# Privacy and anonymity in location- and movement-aware data analysis

**Fosca Giannotti & Dino Pedreschi**
KDD Lab Pisa, Italy
fosca.giannotti@isti.cnr.it
pedre@di.unipi.it

**Tutorial @ PAKDD 2007**

**Nanjing, PRC, 22 May 2007**

# KDD LAB @ Pisa, Italy

- Knowledge Discovery & Data Mining Lab
- Established in 1996 as a joint initiative of
  - **ISTI-CNR** – *Information Science and Technology Institute* of the Italian National Research Council
  - **University of Pisa**

- http://www-kdd.isti.cnr.it

# Acknowledgements

- The tutorialists are grateful to:
    - Maurizio **Atzori** and Francesco **Bonchi**,
        - KDD LAB Pisa, Italy
    - Mohamed F. **Mokbel**
        - University of Minnesota, U.S.A.
    - Useful comments also from Bharat **Bhargava**,
        - Purdue University, U.S.A.

# Plan of the tutorial

- The scenario of ubiquitous computing
  - Analytic opportunities and privacy threats
- Privacy and anonymity: prognosis and therapy
  - In data publishing: attack models and privacy-preserving techniques
  - In data mining: attack models and privacy-preserving data mining techniques
- Privacy and anonymity in Location Based Services
- Preliminary results on privacy and anonymity techniques in mobility data analysis
- Conclusion

# Plan of the tutorial

- **The scenario of ubiquitous computing**
  - **Analytic opportunities and privacy threats**
- Privacy and anonymity: prognosis and therapy
  - In data publishing: attack models and privacy-preserving techniques
  - In data mining: attack models and privacy-preserving data mining techniques
- Privacy and anonymity in Location Based Services
- Preliminary results on privacy and anonymity techniques in mobility data analysis
- Conclusion

# The Wireless Explosion

*Do you use any of these devices ?*

*Do you ever feel that you are tracked?*

# The Wireless Network

- The pervasiveness of mobile and ubiquitous technologies is increasing day after day
  - GSM wireless phone networks
    - 1.5 billions in 2005, still increasing at a high speed
    - Italy:  # mobile phones ≈ # inhabitants
  - GPS and Galileo positioning systems
  - Wi-Fi and Wi-Max wireless networks
  - RFID’s and sensor networks
- miniaturization
- positioning accuracy
    - location technologies capable of providing increasingly better estimate of user location

# Which new opportunities?

- Location based services:
  - A certain service that is offered to the users based on their locations
- Mobility data analysis:
  - Discovering knowledge from the digital traces of our mobile activity to support decision making in mobility related issues.

# Location-based Services: Then

- **Limited to fixed traffic signs**

# Location-based Services: Now

- Location-based traffic reports:
  - ***Range query:*** How many cars in the free way
  - ***Shortest path query***: What is the estimated time travel to reach my destination

- Location-based store finder:
  - ***Range query:*** What are the restaurants within five miles of my location
  - ***Nearest-neighbor query***: Where is my nearest fast (junk) food restaurant

- Location-based advertisement:
  - ***Range query:*** Send E-coupons to all customers within five miles of my store

# Mobility data analysis

- How people move around in the town
  - During the day, during the week, etc.
- Are there typical movement behaviours?
- Are there typical movement behaviours in a certain area at a certain time?
- How frequently people access the network?
- How are people movement habits changing in this area in last decade-year-month-day?
- Are there relations between movements of two areas?
- Are there periodic movements?

# Privacy in Mobility Data and Services

- Trusted/secure storage/Management of Mobility Data
- Privacy in Location Based Services:
  - the right of a user to receive a service without revealing his/her identity
  - Trade-off between quality of service and privacy protection
- Privacy and Anonymity in Mobility Data Analysis
  - Trade-off between privacy protection and analysis opportunities

# Traces and Analytics opportunities

- Our everyday actions leave digital **traces** into the information systems of ICT service providers.
  - web browsing and e-mailing,
  - credit cards and point-of-sale e-transactions, e-banking,
  - electronic administrative transactions and health records,
  - shopping transactions with loyalty cards.

