IEEE Chapter Meeting, Skövde

Data privacy. A briefer.

Vicenç Torra

(vtorra@ieee.org)

May 31st, 2018

Privacy, Information and Cyber-Security Center

SAIL, School of Informatics, University of Skövde, Sweden

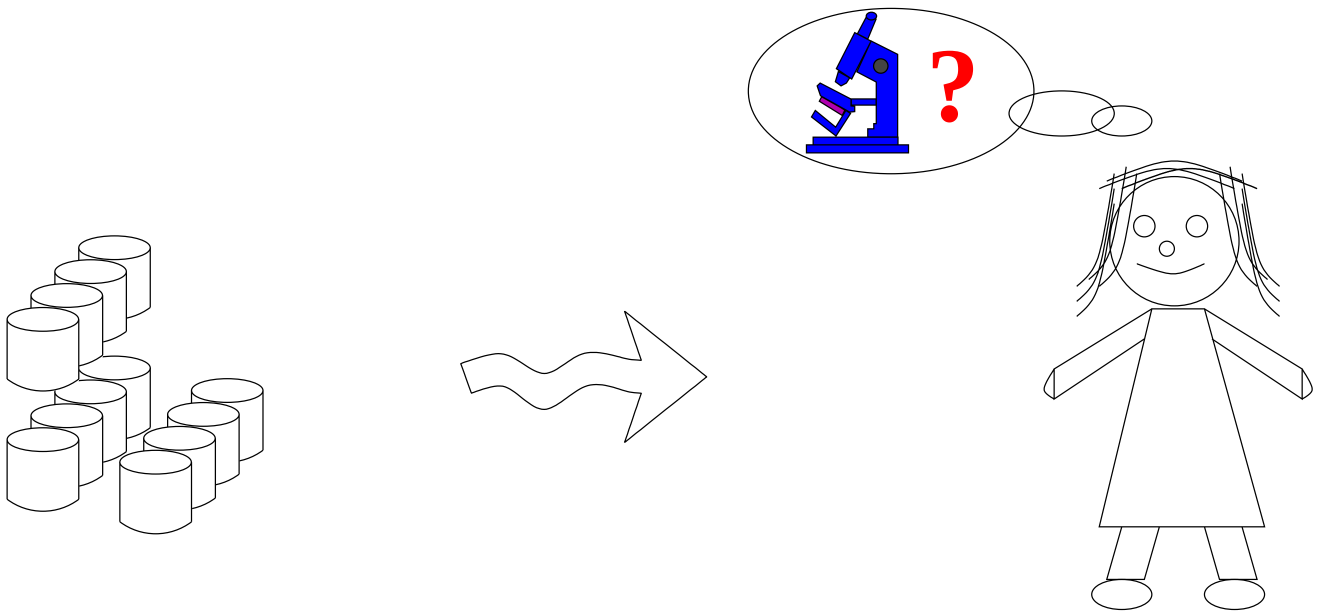
Outline

1. Motivation
2. Privacy models and disclosure risk assessment
3. Data protection mechanisms
4. Disclosure risk: The worst-case scenario
5. Summary

Motivation

Motivation

- Data privacy: (for database)
  - Someone needs to access to data to perform authorized analysis, but access to the data and the result of the analysis should avoid disclosure.



E.g., you are authorized to compute the average stay in a hospital, but maybe you are not authorized to see the length of stay of your neighbor.

Difficulties

- Difficulties: Naive anonymization does not work

  Passenger manifest for the Missouri, arriving February 15, 1882; Port of Boston[1]

  Names, Age, Sex, Occupation, Place of birth, Last place of residence, Yes/No, condition (healthy?)

[1] https://www.sec.state.ma.us/arc/arcgen/genidx.htm

Difficulties

- Difficulties: highly identifiable data
  - (Sweeney, 1997) on USA population
    - ⋆ 87.1% (216 million/248 million) were likely made them unique based on
      5-digit ZIP, gender, date of birth,
    - ⋆ 3.7% (9.1 million) had characteristics that were likely made them unique based on
      5-digit ZIP, gender, Month and year of birth.

Difficulties

- Difficulties: highly identifiable data
  - Data from mobile devices:
    - ⋆ two positions can make you unique (home and working place)
  - AOL[2] and Netflix cases (search logs and movie ratings)

    ⇒ User No. 4417749, hundreds of searches over a three-month period including queries 'landscapers in Lilburn, Ga' ⇒ Thelma Arnold identified!

