Privacy models and disclosure risk

Vicenç Torra

February, 2018

SAIL + PICS, School of Informatics, University of Skövde, Sweden

Outline

Privacy models and disclosure risk assessment

Disclosure risk assessment

Privacy models: What is a privacy model ?

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

Definition:

- A computational definition of privacy

Disclosure risk assessment

Privacy models: What is a privacy model ?

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

Definition:

- A computational definition of privacy

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

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.

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

Privacy models. (selection)

- **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.

Disclosure risk assessment

Privacy model. Secure multiparty computation.

- Several parties want to compute a function of their databases, but only sharing the result.
  - hospital A and hospital B,
  - two independent databases with:

    `age of patient, length of stay in hospital`
- how to compute a regression with all data (both databases)

  $$\text{age} \rightarrow \text{length}$$

  without sharing data?

Disclosure risk assessment

Privacy model. Reidentification privacy.

- Avoid finding a record in a database.
  - hospital A has a database
  - a researcher asks for access to this database
- how to prepare an anonymized database so that the researcher can not find a friend?

Disclosure risk assessment

Privacy model. k-Anonymity.

- Avoid finding a record in a database.
  ... making each record indistinguishable with $k-1$ other records.

Disclosure risk assessment

Privacy model. k-Anonymity.

- Avoid finding a record in a database.
  ... making each record indistinguishable with $k-1$ other records.
  - hospital A has a database
  - a researcher asks for access to this database
- how to prepare an anonymized database so that the researcher can not find a friend?

Disclosure risk assessment

Privacy model. Differential privacy.

- The output of a query to a database should not depend (much) on whether a record is in the database or not.
  - hospital A has a database

    `age of patient, length of stay in hospital`
- how to compute an average length of stay in such a way that the result does not depend (much) on whether we use or not the data of a particular person.

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

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

Disclosure risk assessment

Disclosure risk.

- Function known vs. unknown (ill-defined)
- Identity disclosure vs. Attribute disclosure
- Boolean vs. quantitative measures/models

Disclosure risk assessment

Disclosure risk.

- Function known vs. unknown (ill-defined)
- 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
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
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