Data privacy: Introduction

Vicenç Torra

November, 2020

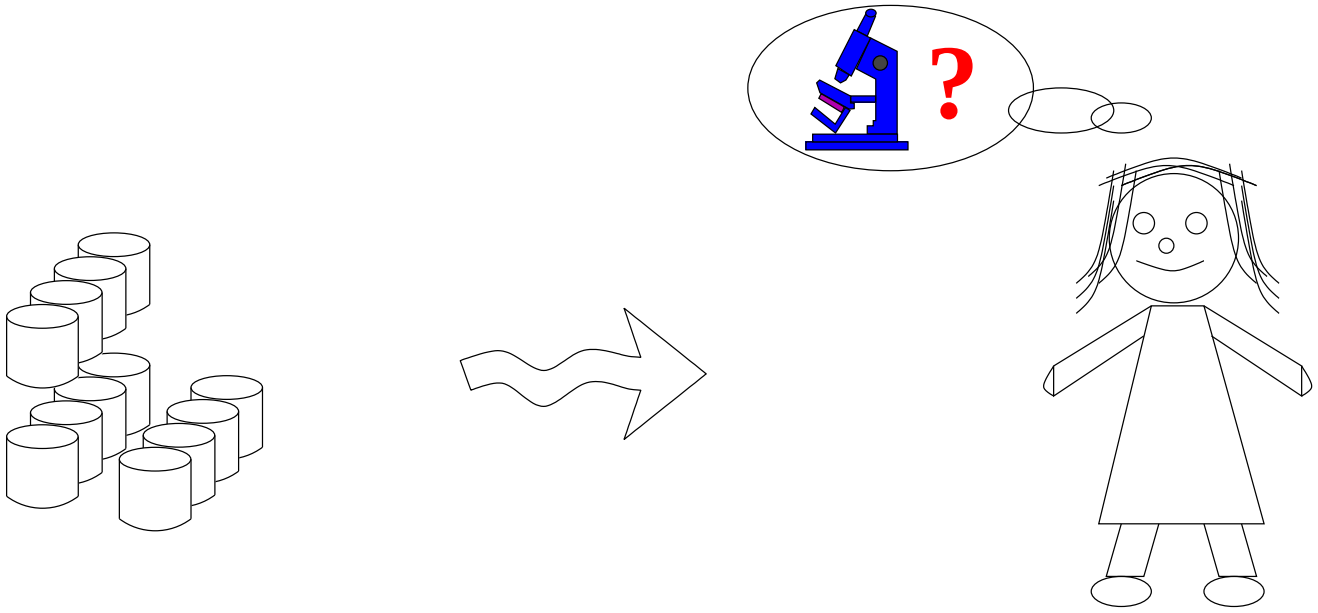
Umeå University, Sweden

Outline

Motivation

Introduction

- Data privacy: core
  - 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.

Introduction

- Problems/difficulties? *Example 1*
  - Q: sickness influenced by studies and commuting distance ?
  - Data: (where students live, what they study, if they got sick)

$$
\begin{aligned}
DB = \{ \quad & (\ Dublin, CS\&SE, no\ ) \\
& (\ Dublin, CS\&SE, yes\ ) \\
& (\ Dublin, \ldots, \ldots\ ) \\
& \ldots \\
& (\ Maynooth, CS\&SE, no\ ) \\
& (\ Maynooth, CS\&SE, no\ ) \\
& (\ Maynooth, CS\&SE, yes\ ) \\
& (\ Maynooth, \ldots, \ldots\ ) \\
& \ldots \\
& (\ Ballyroe^{1}, XXXX, yes\ )
\end{aligned}
$$

  - No "personal data", is this ok ? NO!!