    ⇒ individual users matched with film ratings on the Internet Movie Database.
  - Similar with credit card payments, shopping carts, ...
    (i.e., high dimensional data)

[2] http://www.nytimes.com/2006/08/09/technology/09aol.html

Difficulties

- Difficulties: highly identifiable data
  - Example #1:
    - ⋆ University goal: know how sickness is influenced by studies and by commuting distance
    - ⋆ Data: where students live, what they study, if they got sick
    - ⋆ No "personal data", is this ok ?

Difficulties

- Difficulties: highly identifiable data
  - Example #1:
    - ⋆ University goal: know how sickness is influenced by studies and by commuting distance
    - ⋆ Data: where students live, what they study, if they got sick
    - ⋆ No "personal data", is this ok ?
    - ⋆ NO!!: How many in your degree live in your town ?

Difficulties

- Difficulties: highly identifiable data
  - Example #1:
    - ⋆ University goal: know how sickness is influenced by studies and by commuting distance
    - ⋆ Data: where students live, what they study, if they got sick
    - ⋆ No "personal data", is this ok ?
    - ⋆ NO!!: How many in your degree live in your town ?
  - Example #2:
    - ⋆ Car company goal: Study driving behaviour in the morning
    - ⋆ Data: First drive (GPS origin + destination, time) $\times$ 30 days
    - ⋆ No "personal data", is this ok?

Difficulties

- Difficulties: highly identifiable data
  - Example #1:
    - ⋆ University goal: know how sickness is influenced by studies and by commuting distance
    - ⋆ Data: where students live, what they study, if they got sick
    - ⋆ No "personal data", is this ok ?
    - ⋆ NO!!: How many in your degree live in your town ?
  - Example #2:
    - ⋆ Car company goal: Study driving behaviour in the morning
    - ⋆ Data: First drive (GPS origin + destination, time) × 30 days
    - ⋆ No "personal data", is this ok?
    - ⋆ NO!!!: How many (cars) go from your parking to your university everymorning ? Are you exceeding the speed limit ? Are you visiting a psychiatrisc every tuesday ?

Difficulties

- Data privacy is "impossible", or not ?
  - Privacy vs. utility
  - Privacy vs. security
  - Computationally feasible

Privacy models and disclosure risk assessment

Privacy models

Privacy models: What is a privacy model ?

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

Privacy models

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

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.
- minimize information loss (max. utility) vs. multiobjetive optimization

Privacy models

Privacy models. quite a few *competing* models

- **Secure multiparty computation.** Several parties want to compute a function of their databases, but only sharing the result.
- **Reidentification privacy.** Avoid finding a record in a database.
- **k-Anonymity.** A record indistinguishable with $k-1$ other records.
- **Differential privacy.** The output of a query to a database should not depend (much) on whether a record is in the database or not.
- computational anonymity
- uniqueness
- result privacy
- interval disclosure

Privacy models

Privacy models. quite a few *competing* models

- **Secure multiparty computation.** Several parties want to compute a function of their databases, but only sharing the result.
- **Reidentification privacy.** Avoid finding a record in a database.
- **k-Anonymity.** A record indistinguishable with $k-1$ other records.
- **Differential privacy.** The output of a query to a database should not depend (much) on whether a record is in the database or not.
- computational anonymity
- uniqueness
- result privacy
- interval disclosure

... and combined:

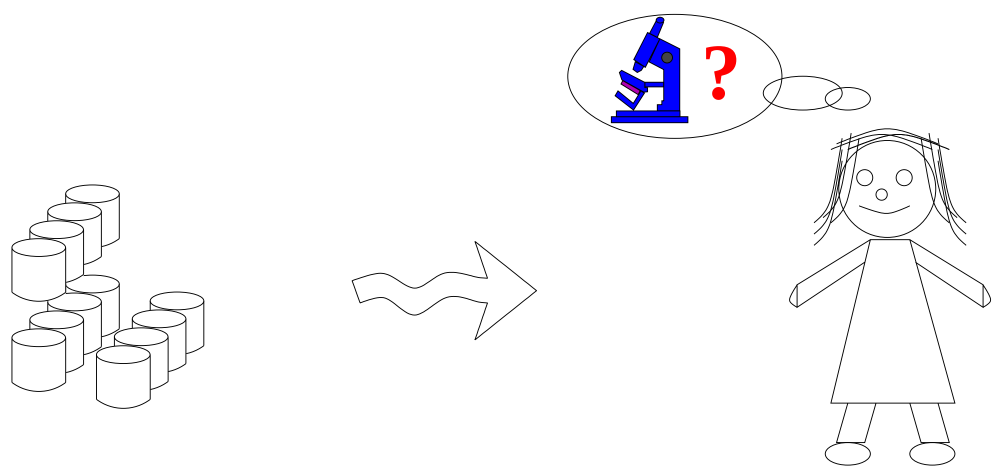
- secure multiparty computation + differential privacy