- Wireless phone networks gather highly informative traces about the human mobile activities in a territory

# 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.

# A paradigmatic example of Mobility Data Analysis: **GeoPKDD**

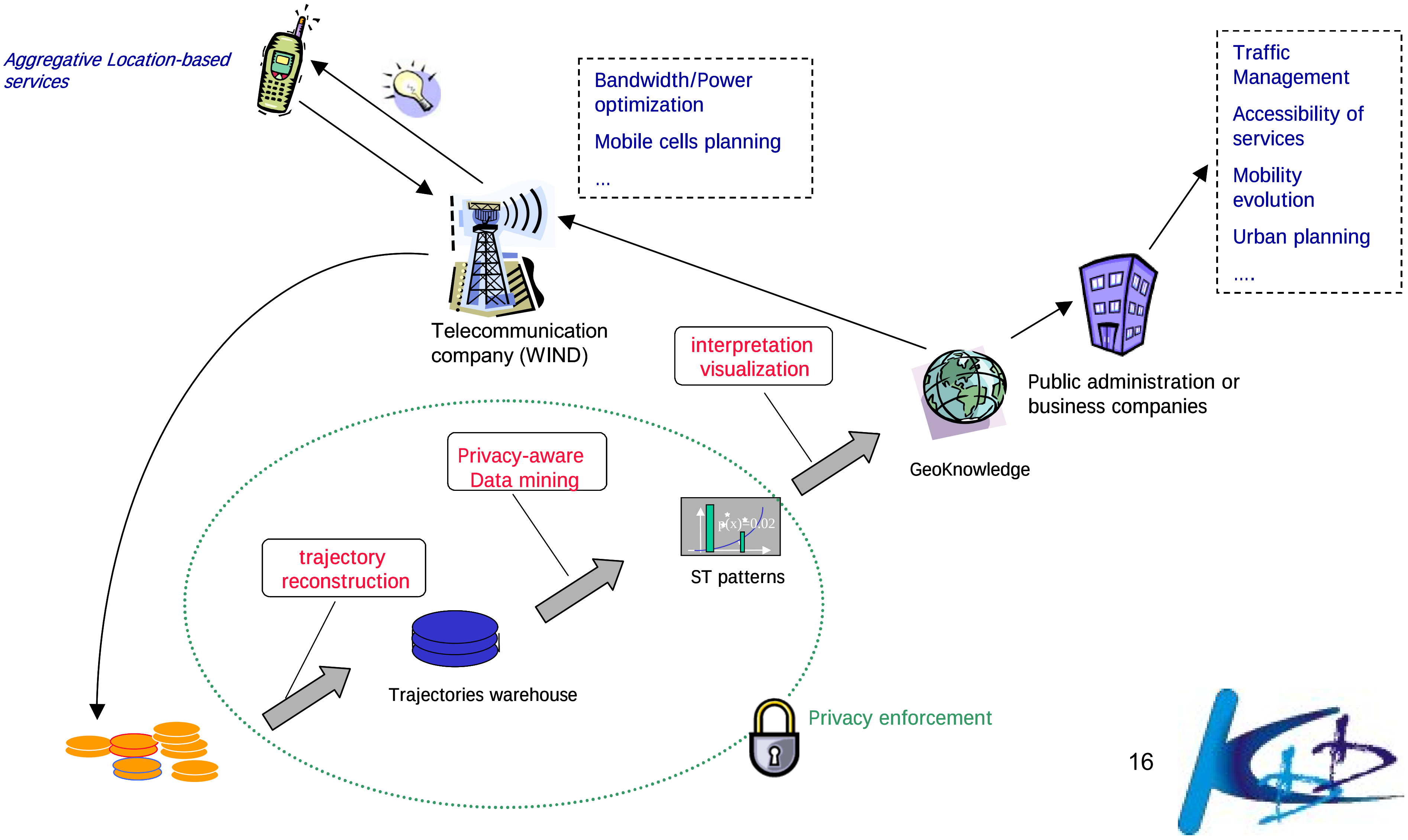
**A European FP7 project**

www.geopkdd.eu

**Geographic Privacy-aware**

**Knowledge Discovery and Delivery**

# Geographic privacy-aware Knowledge Discovery


*Aggregative Location-based services*
Bandwidth/Power optimization
Mobile cells planning
…
Traffic Management
Accessibility of services
Mobility evolution
Urban planning
….
Telecommunication company (WIND)
interpretation visualization
Public administration or business companies
GeoKnowledge
Privacy-aware Data mining
p(x)=0.02
ST patterns
trajectory reconstruction
Trajectories warehouse
Privacy enforcement

