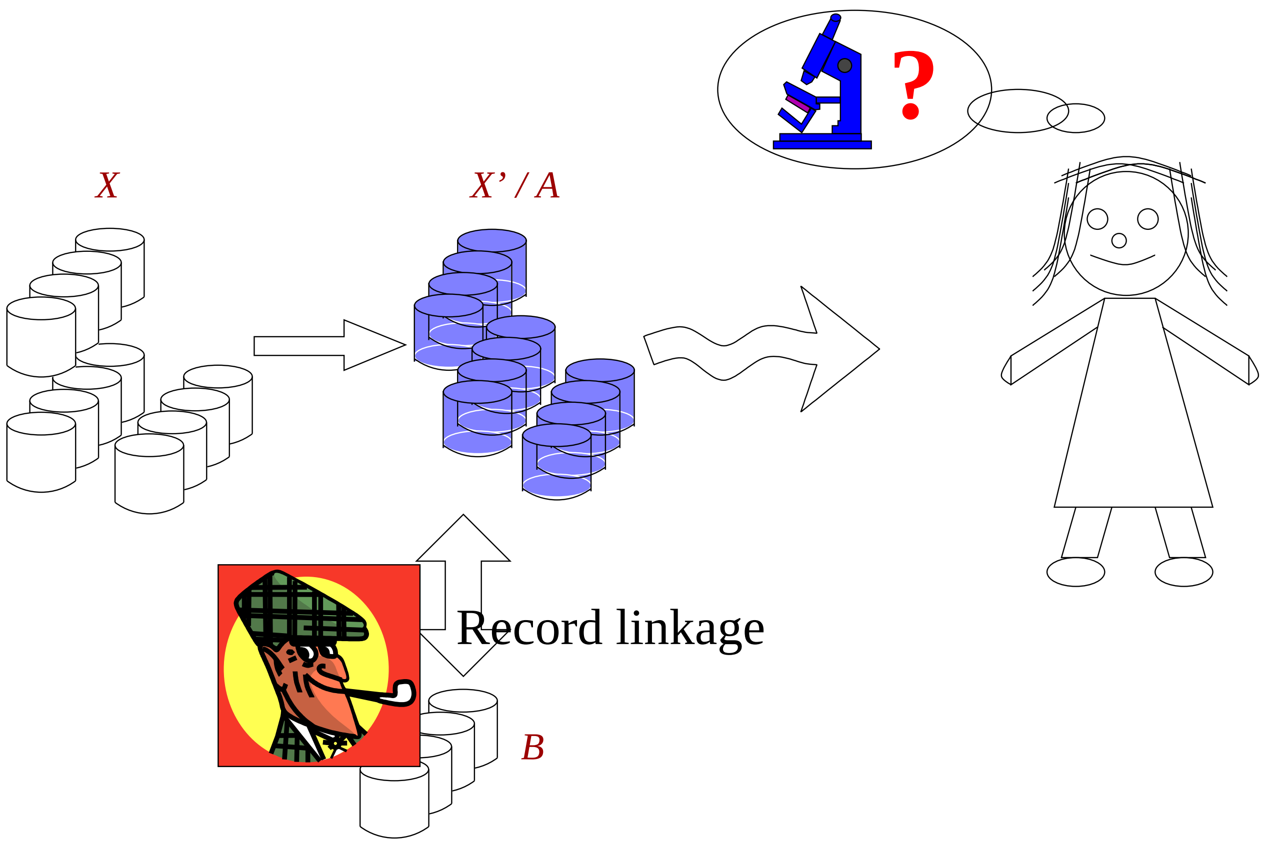
$$P(PU|SU) = \frac{P(PU, SU)}{P(SU)} = \frac{\sum_j I(F_j = 1, f_j = 1)}{\sum_j I(f_j = 1)}$$

  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).

