Privacy and anonymity: prognosis and therapy

Adversarial attacks and privacy-preserving technologies

Dino Pedreschi e Fosca Giannotti

Traces: forget or remember?

- When no longer needed for service delivery, traces can be either forgotten or stored.
  - Storage is cheaper and cheaper.
- But why should we store traces?
  - From business-oriented information – sales, customers, billing-related records, …
  - To finer grained process-oriented information about how a complex organization works.
- Traces are worth being remembered because they may hide precious knowledge about the processes which govern the life of complex economical or social systems.

Direttiva 2002/58/CE (vita privata e comunicazioni elettroniche)

- Art. 6 comma 1. I dati sul traffico relativi agli utenti, trattati e memorizzati dal fornitore di una rete pubblica o di un servizio pubblico di comunicazione elettronica, devono essere **cancellati o resi anonimi** quando non sono più necessari ai fini della trasmissione di una comunicazione, fatti salvi … (fatturazione, servizi a valore aggiunto con acquisizione del consenso, … oltre a finalità dell’autorità giudiziaria)

The Spy and the Historian

- The malicious eyes of the **Spy** – or the detective – aimed at
  - discovering the individual knowledge about the behaviour of a single **person** (or a small group)
  - for **surveillance** purposes.
- The benevolent eyes of the **Historian** – or the archaeologist, or the human geographer – aimed at
  - discovering the collective knowledge about the behaviour of whole **communities**,
  - for the purpose of **analysis**, of understanding the dynamics of these communities, the way they live.

The location privacy problem in mobility data analysis

- The donors of the mobility data are ourselves the citizens,
- Making these data available, even for analytical purposes, would put at risk our own privacy, our right to keep secret
  - the places we visit,
  - the places we live or work at,
  - the people we meet
- How to protect the privacy (guarantee the anonymity) of the donors?

The naive scientist's view (1)

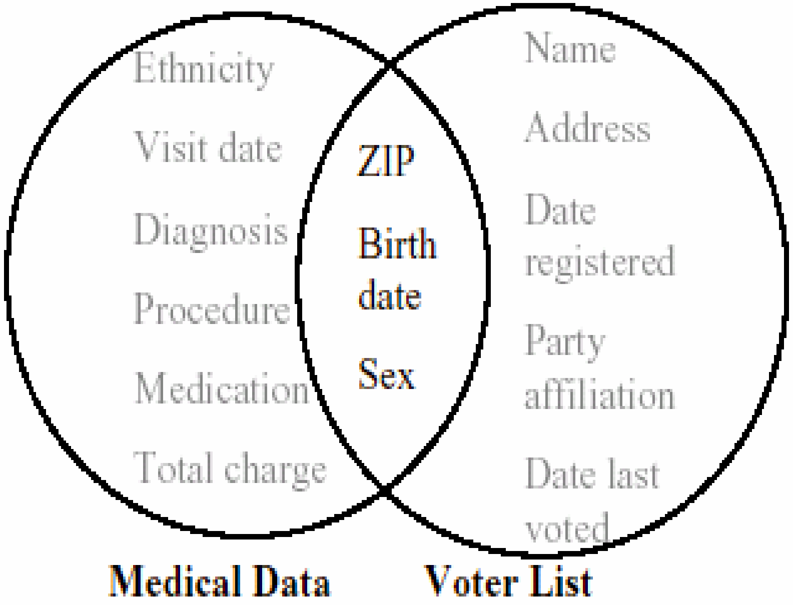
- Knowing the exact identity of individuals is not needed for analytical purposes
  - Anonymous trajectories are enough to reconstruct aggregate movement behaviour, pertaining to groups of people.
- Is this reasoning correct?
- Can we conclude that the analyst runs no risks, while working for the public interest, to inadvertently put in jeopardy the privacy of the individuals?

Unfortunately not!

