Lines of Research on Data Privacy

Vicenç Torra

December, 2019

Hamilton Institute, Maynooth Universtiy, Maynooth, Ireland

Lines of research

- Privacy models
  - Formalization of what privacy means
    - ⋆ Always in search of a good definition (meaningful, usable, programmable, and permit some data uses)
  - Competing privacy models and combination of them
    - ⋆ Privacy for databases: reidentification, k-anonymity, l-diversity, etc
    - ⋆ Privacy with respect to results: differential privacy, local differential privacy, integral privacy
    - ⋆ Privacy about some particular inferences: result privacy,
    - ⋆ Privacy in multiparty computation: multiparty computation,
    - ⋆ Combination of models: multiparty computation + differential privacy,
    - ⋆ Translation of models to new types of data
      - – What is k-anonymity in graphs/social media?

Lines of research

- Implementing privacy models (I)
  - for each privacy model $\times$ each type of data $\times$ each data use
  - E.g., for k-anonymity
    - $\star$ Methods to achieve k-anonymity for numerical databases (standard file that can be stored in memory)
    - $\star$ Methods to achieve k-anonymity for streaming data (real-time)
    - $\star$ Methods to achieve k-anonymity for big databases (efficiency in computation)
  - New algorithms to improve previous ones
  - E.g., improve with respect to data utility, speed, or disclosure risk.
    - $\star$ best (high) utility for the same risk
    - $\star$ best (low) risk for the same utility
    - $\star$ faster for the same risk and utility

Lines of research

- Implementing privacy models (II)
  - Centralized vs. distributed databases.
    - ⋆ Centralized
      ⇒ One company, one database, one *technician*
    - ⋆ Distributed low scale
      ⇒ n companies, n databases, on agreement
      ⇒ they decide to compute $f(X_1, ..., X_n)$
      (multiparty computation model)
    - ⋆ Highly distributed applications
      ⇒ One company, milions of users (mobiles, cars)
      ⋆ how data is transferred in a safe/private way to the company
      ⋆ federated learning: data is not transferred, models are

Lines of research

- Implementing privacy models (IIIa).
  Data driven methods ⇒ Different types of data
  - Standard databases
  - Streaming data (real time)
  - Large static databases (*just big*)
  - Dynamic data (databases change, and need of multiple releases)
    ⇒ multiple releases is problematic as inference can take advantage of the same record published several times with slightly different information
  - Social networks and graphs
    ⇒ *individuals* are not *independent*, links relate *individuals*. Masking a person is not enough.
  - Search logs
    ⇒ records are not independent

Lines of research

- Implementing privacy models (IIIb).
  Data driven methods ⇒ Different types of data
  - NoSQL databases
    - ⋆ Textual documents
      - – Sanitization of documents. Detection of sensitive words. However, other elements of the text (combinations of them) may lead to disclosure. E.g., in health care.
    - ⋆ Scanned textual documents (handwritten documents)
    - ⋆ Sound / voice
    - ⋆ Images (people, places, etc.)
    - ⋆ Video

Lines of research

- Attacks, analysis of disclosure, and uniqueness
  - Attacking a database
    $\Rightarrow$ with side information, or internal attacks (individuals in the database use their information to infer about others)
  - Transparency attacks
    $\Rightarrow$ use information on how data has been protected
  - Attacks to the models
    $\Rightarrow$ Membership attacks
  - Fingerprinting of browsers and computers
  - Inferences from data
    $\Rightarrow$ hidden inferences as gender or political affiliation from other variables (or connections in a social network)

Lines of research

Venues for Data Privacy: (Security, Cryptography, Databases, Statistics, Official Statistics)

- Journals
  - Transactions on Data Privacy (Open Access)
  - Journal of Privacy and Confidentiality (CMU, Open Access, from 2009)
  - IEEE Security and Privacy (IEEE, from 2003)
  - ACM Transactions on Privacy and Security (ACM)
- Conferences
  - S&P IEEE Symposium on Security and Privacy. Also Euro S&P.
  - PETS - Privacy Enhancing Technologies
  - WPES - Workshop On Privacy In The Electronic Society
  - PST - Conference on Privacy, Security and Trust
  - PSD - Privacy in Statistical Databases

Lines of research

Our research on data privacy (I)

- Disclosure risk.
  - Transparency attacks and worst-case analysis (using machine learning)
- Metrics for comparing data protection methods: information loss and data utility

Lines of research

Our research on data privacy (II)

- Data protection (masking methods) for centralized databases
  - *Classical* databases. Microagregation, rank swapping, PRAM
  - No-SQL databases.
    - ⋆ Documents (document sanitization, indexes for documents)
    - ⋆ Graphs and online social networks (masking methods for k-anonymity and reidentification)
  - Privacy models: integral privacy

Thank you

http://www.ppdm.cat/dp/