Data Mining2

Fosca Giannotti and Mirco Nanni
Pisa KDD Lab, ISTI-CNR & Univ. Pisa
http://www-kdd.isti.cnr.it/

DIPARTIMENTO DI INFORMATICA - Università di Pisa
anno accademico 2013/2014

Privacy: Regulations and and Privacy Aware Data Mining

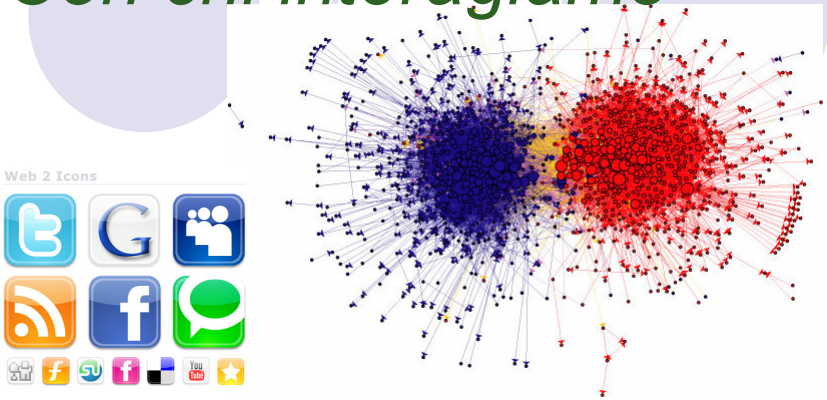
Plan of the Talk

- Privacy Constraints Sources:
  - EU rules
  - US rules
  - Safe Harbor Bridge
  - The new EU regulation (from2015)
- Privacy Constraints Types:
  - Individual (+ k-anonymity)
  - Collection (Corporate privacy)
  - Result limitation
- Classes of solutions
  - Brief State of the Art of PPDM
    - Knowledge Hiding
    - Data Perturbation and Obfuscation
    - Distributed Privacy Preserving Data Mining
    - Privacy-aware Knowledge Sharing
  - Privacy-by-design
    - Examples from mobility data-minin

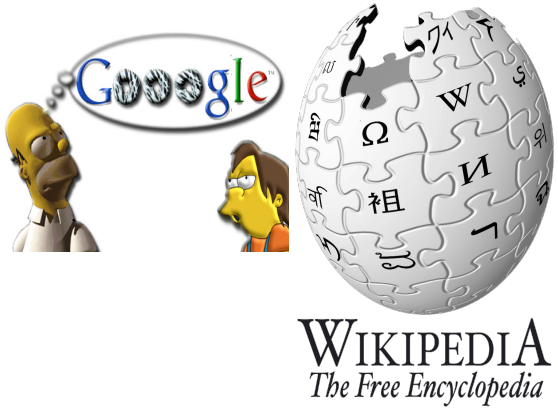

IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban dictionary
1,600+ READS ON Scribd.
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
20,000+ NEW POSTS ON tumblr.
13,000+ iPhone APPLICATIONS DOWNLOADED
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
New Craigslist Ads
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
QUESTIONS ASKED ON THE INTERNET...
100+ Answers.com
40+ YAHOO! ANSWERS
600+ NEW VIDEOS
25+ HOURS TOTAL DURATION
70+ DOMAINS REGISTERED
60+ NEW BLOGS
1,500+ BLOG POSTS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
Google
Google Search
1,700+ Firefox DOWNLOADS
695,000+ facebook STATUS UPDATES
50+ WordPress DOWNLOADS
6,600+ NEW PICTURES ARE UPLOADED ON flickr
125+ PLUGIN DOWNLOADS
79,364 WALL POSTS
510,040 COMMENTS
GO-Globe.com

Cosa compriamo

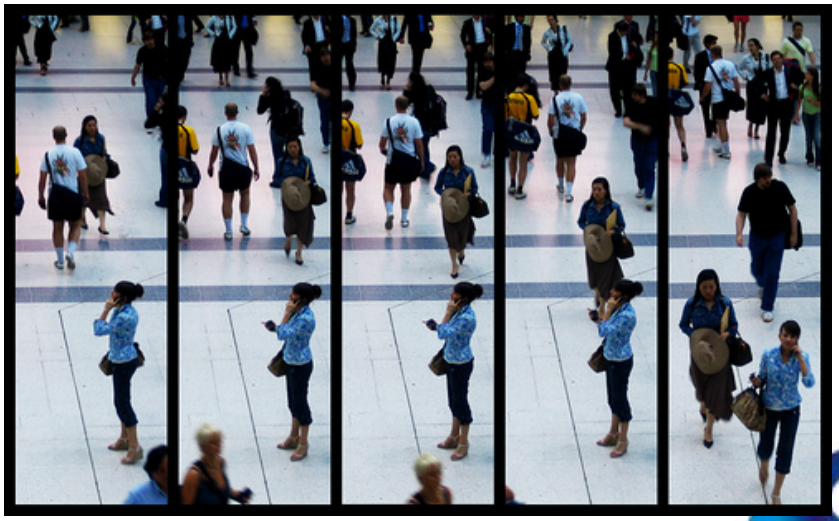
Con chi interagiamo



Cosa cerchiamo



Dove andiamo



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.

Data Scientist

- … a new kind of professional has emerged, the **data scientist**, who combines the skills of **software programmer, statistician and storyteller/artist** to extract the nuggets of gold hidden under mountains of data.

- *Hal Varian, Google's chief economist, predicts that the job of statistician will become the "sexiest" around. Data, he explains, are widely available; what is scarce is the ability to extract wisdom from them.*
- *Data Scientist has also the responsibility of being **ethically** correct*

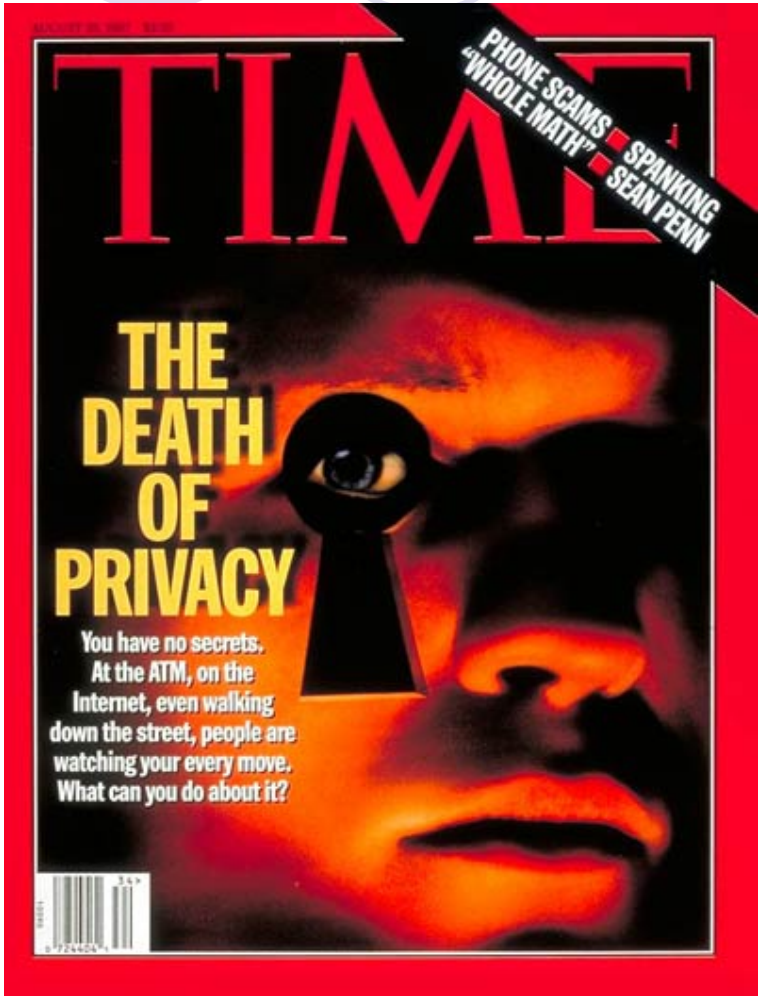
Definition of privacy

What is privacy?

Global Attention to Privacy

- Time (August 1997)
  - The Death of Privacy
- The Economist (May 1999)
  - The End of Privacy
- The European Union (October 1998)
  - Directive on Privacy Protection
- The European Union (January2012)
  - Proposal for new Directive on Privacy Protection
- New deal on personal data : World Economic Forum 2010-2013

Time: The Death of Privacy



- Invasion of privacy
  - Our right to be left alone has disappeared, bit by bit, in little brotherly steps.
  - Still, we've got something in return, and it's not all bad

Definition of privacy

What is privacy?

European legislation for protection of personal data

- European directives:
  - Data protection directive (95/46/EC) and proposal for a new EU directive (25 Jan 2012)
  - http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm
  - ePrivacy directive (2002/58/EC) and its revision (2009/136/EC)

EU: Personal Data

- *Personal data* is defined as any information relating to an identity or *identifiable* natural person.

- An *identifiable person* is one who can be identified, *directly or indirectly*, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.

EU: Processing of Personal Data

- The *processing of personal data* is defined as any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as:
  - collection,
  - recording,
  - organization,
  - storage,
  - adaptation or alteration,
  - retrieval,
  - consultation,
  - use,
  - disclosure by transmission,
  - dissemination,
  - alignment or combination,
  - blocking,
  - erasure or destruction.

EU Privacy Directive requires:

- That personal data must be processed fairly and lawfully
- That personal data must be accurate
- **That data be collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes**
- That personal data is to be kept in the form which permits identification of the subject of the data for no longer than is necessary for the purposes for which the data was collected or for which it was further processed
- That subject of the data must have given his unambiguous consent to the gathering and processing of the personal data
- If consent was not obtained from the subject of the data, that personal data be processed for the performance of a contract to which the subject of the data is a party
- That processing of personal data revealing **racial or ethnical origin, political opinions, religious or philosophical beliefs, trade union membership, and the processing of data concerning health or sex life is prohibited**

EU Privacy Directive

- Personal data is any information that can be traced directly or *indirectly* to a specific person
- Use allowed if:
  - Unambiguous consent given
  - Required to perform contract with subject
  - Legally required
  - Necessary to protect vital interests of subject
  - In the public interest, or
  - Necessary for legitimate interests of processor and doesn't violate privacy
- Some uses specifically proscribed (sensitive data)
  - Can't reveal racial/ethnic origin, political/religious beliefs, trade union membership, health/sex life

Anonymity according to 1995/46/EC

- The principles of protection must apply to any information concerning an identified or identifiable person;
- To determine whether a person is identifiable, account should be taken of *all the means likely reasonably to be used* either by the controller or by any other person to identify the said person;
- **The principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable;**

Privacy by design principle

- In many cases (e.g., all previous questions!), it is possible to reconcile the dilemma between privacy protection and knowledge sharing
  - Make data **anonymous** with reference to social mining goals
  - **Use anonymous data to extract knowledge**
  - Only a little loss in data quality often earns a strong privacy protection

ePrivacy Directive

- GOAL:
  - the protection of natural and legal persons w.r.t. the processing of personal data in connection with the provision of publicly available electronic communications services in public communications networks.