Identity disclosure (database protection)

Identity disclosure

- **Privacy from re-identification**. Identity disclosure. Scenario:
  - A: File with the protected data set
  - B: File with the data from the intruder (subset of original X)



Identity disclosure

- **Privacy from re-identification**. Identity disclosure.
  - A: File with the protected data set
  - B: File with the data from the intruder (subset of original X)

How to establish the correct links between the two files?
Record linkage algorithms (used in e.g. database integration)

Identity disclosure

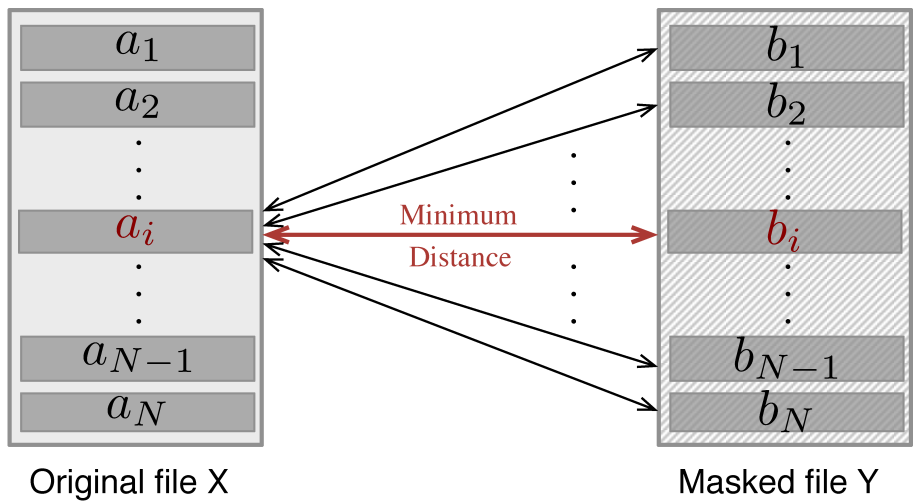
- **Privacy from re-identification**. Identity disclosure.
  - A: File with the protected data set
  - B: File with the data from the intruder (subset of original X)

  How to establish the correct links between the two files?
  Record linkage algorithms (used in e.g. database integration)

- Two main types.
  - Distance-based record linkage
  - Probabilistic record linkage

Identity disclosure

- Distance-based record linkage: $d(a, b)$ with $a \in A$ and $b \in B$.
  - Assign to the record at a minimum distance, ideally an intruder wants for a record i: $d(a_i, b_j) \geq d(a_i, b_i)$ for all j
    but due to masking we expect this does not happen



Identity disclosure

- **Algorithm** Distance-based record linkage
  - **Input** A: file; B: file
  - **Output** LP: linked pairs; NP: non-linked pairs
  - **For** $a \in A$
  - b' $=$ arg $\min_{b \in B} d(a, b)$
  - $LP = LP \cup (a, b')$
  - **for** $b \in B$ **such that** $b \neq b'$
  - $NP := NP \cup (a, b)$
  - **end for**
  - **end for**
  - **Return** (LP, NP)

Identity disclosure

- Probabilistic record linkage: $d(a, b)$ with $a \in A$ and $b \in B$.
  - Classification of pairs of records (a, b) in 3 classes
    Linked pair, non-linked, clerical pair
  - How?
    - ⋆ For each pair (a, b), an 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
    - ⋆ Classification using thresholds

Identity disclosure

- Probabilistic record linkage: $d(a, b)$ with $a \in A$ and $b \in B$.
  - Computation of
    $P(coincidence|Matching)$ and
    $P(coincidence|Unmatching)$:

Identity disclosure

- Probabilistic record linkage: $d(a, b)$ with $a \in A$ and $b \in B$.
  - Computation of
    $P(coincidence|Matching)$ and
    $P(coincidence|Unmatching)$:
    - Using EM algorithm

Identity disclosure

- Probabilistic record linkage: $d(a, b)$ with $a \in A$ and $b \in B$.
  - Computation of
    $P(coincidence|Matching)$ and
    $P(coincidence|Unmatching)$:
    - ⋆ Using EM algorithm
  - Computation of thresholds

Identity disclosure

- Probabilistic record linkage: $d(a, b)$ with $a \in A$ and $b \in B$.
  - Computation of
    $P(coincidence|Matching)$ and
    $P(coincidence|Unmatching)$:
    - ⋆ Using EM algorithm
  - Computation of thresholds
    - ⋆ From the probabilities of false positive/negative
      $P(Linkedpair|Unmatching)$
      $P(Nonlinkedpair|Matching)$

Identity disclosure

A scenario for identity disclosure. Reidentification

- Flexible scenario for identity disclosure
  - A protected file using a masking method
  - B (intruder's) is a subset of the original file.