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

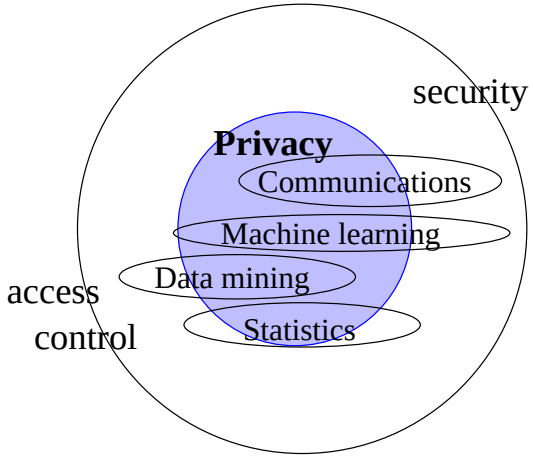
Introduction

- Problems/difficulties? *Example 2*
  - Q: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town?
  - Mean income is not "personal data", is this ok ? NO!!:
  - Example[2]: 1000 2000 3000 2000 1000 6000 2000 10000 2000 4000 $\Rightarrow$ mean = 3300
  - Adding Ms. Rich's salary 100,000 Eur/month: mean = 12090,90 !
    (a extremely high salary changes the mean significantly)
    $\Rightarrow$ We infer Ms. Rich from Town was attending the unit

[2] Average wage in Ireland (2018): 38878 $\Rightarrow$ monthly 3239 Eur
https://www.frsrecruitment.com/blog/market-insights/average-wage-in-ireland/

Introduction

- A personal view of core and boundaries of data privacy: core
  - data uses / rellevant techniques
    - ⋆ Data to be used for data analysis
      - ⇒ statistics, machine learning, data mining
      - ⇒ compute indices, find patterns, build models
    - ⋆ Data is transmitted
      - ⇒ communications



- Someone needs to access to data to perform authorized analysis, but access to the data and the result of the analysis should avoid disclosure.

Introduction

- A personal view of core and boundaries of data privacy: boundaries
  - Database in a computer or in a removable device

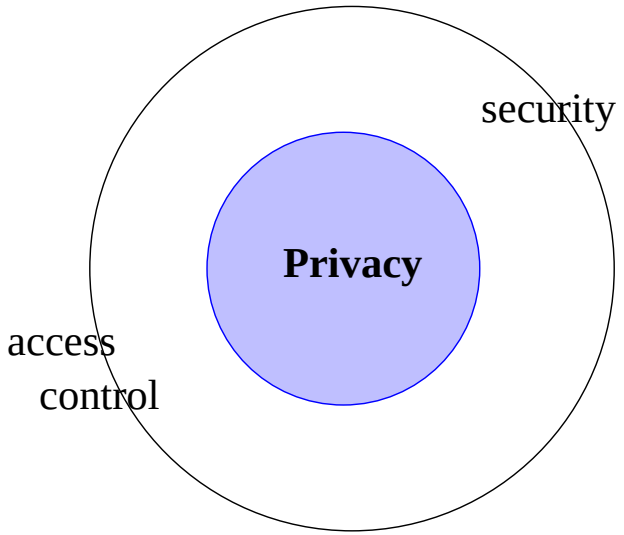
    $\Rightarrow$ access control to avoid unauthorized access

    $\Longrightarrow$ Access to address (admissions), Access to blood test (admissions?)
  - Data is transmitted

    $\Rightarrow$ security technology to avoid unauthorized access

    $\Longrightarrow$ Data from blood glucose meter sent to hospital. Network sniffers

    Transmission is sensitive: Near miss/hit report to car manufacturers



Motivation

- Legislation.
  - Privacy a fundamental right. (Ch. 1.1)
    - ⋆ Universal Declaration of Human Rights (UN). European Convention on Human Rights (Council of Europe). General Data Protection Regulation - GDPR (EU). National regulations.
  - Enforcement (GDPR)
    - ⋆ Obligations with respect to data processing
    - ⋆ Requirement to report personal data breaches
    - ⋆ Grant individual rights (to be informed, to access, to rectification, to erasure, ...)
- Companies own interest.
  - Competitors can take advantage of information.
- Avoiding privacy breach. Several well known cases.