- Hiding identities is not enough.
- In certain cases, it is possible to reconstruct the exact identities from the released data, even when identities have been removed and replaced by pseudonyms.
- A famous example of re-identification by L. Sweeney

Re-identifying “anonymous” data (Sweeney ’01)

- Dataset #1: medical records from the US Nat. Association of the Health Data Organizations, made available to research institutes – believed anonymous!
- Dataset #2: voter registration list for Cambridge Massachusetts
  - 54,805 people



87% unique US-wide with (ZIP + birth date + Sex)!!!

Private Information in Publicly Available Data

Date of Birth	Zip Code	Allergy	History of Illness
03-24-79	07030	Penicillin	Pharyngitis
08-02-57	07028	No Allergy	Stroke
11-12-39	07030	No Allergy	Polio
08-02-57	07029	Sulfur	Diphtheria
08-01-40	07030	No Allergy	Colitis

Medical Research Database

Sensitive Information

Linkage attack: Link Private Information to Person

Quasi-identifiers

Date of Birth	Zip Code	Allergy	History of Illness
03-24-79	07030	Penicillin	Pharyngitis
08-02-57	07028	No Allergy	Stroke
11-12-39	07030	No Allergy	Polio
08-02-57	07029	Sulfur	Diphtheria
08-01-40	07030	No Allergy	Colitis

Victor is the only person born 08-02-57 in the area of 07028… Ha, he has a history of stroke!

Sweeney's experiment

- Consider the governor of Massachusetts:
  - only 6 persons had his birth date in the joined table (voter list),
  - only 3 of those were men,
  - and only … 1 had his own ZIP code!
- The medical records of the governor were uniquely identified from legally accessible sources!

The naive scientist’s view (2)

- Why using quasi-identifiers, if they are dangerous?
- A brute force solution: replace identities or quasi-identifiers with totally unintelligible codes
- Aren’t we safe now?
- No! Two examples:
  - The AOL August 2006 crisis
  - Movement data

A face is exposed for AOL searcher no. 4417749 [New York Times, August 9, 2006]

- No. 4417749 conducted hundreds of searches over a three months period on topics ranging from “numb fingers” to “60 single men” to “dogs that urinate on everything”.
- And search by search, click by click, the identity of AOL user no. 4417749 became easier to discern. There are queries for “landscapers in Lilburn, Ga”, several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnet county georgia”.

A face is exposed for AOL searcher no. 4417749 [New York Times, August 9, 2006]

- It did not take much investigating to follow this **data trail** to Thelma Arnold, a 62-year-old widow of Lilburn, Ga, who loves her three dogs. “Those are my searches,” she said, after a reporter read part of the list to her.
- Ms. Arnold says she loves online research, but the disclosure of her searches has left her disillusioned. In response, she plans to drop her AOL subscription. “We all have a right to privacy,” she said, “Nobody should have found this all out.”
- http://data.aolsearchlogs.com

Mobility data example: spatio-temporal linkage

- [Jajodia et al. 2005]
- An anonymous trajectory occurring every working day from location A in the suburbs to location B downtown during the morning rush hours and in the reverse direction from B to A in the evening rush hours can be linked to
  - the persons who live in A and work in B;
- If locations A and B are known at a sufficiently fine granularity, it possible to identify specific persons and unveil their daily routes
  - Just join phone directories
- In mobility data, positioning in space and time is a powerful quasi identifier.

The naive scientist’s view (3)

- In the end, it is not needed to disclose the data: the (trusted) analyst only may be given access to the data, in order to produce knowledge (mobility patterns, models, rules) that is then disclosed for the public utility.
- Only **aggregated information is published**, while **source data are kept secret**.
- Since aggregated information concerns **large** groups of individuals, we are tempted to conclude that its disclosure is safe.

Wrong, once again!