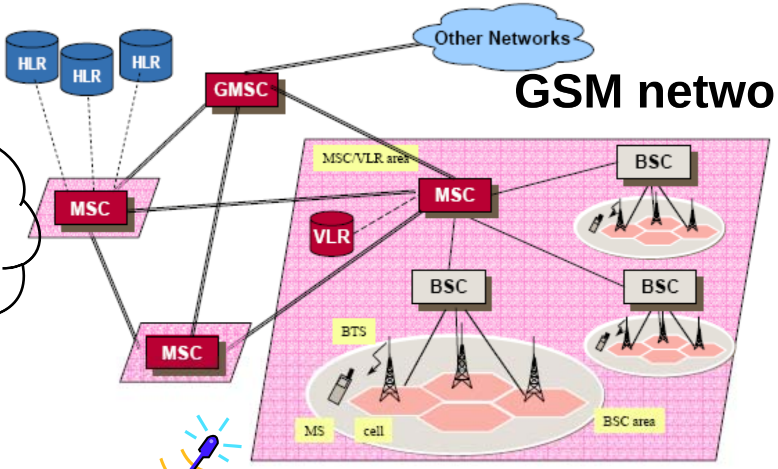
# The GeoPKDD scenario

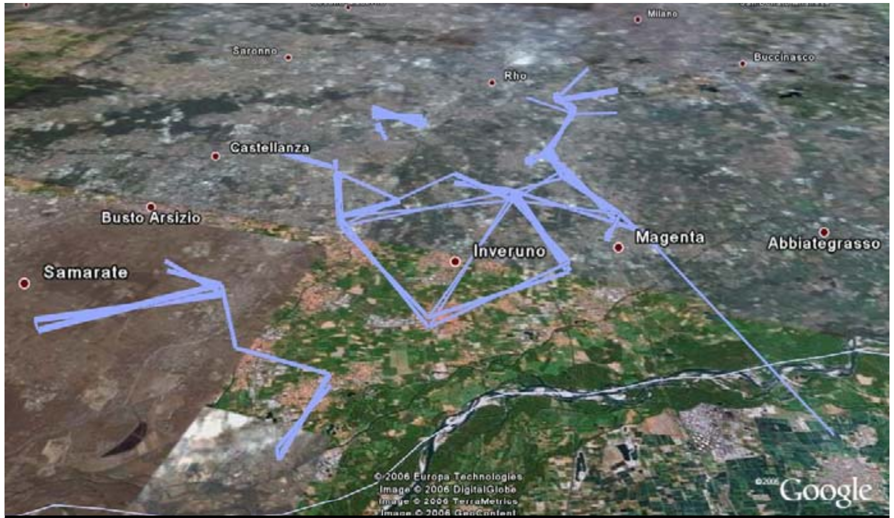
- From the analysis of the traces of our mobile phones it is possible to reconstruct our mobile behaviour, the way we collectively move
- This knowledge may help us improving decision-making in mobility-related issues:
  - Planning traffic and public mobility systems in metropolitan areas;
  - Planning physical communication networks
  - Localizing new services in our towns
  - Forecasting traffic-related phenomena
  - Organizing logistics systems
  - Avoid repeating mistakes
  - Timely detecting changes.

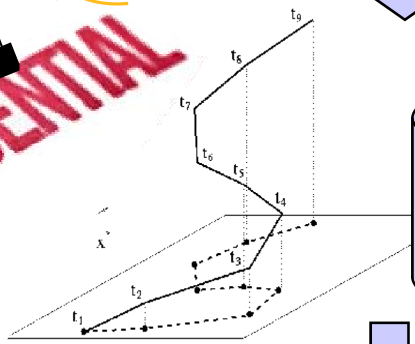
# Mobility Manager Office

Sustainable Mobility?