Topics related to (mobility) Data Mining

- **Location data**
  - any data processed indicating the **geographic position** of the terminal equipment of a user of a publicly available electronic communications service
- **Traffic Data**
  - any data processed for the purpose of the **conveyance of a communication** on an electronic communications network or for the billing thereof
- **Value added Services**
  - **any service** which requires the processing of traffic data or location data other than traffic data **beyond** what is necessary for the **transmission** of a communication or the **billing** thereof
    - **Examples**: *route guidance, traffic information, weather forecasts and tourist information.*

Location/Traffic Data Anonymization

- **Location data** and **Traffic data** must be erased or made **anonymous** when it is no longer needed for the purpose of the transmission of a communication and the billing

- **Location/Traffic Data anonymization** for providing Value added Services

EU Directive (95/46/EC) and new Proposal

- GOALS:
  - protection protection of individuals with regard to the processing of personal data
  - the free movement of such data

New Elements in the EU Proposal

- Principle of Transparency
- Data Portability
- Right of Oblivion
- Profiling
- Privacy by Design

Transparency & Data Portability

- **Transparency:**
  - Any information addressed to the public or to the data subject should be easily accessible and easy to understand
- **Data Portability:**
  - The right to transmit his/her personal data from an automated processing system, into another one

Oblivion & Profiling

- **Right to Oblivion:**
  - The data subject shall have the right to obtain the erasure of his/her personal data and the abstention from further dissemination of such data
- **Profiling:**
  - The right not to be subject to a measure which is based on profiling by means of automated processing

Privacy by Design

- The controller shall implement appropriate technical and organizational measures and procedures in such a way that the data processing
  - will meet the requirements of this Regulation
  - will ensure the protection of the rights of the data subject

Privacy by Design in Data Mining

- Design frameworks
  - to counter the threats of privacy violation
  - without obstructing the knowledge discovery opportunities of data mining technologies
- Trade-off between privacy quantification and data utility

Privacy by Design in Data Mining

- *The framework is designed with assumptions about*
  - *The **sensitive data** that are the subject of the analysis*
  - *The **attack model**, i.e., the knowledge and purpose of a malicious party that wants to discover the sensitive data*
  - *The **target analytical questions** that are to be answered with the data*

- *Design a privacy-preserving framework able to*
  - *transform the data into an anonymous version with a **quantifiable privacy guarantee***
  - *guarantee that the analytical questions can be answered correctly, within a **quantifiable** approximation that specifies the **data utility***

Plan of the Talk

- Privacy Constraints Sources:
  - EU rules
  - US rules
  - Safe Harbor Bridge
  - The new EU regulation (from2015)
- Privacy Constraints Types:
  - Individual (+ k-anonymity)
  - Collection (Corporate privacy)
  - Result limitation
- Classes of solutions
  - Brief State of the Art of PPDM
    - Knowledge Hiding
    - Data Perturbation and Obfuscation
    - Distributed Privacy Preserving Data Mining
    - Privacy-aware Knowledge Sharing

Opportunities and threats

- Knowledge may be discovered from the traces left behind by mobile users in the information systems of wireless networks.
- Knowledge, in itself, is neither good nor bad.
- What knowledge to be searched from digital traces? For what purposes?
- Which **eyes** to look at these traces with?

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 naive scientist's view (1)

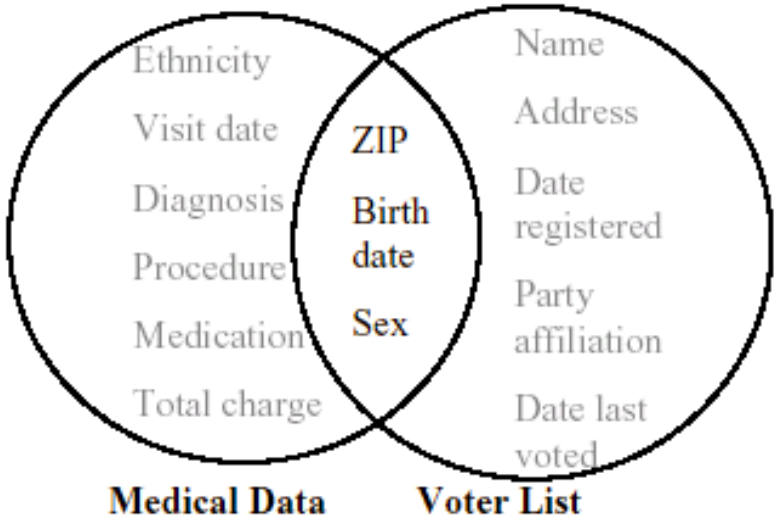
- Knowing the exact identity of individuals is not needed for analytical purposes
  - Anonymous trajectories are enough to reconstruct aggregate movement behaviour, pertaining to groups of people.
- Is this reasoning correct?
- Can we conclude that the analyst runs no risks, while working for the public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

- She purchased the voter registration list for Cambridge Massachusetts
  - 54,805 people



- 69% unique on postal code and birth date
- 87% US-wide with all three (ZIP + birth date + Sex)

- Solution: *k*-anonymity
  - Any combination of values appears at least *k* times
- Developed systems that guarantee k-anonymity
  - Minimize distortion of results

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 [Atzori 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 publishing : K-Anonymity

Motivation: 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

Security Threat: May 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!

k-Anonymity [SS98]: Eliminate Link to Person through 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
08-02-57	0702*	No Allergy	Stroke
*	07030	No Allergy	Polio
08-02-57	0702*	Sulfur	Diphtheria
	07030	No Allergy	Colitis

Which of them is Victor's record?
Confusing…

k-anonymity – Problem Definition

- Input: Database consisting of n rows, each with m attributes drawn from a finite alphabet.
- Assumption: the data owner knows/indicates which of the m attributes are *Quasi-Identifiers.*
- Goal: trasform the database in such a way that is K-anonymous w.r.t. a given *k,* and the QIs.
- How: By means of generalization and suppression.
- Objective: Minimize the distortion.
- Complexity: NP-Hard.
- A lot of papers on k-anonymity in 2004-2006

(SIGMOD, VLDB, ICDE, ICDM)

Privacy Preserving Data Mining: Short State of the Art

Privacy Preserving Data Mining

- Very Short Definition:

"the study of data mining side-effects on privacy"

- A Bit Longer Definition:

  "the study of how to produce valid mining models and patterns without disclosing private information"

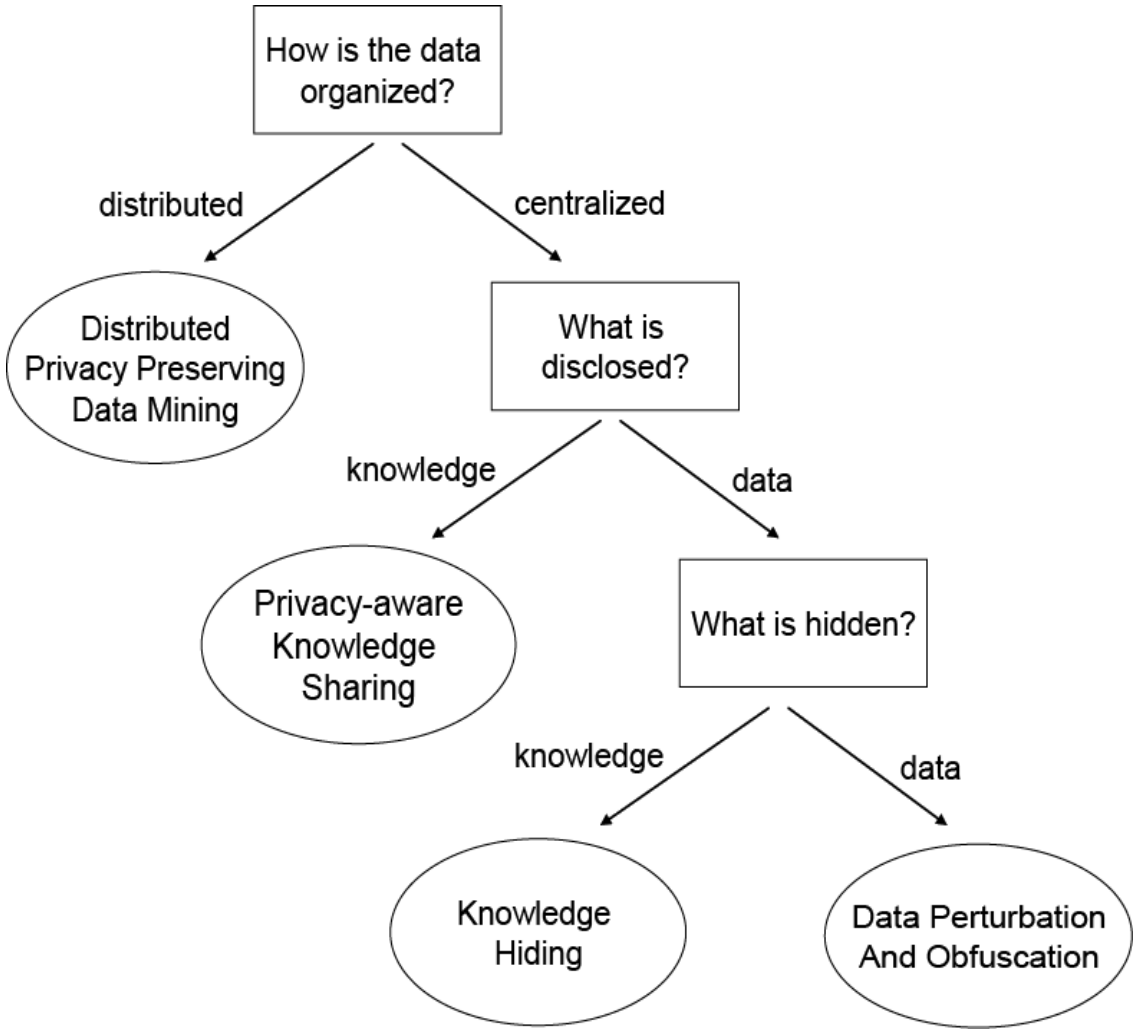
  - *Requires to define what is "private"…*
  - *Many different definitions…*
  - *… many different aproaches to Privacy Preserving Data Mining*

Privacy Preserving Data Mining

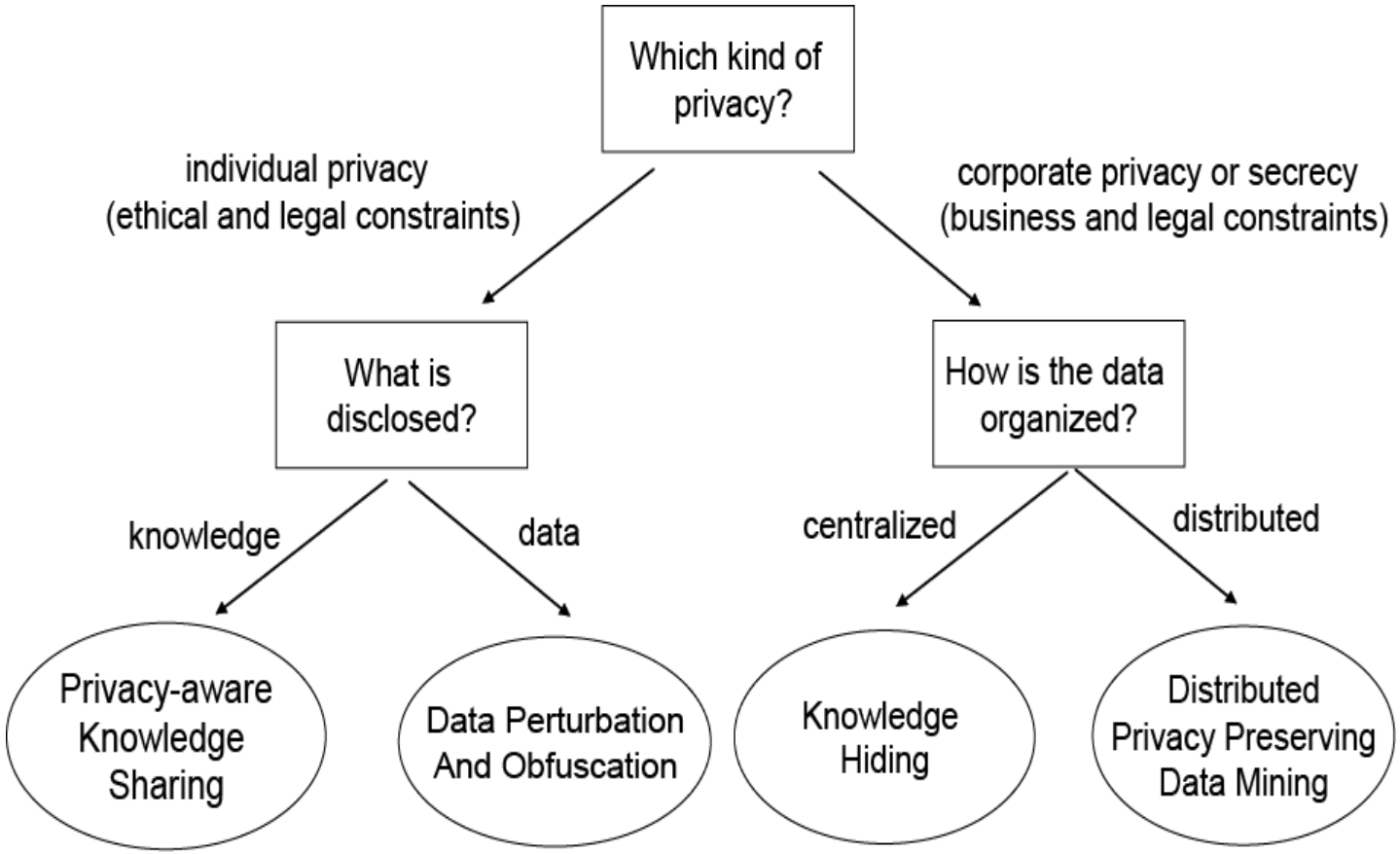
- We identify 4 main approaches, distinguished by the following questions:
  - *what is disclosed/published/shared?*
  - *what is hidden?*
  - *how is the data organized? (centralized or distributed)*

1. Knowledge Hiding
2. Data Perturbation and Obfuscation
3. Distributed Privacy Preserving Data Mining
4. Privacy-aware Knowledge Sharing

A taxonomy tree…



And another one…



Knowledge Hiding

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 less as possible.