- Two reasons (at least)
- For **movement patterns**, which are sets of trajectories, the control on space granularity may allow us to re-identify a small number of people
  - Privacy (anonymity) **measures** are needed!
- From **rules** with high support (i.e., concerning many individuals) it is sometimes possible to deduce new rules with very limited support, capable of identifying precisely one or few individuals

An example of rule-based linkage [Bonchi et al. 2005]

- **Age = 27 and**
  ZIP = 45254 and
  Diagnosis = HIV ⇒ Native Country = USA
  [sup = 758, conf = 99.8%]
- Apparently a safe rule:
  - **99.8% of 27-year-old people from a given geographic area that have been diagnosed an HIV infection, are born in the US.**
- But we can derive that only the 0.2% of the rule population of 758 persons are 27-year-old, live in the given area, have contracted HIV and **are not born in the US**.
  - **1 person only! (without looking at the source data)**
- The triple Age, ZIP code and Native Country is a quasi-identifier, and it is possible that in the demographic list there is only one 27-year-old person in the given area who is not born in the US (as in the governor example!)

Moral: protecting privacy when disclosing information is not trivial

- Anonymization and aggregation do not necessarily put ourselves on the safe side from attacks to privacy
- For the very same reason the problem is scientifically attractive – besides socially relevant.
- As often happens in science, the problem is to find an optimal trade-off between two conflicting goals:
  - obtain **precise, fine-grained** knowledge, useful for the analytic eyes of the Historian;
  - obtain **imprecise, coarse-grained** knowledge, useless for the sharp eyes of the Spy.

Privacy-preserving data publishing and mining

- Aim: guarantee anonymity by means of controlled transformation of data and/or patterns
  - little distortion that avoids the undesired side-effect on privacy while preserving the possibility of discovering useful knowledge.
- An exciting and productive research direction.

Privacy Preserving Data Analysis and Mining

- 4 main approaches, distinguished by the following questions:
  - *what is disclosed/published/shared?*
  - *what is hidden?*
  - *how?*

"Individual" Privacy

1. Secure Data Publishing
2. Secure Knowledge Publishing
3. Distributed Data Hiding
4. Knowledge Hiding

"Corporate" Privacy (or "Secrecy")

A very short State of the Art in PPDM

Secure Data Publishing

Secure data publishing

- What is disclosed?
  - the data (modified: generalized, randomized, …)
- What is hidden?
  - the real data
- How?
  - by perturbating the data in such a way that it is not possible the identification of original database rows (individual privacy), but it is still possible to extract **valid** knowledge (models and patterns).

- A.K.A. ***“distribution reconstruction” or “k-anonymization”***

k-Anonymity [Samarati, Sweeney 98]: Eliminate Link to Person by generalizing Quasi-identifiers

Date of Birth	Zip Code	Allergy	History of Illness
*	07030	Penicillin	Pharyngitis
08-02-57	0702*	No Allergy	Stroke
*	07030	No Allergy	Polio
08-02-57	0702*	Sulfur	Diphtheria
*	07030	No Allergy	Colitis

k(=2 in this example)-anonymous table

Property of *k*-anonymous table

- Each value of quasi-identifier attributes appears ≥ k times in the table (or it does not appear at all)

⇒ Each row of the table is hidden in ≥ *k* rows

⇒ Each person involved is hidden in ≥ *k* peers

k-Anonymity Protects Privacy

Date of Birth	Zip Code	Allergy	History of Illness
08-02-57	0702*	No Allergy	Stroke
*	07030	No Allergy	Polio
	07030	No Allergy	Colitis

08-02-57	0702*	No Allergy	Stroke

08-02-57	0702*	Sulfur	Diphtheria

Which of them is Victor's record? Confusing…

Secure data publishing biblio

- D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In Proceedings of PODS, 2001.

- R. Agrawal and R. Srikant. Privacy-preserving data mining. In Proceedings of SIGMOD 2000.

- W. Du and Z. Zhan. Using randomized response techniques for privacy-preserving data mining. In Proceedings of SIGKDD 2003.

- A. Evfimievski. Randomization in privacy preserving data mining. SIGKDD Explor. Newsl., 4(2), 2002.