Identity disclosure

A scenario for identity disclosure. Reidentification

- Flexible scenario for identity disclosure
  - A protected file using a masking method
  - B (intruder's) is a subset of the original file.
    $\rightarrow$ intruder with information on only some individuals

Identity disclosure

A scenario for identity disclosure. Reidentification

- Flexible scenario for identity disclosure
  - A protected file using a masking method
  - B (intruder's) is a subset of the original file.
    $\rightarrow$ intruder with information on only some individuals
    $\rightarrow$ intruder with information on only some characteristics

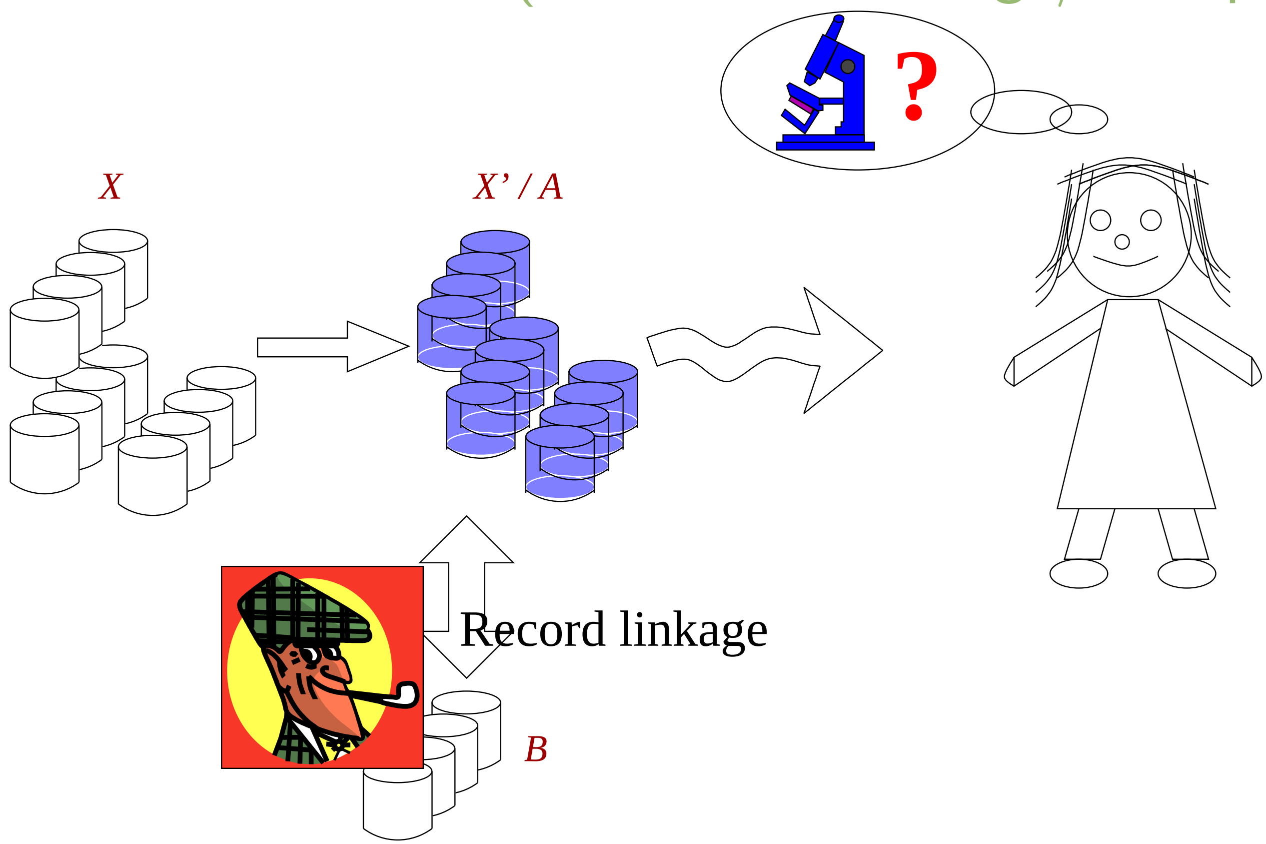
Identity disclosure

A scenario for identity disclosure. Reidentification

- Flexible scenario for identity disclosure
  - A protected file using a masking method
  - B (intruder's) is a subset of the original file.
    $\rightarrow$ intruder with information on only some individuals
    $\rightarrow$ intruder with information on only some characteristics
  - But also,
    - $\star$ B with a schema different to the one of A (different attributes)
    - $\star$ Other scenarios. E.g., synthetic data
    - $\star$ Other type of data: graph data
      (reidentifying people in a social network)