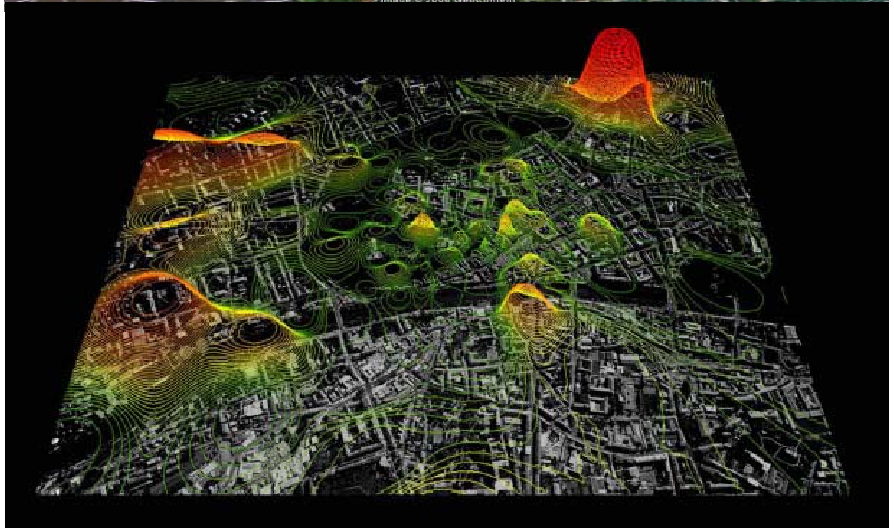
# GSM network



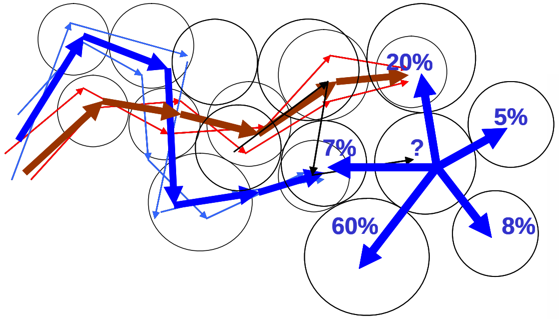


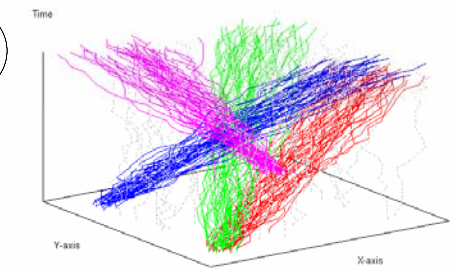


Position logs

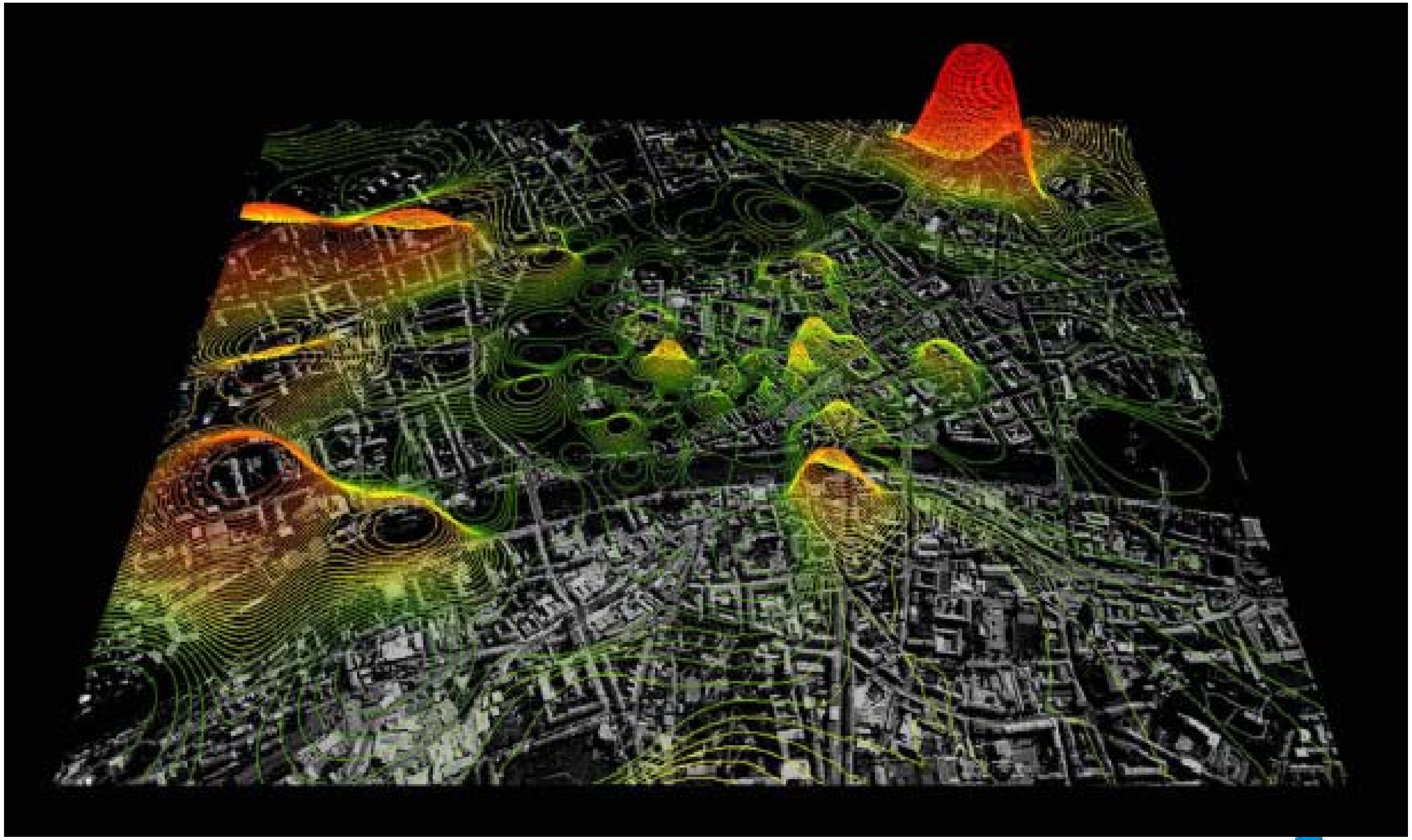


# Mobility Models



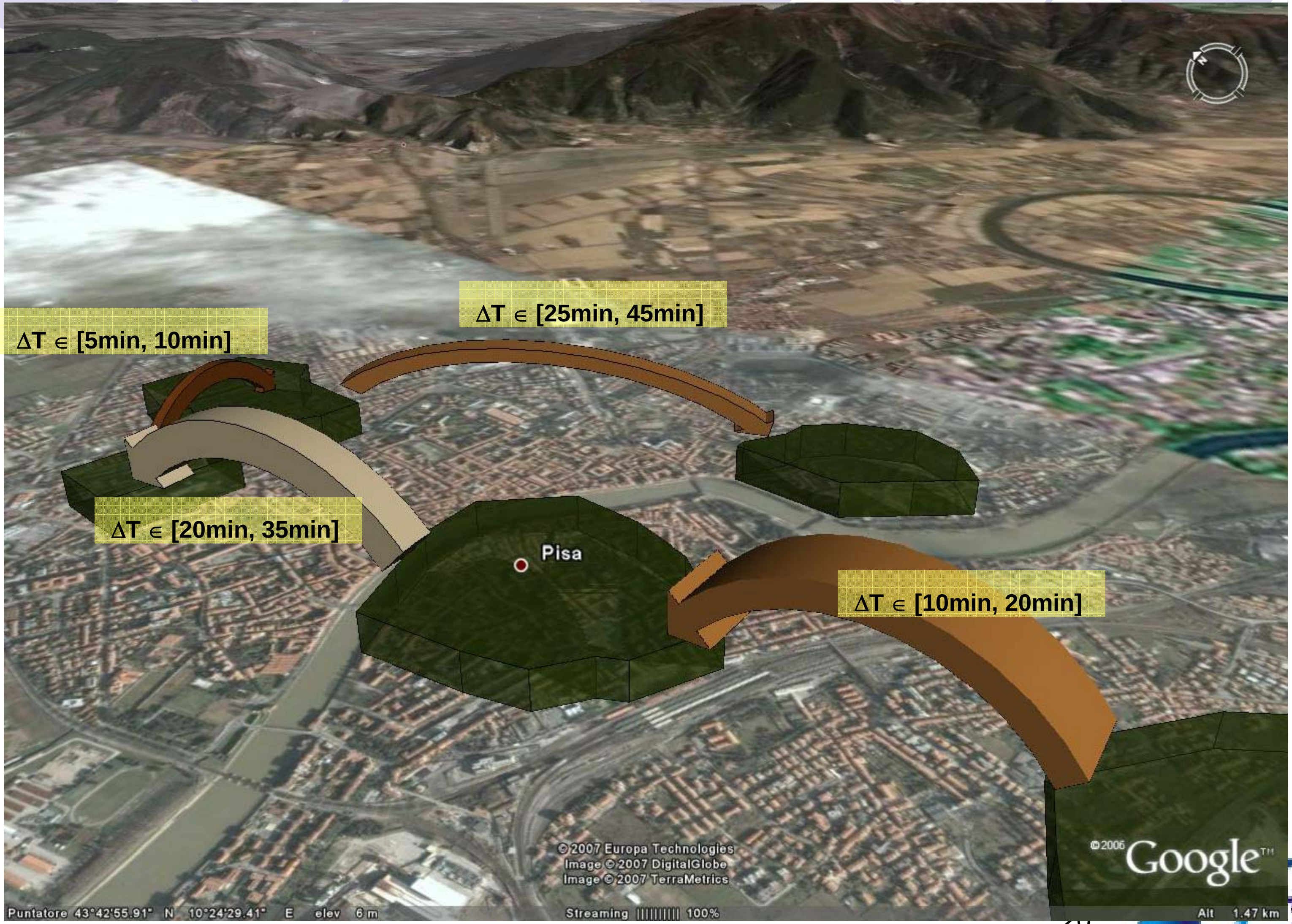


# Real-time density estimation in urban areas



The senseable project: http://senseable.mit.edu/grazrealtime/

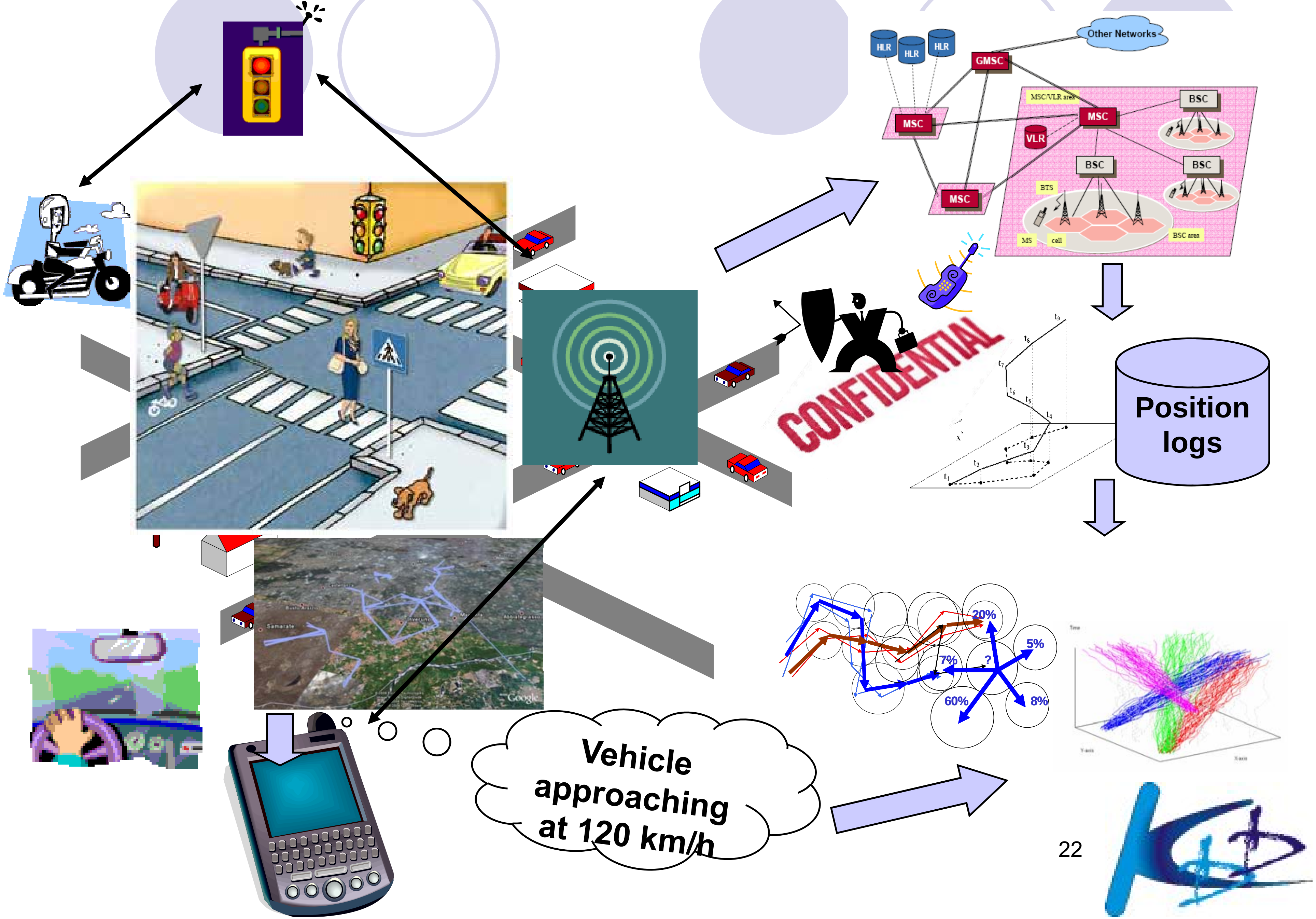
# Mobility patterns in urban areas



# From wireless networks to Ubi Comp environments

- Log data from mobile phones, i.e. sampling  of localization points in the GSM/UMTS network.
- Log data from GPS-equipped devices
- Log data from
  - peer-to-peer mobile networks
  - intelligent transportation environments
  - ad hoc sensor networks, RFIDs
- Increasing precision and pervasiveness

# Intelligent Transportation & Infomobility



Position logs

Vehicle approaching at 120 km/h

