INFUS 2020

Fuzzy meets privacy: a short overview

Vicenç Torra

July, 2020

University of Umeå (Sweden)

A kind of justification

- Research on data privacy since year 2000
- We have applied fuzzy sets theory in some research problems
  - Where fuzzy sets theory can be used?

A kind of justification

- Data privacy can be seen from different perspectives (social, legal, etc)
  - Technological perspective
  - Data to be used for machine and statistical learning (data analytics)

A kind of justification

- Data privacy can be seen from different perspectives (social, legal, etc)
  - Technological perspective
  - Data to be used for machine and statistical learning (data analytics)

- In this framework, fuzzy set theory as
  - one of the tools for data analytics, but also
  - one of the tools related to data protection

Outline

1. A data privacy context
   (a) A data privacy problem? Why?
   (b) Privacy models
   (c) Masking methods
2. Fuzzy sets in data masking
3. Summary

Introduction

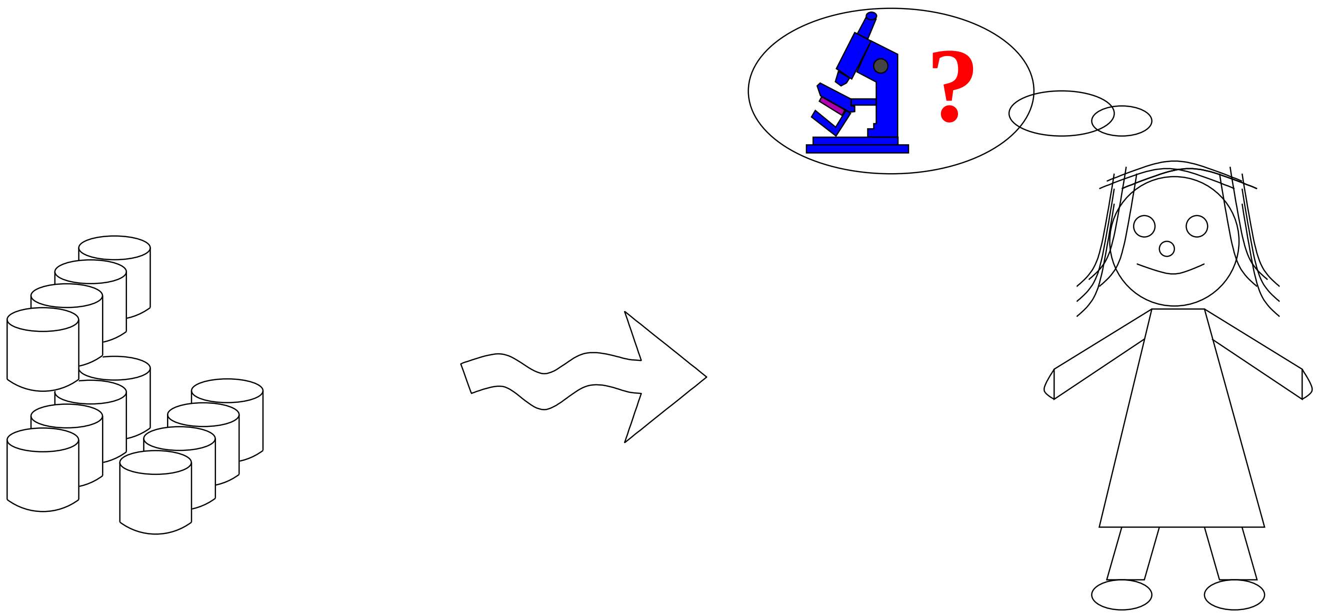
A data privacy context:

A data privacy problem? Why?

(examples of disclosure)

A data privacy problem? Why?

Data privacy in context. A researcher wants to analyze data



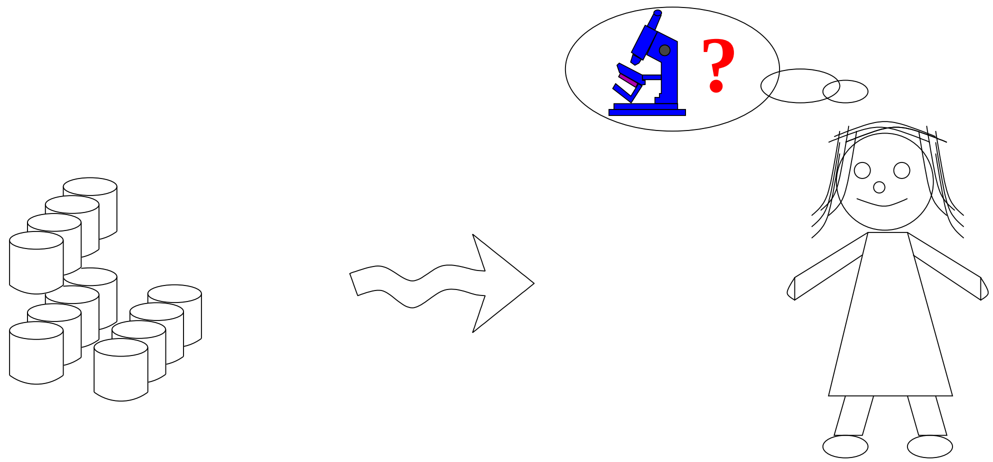
DB = {(Aylin, Age = 40, Street=Maçka caddesi İstanbul, salary=147000 TRY/TL), ...}

A data privacy problem? Why?

Data privacy in context. A researcher wants to analyze data

- Two main scenarios in which disclosure can take place

1. Disclosure from the data themself
2. Disclosure from the computation, query, data analysis

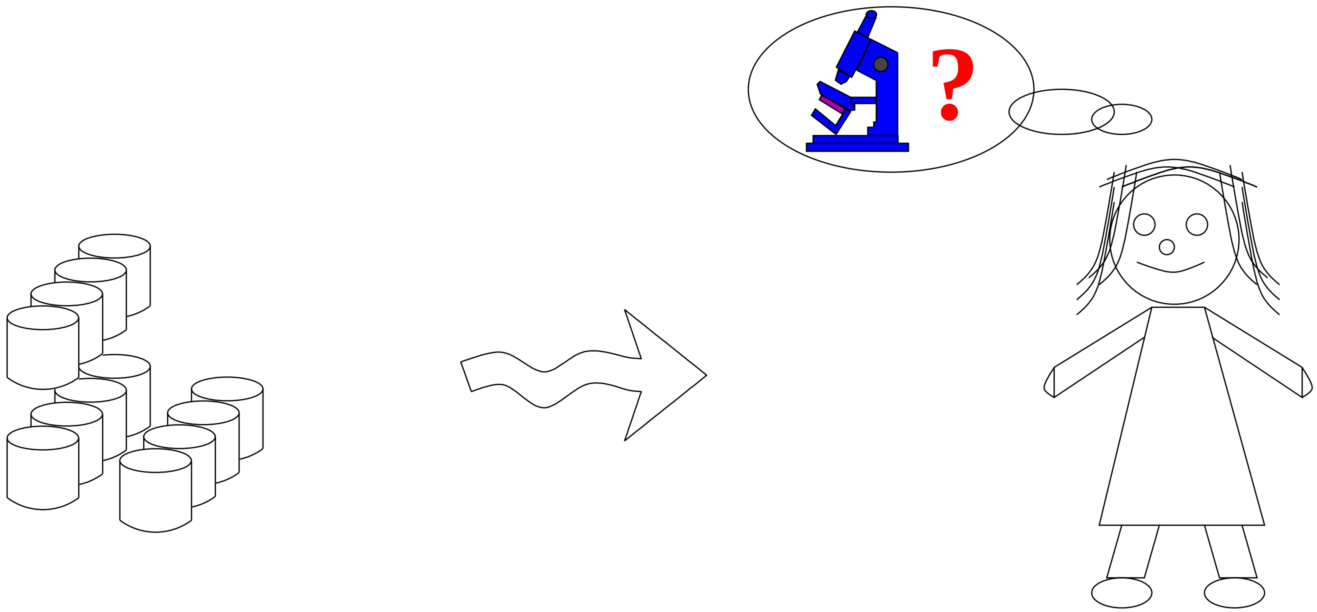


A data privacy problem? Why?