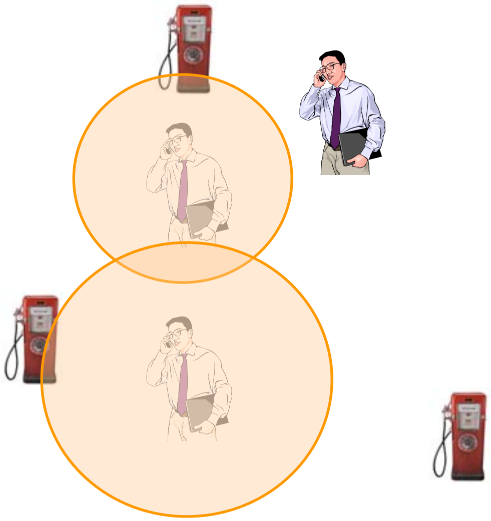
- A. Evfimievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In Proceedings of PODS 2003.

- A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In Proceedings of SIGKDD 2002.

Concepts for Location Privacy

Location Perturbation

- The user location is represented with a wrong value
- The privacy is achieved from the fact that the reported location is false
- The accuracy and the amount of privacy mainly depends on how far the reported location form the exact location



Concepts for Location Privacy

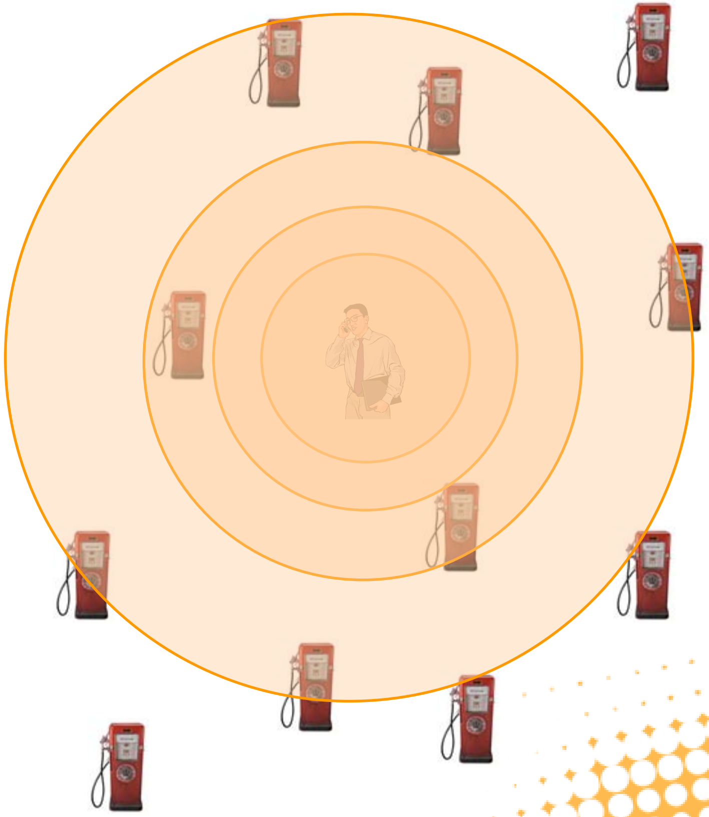
Spatial Cloaking

- Location *cloaking*, location *blurring*, location *obfuscation*

  - The user exact location is represented as a region that includes the exact user location

  - An adversary does know that the user is located in the *cloaked* region, but has no clue where the user is exactly located