Data protection mechanisms: Masking methods

Data protection mechanisms

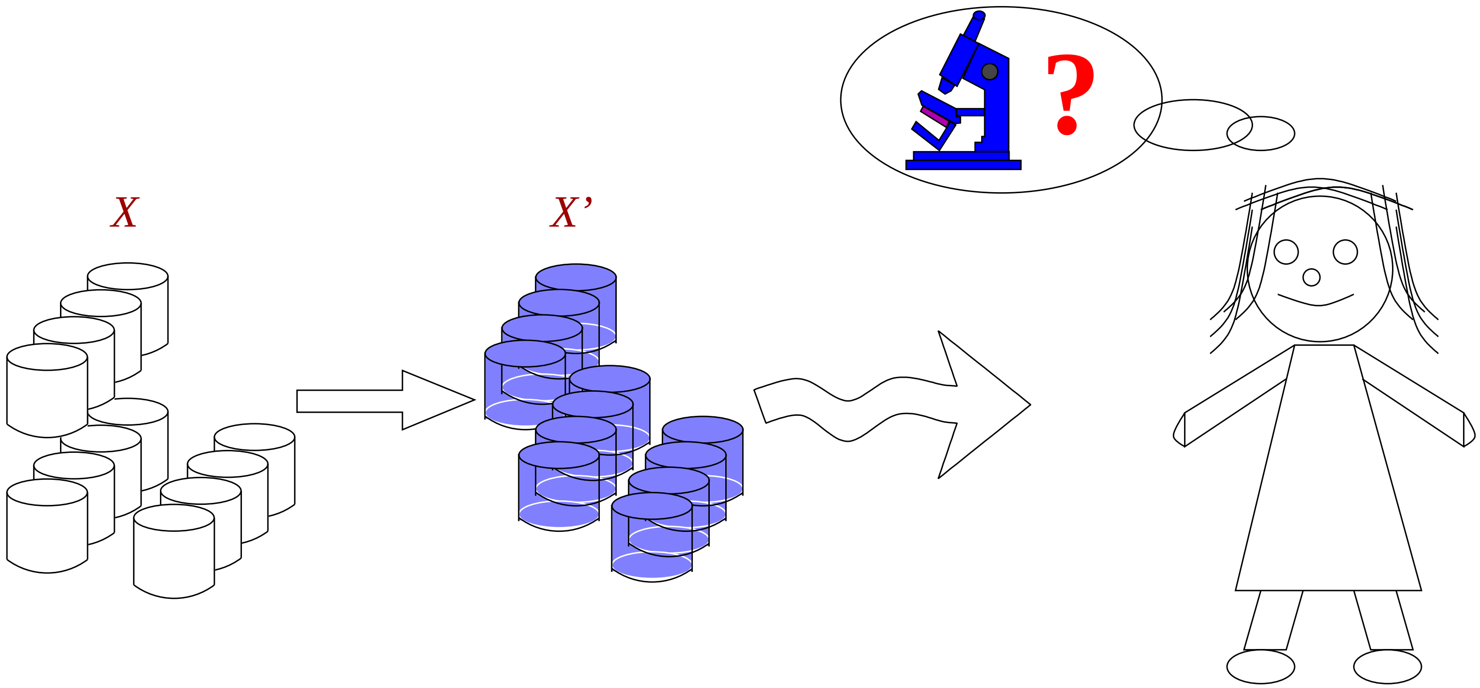
- **Focus** on respondent privacy (in databases)
- **Classification** w.r.t. knowledge on the computation of a third party
  - Data-driven or general purpose (*analysis not known*)
    → anonymization methods / masking methods
  - Computation-driven or specific purpose (*analysis known*)
    → cryptographic protocols, differential privacy
  - Result-driven (*analysis known: protection of its results*)