Motivation

- Privacy and society
  - Not only a computer science/technical problem
    - ⋆ Social roots of privacy
    - ⋆ Multidisciplinary problem
  - Social, legal, philosophical questions
  - Culturally relative?
    I.e., the importance of privacy is the same among all people ?
  - Are there aspects of life which are inherently private or just conventionally so?

Motivation

- Privacy and society. Is this a new problem? Yes and not
  - No side. See the following:

    Instantaneous photographs and newspaper enterprise have invaded the sacred precincts of private and domestic life; and numerous mechanical devices threaten to make good the prediction that "what is whispered in the closet shall be proclaimed from the house-tops." (...)
    Gossip is no longer the resource of the idle and of the vicious, but has become a trade, which is pursued with industry as well as effrontery (...) To occupy the indolent, column upon column is filled with idle gossip, which can only be procured by intrusion upon the domestic circle.
    (S. D. Warren and L. D. Brandeis, 1890)
  - Yes side: big data, storage, surveillance/CCTV, RFID, IoT

Motivation

- Technical solutions
  - Statistical disclosure control (SDC)
  - Privacy preserving data mining (PPDM)
  - Privacy enhancing technologies (PET)
- Socio-technical aspects
  - Technical solutions are not enough
  - Implementation/management of solutions for achieving data privacy need to have a holistic perspective of information systems
  - E.g., employees and customers: how technology is applied

Difficulties

Difficulties

- Difficulties: Naive anonymization does not work

  Passenger manifest for the Missouri, arriving February 15, 1882; Port of Boston[3]

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

LIST OF PASSENGERS.

REPORT AND LIST OF PASSENGERS taken on board the S.S. Hooper at London.

NAMES	AGE	SEX	OCCUPATION	PLACE OF BIRTH	Last Place of Residence	CONDITION
[illegible]	11	Male	Gentleman	Aberdeen Eng	Aberdeen Eng	Yes		Healthy
[illegible]	24	"	Lawyer's Clerk	Madras India	Scotland Eng		No	"
[illegible]	21	"	Tailor	London	London		No	"
[illegible]	26	"	Surgeon	London	London		No	"
[illegible]	28	"	Gentleman	Boston U.S.	Boston U.S.	Yes		"
[illegible]	30	"	"	Scotland	Boston U.S.	Yes		"
[illegible]	16	"	"	London (?)	London	Yes		"
[illegible]	44	"	"	Down Eng.	Down Eng.	Yes		"
[illegible]	28	"	"	Aberdeen Scotland	Boston U.S.	Yes		"
[illegible]	28	"	Cattle Dealer	Boston U.S.	Boston U.S.	Yes		"

[3] 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 and high dimensional data
  - Data from mobile devices:
    $\Rightarrow$ two positions can make you unique (home and working place)
  - AOL[4] and Netflix cases (search logs and movie ratings)
    $\Rightarrow$ User No. 4417749, hundreds of searches over a three-month period including queries 'landscapers in Lilburn, Ga'
    $\longrightarrow$ Thelma Arnold identified!
    $\Rightarrow$ individual users matched with film ratings on the Internet Movie Database.
  - Similar with credit card payments, shopping carts, ...

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

Difficulties

- Difficulties: highly identifiable data and high dimensional data
  - Ex1: Sickness influenced by studies and commuting distance ?
  - Ex2: Mean income of admitted to hospital unit (e.g., psychiatric unit) for a given Town?
  - Ex3: Driving behavior in the morning
    - ⋆ Automobile manufacturer uses (data from vehicles)
    - ⋆ Data: First drive after 6:00am
      (GPS origin + destination, time) × 30 days
    - ⋆ No "personal data", is this ok?: NO!!!:
    - ⋆ How many cars from your home to your work?
      Are you exceeding the speed limit? Are you visiting a psychiatric clinic every tuesday?