Knowledge Hiding: Association Rules

- This approach can be instantiated to association rules as follows:
  - *D* source database;
  - *R* a set of association rules that can be mined from *D;*
  - R_h a subset of *R* which must be hidden.

  - Problem: how to transform *D* into *D'* (the database we are going to disclose) in such a way that *R/* R_h can be mined from *D'.*

Knowledge Hiding

- 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.
- O. Abul, M. Atzori, F. Bonchi, F. Giannotti: *Hiding Sequences.* ICDE Workshops 2007

Hiding association rules by using confidence and support

E. Dasseni, V. S. Verykios,
A. K. Elmagarmid, and E. Bertino

Scenario

User

Data Mining

Association Rules

Hide Sensitive Rules

Changed Database

Association Rule Discovery

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of literals, called items.

A set of items $X \subset I$ is called an itemset.

Let D be a set of transactions, where each transaction T is an itemset such that $T \subseteq I$.

A transaction T contains an itemset X, if $X \subseteq T$.

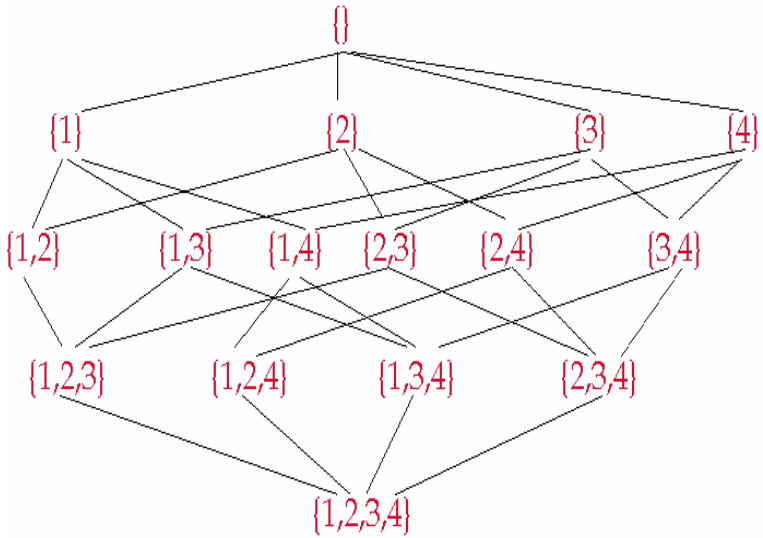
Knowledge Hiding

Consider a transactional database **D** involving a set of transactions **T**. Each transaction involves some items from the set **I** = {1,2,3,4}.

Association Rule Mining is the data mining process involving the identification of sets of items (a.k.a. itemsets) that frequently co-occur in the set of transactions T (a.k.a. frequent itemset mining), and constructing rules among them that hold under certain levels of support and confidence.

The whole set of potentially frequent itemsets involving 4 items is demonstrated in the lattice structure shown below. The original database **D** is also presented.

D	{1}	{2}	{3}	{4}
T1	1	1	0	0
T2	0	1	0	1
T3	1	0	1	1
T4	1	0	0	1
T5	1	1	0	0
T6	0	1	1	0
T7	0	0	1	0



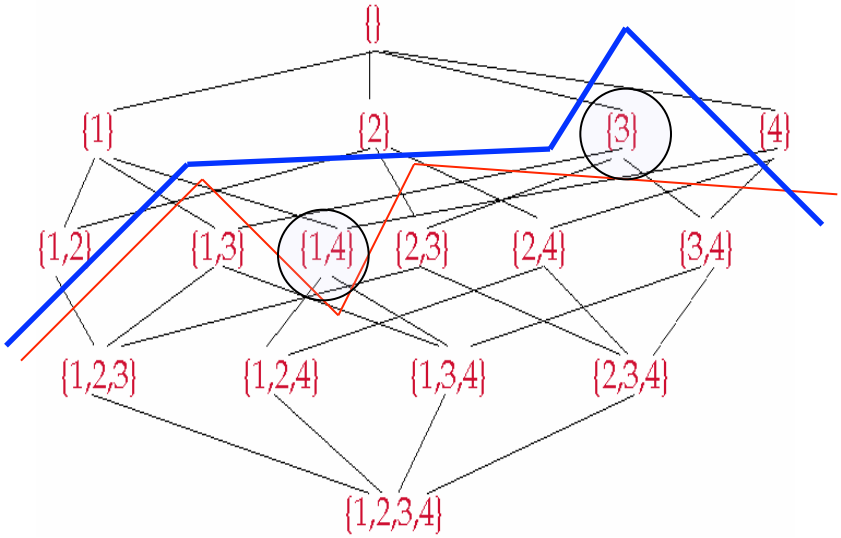
Suppose that we set the *minimum support count* to 2. Then, the following itemsets are said to be *frequent*:

itemset	support
{1}	4
{2}	4
{3}	3
{4}	3
{1,2}	2
{1,4}	2

We separate the *frequent* from the *infrequent* itemsets in the lattice, using a *borderline* (red color).

Now, suppose that itemsets {3} and {1,4} are *sensitive*, meaning that they contain knowledge which the owner of the data wants to keep private!

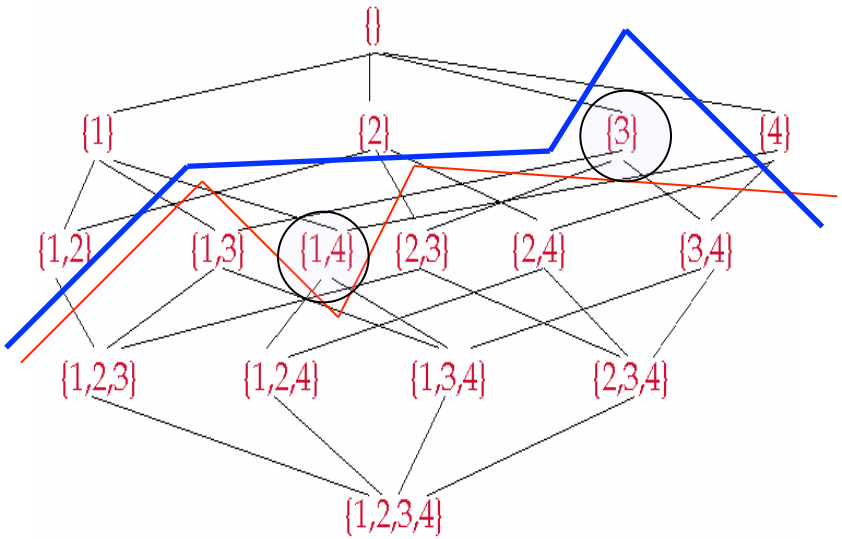
To do so, one needs to make sure that no rules will be produced by Apriori that contain *any* of these item sets.



The new – *ideal borderline* is shown in the lattice in blue color.

In order to hide all sensitive rules, the *supporting* sensitive itemsets need to be made infrequent in D. This is accomplished through *data sanitization*, by selectively altering transactions in D that support these itemsets.

D	{1}	{2}	{3}	{4}
T1	1	1	0	0
T2	0	1	0	1
T3	?	0	?	?
T4	?	0	0	?
T5	1	1	0	0
T6	0	1	?	0
T7	0	0	?	0



An intermediate form of the database is shown above, where all transactions supporting sensitive item sets {3} and {1,4} have the corresponding '1's turned into '?'. Some of these '?' will later on be turned into zeros, thus reducing the support of the sensitive item sets.

Heuristics exist to properly select which of the above transactions, namely {T3, T4, T6, T7} will be *sanitized*, to which *extent* (meaning how many items will be affected) and in which relative *order*, to ensure that the resulting database no longer allows the identification of the sensitive item sets (hence the production of sensitive rules) at the same support threshold.

Knowledge Hiding

- Heuristics do not guarantee (in any way) the identification of the best possible solution. However, they are usually fast, generally computationally inexpensive and memory efficient, and tend to lead to good overall solutions.

- An important aspect in knowledge hiding is that a solution always exists! This means that whichever itemsets (or rules) an owner wishes to hide prior sharing his/her data set with others, there is an applicable database D' that will allow this to happen. The easiest way to see that is by turning all '1's to '0's in all the 'sensitive' items of the transactions supporting the sensitive itemsets.

- Since a solution always exists, the target of knowledge hiding algorithms is to successfully hide the sensitive knowledge while minimizing the impact the sanitization process has on the non-sensitive knowledge!

- Several heuristics can be found in the scientific literature that allow for efficient hiding of sensitive itemsets and rules.

Data Perturbation and Obfuscation

Data Perturbation and Obfuscation

- What is disclosed?
  - the data (modified somehow)
- 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** intensional knowledge (models and patterns).
  - A.K.A. ***"distribution reconstruction"***

Data Perturbation and Obfuscation

- R. Agrawal and R. Srikant. Privacy-preserving data mining. In Proceedings of SIGMOD 2000.
- D. Agrawal and C. C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In Proceedings of PODS, 2001.
- W. Du and Z. Zhan. Using randomized response techniques for privacy-preserving data mining. In Proceedings of SIGKDD 2003.
- 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.
- Kun Liu, Hillol Kargupta, and Jessica Ryan. Random Projection-based Multiplicative Perturbation for Privacy Preserving Distributed Data Mining. IEEE Transactions on Knowledge and Data Engineering (TKDE), VOL. 18, NO. 1.
- K. Liu, C. Giannella and H. Kargupta. An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining. In Proceedings of PKDD'06