Figure. Basic model (multiple/dynamic databases + multiple *people*)

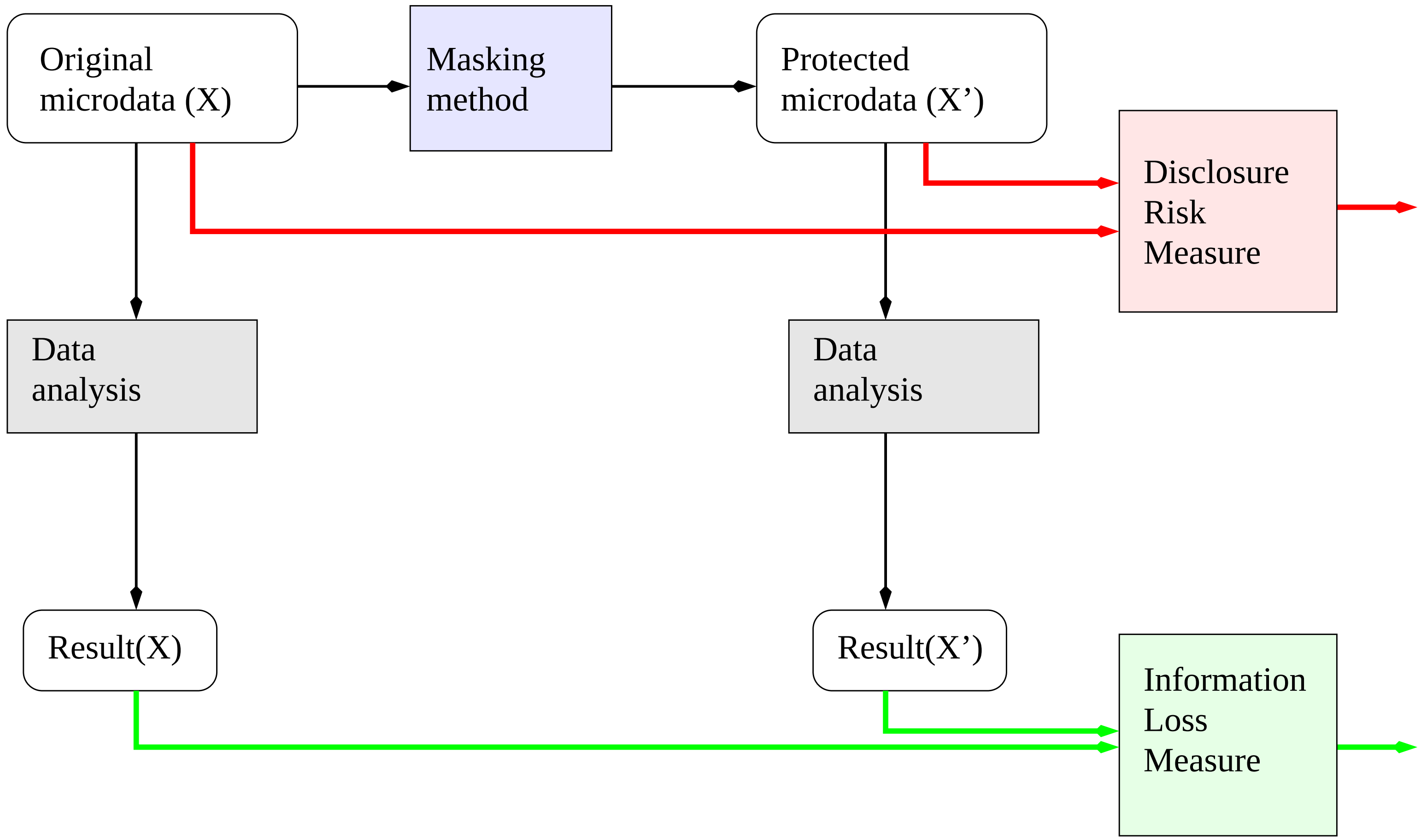


Masking methods

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



Masking methods: questions



Research questions I: Masking methods

Masking methods (anonymization methods).