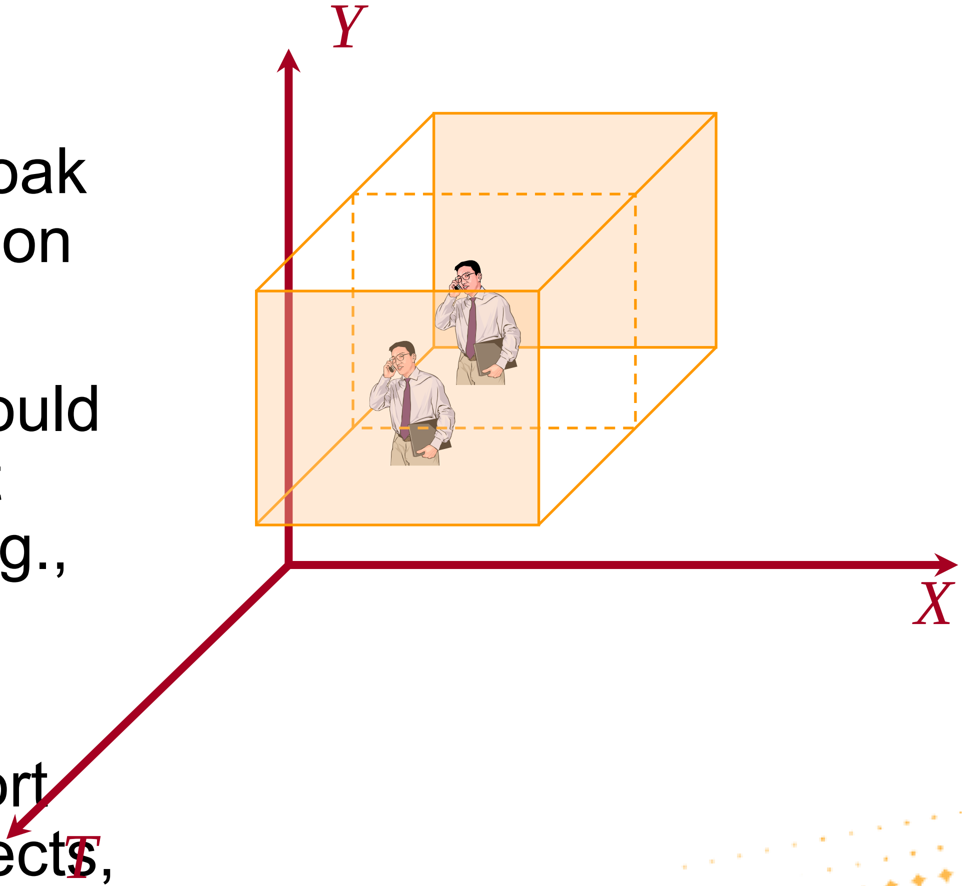
  - The area of the *cloaked* region achieves a trade-off between the user privacy and the service



Concepts for Location Privacy

Spatio-temporal Cloaking

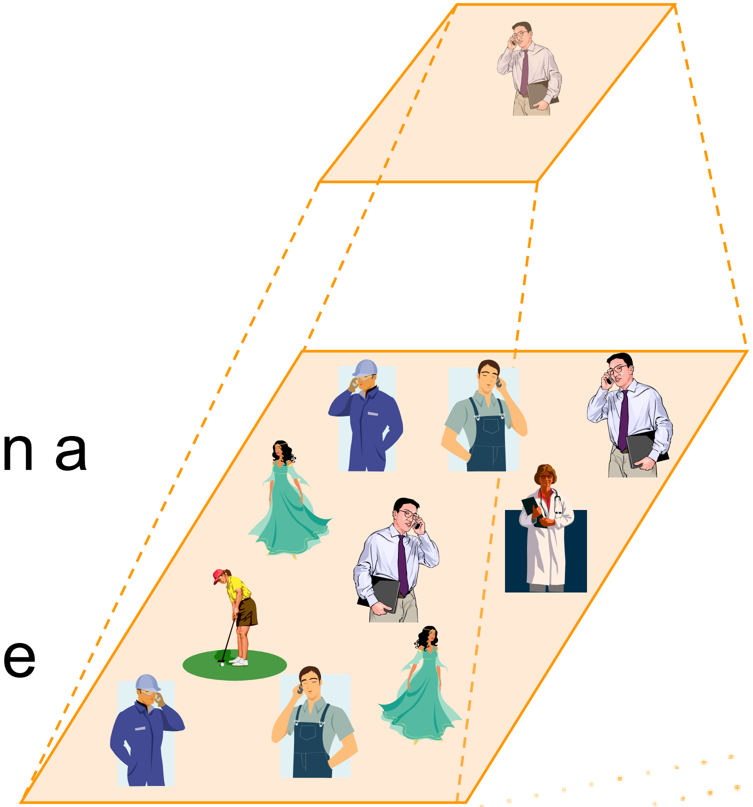
- In addition to spatial cloaking the user information can be delayed a while to cloak the temporal dimension
- Temporal cloaking could tolerate asking about stationary objects (e.g., gas stations)
- Challenging to support querying moving objects, e.g., what is my nearest gas station



Concepts for Location Privacy

k-anonymity

- The *cloaked* region contains at least *k* users
- The user is indistinguishable among other *k* users
- The cloaked area largely depends on the surrounding environment.
- A value of *k* =100 may result in a very small area if a user is located in the stadium or may result in a very large area if the user in the desert.