- Disclosure from the data themself
- Example of trivial disclosure: we learn Aylin salary:

A data privacy problem? Why?

- Disclosure from the data themself
- Example of trivial disclosure: we learn Aylin salary:



DB = {(Aylin, Age = 40, Street=Maçka caddesi İstanbul, salary=147000 TRY/TL), ...}

A data privacy problem? Why?

- Disclosure from data (non-trivial): learning Burcu's sickness

A data privacy problem? Why?

- Disclosure from data (non-trivial): learning Burcu's sickness
  - Q: sickness influenced by studies & commuting distance?
  - University *protects* data and supplies only:

    (where students live, what they study, if they got sick)

A data privacy problem? Why?

- Disclosure from data (non-trivial): learning Burcu's sickness
  - Q: sickness influenced by studies & commuting distance?
  - University *protects* data and supplies only:
    (where students live, what they study, if they got sick)
  - No "personal data",
    $DB = \{$(İstanbul, CS, No), (İstanbul, CS, No),
    (İstanbul, CS, Yes), (Konak (İzmir), CS, No), . . . ,
    (İstanbul, BA MEDIA STUDIES, No)
    (İstanbul, BA MEDIA STUDIES, Yes), . . . $\}$

A data privacy problem? Why?

- Disclosure from data (non-trivial): learning Burcu's sickness
  - Q: sickness influenced by studies & commuting distance?
  - University *protects* data and supplies only:
    (where students live, what they study, if they got sick)
  - No "personal data",
    $DB = \{$(İstanbul, CS, No), (İstanbul, CS, No),
    (İstanbul, CS, Yes), (Konak (İzmir), CS, No), . . . ,
    (İstanbul, BA MEDIA STUDIES, No)
    (İstanbul, BA MEDIA STUDIES, Yes), . . . $\}$
    This is NOT ok!!:

A data privacy problem? Why?

- Disclosure from data (non-trivial): learning Burcu's sickness
  - Q: sickness influenced by studies & commuting distance?
  - University *protects* data and supplies only:
    (where students live, what they study, if they got sick)
  - No "personal data",
    $DB = \{$(İstanbul, CS, No), (İstanbul, CS, No),
    (İstanbul, CS, Yes), (Konak (İzmir), CS, No), . . . ,
    (İstanbul, BA MEDIA STUDIES, No)
    (İstanbul, BA MEDIA STUDIES, Yes), . . . $\}$
    This is NOT ok!!:
  - E.g., only one student (Burcu) on anthropology in Alaçatı:
    (Alaçatı(İzmir), Anthropology, Yes)

A data privacy problem? Why?

- Disclosure from data (non-trivial): learning Burcu's sickness
  - Q: sickness influenced by studies & commuting distance?
  - University *protects* data and supplies only:
    (where students live, what they study, if they got sick)
  - No "personal data",
    $DB = \{$(İstanbul, CS, No), (İstanbul, CS, No),
    (İstanbul, CS, Yes), (Konak (İzmir), CS, No), . . . ,
    (İstanbul, BA MEDIA STUDIES, No)
    (İstanbul, BA MEDIA STUDIES, Yes), . . . $\}$
    This is NOT ok!!:
  - E.g., only one student (Burcu) on anthropology in Alaçatı:
    (Alaçatı(İzmir), Anthropology, Yes)

$\Rightarrow$ 1. We learn that our friend is in the database

A data privacy problem? Why?

- Disclosure from data (non-trivial): learning Burcu's sickness
  - Q: sickness influenced by studies & commuting distance?
  - University *protects* data and supplies only:

    (where students live, what they study, if they got sick)
  - No "personal data",

    $DB = \{$(İstanbul, CS, No), (İstanbul, CS, No),
    (İstanbul, CS, Yes), (Konak (İzmir), CS, No), . . . ,
    (İstanbul, BA MEDIA STUDIES, No)
    (İstanbul, BA MEDIA STUDIES, Yes), . . . $\}$

    This is NOT ok!!:
  - E.g., only one student (Burcu) on anthropology in Alaçatı:

    (Alaçatı(İzmir), Anthropology, Yes)

$\Rightarrow$ 1. We learn that our friend is in the database

$\Rightarrow$ 2. We learn that our friend is sick !!

A data privacy problem? Why?

- Disclosure from the computation (data analysis)
- Example of trivial disclosure: Ms. Rich attends pyshiatric unit.

A data privacy problem? Why?

- Disclosure from the computation (data analysis)
- Example of trivial disclosure: Ms. Rich attends pyshiatric unit.
  - Q: Mean income of admitted to psychiatric unit given Town=Alaçatı?

A data privacy problem? Why?

- Disclosure from the computation (data analysis)

- Example of trivial disclosure: Ms. Rich attends pyshiatric unit.
  - Q: Mean income of admitted to psychiatric unit given Town=Alaçatı?
  - Monthly salaries when Town=Alaçatı:

$$800\ 1000\ 700\ 900\ 1000\ 800\ 600\ 800\ 1200\ 1400 \Rightarrow \text{mean} = 920$$

A data privacy problem? Why?

- Disclosure from the computation (data analysis)
- Example of trivial disclosure: Ms. Rich attends pyshiatric unit.
  - Q: Mean income of admitted to psychiatric unit given Town=Alaçatı?
  - Monthly salaries when Town=Alaçatı:

    800 1000 700 900 1000 800 600 800 1200 1400 $\Rightarrow$ mean $= 920$
  - Mean seems fine, no "personal data" (aggregate), is this ok ?

A data privacy problem? Why?

- Disclosure from the computation (data analysis)
- Example of trivial disclosure: Ms. Rich attends pyshiatric unit.
  - Q: Mean income of admitted to psychiatric unit given Town=Alaçatı?
  - Monthly salaries when Town=Alaçatı:

    $$800\ 1000\ 700\ 900\ 1000\ 800\ 600\ 800\ 1200\ 1400 \Rightarrow \text{mean} = 920$$
  - Mean seems fine, no "personal data" (aggregate), is this ok ? NO!!:

A data privacy problem? Why?