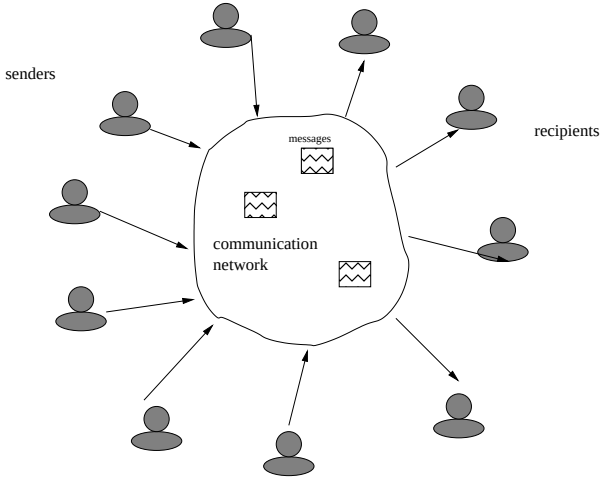
Difficulties

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

Terminology

Terminology

- Terminology using as framework a communication network with senders (actors) and receivers (actees)



- Attacker, adversary, intruder
  - the set of entities working against some protection goal
  - increase their knowledge (e.g., facts, probabilities, . . .) on the items of interest (IoI) (senders, receivers, messages, actions)

Terminology

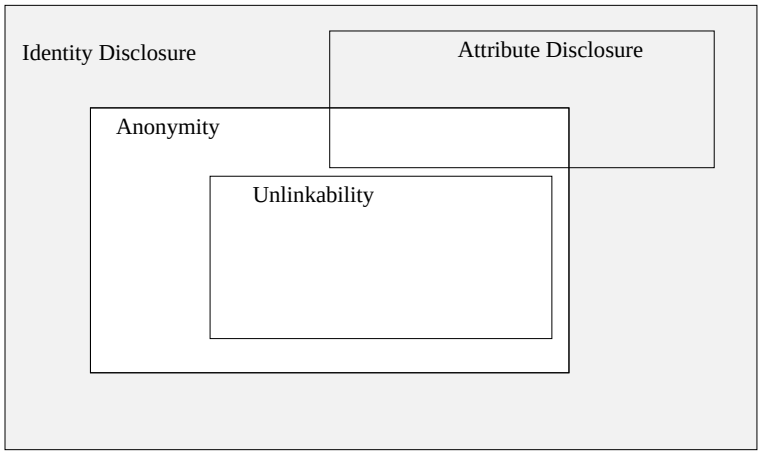
- Anonymity set. Anonymity of a subject means that the subject is not identifiable within a set of subjects, the anonymity set. Not distinguishable!

- Unlinkability. Unlinkability of two or more IoI, the attacker cannot sufficiently distinguish whether these IoIs are related or not.
  $\Rightarrow$ Unlinkability with the sender implies anonymity of the sender.

  - Linkability but anonymity. E.g., an attacker links all messages of a transaction, due to timing, but all are encrypted and no information can be obtained about the subjects in the transactions: anonymity not compromised.
    (region of the anonymity box outside unlinkability box)

Terminology

- Examples of anonymity in communications (definition of IoI):
  - **Sender anonymity.** No link between a message and the sender.
  - **Recipient anonymity.** No link between a message and the receiver.
  - **Relationship anonymity.** No link between a message and both sender and receiver.



Terminology

- Disclosure. Attackers take advantage of observations to improve their knowledge on some confidential information about an IoI.
  $\Rightarrow$ SDC/PPDM: Observe DB, Δ knowledge of a particular subject (the respondent in a database)
  - Identity disclosure (entity disclosure). Linkability. Finding Mary in the database.
  - Attribute disclosure. Increase knowledge on Mary's salary.
    also: learning that someone is in the database, although not found.