10-anonymity

Secure Knowledge Publishing

Information Society
Technologies

Secure Knowledge Publishing

- What is disclosed?
  - the intentional knowledge (i.e. rules/patterns/models)
- What is hidden?
  - the source data

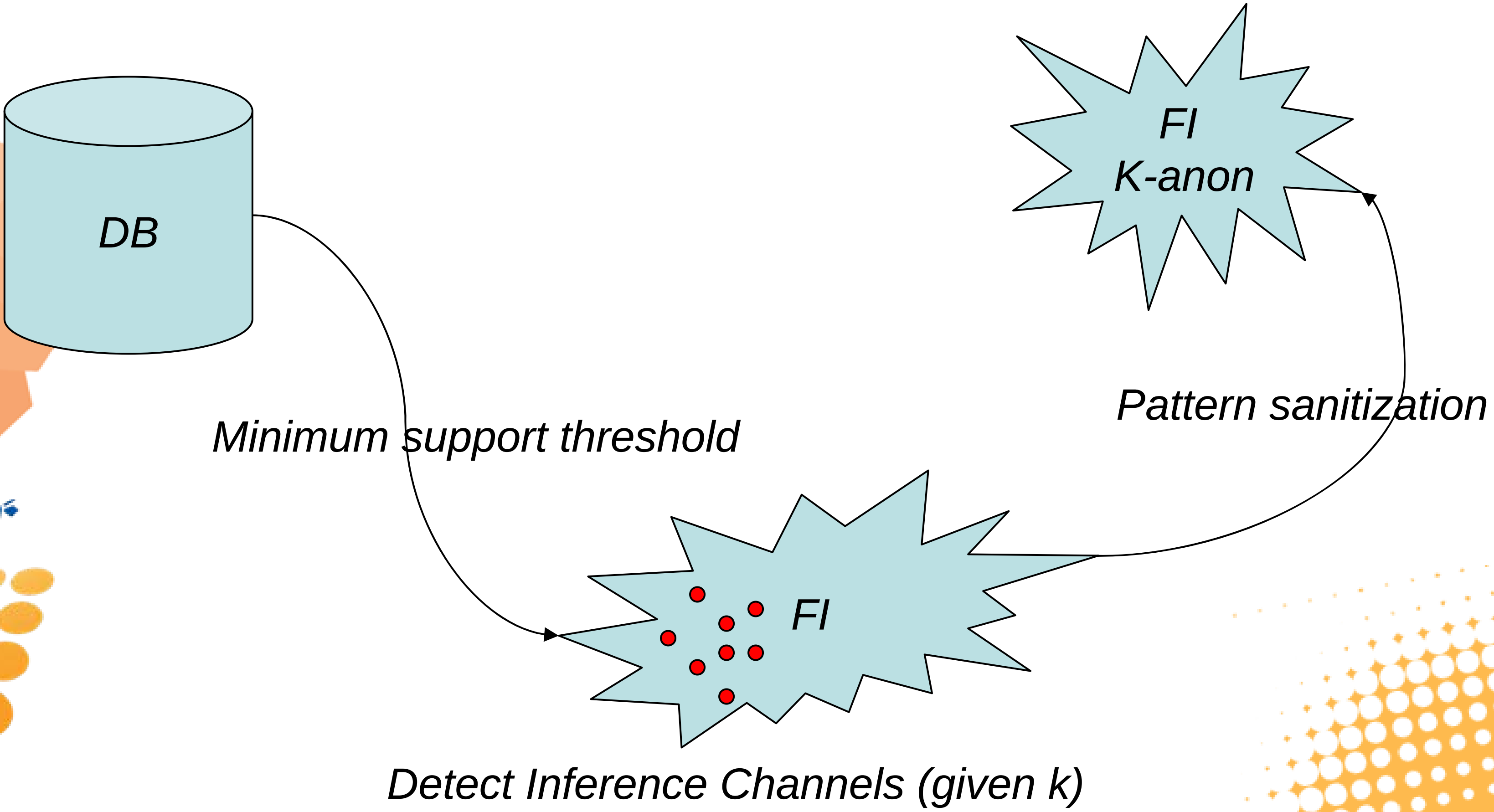
- The central question:
  “do the data mining results themselves violate privacy?”