- Disclosure from the computation (data analysis)
- Example of trivial disclosure: Ms. Rich attends pyshiatric unit.
  - Q: Mean income of admitted to psychiatric unit given Town=Alaçatı?
  - Monthly salaries when Town=Alaçatı:

    $$800\ 1000\ 700\ 900\ 1000\ 800\ 600\ 800\ 1200\ 1400 \Rightarrow \text{mean} = 920$$
  - Mean seems fine, no "personal data" (aggregate), is this ok ? NO!!:
  - Adding Ms. Rich's salary 70,000:

    $$800\ 1000\ 700\ 900\ 1000\ 800\ 600\ 800\ 1200\ 1400\ 70000$$
    $$\Rightarrow \text{mean} = 7200{,}00\ !!$$

  (a extremely high salary changes the mean significantly)

A data privacy problem? Why?

- Disclosure from the computation (data analysis)
- Example of trivial disclosure: Ms. Rich attends pyshiatric unit.
  - Q: Mean income of admitted to psychiatric unit given Town=Alaçatı?
  - Monthly salaries when Town=Alaçatı:

$$800\ 1000\ 700\ 900\ 1000\ 800\ 600\ 800\ 1200\ 1400 \Rightarrow \text{mean} = 920$$

  - Mean seems fine, no "personal data" (aggregate), is this ok ? NO!!:
  - Adding Ms. Rich's salary 70,000:

$$800\ 1000\ 700\ 900\ 1000\ 800\ 600\ 800\ 1200\ 1400\ 70000$$
$$\Rightarrow \text{mean} = 7200{,}00 \text{ !!}$$

  (a extremely high salary changes the mean significantly)

  $\Rightarrow$ We infer Ms. Rich from Town was attending the unit

Introduction

A data privacy context:
Privacy models
(how to solve this?: provide a definition)

Computational definitions of privacy?

- How to solve this?
  - Provide a definition !!

Computational definitions of privacy?

- How to solve this?
  - Provide a definition !!
  - Well, not one, there are lots of them:
  - Privacy models:
    computational definitions of privacy

Computational definitions of privacy?

- Privacy models: computational definitions of privacy
  - Definitions? Why many?
  - Different focuses. E.g.,
    - ⋆ Disclosure from data
    - ⋆ Disclosure from computation, query, data analysis

Computational definitions of privacy?

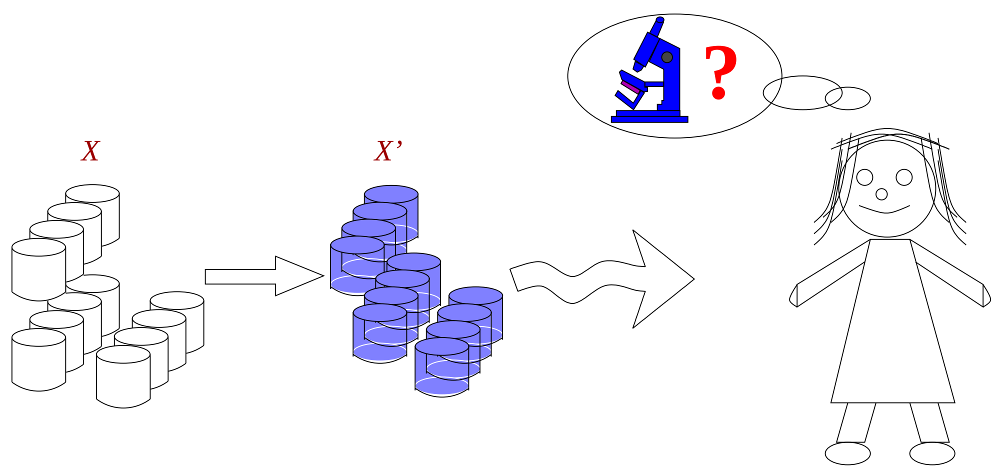
Privacy models. A computational definition for privacy. Examples.

- **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.

Computational definitions of privacy?

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.



Computational definitions of privacy?

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	İstanbul	30	Heart attack
ABD	İstanbul	32	Cancer
COL	İstanbul	33	Cancer
GHE	Konak (İzmir)	62	AIDS
CIO	Alaçatı(İzmir)	65	AIDS
HYU	Konak (İzmir)	60	Heart attack

Published DB X':

——	City	Age	Illness
—	İstanbul	30	Cancer
—	İstanbul	30	Cancer
—	İstanbul	30	Cancer
—	İzmir	60	AIDS
—	İzmir	60	AIDS
—	İzmir	–	——

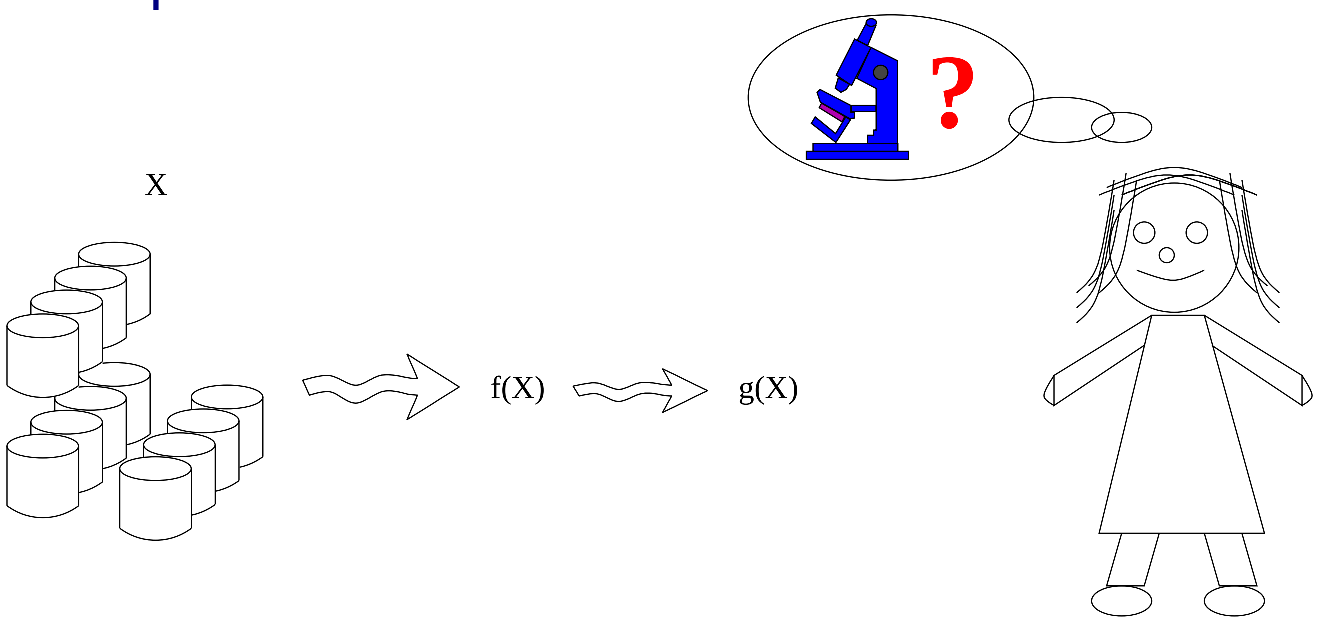
Computational definitions of privacy?

- Difficulties
  - Naive anonymization does not work,
    highly identifiable data, high dimensional data
  - Anonyimization causes information loss
- Examples of successful reidentification attacks
  - Sweeney analysis of USA population,
  - data from mobile data (home + work reidentifies a person),
  - shopping cards
    (high dimensional, large number of shopping elements),
  - film ratings (high dimensional, large number of film)

Computational definitions of privacy?