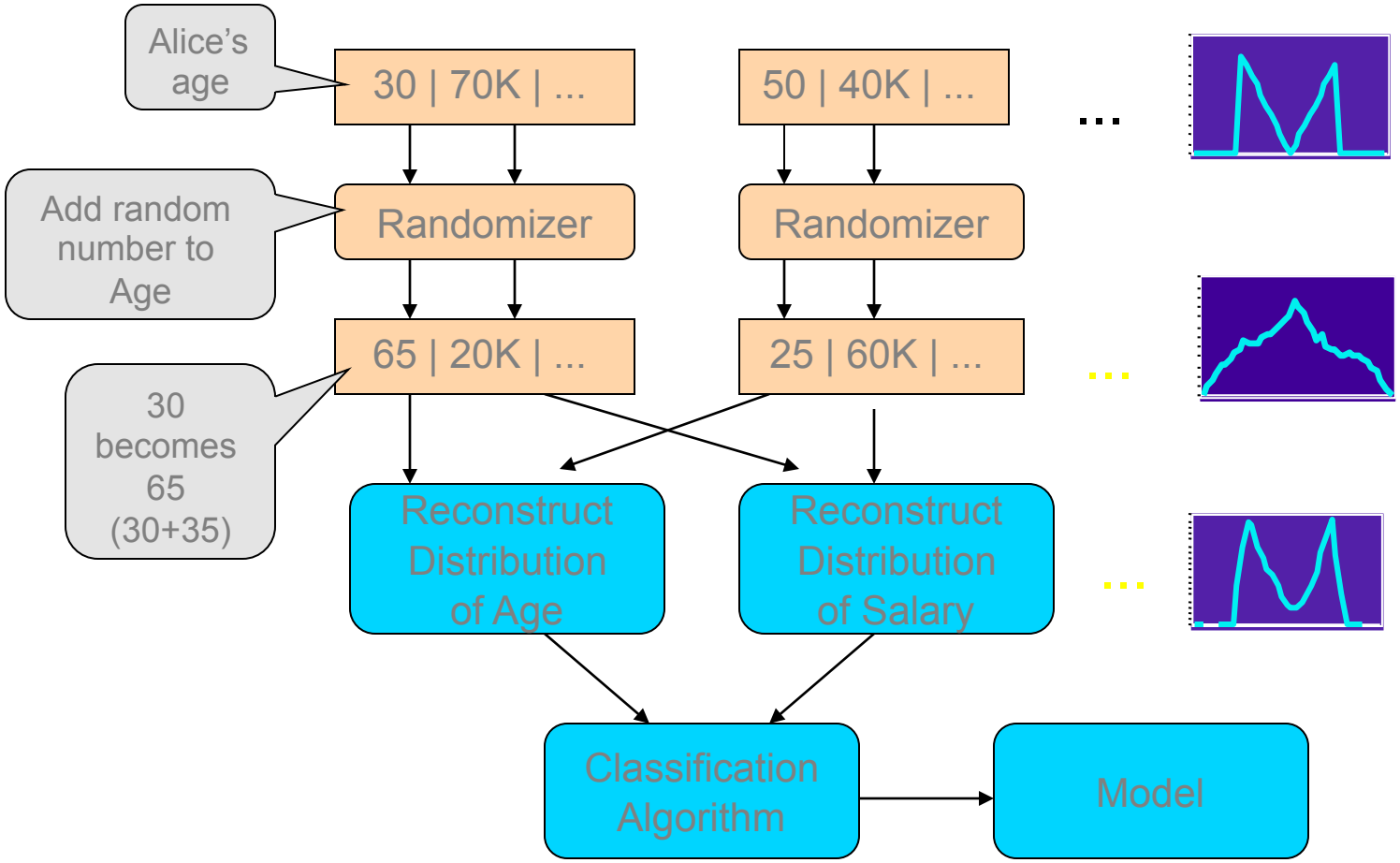
Data Perturbation and Obfuscation

- This approach can be instantiated to association rules as follows:
  - *D* source database;
  - *R* a set of association rules that can be mined from *D;*
  - Problem: define two algorithms *P* and M_P such that
    - *P(D)* = *D'* where *D'* is a database that do not disclose any information on singular rows of *D*;
    - $M_P(D') = R$

Decision Trees
Agrawal and Srikant '00

- Assume users are willing to
  - Give true values of certain fields
  - Give modified values of certain fields
- Practicality
  - 17% refuse to provide data at all
  - 56% are willing, as long as privacy is maintained
  - 27% are willing, with mild concern about privacy
- Perturb Data with Value Distortion
  - User provides x_i+r instead of x_i
  - r is a random value
    - Uniform, uniform distribution between $[-\alpha, \alpha]$
    - Gaussian, normal distribution with $\mu = 0, \sigma$

Randomization Approach Overview



Reconstruction Problem

- Original values $x_1, x_2, \ldots, x_n$
  - from probability distribution X (unknown)
- To hide these values, we use $y_1, y_2, \ldots, y_n$
  - from probability distribution Y
- Given
  - $x_1+y_1, x_2+y_2, \ldots, x_n+y_n$
  - the probability distribution of Y

Estimate the probability distribution of X.

Intuition (Reconstruct single point)

- Use Bayes' rule for density functions

Intuition (Reconstruct single point)

- Use Bayes' rule for density functions

Reconstructing the Distribution

- Combine estimates of where point came from for all the points:
  - Gives estimate of original distribution.

$$f_X = \frac{1}{n}\sum_{i=1}^{n}\frac{f_Y((x_i + y_i) - a) f_X^j(a)}{\int_{-\infty}^{\infty} f_Y((x_i + y_i) - a) f_X^j(a)}$$

Reconstruction: Bootstrapping

f_X^0 := Uniform distribution

j := 0 // Iteration number

repeat

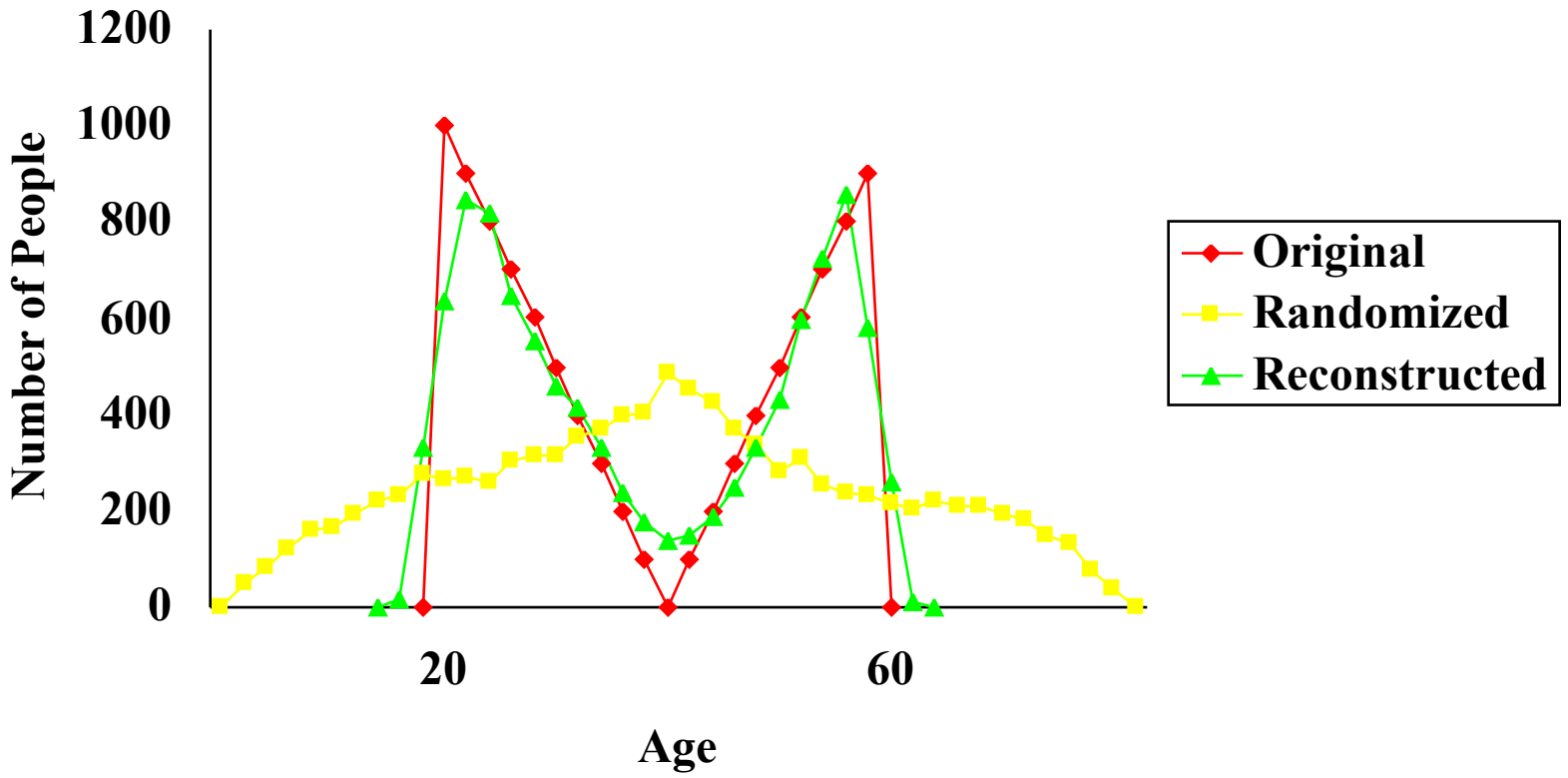
$f_X^{j+1}(a) := \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y((x_i + y_i) - a) f_X^j(a)}{\int_{-\infty}^{\infty} f_Y((x_i + y_i) - a) f_X^j(a)}$ (Bayes' rule)

j := j+1

until (stopping criterion met)

- Converges to maximum likelihood estimate.

Works well


1200
1000
800
600
400
200
0
Number of People
20
60
Age
Original
Randomized
Reconstructed

Recap: Why is privacy preserved?

- Cannot reconstruct individual values accurately.
- Can only reconstruct distributions.

Distributed Privacy Preserving Data Mining

Distributed Privacy Preserving Data Mining

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

Distributed Privacy Preserving Data Mining

- 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.

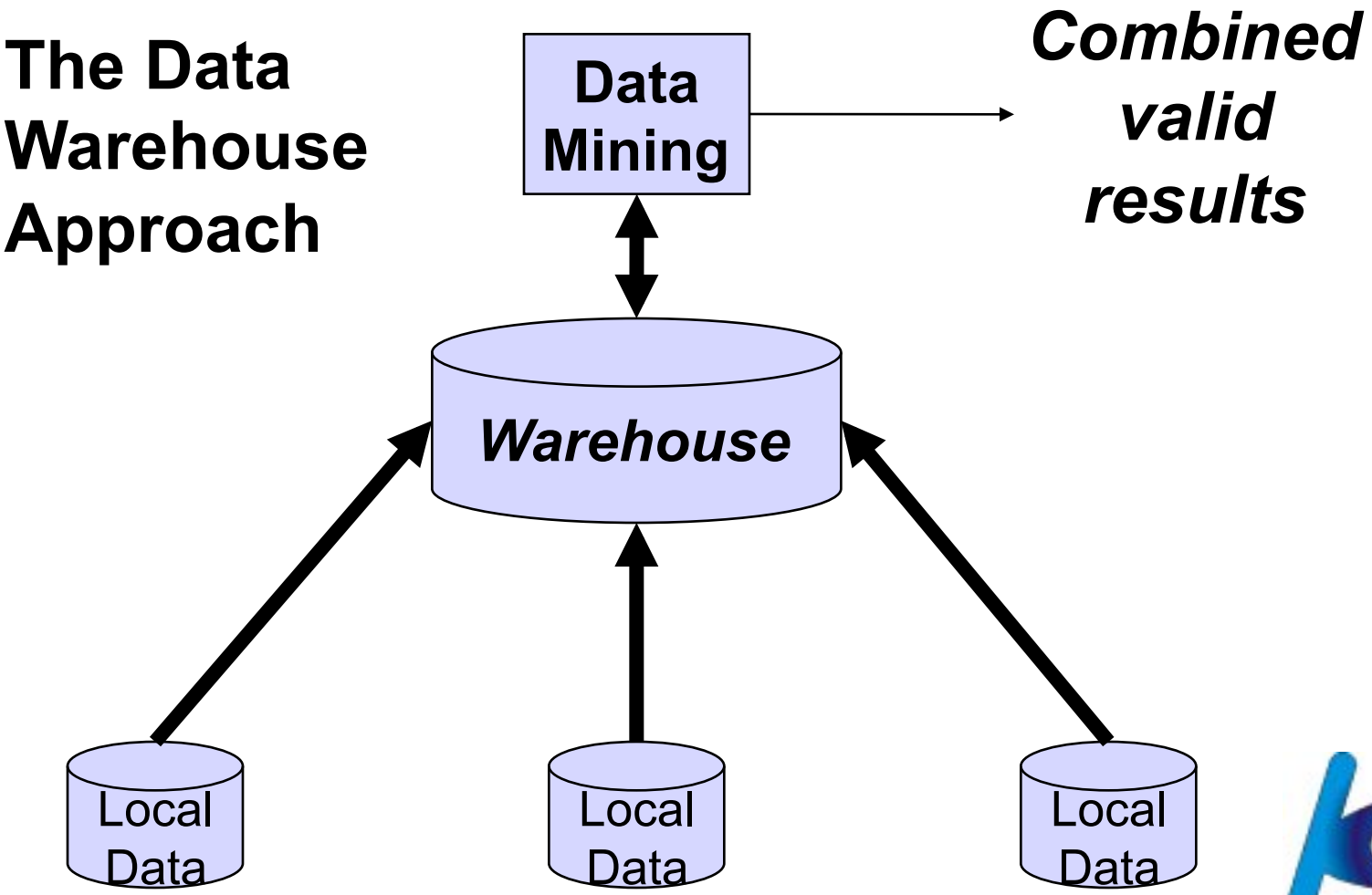
Distributed Privacy Preserving Data Mining

- This approach can be instantiated to association rules in two different ways corresponding to two different data partitions: vertically and horizontally partitioned data.