- Focus on **individual privacy**: the individuals whose data are stored in the source database being mined.

Secure Knowledge Publishing biblio

- M. Kantarcioglu, J. Jin, and C. Clifton. When do data mining results violate privacy? In Proceedings of the tenth ACM SIGKDD, 2004.
- S. R. M. Oliveira, O. R. Zaiane, and Y. Saygin. Secure association rule sharing. In Proc.of the 8th PAKDD, 2004.
- P. Fule and J. F. Roddick. Detecting privacy and ethical sensitivity in data mining results. In Proc. of the 27° conference on Australasian computer science, 2004.
- Atzori, Bonchi, Giannotti, Pedreschi. K-anonymous patterns. In PKDD and ICDM 2005. Also VLDB Journal, to appear.

The scenario



DB
Minimum support threshold
FI
Detect Inference Channels (given k)
Pattern sanitization
FI
K-anon

Distributed Data Hiding

Information Society
Technologies

Distributed Data Hiding

- Objective?
  - computing a valid mining model from several distributed datasets, where each party (data owner) does not communicate its data to the other parties involved in the computation.
- How?
  - cryptographic techniques
- A.K.A. “*Secure Multiparty Computation*”

Distributed Data Hiding biblio

- C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y.Zhu. Tools for privacy preserving distributed data mining. SIGKDD Explor. Newsl., 4(2), 2002.
- M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. In SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD’02), 2002.
- B. Pinkas. Cryptographic techniques for privacy-preserving data mining. SIGKDD Explor. Newsl., 4(2), 2002.
- J. Vaidya and C. Clifton. Privacy preserving association rule mining in vertically partitioned data. In Proceedings of ACM SIGKDD 2002.

Knowledge Hiding

Information Society
Technologies

Knowledge Hiding

- What is disclosed?
  - the data (modified somehow)
- What is hidden?
  - some “sensitive” knowledge (i.e. secret rules/patterns)
- How?
  - usually by means of data **sanitization**
    - the data which we are going to disclose is modified,
    - in such a way that the sensitive knowledge can non longer be inferred,
    - while the original database is modified as little as possible.

Knowledge Hiding biblio

- E. Dasseni, V. S. Verykios, A. K. Elmagarmid, and E. Bertino. *Hiding association rules by using confidence and support*. In Proceedings of the 4th International Workshop on Information Hiding, 2001.
- Y. Saygin, V. S. Verykios, and C. Clifton. *Using unknowns to prevent discovery of association rules*. SIGMOD Rec., 30(4), 2001.
- S. R. M. Oliveira and O. R. Zaiane. *Protecting sensitive knowledge by data sanitization*. In Third IEEE International Conference on Data Mining (ICDM’03), 2003.

PPDM research strives for a win-win situation

- Obtaining the advantages of collective mobility knowledge without disclosing inadvertently any individual mobility knowledge.
- This result, if achieved, may have an impact on
  - laws and jurisprudence,
  - the social acceptance of ubiquitous technologies.
- This research must be tackled in a multi-disciplinary way: the opportunities and risks must be shared by social analysts, jurists, policy makers, concerned citizens.

Mobility data are a public good

- After all, mobility data are produced by people, as an effect of our own living
- The research community should promote policy makers’ awareness of the potential benefits of mobility data that can be collected by wireless networks