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.



Computational definitions of privacy?

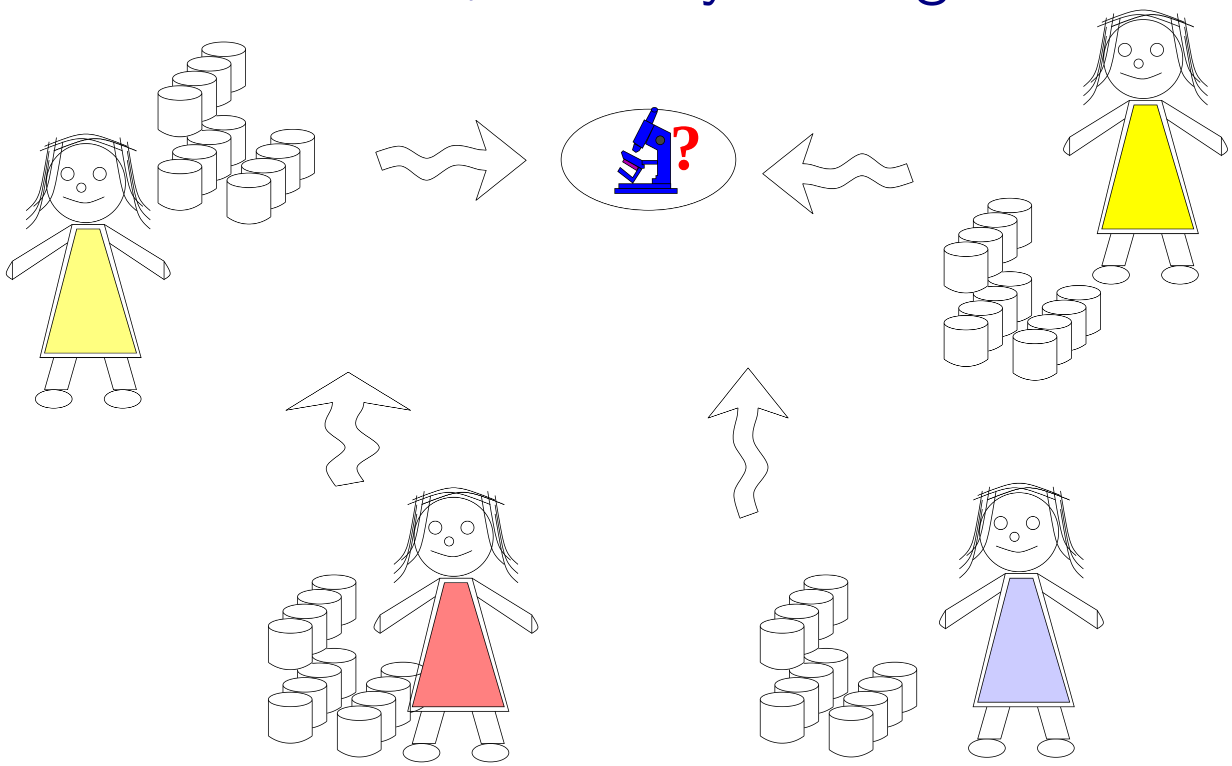
- Difficulties.
  - A simple function can give information on who is in the database
  - Modifying the function may lead to high information loss
  - Function-dependent solution

  - E.g., mean salary,
  - if *mean* outcome is not affected by a single person, is it useful?

Computational definitions of privacy?

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.



Computational definitions of privacy?

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!)

Computational definitions of privacy?

- Difficulties
  - Distributed approach (no trusted-third party) – computational cost of solutions
  - Function-dependent solution

Introduction

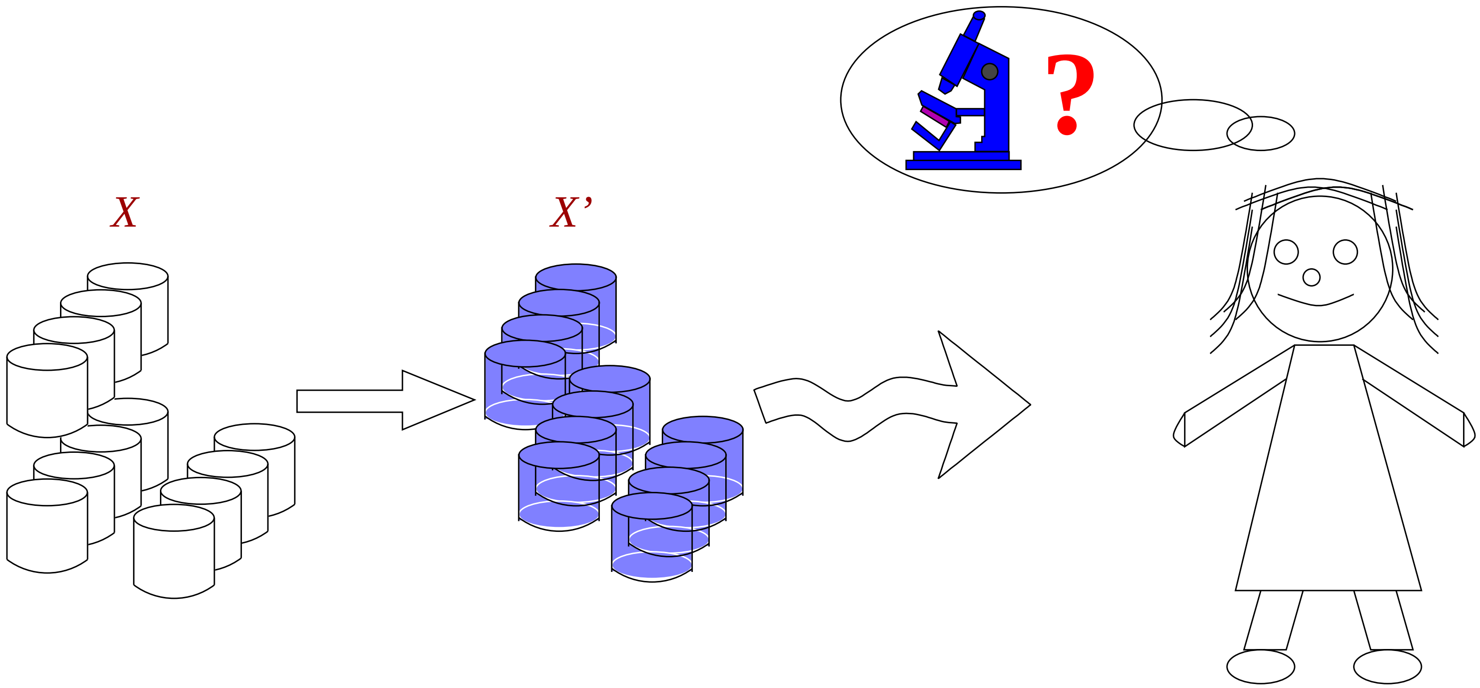
A data privacy context:

Masking methods

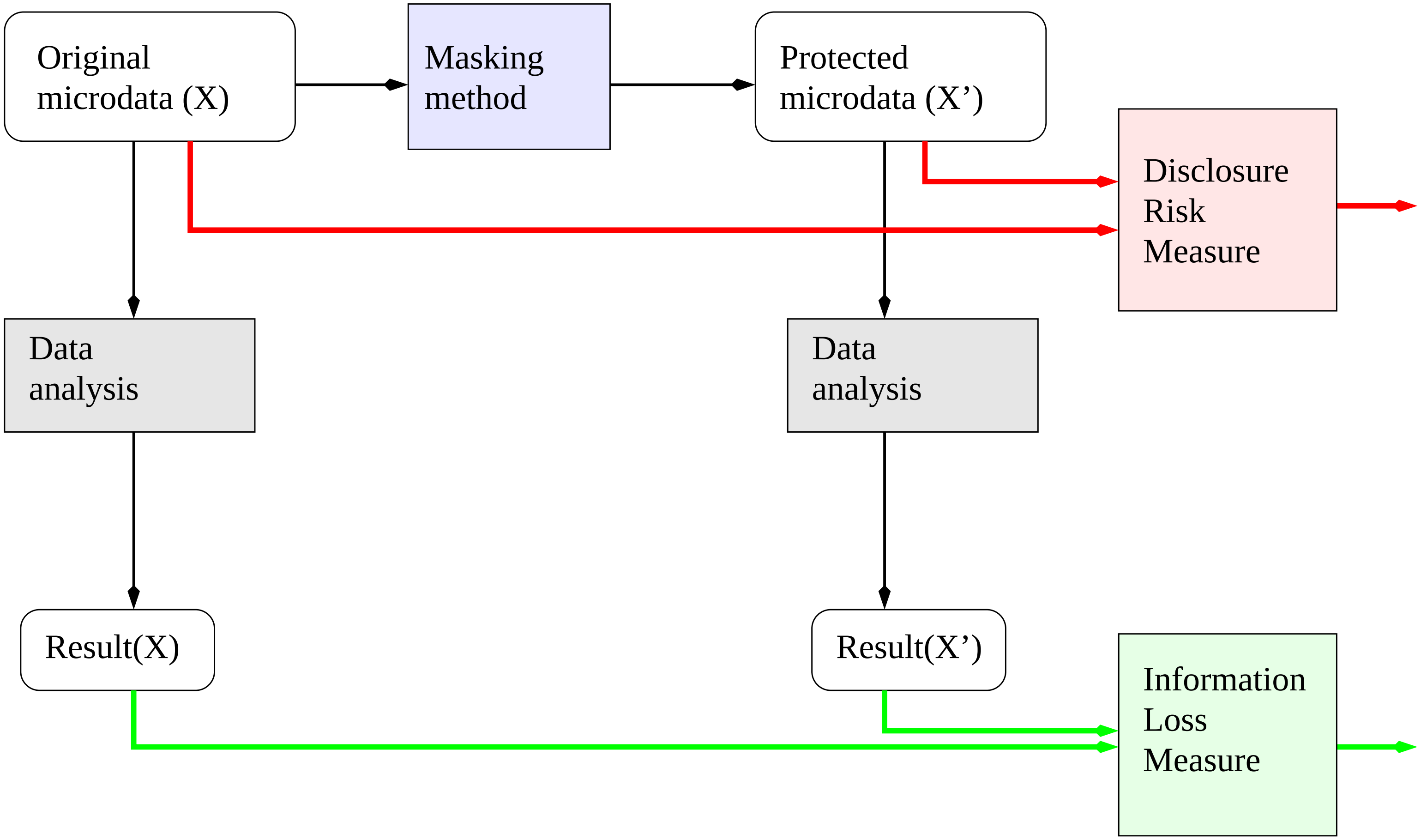
(to protect a database against reidentification)

Masking methods

Anonymization/masking method: Given a data file X compute a file X' with data of *less quality*.



Research questions



Masking: Less quality (information loss) less risk (disclosure risk)

$X' = \rho(X)$: $IL_f(X, X') = divergence(f(X), f(X'))$,

$\text{DR}_X(X')$=recordLinkage(X, X')

Introduction

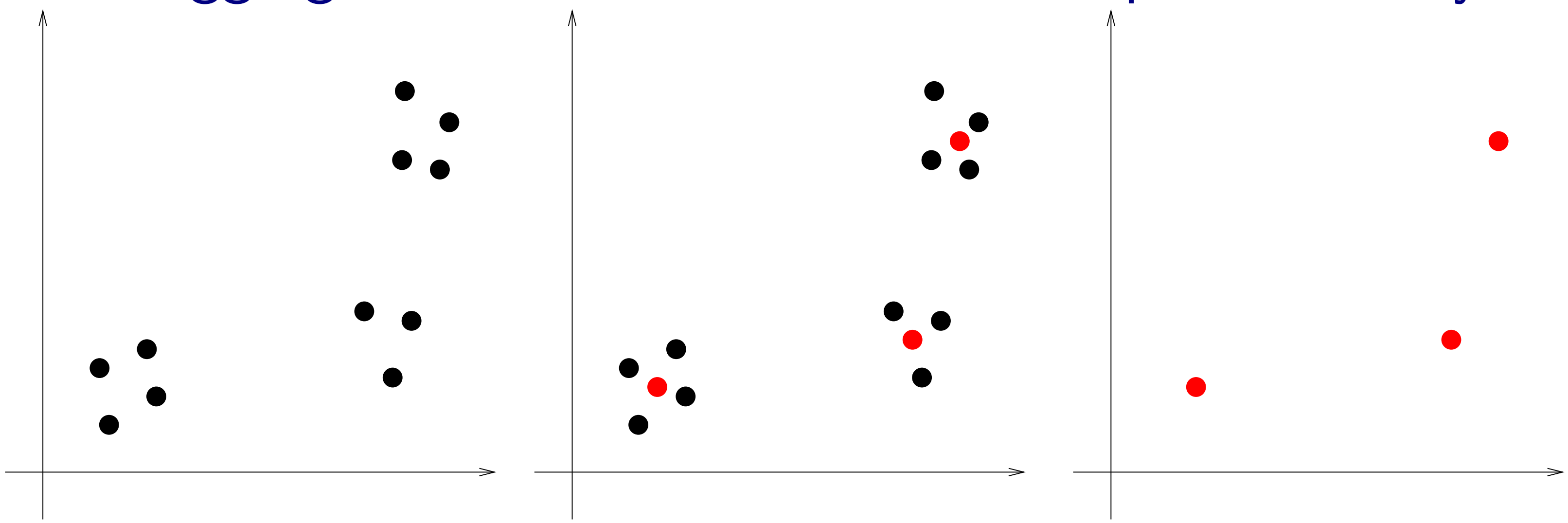
Fuzzy sets in data masking

Introduction

Fuzzy sets in data masking
Fuzzy sets based microaggregation
(clustering-based masking method)