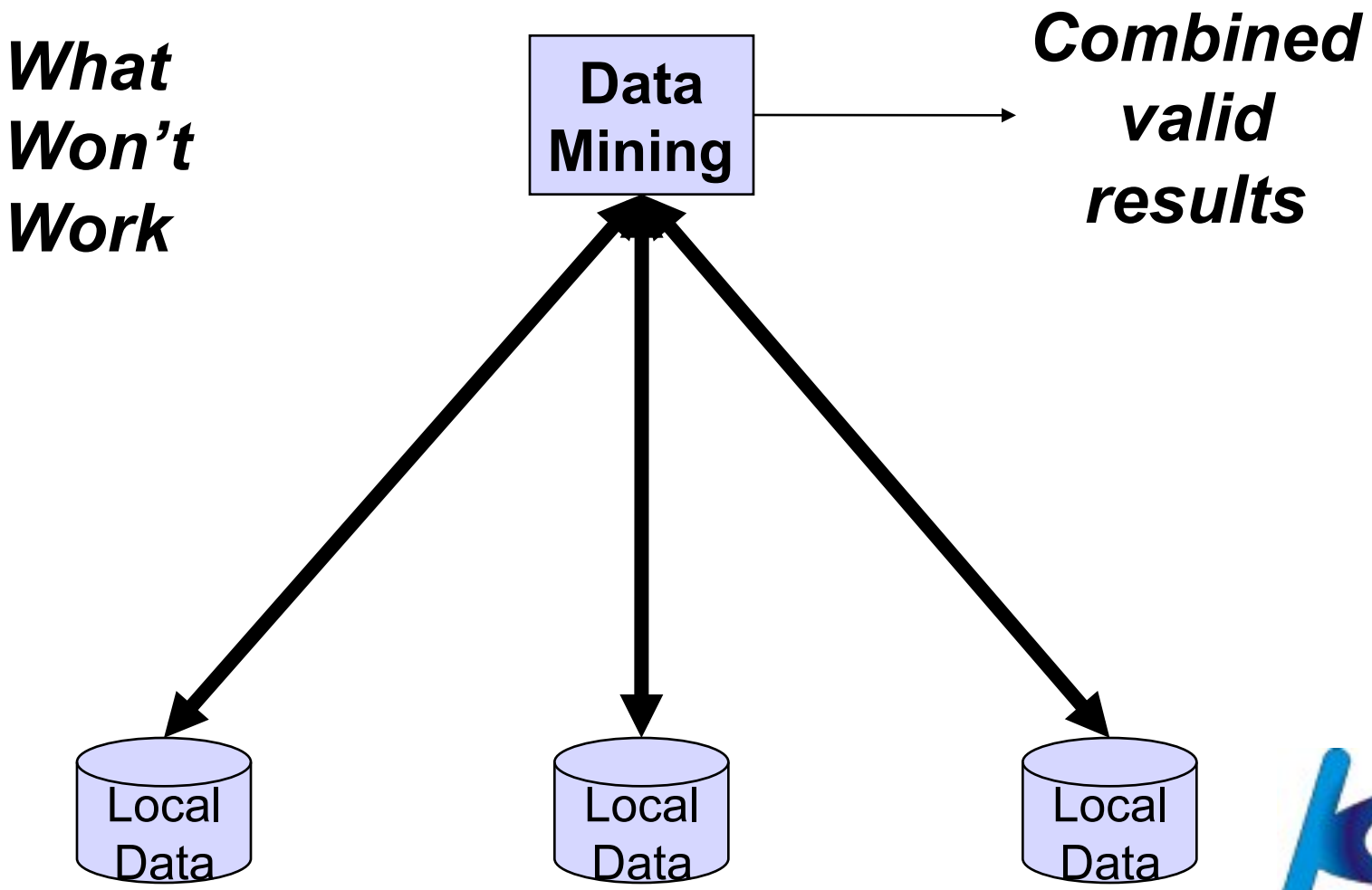
1. Each site *s* holds a portion *Is* of the whole vocabulary of items *I*, and thus each itemset is split between different sites. In such situation, the key element for computing the support of an itemset is the“secure” scalar product of vectors representing the subitemsets in the parties.

2. The transactions of *D* are partitioned in *n* databases *D1, . . . ,Dn*, each one owned by a different site involved in the computation. In such situation, the key elements for computing the support of itemsets are the “secure”union and “secure” sum operations.

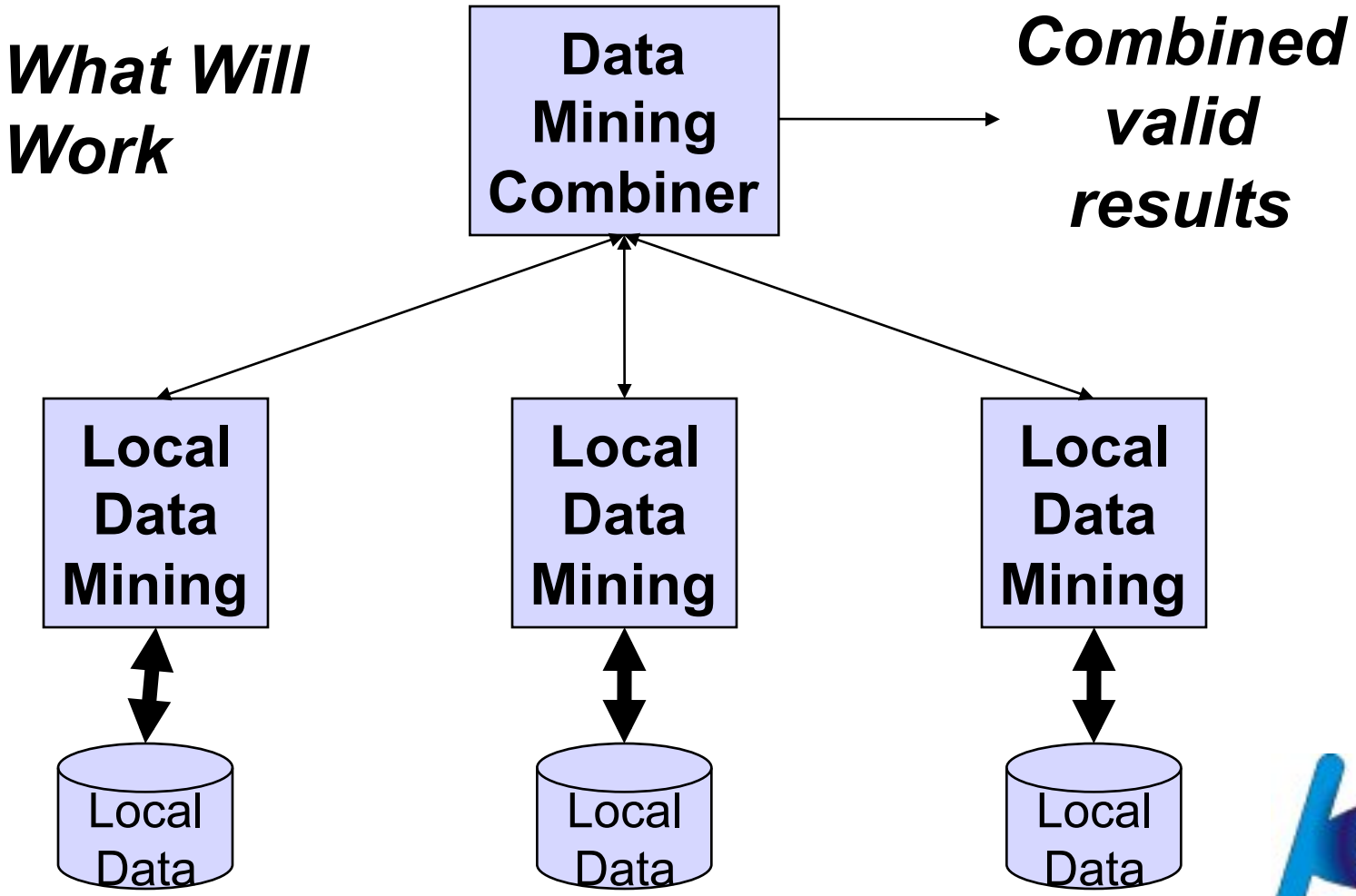
Distributed Data Mining: The "Standard" Method



Private Distributed Mining: What is it?



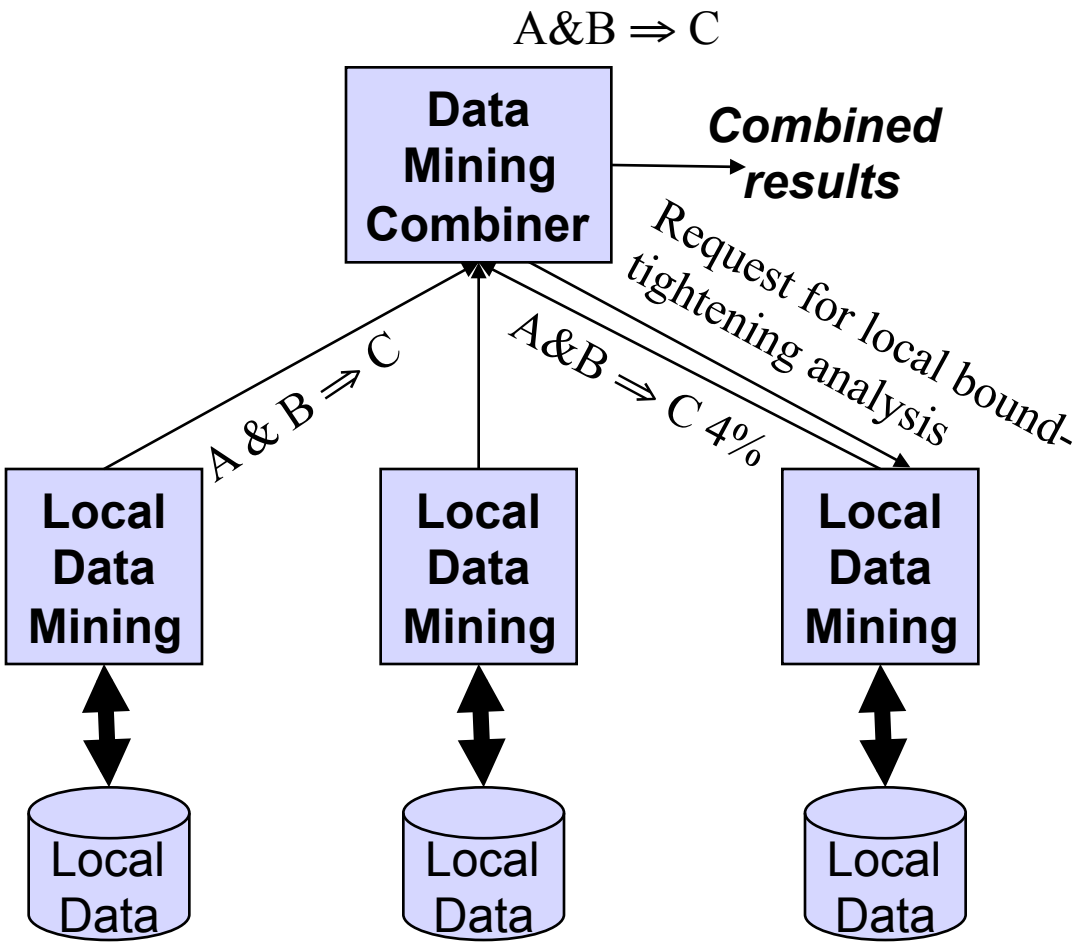
Private Distributed Mining: What is it?