# Privacy in GeoPKDD

- How to design Data Analysis methods that, **by construction**, meet the the privacy constraints?
- How to develop trustable data mining technology capable of producing
  - ***provably/measurably*** privacy-preserving patterns
  - which may be safely distributed

# Scientific Privacy Issues in GeoPKDD

- Is there any specific challenge/risk/opportunity in the context of ST data?
  - New threats from traces analysis: learning who you are from where you have been (Malin et al 2003)
  - Space and Time in a trajectory can act as quasi-identifiers (Bettini and Jajodia 2005)
- How to formalize privacy measures over Spatio-Temporal data and Spatio-Temporal patterns?
  - E.g., anonimity threshold on clusters of individual trajectories

# Ethical, Legal and Societal Privacy Issues in GeoPKDD

- Harmonization with national privacy regulations and authorities – **privacy observatory**
- Brings together
  - GeoPKDD technologists,
  - representatives of the national and European privacy authorities,
  - non-governmental privacy-related associations

# Goals of the Privacy Observatory

1. Implement the privacy regulations into the GeoPKDD methods and tools
2. Suggest refinements of the regulations made possible by new privacy preserving analysis techniques
3. Foster inter-disciplinary dialogue and disseminate key issues to broad audience

# Industrial value of privacy technologies

- Several social studies report that users become more aware about their privacy and may end up not using any LBS
- Hence, trustability is both a social and a business value, to foster the large-scal deployment of LBS and ubiquitous computing

# “To Report or Not To Report:” Tension between Personal Privacy and Public Responsibility

An info tech company will typically lose between ten and one hundred times more money from shaken consumer confidence than the hack attack itself represents if they decide to prosecute the case.

Mike Rasch, VP Global Security, testimony before the Senate Appropriations Subcommittee, February 2000 reported in The Register and online testimony transcript

# 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 privacy problem

- 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
  - ○ ...

# Plan of the tutorial

- The scenario of ubiquitous computing
  - Analytic opportunities and privacy threats
- **Privacy and anonymity: prognosis and therapy**
  - **In data publishing: attack models and privacy-preserving techniques**
  - In data mining: attack models and privacy-preserving data mining techniques
- Privacy and anonymity in Location Based Services
- Preliminary results on privacy and anonymity techniques in mobility data analysis
- Conclusion