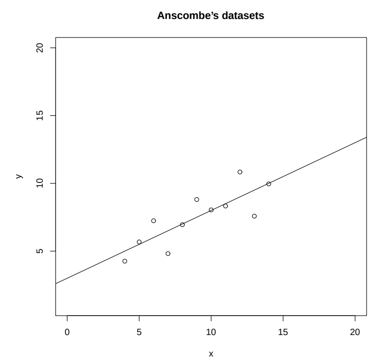
$$X' = \rho(X)$$

- Perturbative. (less quality=erroneous data)
  E.g. noise addition/multiplication, microaggregation, rank swapping
- Non-perturbative. (less quality=less detail)
  E.g. generalization, suppression
- Synthetic data generators. (less quality=not real data)
  E.g. (i) model from the data; (ii) generate data from model

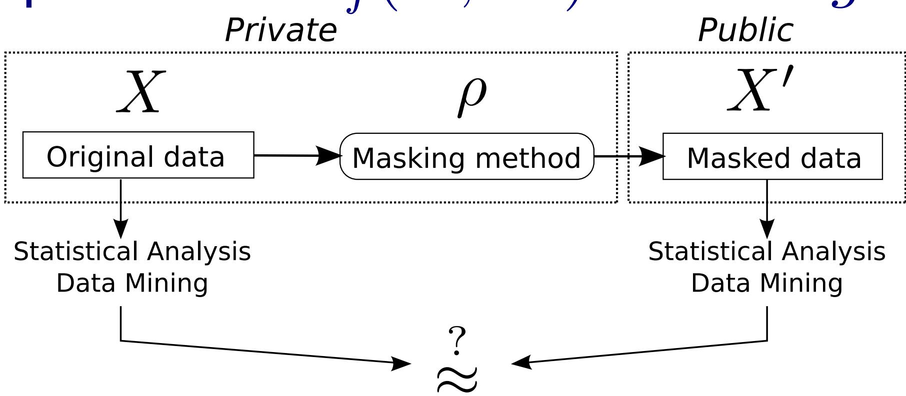
Research questions II: Information loss/Utility

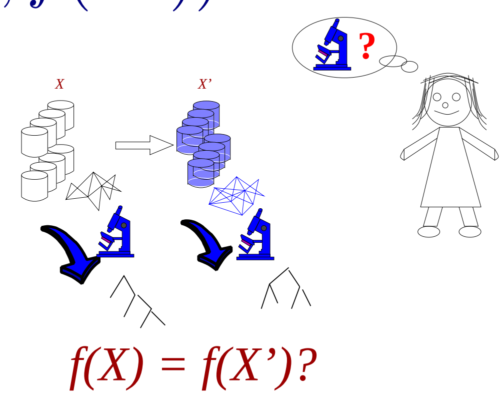
Information loss measures. Compare X and X' w.r.t. analysis (f)

- f: generic vs. specific (data uses). E.g. regression



- Comparison: $IL_f(X, X') = divergence(f(X), f(X'))$

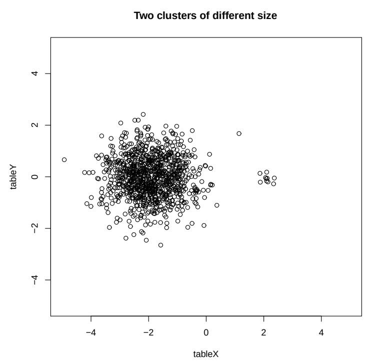




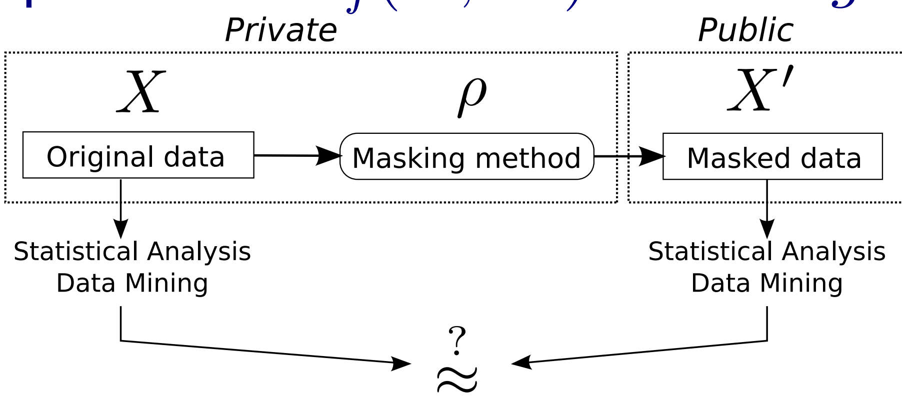
Research questions II: Information loss/Utility