Example: *Association Rules*

- Assume data is horizontally partitioned
  - Each site has complete information on a set of entities
  - Same attributes at each site
- If goal is to avoid disclosing entities, problem is easy
- Basic idea: Two-Phase Algorithm
  - First phase: Compute candidate rules
    - Frequent globally ⇒ frequent at some site
  - Second phase: Compute frequency of candidates

Association Rules in Horizontally Partitioned Data



Privacy-aware Knowledge Sharing

Privacy-aware Knowledge Sharing

- 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.

Privacy-aware Knowledge Sharing

- 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, The VLDB Journal (accepted for publication).
- A. Friedman, A. Schuster and R. Wolff. *k*-Anonymous Decision Tree Induction. In Proc. of PKDD 2006.

Privacy-aware Knowledge Sharing

- Association Rules can be dangerous…

Example

$$a_1 \wedge a_2 \wedge a_3 \Rightarrow a_4 \quad [sup = 80,\ conf = 98.7\%]$$

$$sup(\{a_1, a_2, a_3\}) = \frac{sup(\{a_1, a_2, a_3, a_4\})}{conf} \approx \frac{80}{0.987} = 81.05$$

In other words, we know that there is just one individual for which the pattern $a_1 \wedge a_2 \wedge a_3 \wedge \neg a_4$ holds.

- How to solve this kind of problems?

Privacy-aware Knowledge Sharing

- Association Rules can be dangerous…

Age = 27, Postcode = 45254, Christian ⇒ American
(support = 758, confidence = 99.8%)

Age = 27, Postcode = 45254 ⇒ American
(support = 1053, confidence = 99.9%)

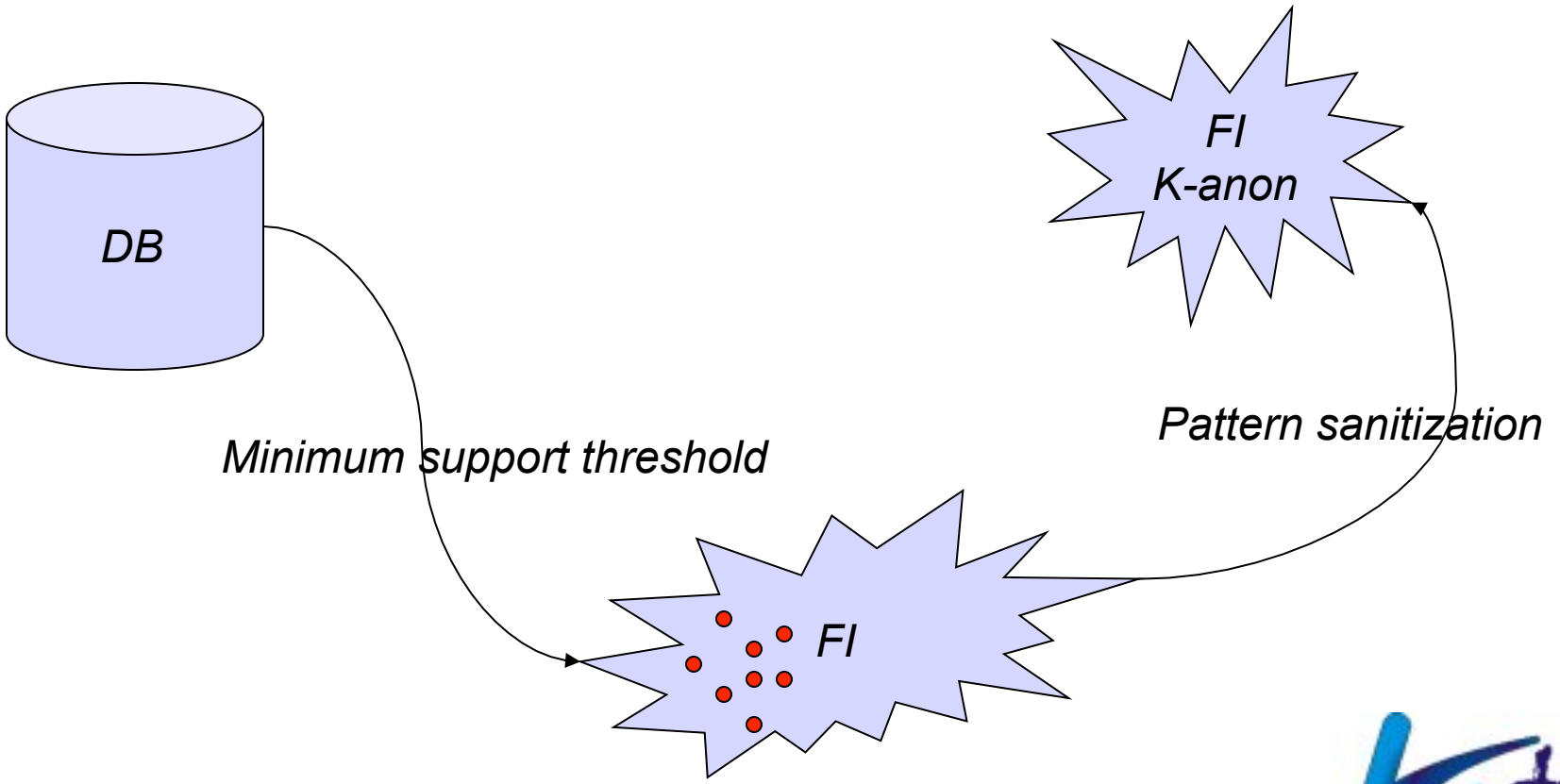
Since *sup(rule) / conf(rule)* = *sup(head)* we can derive:

Age = 27, Postcode = 45254, not American ⇒ Christian
(support = 1, confidence = 100.0%)

This information refers to my France neighbor…. he is Christian!
(and this information was clearly not intended to be released as it links public information regarding few people to sensitive data!)

- How to solve this kind of problems?

The scenario



Detecting Inference Channels

- See Atzori et al. K-anonymous patterns

$$p = i_1 \wedge \cdots \wedge i_m \wedge \neg a_1 \wedge \cdots \wedge \neg a_n$$

$$sup_{\mathcal{D}}(p) = \sum_{I \subseteq X \subseteq J} (-1)^{|X \setminus I|} sup_{\mathcal{D}}(X) \qquad f_I^J(\mathcal{D})$$

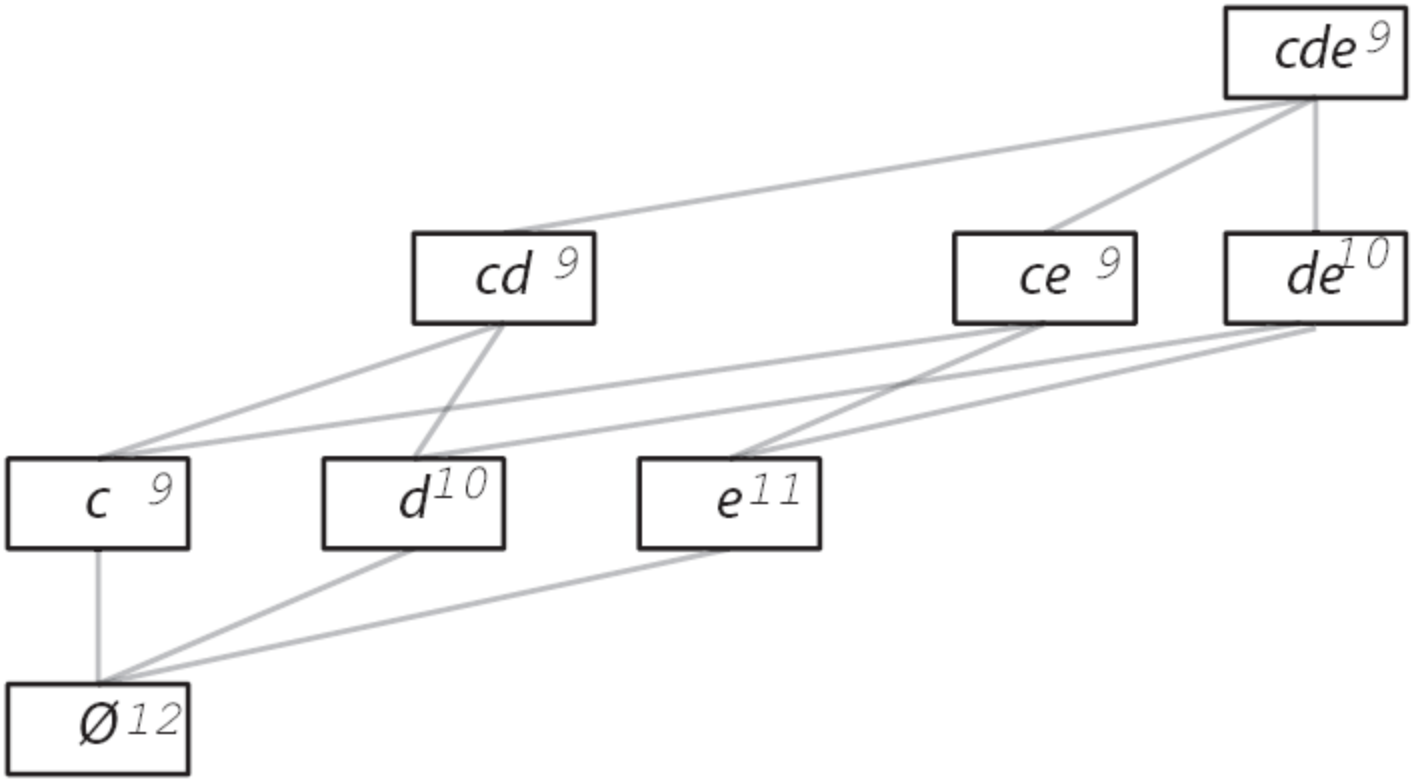
$$I = \{i_1, \ldots, i_m\} \qquad J = I \cup \{a_1, \ldots, a_n\}$$

- ✓ inclusion-exclusion principle used for support inference
- ✓ support inference as key attacking technique

- ✓ inference channel: $\{\langle X, sup_{\mathcal{D}}(X)\rangle \mid I \subseteq X \subseteq J\}$
  such that: $0 < f_I^J(\mathcal{D}) < k$