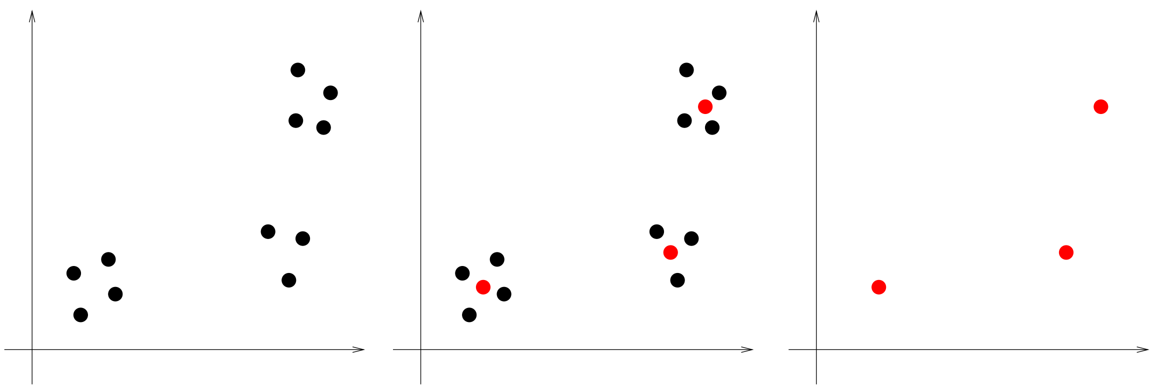
Fuzzy sets based microaggregation

- Microaggregation: small clusters; then, replace data by cluster centers



Fuzzy sets based microaggregation

- Microaggregation:
  - Privacy: each cluster at least k records
  - Utility: small clusters to have low information loss

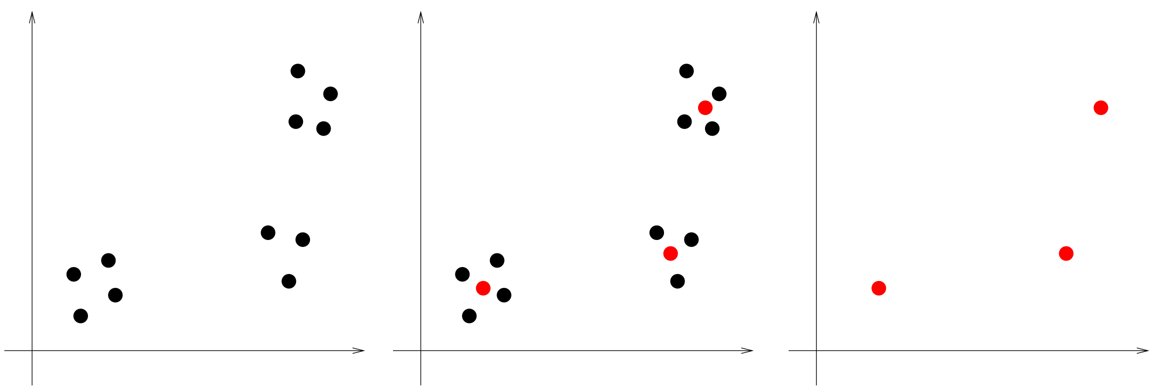


Fuzzy sets based microaggregation

- Microaggregation: Implementation
  - Build clusters
  - Define cluster representatives
  - Replace records by cluster representatives

Fuzzy sets based microaggregation

- Microaggregation:
  - Privacy: each cluster at least k records
  - Utility: small clusters to have low information loss
  - If $k = 1$, one cluster = one record. No loss, maximum risk
  - If $k = |X|$, only one cluster $= X$. Maximum loss, no risk



Fuzzy sets based microaggregation

- Microaggregation: Formalization in terms of error minimization

$$
\begin{aligned}
\text{Minimize} \quad & SSE = \sum_{i=1}^{c} \sum_{x \in X} \chi_i(x)(d(x, p_i))^2 \\
\text{Subject to} \quad & \sum_{i=1}^{c} \chi_i(x) = 1 \text{ for all } x \in X \\
& 2k \geq \sum_{x \in X} \chi_i(x) \geq k \text{ for all } i = 1, \ldots, c \\
& \chi_i(x) \in \{0, 1\}
\end{aligned} \tag{1}
$$

- Similar to c-means but with constraints on number of records in clusters

Introduction

Fuzzy sets in data masking
Why fuzzy sets based microaggregation?
(the transparency principle)

Transparency

- The transparency principle in data privacy[1]

  Given a privacy model, a masking method should be compliant with this privacy model even if everything about the method is public knowledge. (Torra, 2017, p17)

[1] Similar to the Kerckhoffs's principle (Kerckhoffs, 1883) in cryptography: a cryptosystem should be secure even if everything about the system is public knowledge, except the key

Transparency

- The transparency principle in data privacy[1]
  Given a privacy model, a masking method should be compliant with this privacy model even if everything about the method is public knowledge. (Torra, 2017, p17)

- Transparency a requirement of Trustworthy AI. Related to three elements: traceability, explicability (why decisions are made), and comunication (distinguish AI systems from humans). Transparency in data privacy relates to traceability.

[1] Similar to the Kerckhoffs's principle (Kerckhoffs, 1883) in cryptography: a cryptosystem should be secure even if everything about the system is public knowledge, except the key

Fuzzy sets based microaggregation

- Transparency
  - DB is published: give details on how data has been produced. Description of any data protection process and parameters
  - Positive effect on data utility. Use information in data analysis.
  - Negative effect on risk. Intruders use the information to attack.

Fuzzy sets based microaggregation

- Transparency
  - DB is published: give details on how data has been produced. Description of any data protection process and parameters
  - Positive effect on data utility. Use information in data analysis.
  - Negative effect on risk. Intruders use the information to attack.

In microaggregation.

- An intruder can infer in which cluster is a record
- If different variables are microaggregated independently, intersection attacks can lead to reidentification

Fuzzy sets based microaggregation

- Transparency.
  - An intruder can infer in which cluster is a record
  - If different variables are microaggregated independently, intersection attacks can lead to reidentification

Fuzzy sets based microaggregation

- Transparency.
  - An intruder can infer in which cluster is a record
  - If different variables are microaggregated independently, intersection attacks can lead to reidentification

- Fuzzy clustering can *fuzzify* membership to clusters

- A fuzzy approach can reduce disclosure risk

Introduction

Fuzzy sets in data masking
Fuzzy microaggregation: definition
(using fuzzy clustering)

Fuzzy sets based microaggregation

- Consider fuzzy c-means, the usual algorithm for fuzzy clustering
  $\Rightarrow$ to achieve fuzzy assignment of elements to clusters

Fuzzy sets based microaggregation

- Consider fuzzy c-means, the usual algorithm for fuzzy clustering
  $\Rightarrow$ to achieve fuzzy assignment of elements to clusters

Step 1: Generate an initial U and V

Step 2: Solve $min_{U \in M} J(U,V)$ computing:

$$u_{ij} = \left(\sum_{r=1}^{c} \left(\frac{||x_j - v_i||^2}{||x_j - v_r||^2} \right)^{\frac{1}{m-1}} \right)^{-1}$$

Step 3: Solve $min_V J(U,V)$ computing:

$$v_i = \frac{\sum_{j=1} n (u_{ij})^m x_j}{\sum_{j=1}^{n} (u_{ij})^m}$$

Step 4: If the solution does not converge, go to step 2; otherwise, stop

Fuzzy sets based microaggregation

- Implement fuzzy microaggregation with parameters c, m_1, and m_2 as:

  Step 1: Apply FCM with given c and a given $m := m_1$

  Step 2: For each x_j in X, compute memberships to all clusters $i = 1, \ldots, c$ for a given m_2:

$$u_{ij} = \left(\sum_{r=1}^{c} \left(\frac{||x_j - v_i||^2}{||x_j - v_r||^2} \right)^{\frac{1}{m_2 - 1}} \right)^{-1}$$

  Step 3: For each x_j determine a random value $\chi \in [0, 1]$ using a uniform distribution in [0,1], and
  assign x_j to cluster according probability distr. $u_{1j}, \ldots, u_{cj}$
  Formally, given χ select the ith cluster satisfying $\sum_{k<i} u_{kj} < \chi < \sum_{k \leq i} u_{kj}$

Fuzzy sets based microaggregation

- Properties:

1. The larger the m_1, the larger IL (information loss)
   Clusters collide, all protected data collapses to $v_i = v_j = \bar{X}$.
2. The larger the m_2, the larger IL.
   All memberships tend to $u_{ij} = 1/c$.
   Any record can be replaced by any cluster center.
   All clusters, same size. If $c = |X|/k+$, (probabilistically) k-anonymity
3. The smaller the number of clusters c, the larger IL
   Minimum IL with $c = |X|$, Maximum IL with $c = 1$.