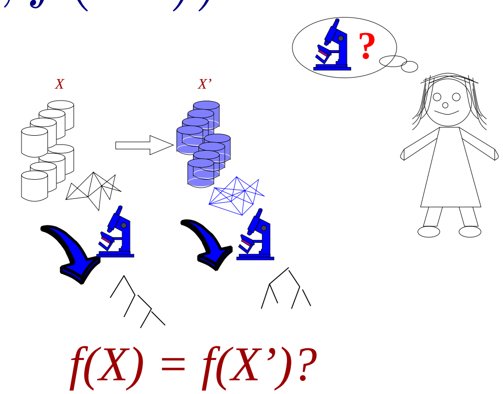
Information loss measures. Compare X and X' w.r.t. analysis (f)

- f: generic vs. specific (data uses). E.g. clustering



- Comparison: $IL_f(X, X') = divergence(f(X), f(X'))$

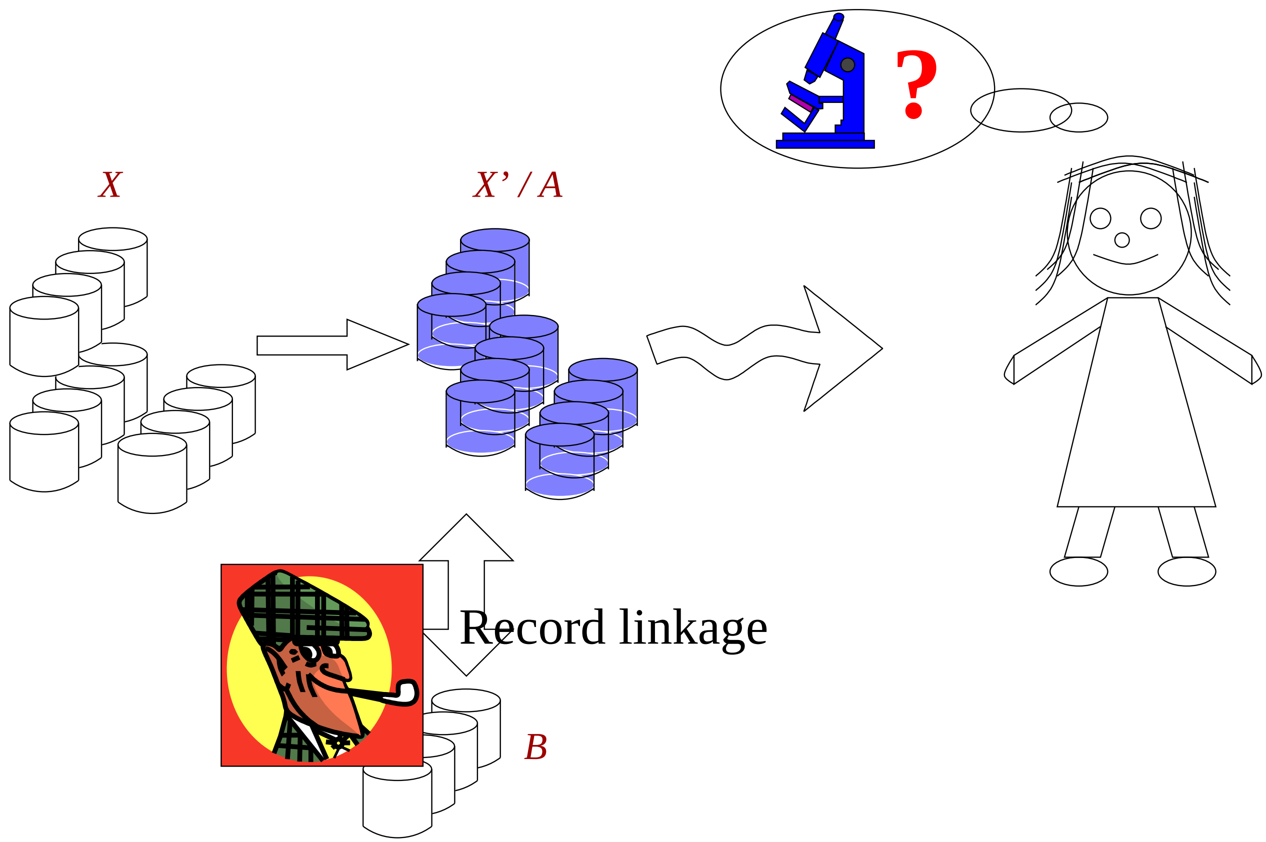




Research questions II: Information loss

Disclosure risk. One of the privacy models: reidentification (identity disclosure)