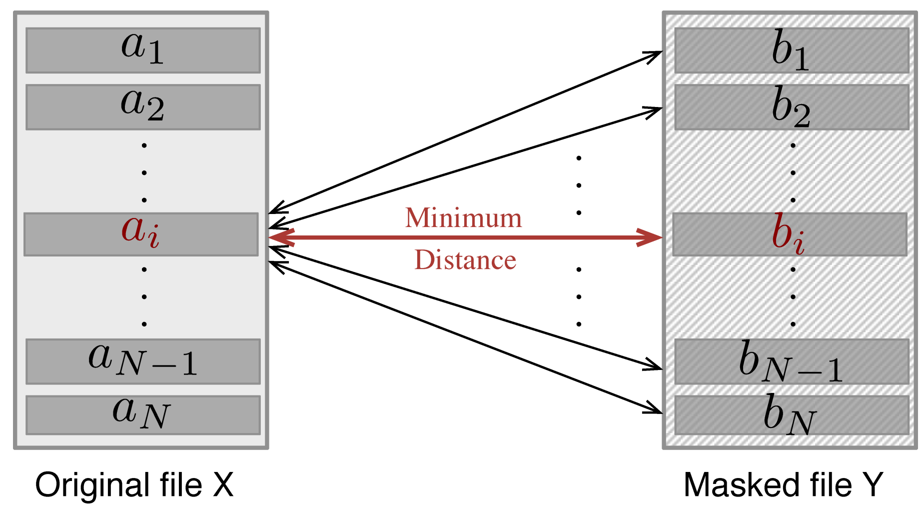
Identity disclosure

- **Privacy from re-identification**. Worst-case scenario (maximum knowledge) to give upper bounds of risk:
  - transparency attacks (information on how data has been protected)
  - largest data set (original data)
  - best re-identification method (best record linkage/best parameters)



Identity disclosure

- **Privacy from re-identification**. Worst-case scenario.
  - ML for distance-based record linkage parameters. (A and B aligned)
    - ⋆ Goal: as many correct reidentifications as possible: for each record i: $d(a_i, b_j) \geq d(a_i, b_i)$ for all j



- $d(a_i, b_j)$ as average/sum of attribute/variable distances

$$\mathbb{C}_p(\mathit{diff}_1(a_i, b_j), \ldots, \mathit{diff}_n(a_i, b_j))$$

Identity disclosure

- **Privacy from re-identification**. Worst-case scenario.
  - ML for distance-based record linkage parameters. (A and B aligned)
    - ⋆ Goal: as many correct reidentifications as possible. But,
      if error for a_i: $K_i = 1$ and $d(a_i, b_j) + CK_i \geq d(a_i, b_i)$ for all j
      where d is an aggregated distance $d(a, b) = \mathbb{C}_p(\mathit{diff}_1, \ldots, \mathit{diff}_n)$:
    - ⋆ Formally,

$$\mathbb{C}_p(\mathit{diff}_1(a_i, b_j), \ldots, \mathit{diff}_n(a_i, b_j)) + CK_i \geq \mathbb{C}_p(\mathit{diff}_1(a_i, b_i), \ldots, \mathit{diff}_n(a_i, b_i))$$

Identity disclosure

- **Privacy from re-identification**. Worst-case scenario.
  - ML for distance-based record linkage parameters. (A and B aligned)
  - Goal: as many correct reidentifications as possible.
  - Minimize K_i: minimize the number of records a_i that fail
- Formalization:

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

$$Subject\ to:$$

$$\mathbb{C}_p(\mathit{diff}_1(a_i, b_j), \ldots, \mathit{diff}_n(a_i, b_j)) -$$
$$- \mathbb{C}_p(\mathit{diff}_1(a_i, b_i), \ldots, \mathit{diff}_n(a_i, b_i)) + CK_i > 0$$
$$K_i \in \{0, 1\}$$

Additional constraints according to $\mathbb{C}$

Identity disclosure

- **Privacy from re-identification**. Worst-case scenario.
  - ML for distance-based record linkage parameters. (A and B aligned)
  - The case of the weighted mean ($\mathbb{C} = WM$)/Weighted Euclidean
  - Formalization:

$$d^2(a,b) = WM_p(\mathit{diff}_1(a,b), \ldots, \mathit{diff}_n(a,b))$$

with arbitrary vector $p = (p_1, \ldots, p_n)$ and
$\mathit{diff}_i(a,b) = ((a_i - \bar{a}_i)/\sigma(a_i) - (b_i - \bar{b}_i)/\sigma(b_i))^2$