Picture of an inference channel

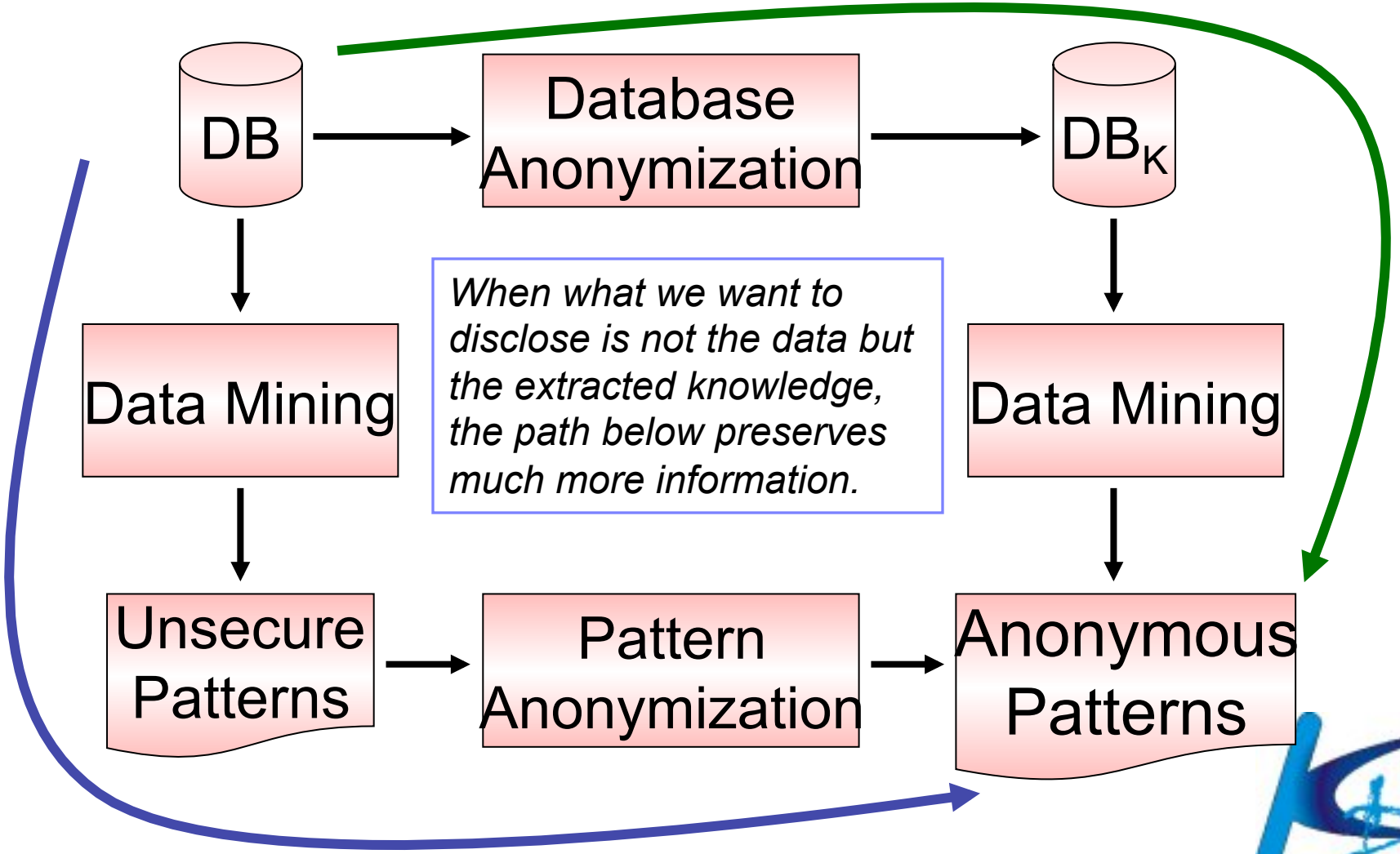
$$sup_{\mathcal{D}}(\mathcal{C}_{\emptyset}^{cde}) = f_{\emptyset}^{cde}(\mathcal{D}) = sup_{\mathcal{D}}(\emptyset) - sup_{\mathcal{D}}(c) - sup_{\mathcal{D}}(d) - sup_{\mathcal{D}}(e) + sup_{\mathcal{D}}(cd) + sup_{\mathcal{D}}(ce) + sup_{\mathcal{D}}(de) - sup_{\mathcal{D}}(cde) = 12 - 9 - 10 - 11 + 9 + 9 + 10 - 9 = 1.$$



Blocking Inference Channels

- Two patterns sanitization algorithms proposed: Additive (ADD) and Suppressive (SUP)
- ADD and SUP algorithms block anonymity threats, by merging inference channels and then modifying the original support of patterns. ADD increments the support of infrequent patterns, while SUP suppresses the information about infrequent data.
- ADD: for each inference channel $\mathcal{C}_I^J$ the support of *I* is increased to obtain $f_I^J > k$. The support of all its subsets is increased accordingly, in order to mantain database compatibility.
- *Property: ADD maintain the exactly same set of frequent itemsets, with just some slightly changed support.*

Privacy-aware Knowledge Sharing



The reform of EC data protection directive

- New proposed directive submitted to European Parliament on Jan 25, 2012, approval process expected to complete within 2 years
- http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm
- Topics related the new deal on data:
  - Data portability
  - Right to oblivion
  - Profiling and automated decision making
  - Privacy-by-design

Privacy by design principle

- In many cases (e.g., all previous questions!), it is possible to reconcile the dilemma between privacy protection and knowledge sharing
  - Make data **anonymous** with reference to social mining goals
  - **Use anonymous data to extract knowledge**
  - Only a little loss in data quality often earns a strong privacy protection

Privacy by Design in Mobility Atlas

A. Monreale, G. Andrienko, N. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo

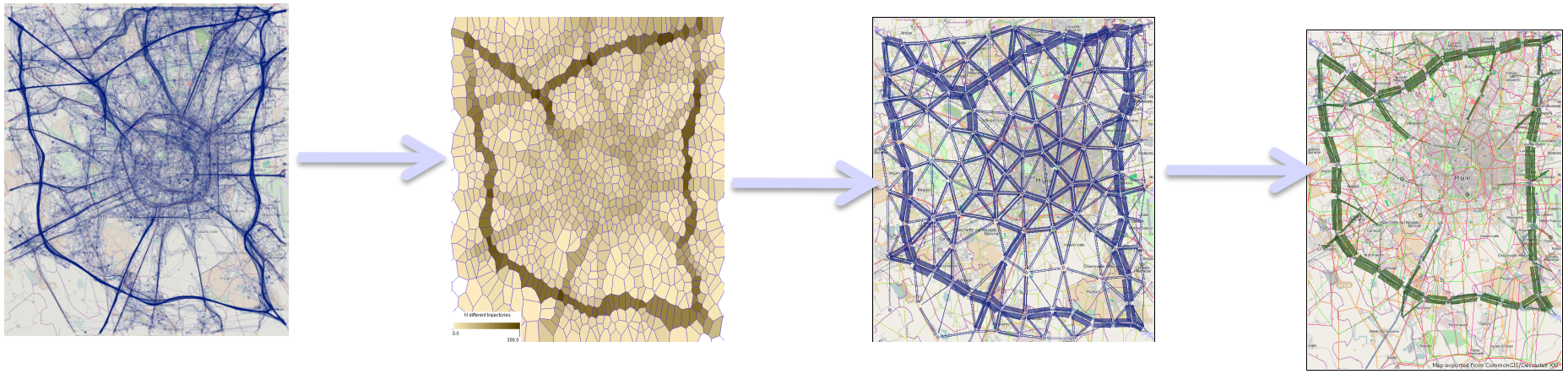
The Journal Transactions on Data Privacy, 2010

Knowledge Discovery and Delivery Lab
(ISTI-CNR & Univ. Pisa)
www-kdd.isti.cnr.it

Privacy-Preserving Framework

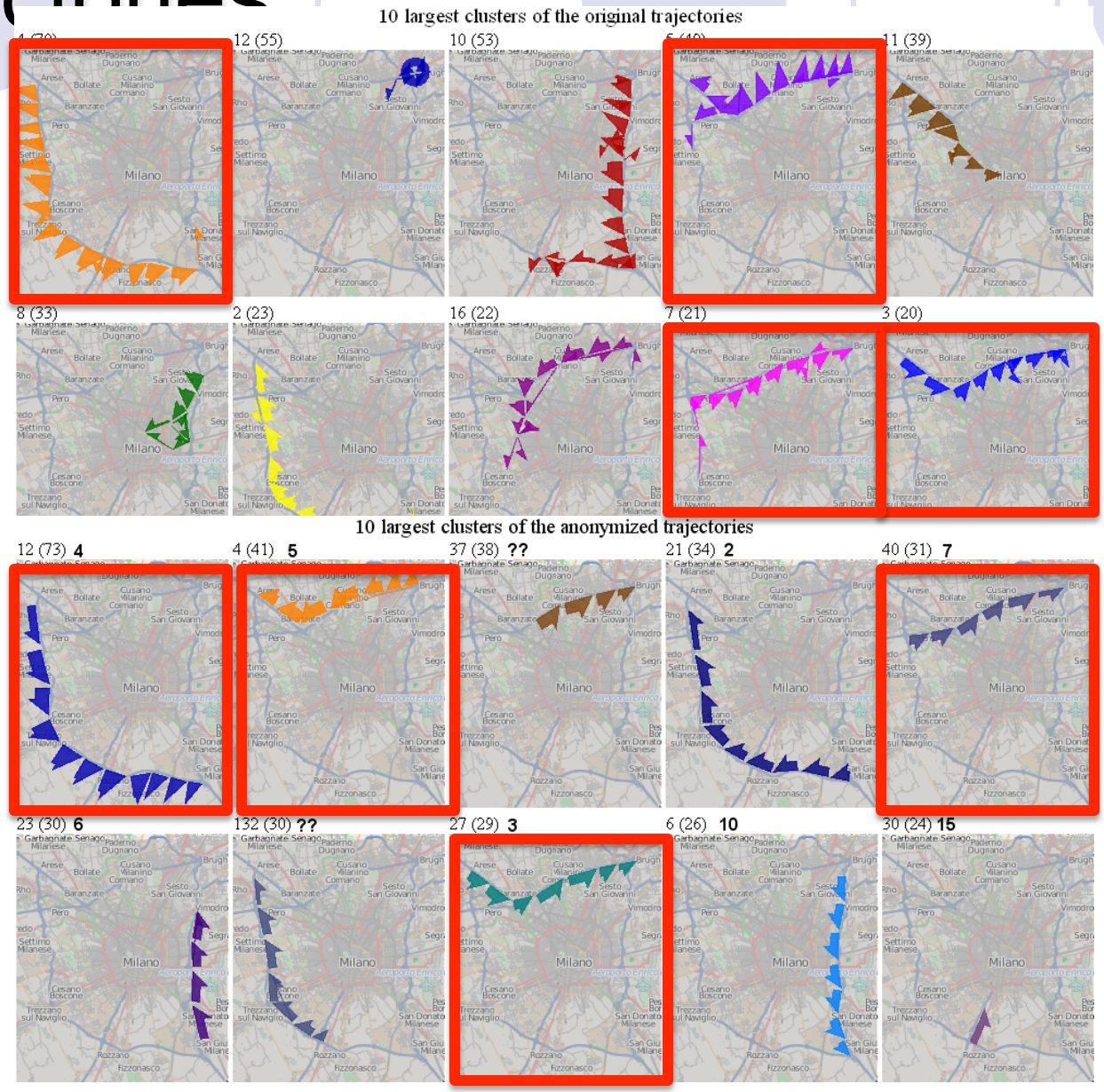
- Anonymization of movement data while preserving clustering
- **Trajectory Linking Attack**: the attacker
  - knows some points of a given trajectory
  - and wants to infer the whole trajectory
- **Countermeasure**: method based on
  - **spatial generalization** of trajectories
  - **k-anonymization** of trajectories

Trajectory Anonymization



- Given a trajectoy dataset
  1. Partition of the territory into Voronoi cells
  2. Transform trajectories into sequence of cells
  3. Ensure k-anonymity:
     - For each generalized trajectory there exist at least others k-1 different people with the same trajectory? If not transform data in similar ones.

Clustering on Anonymized Trajectories



Probability of re-identification: k=16

Known Positions	Probability of re-identification
1 position	98% trajectories have a P <= 0.03 (K=30)
2 positions	98% of trajectories have a P <= 0.05 (K=20)
4 positions	99% of trajectories have a P <= 0.06 (K=17)
…..	