Terminology

- Disclosure. Discussion.
  - Identity disclosure. Avoid.
  - Attribute disclosure. A more complex case. Some attribute disclosure is expected in data mining.

    At the other extreme, any improvement in our knowledge about an individual could be considered an intrusion. The latter is particularly likely to cause a problem for data mining, as the goal is to improve our knowledge. (J. Vaidya et al., 2006, p. 7.

Terminology

- Identity disclosure vs. attribute disclosure
  - Usually, identity disclosure implies attribute disclosure
    Find record $(HYU, Tarragona, 58)$, learn variable (*Heart Attack*)

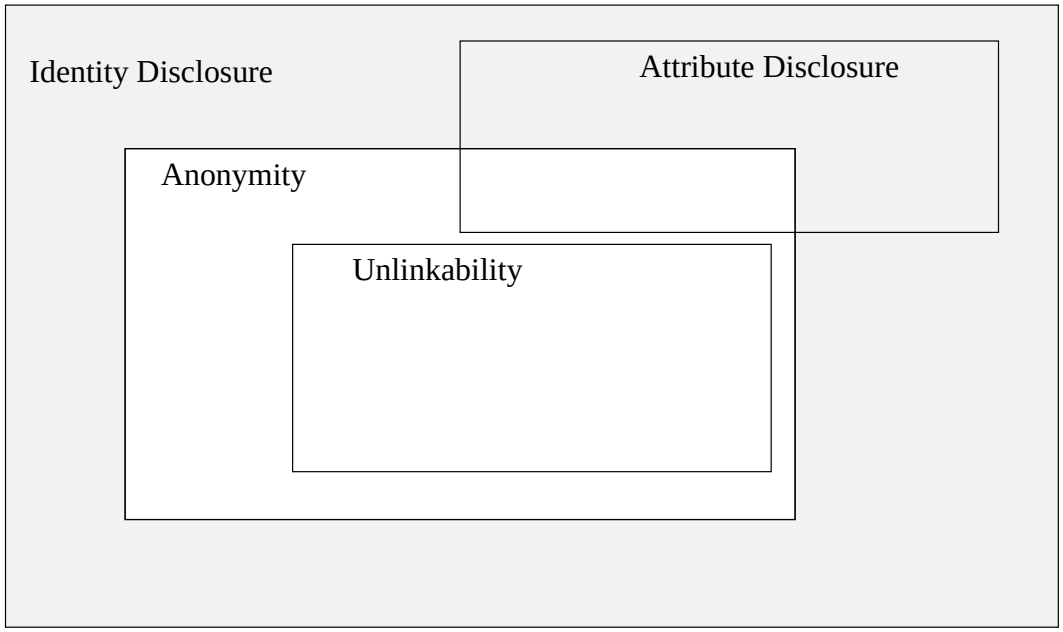
Respondent	City	Age	Illness
ABD	Barcelona	30	Cancer
COL	Barcelona	30	Cancer
GHE	Tarragona	60	AIDS
CIO	Tarragona	60	AIDS
HYU	Tarragona	58	Heart attack

  - Identity disclosure without attribute disclosure. Use all attributes
  - Attribute disclosure without identity disclosure. k-anonymity
    $(ABD, Barcelona, 30)$ not reidentified but learn *Cancer*

Respondent	City	Age	Illness
ABD	Barcelona	30	Cancer
COL	Barcelona	30	Cancer
GHE	Tarragona	60	AIDS
CIO	Tarragona	60	AIDS

Terminology

- Identity disclosure and anonymity are exclusive.
  - Identity disclosure implies non-anonymity
  - Anonymity implies no identity disclosure.



Terminology

- Undetectability and unobservability
  - Undetectability of an IoI. The attacker cannot sufficiently distinguish whether IoI exists or not.
    E.g. Intruders cannot distinguish messages from random noise
    $\Rightarrow$ Steganography
  - Unobservability of an IoI means
    - $\star$ undetectability of the IoI against all subjects uninvolved in it and
    - $\star$ anonymity of the subject(s) involved in the IoI even against the other subject(s) involved in that IoI.