Identity disclosure

- **Privacy from re-identification**. Worst-case scenario.
  - ML for distance-based record linkage parameters. (A and B aligned)
  - The case of the weighted mean ($\mathbb{C} = WM$)
  - Formalization:

$$
\begin{aligned}
Minimize & \sum_{i=1}^{N} K_i \\
Subject\ to: & WM_p(\mathit{diff}_1(a_i, b_j), \dots, \mathit{diff}_n(a_i, b_j)) - \\
& \quad - WM_p(\mathit{diff}_1(a_i, b_i), \dots, \mathit{diff}_n(a_i, b_i)) + C\ K_i > 0 \\
& K_i \in \{0, 1\} \\
& \sum_{i=1}^{n} p_i = 1 \\
& p_i \geq 0
\end{aligned}
$$

Identity disclosure

- **Privacy from re-identification**. Worst-case scenario.
  - ML for DBRL parameters: Distances considered $\mathbb{C}$
    - Weighted mean.
      Weights: importance to the attributes
      Parameter: weighting vector $n = \#$ attributes

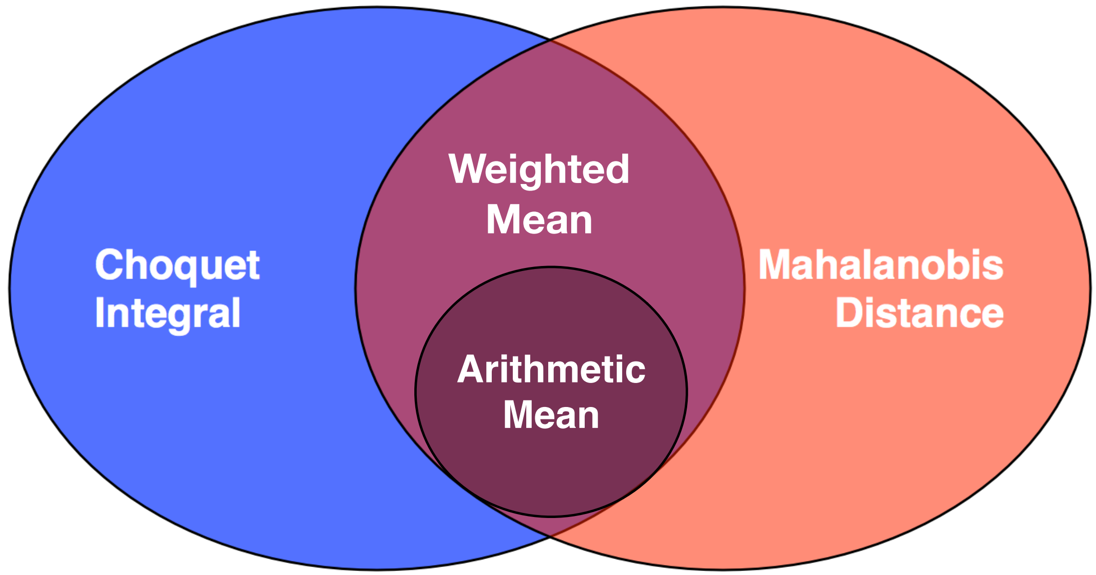
Identity disclosure

- **Privacy from re-identification**. Worst-case scenario.
  - ML for DBRL parameters: Distances considered $\mathbb{C}$
    - ⋆ Weighted mean.
      Weights: importance to the attributes
      Parameter: weighting vector $n =\#$ attributes
    - ⋆ OWA - linear combination of order statistics (weighted):
      Weights: to discard lower or larger distances
      Parameter: weighting vector $n =\#$ attributes
    - ⋆ Bilinear form - generalization of Mahalanobis distance
      Weights: interactions between pairs of attributes
      Parameter: square matrix: $n \times n$ ($n =\#$ attributes)
    - ⋆ Choquet integral.
      Weights: interactions of sets of attributes ($\mu : 2^X \to [0, 1]$)
      Parameter: non-additive measure: $2^n - 2$ ($n =\#$ attributes)

Identity disclosure

Distances used in record linkage based on aggregation operators

- Graphically



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

k-Anonymity
(a privacy model)

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 $QI_{RT} = \{$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 algorithms 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 algorithms 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 algorithms 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.