Introduction

Fuzzy sets in data masking

Other uses of fuzzy set theory

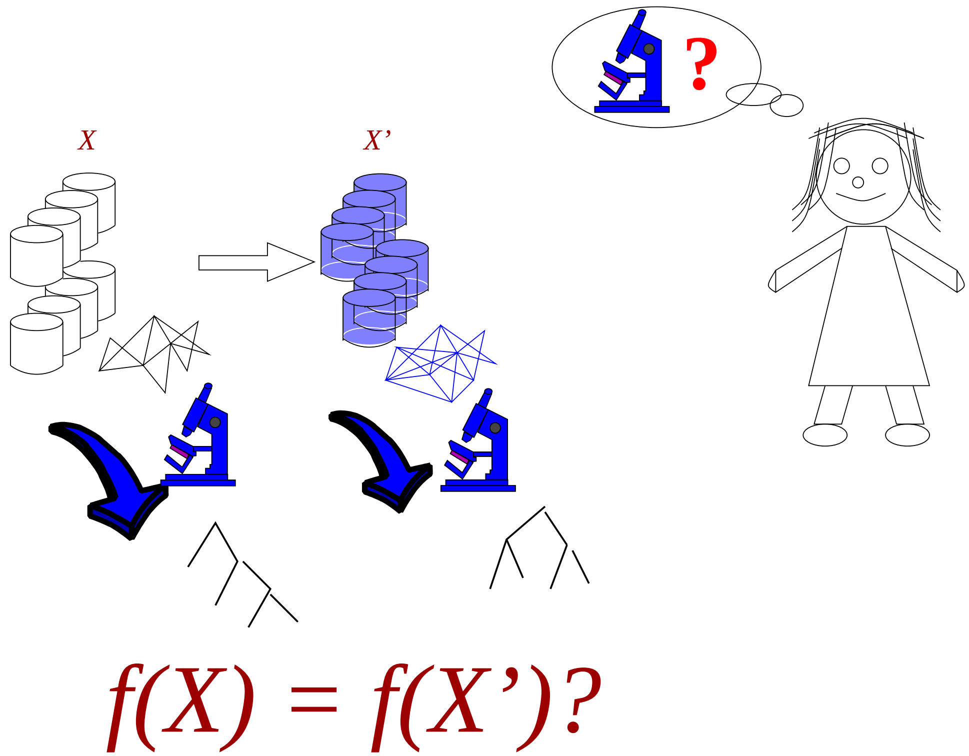
(in IL and DR)

(in information loss and disclosure risk)

Fuzzy in IL + DR

- Fuzziness in Information loss.
  - Compare X and X' w.r.t. analysis (f)

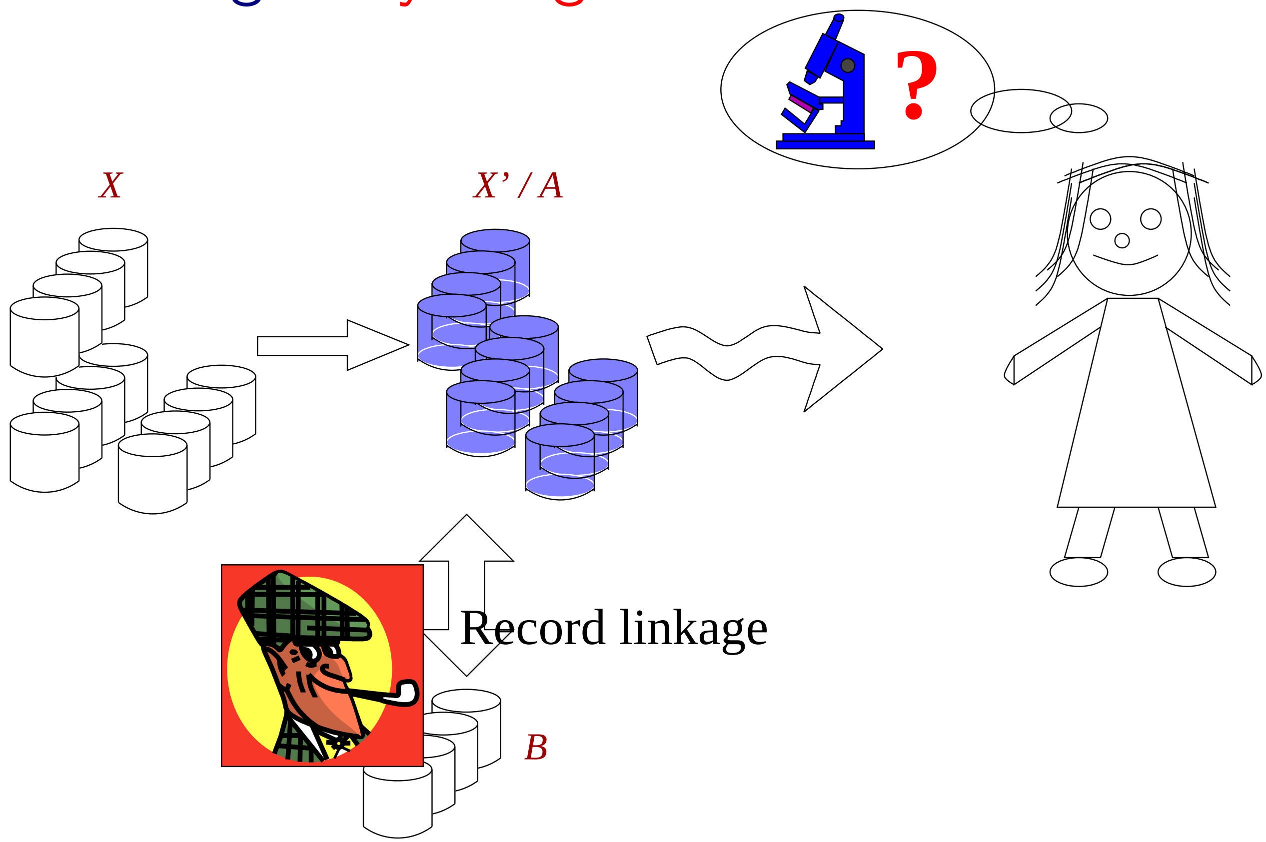
$$IL_f(X, X') = divergence(f(X), f(X'))$$



  - f is fuzzy clustering. Extensive work with S. Miyamoto and Y. Endo.
  - Difficulty: How to compare fuzzy clusters? (fuzzy clust. suboptimal)