Unobservability pressumes undetectability but at the same time it also pressumes anonymity in case the items are detected by the subjects involved in the system. From this definition, it is clear that unobservability implies anonymity and undetectability.

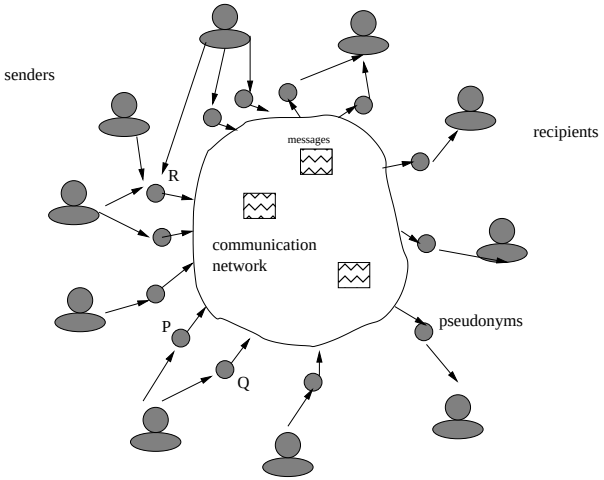
Terminology

- Pseudonyms and identity
  - Pseudonym. An identifier of a subject other than one of the subject's real names.

    Pseudonymising is defined as the replacing of the name or other identifiers by a number in order to make the identification of the data subject impossible or substantially more difficult. (Federal Data Protection Act, Germany, 2001)

    ⋆ 1:1, 1:n, n:1 relationship.

    ⋆ Model a range between anonymity (no linkability) to accountability (maximum linkability)



Terminology

- Pseudonyms and identity
  - Identity. Any subset of attribute values of an individual person which sufficiently identifies this individual person within any set of persons. So usually there is no such thing as "the identity", but several of them.
  - Roles are defined as the set of actions that users (people) are allowed to perform.
  - Each partial identity represents the person in a specific context or role.

Transparency

Transparency

- 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.

- The transparency principle in data privacy[5]

  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)

[5] 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

Privacy by design

Privacy by design

- Privacy by design (Cavoukian, 2011)
  - Privacy "must ideally become an organization's default mode of operation" (Cavoukian, 2011) and thus, not something to be considered a posteriori. In this way, privacy requirements need to be specified, and then software and systems need to be engineered from the beginning taking these requirements into account.
  - *In the context of developing IT systems, this implies that privacy protection is a system requirement that must be treated like any other functional requirement. In particular, privacy protection (together with all other requirements) will determine the design and implementation of the system* (Hoepman, 2014)

Privacy by design

- Privacy by design principles (Cavoukian, 2011)

1. Proactive not reactive; Preventative not remedial.
2. Privacy as the default setting.
3. Privacy embedded into design.
4. Full functionality – positive-sum, not zero-sum.
5. End-to-end security – full lifecycle protection.
6. Visibility and transparency – keep it open.
7. Respect for user privacy – keep it user-centric.

Summary

Terminology

- Concepts
  - What is data privacy?
  - Multidisciplinary problem and socio-technical aspects to be considered
  - Difficulties of data privacy: naive annonymization does not work
  - Linkability and anonymity set
  - Identity and attribute disclosure
  - Transparency
  - Privacy by design

References

References

- V. Torra (2017) Data privacy, Springer.

- V. Torra, G. Navarro-Arribas (2016) Big Data Privacy and Anonymization, Privacy and Identity Management 15-26
https://doi.org/10.1007/978-3-319-55783-0_2 (open access)

- V. Torra, G. Navarro-Arribas (2014) Data privacy, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 4:4 269-280
https://doi.org/10.1002/widm.1129