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



Disclosure risk: The worst-case scenario

Disclosure Risk

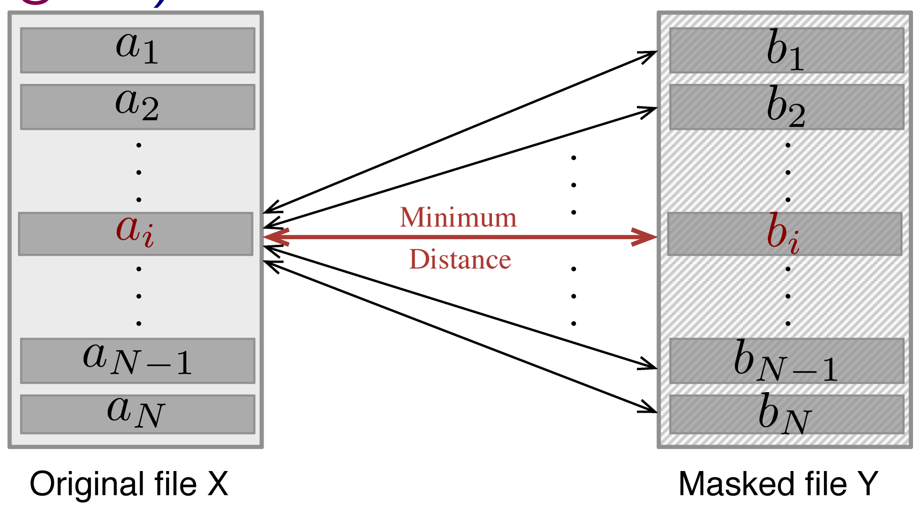
Disclosure risk (DR)

- The worst-case scenario
  - DR using the largest data set: original file
  - DR using the best reidentification method: optimal attacks (ML in reidentification)
  - DR under the transparency principle: transparency attacks

Optimal attacks

Machine Learning for distance-based record linkage

- Supervised approach: maximize the number of correct links.
- Use: Metric learning
- Goal (A and B aligned)



Transparency

Transparency.

- "the release of information about processes and even parameters used to alter data" (Karr, 2009).

Transparency principle. (similar to the Kerckhoffs's principle in cryptography)

- "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, p. 17)

Transparency

Effect.

- Information Loss. Positive effect, less loss/improve inference
  E.g., noise addition $\rho(X) = X + \epsilon$ where ϵ s.t.
  $E(\epsilon) = 0$ and $Var(\epsilon) = kVar(X)$

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

- Disclosure Risk. Negative effect, larger risk
  - Attack to single-ranking microaggregation (Winkler, 2002)
  - Formalization of the transparency attack (Nin, Herranz, Torra, 2008)
  - Attacks to microaggregation and rank swapping (Nin, Herranz, Torra, 2008)

  $\Rightarrow$ Transparency aware masking methods

Summary

Summary

Summary

- Short introduction to data privacy (focus on databases)
- Worst-case scenario and transparency

Thank you

References

References.

- Worst-case scenario
  - D. Abril, G. Navarro-Arribas, V. Torra, Supervised Learning Using a Symmetric Bilinear Form for Record Linkage, Information Fusion 26 (2015) 144-153.
- Transparency attacks and transparency aware methods
  - J. Nin, J. Herranz, V. Torra, On the Disclosure Risk of Multivariate Microaggregation, Data and Knowledge Engineering, 67 (2008) 399-412.
  - J. Nin, J. Herranz, V. Torra, Rethinking Rank Swapping to Decrease Disclosure Risk, Data and Knowledge Engineering, 64:1 (2008) 346-364.
  - V. Torra, Fuzzy microaggregation for the transparency principle, J. Applied Logic 23 (2017) 70-80.
- Book
  - V. Torra, Data Privacy: Foundations, New Developments and the Big Data Challenge, Springer, 2017.