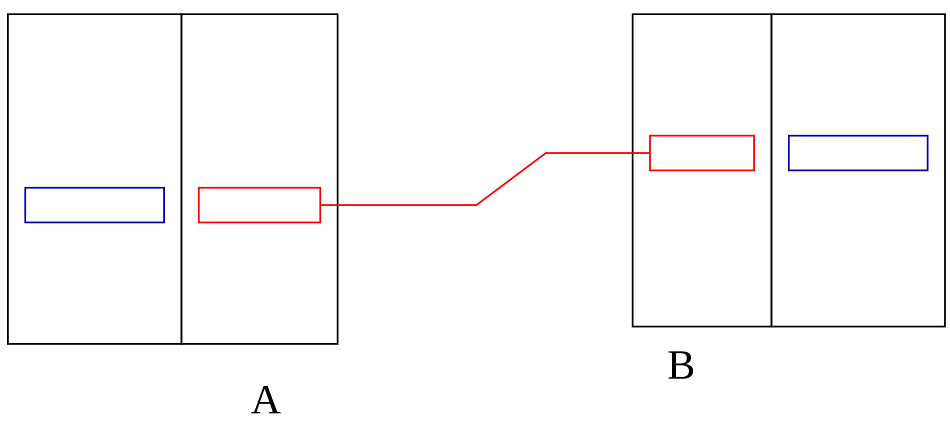
Fuzzy in IL + DR

- Fuzziness in disclosure risk assessment.
  - Link databases using fuzzy integrals based distances



Fuzzy in IL + DR

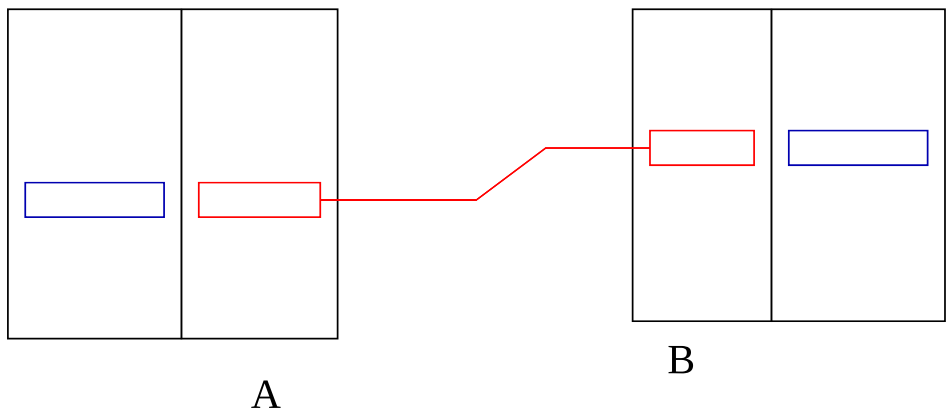
- Distance based record linkage: $d(A_i, B_i)$



- Find the *nearest* record (*nearest* in terms of a distance)
- Formally, 2 sets of vectors
  $A_i = (a_1, \ldots, a_N)$,
  (a_i protected version of b_i)
  $B_i = (b_1, \ldots, b_N)$
- $V_k(a_i)$: kth variable, ith record
- Distance $d(V_k(a_i), V_k(b_j))$ for all pairs (a_i, b_j).

Fuzzy in IL + DR

- Distance based record linkage: $d(A_i, B_i)$



- Find the *nearest* record (*nearest* in terms of a distance)
- Formally, 2 sets of vectors $A_i = (a_1, \ldots, a_N)$, (a_i protected version of b_i) $B_i = (b_1, \ldots, b_N)$
- $V_k(a_i)$: kth variable, ith record
- Distance $d(V_k(a_i), V_k(b_j))$ for all pairs (a_i, b_j).

- Distance based on aggregation functions $\mathbb{C}$
  E.g., $\mathbb{C} = CI$ (Choquet integral)

- Worst-case scenario: learn weights/fuzzy measure
  $\rightarrow$ Optimization problem

Fuzzy in IL + DR

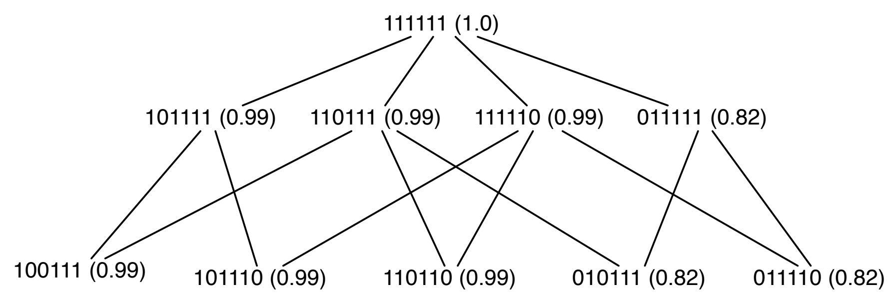
- Case $\mathbb{C} = WM$:

$$
\begin{aligned}
&Minimise \quad && \sum_{i=1}^{N} K_i \\
&Subject\ to: && \\
& && \sum_{k=1}^{N} p_i(d(V_k(a_i), V_k(b_j)) - d(V_k(a_i), V_k(b_i))) + CK_i > 0 \\
& && K_i \in \{0, 1\} \\
& && \sum_{i=1}^{N} p_i = 1 \\
& && p_i \geq 0
\end{aligned}
$$

- Similar with $\mathbb{C} = CI$ (Choquet integral)
- Extensive work comparing different scenarios and $\mathbb{C}$.

Fuzzy in IL + DR

- Results give:
  - number reidentifications in the worst-case scenario
  - Importance of weights (or sets of weights in fuzzy measures)
- Examples:
  - Choquet integral (



  - Weighted Mean (WM):
    - ⋆ V_1 0.016809573957189, V_2 0.00198841786482128, V_3 0.0045292377707479
    - ⋆ V_4 0.138812880222131, V_5 0.835523953314578, V_6 0.0023359368705328

Summary

Summary

- Outline the use of fuzzy methods in database privacy
  - Data protection
  - Information loss measures
  - Disclosure risk
- Fuzzy in other models: multiparty computation and differential privacy
- Research directions related to fuzzy set theory
  - Constraints on data (e.g., net + tax = gross), fuzzy microaggregation
  - Hesitant fuzzy clustering (e.g., several cluster centers x cluster)

References

References

- V. Torra, G. Navarro-Arribas (2020) Fuzzy meets privacy: a short overview, Proc. INFUS 2020.

- V. Torra (2017) Data privacy: Foundations, New Developments and the Big Data Challenge, Springer.

- V. Torra (2017) Fuzzy microaggregation for the transparency principle. J. Appl. Log. 23: 70-80.

- D. Abril, V. Torra, G. Navarro-Arribas (2015) Supervised learning using a symmetric bilinear form for record linkage. Inf. Fusion 26: 144-153.

- V. Torra, Fuzzy clustering-based microaggregation to achieve probabilistic k-anonymity for data with constraints, J. Intelligent and Fuzzy Systems, in press.

- M. Inuiguchi, H. Ichida, V. Torra, Data anonymization with imprecise rules and its performance evaluations, J. Ambient Int. Humanized Computing, in press.

Thank you