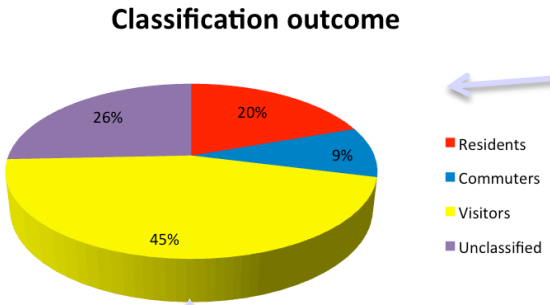
Privacy by Design in Mobile phone socio-meters Analysis

A. Monreale, F. Giannotti, D. Pedreschi, S. Rinzivillo

IEEE Big Data Conference, 2013

Knowledge Discovery and Delivery Lab
(ISTI-CNR & Univ. Pisa)
www-kdd.isti.cnr.it

Privacy-Aware socio-meter



***Output**: quantification of profiles (**safe!!**)*

Profiles computed on anonymous data

Clustering for Building profiles

Aggregated call activities assuring anonymity computed by the Telco Operator

User 1 *User 2* ..... *User n-1* *User n*

Attack risk

- Suppose I am an analyst workining on GSM data of 10 Millions of users with access to their call profiles
- I know where my boy-friend has been the last 4 weeks, can I guess that he has been to Pisa during this week-end?

Probability of re-identification for 4 weeks (10 Mil users GSM)

Observe the User for 4 weeks	Probability of re-identification
1% users	P <=0,0001(K=10000)
3% users	P <= 0.0003 (K=300000)
70% users	P <= 0.00000025 (K=40000000)
…..	

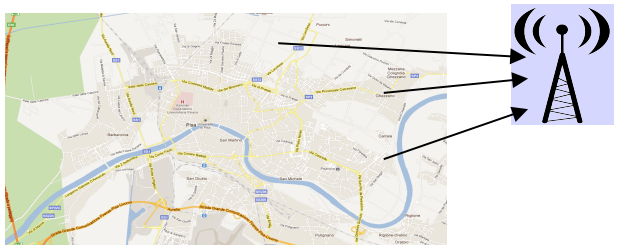
Privacy by Design in Social Borders

A. Monreale, F.Pratesi, D. Pedreschi, S. Rinzivillo

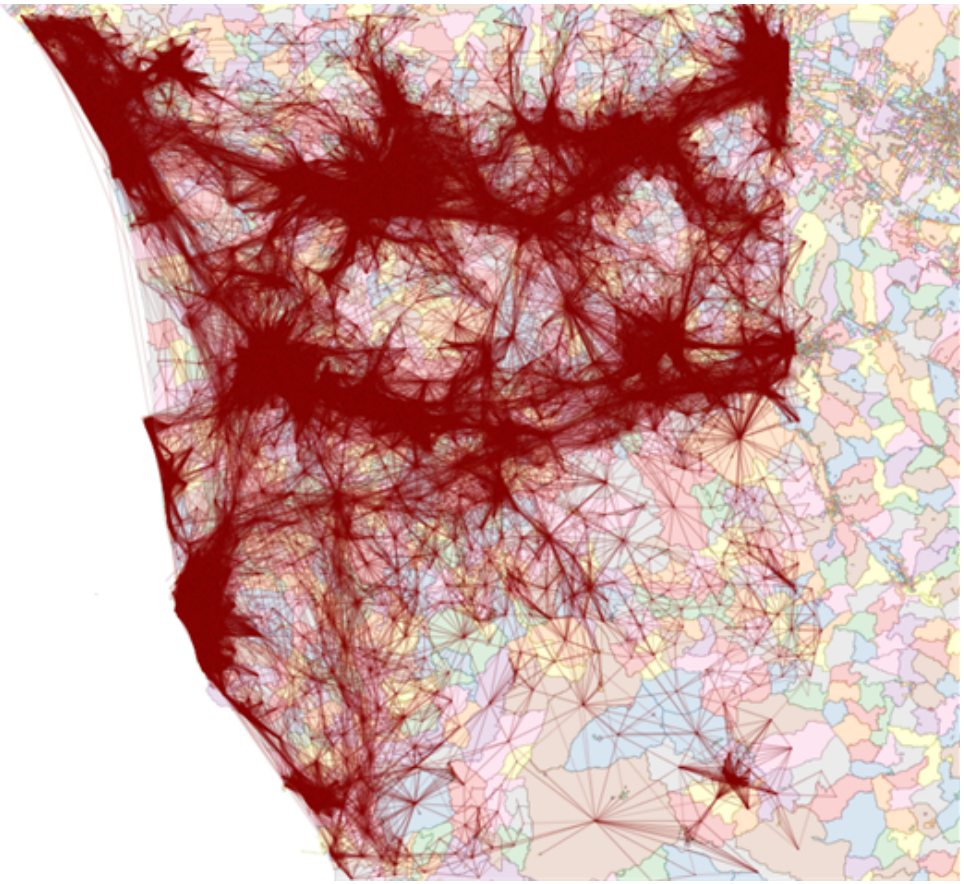
Knowledge Discovery and Delivery Lab
(ISTI-CNR & Univ. Pisa)
www-kdd.isti.cnr.it

Distributed Scenario

Vehicles have on board GPS devices
that collect the sequence of positions
composing ***trajectory****,*
that can be transmitted
(after a ***generalization*** *step)*
to the coordinator



The coordinator uses
the collected user mobility data
to compute an ***aggregation*** *describi*
the traffic flows in a city



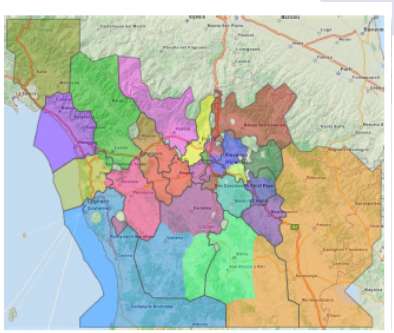
Privacy-Preserving Framework

- Distributed Randomization of individual OD matrix from GPS data while preserving global traffic flow
- **Linking Attack**: the attacker
  - wants to infer the movements from an area to another area of a specific user
- **Countermeasure**:
  - Each user constructs the own OD matrix and sends to a third party the randomized version
  - The data collector aggregates the individual OD matrix for computing the global OD matrix
  - The noise added locally from each user globally cancels each other

Privacy-Preserving Framework

- Distributed Randomization of individual OD matrix from GPS data while preserving global traffic flow
- **Linking Attack**: the attacker
  - wants to infer the movements from an area to another area of a specific user
- **Countermeasure** based on **Differential Privacy**
  - Each user constructs the own OD matrix and sends to a third party the randomized version
  - The data collector aggregates the individual OD matrix for computing the global OD matrix
  - The noise added locally from each user globally cancels each other

Privacy-aware Analytical Process



Output: *Borders*

Individual privacy is protected

Compute a global OD matrix

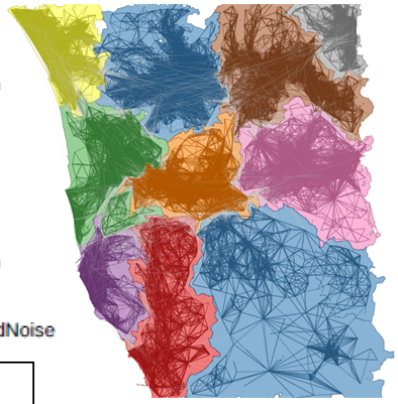
Constructs a local randomized OD Matrix



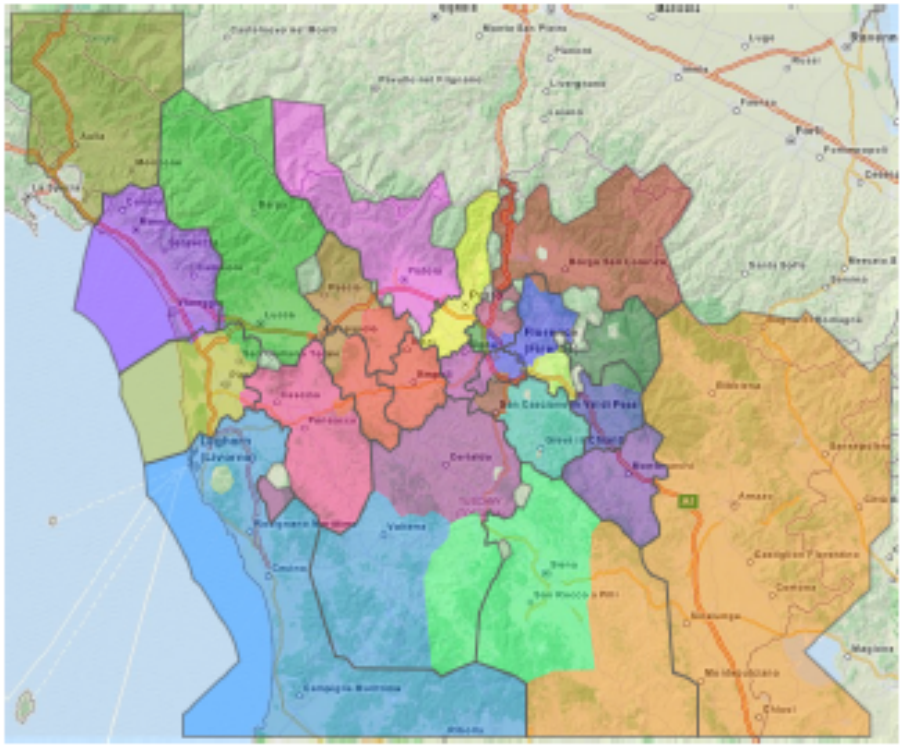


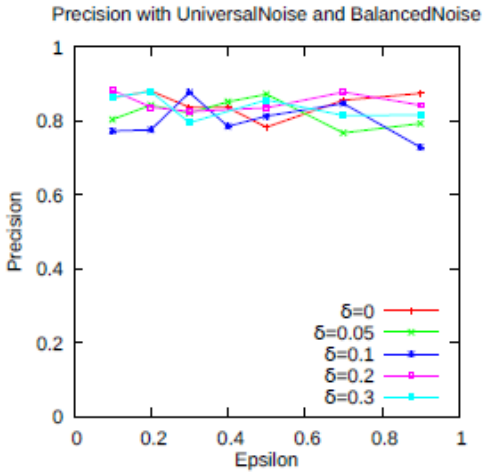
.....

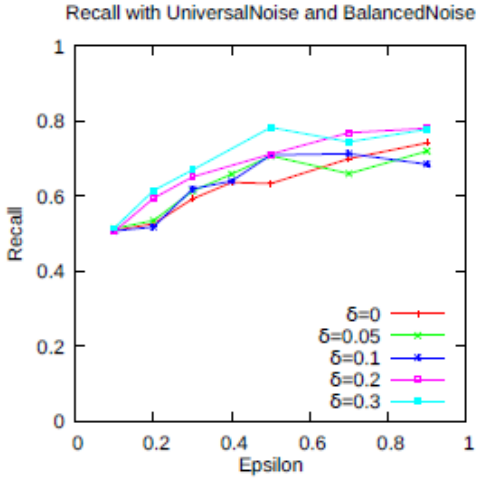
Mobility Analysis: Borders



ε=0.2







Welcome to Privacy Observatory Magazine

Privacy Observatory Magazine is an electronic journal developed under the European project MODAP. This journal combines scientific and legal expertise with a focus on issues related to data privacy and data protection . » read more

Sign in to Modap Social Network for staying up to date with the news from the project!

» SECTIONS

- » Technology (2)
- » Law & Policy (1)
- » Industry & Business (1)
- » Case Study (0)

» TAGS

business economic approaches to data privacy gps guidelines industry iphone map michele coscia privacy market terms vehicular communication webmaster

» CURRENT ISSUE

Editorial

Today, we are faced with unprecedented opportunities of sensing and storing detailed data of human activities at a society-wide scale, because most of our activities are mediated by the ICT's.
» read more

» SEARCH

Search...

Local Web

» Advanced Search

» LATEST NEWS

» The Pros & Cons Frictionless Sharing

» Dissecting Case 014 Exhibit B

» Lawmakers Seek FTC Pro of Facebook Post-Log C Tracking