# The naive scientist’s view (1)

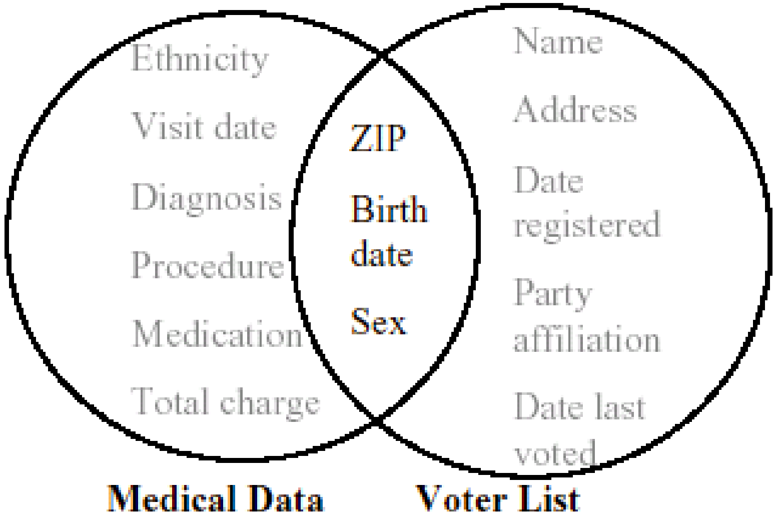
- Knowing the exact identity of individuals is not needed for analytical purposes
  - Anonymous trajectories are enough to reconstruct aggregate movement behaviour, pertaining to groups of people.
- Is this reasoning correct?
- Can we conclude that the analyst runs no risks, while working for the public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

# Plan of the tutorial

- The scenario of ubiquitous computing
  - Analytic opportunities and privacy threats
- **Privacy and anonymity: prognosis and therapy**
  - In data publishing: attack models and privacy-preserving techniques
  - **In data mining: attack models and privacy-preserving data mining techniques**
- Privacy and anonymity in Location Based Services
- Preliminary results on privacy and anonymity techniques in mobility data analysis
- Conclusion

# Privacy Preserving Data Mining: Condensed State of the Art

# Privacy Preserving Data Mining

- Very Short Definition:

*“the study of data mining side-effects on privacy”*

- A Bit Longer Definition:

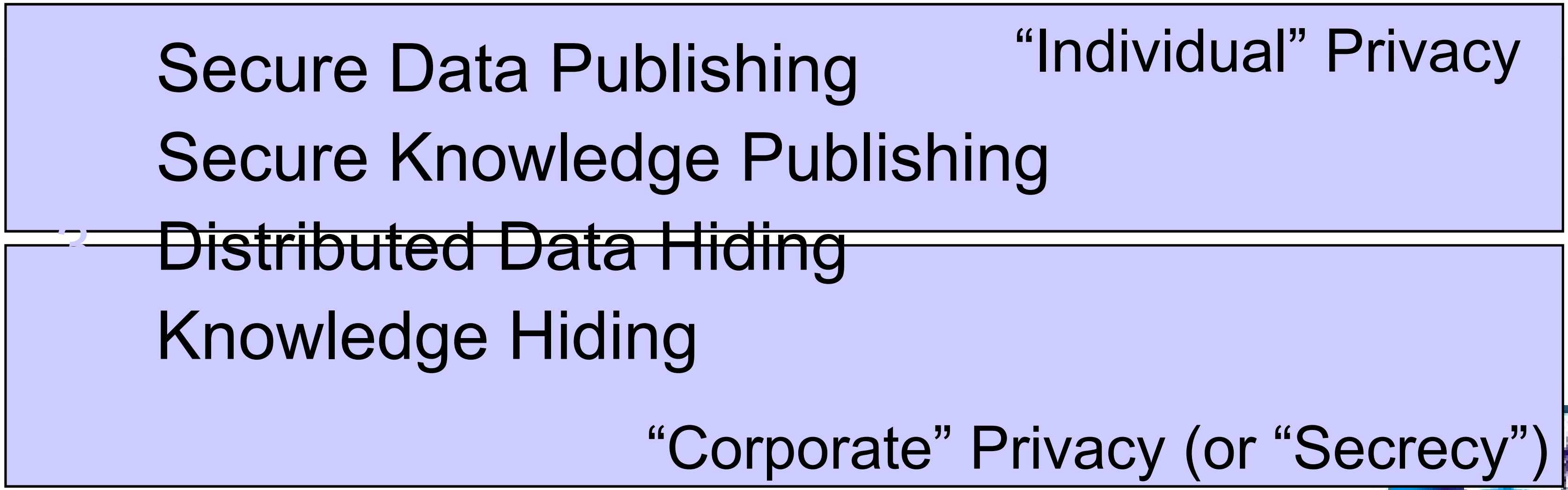
  *“the study of how to produce valid mining models and patterns without disclosing private information”*

  - *Requires to define what is “private”…*
  - *Many different definitions…*
  - *… many different aproaches to*

*Privacy Preserving Data Mining*

# Privacy Preserving Data Analysis and Mining

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



Secure Data Publishing

Secure Knowledge Publishing

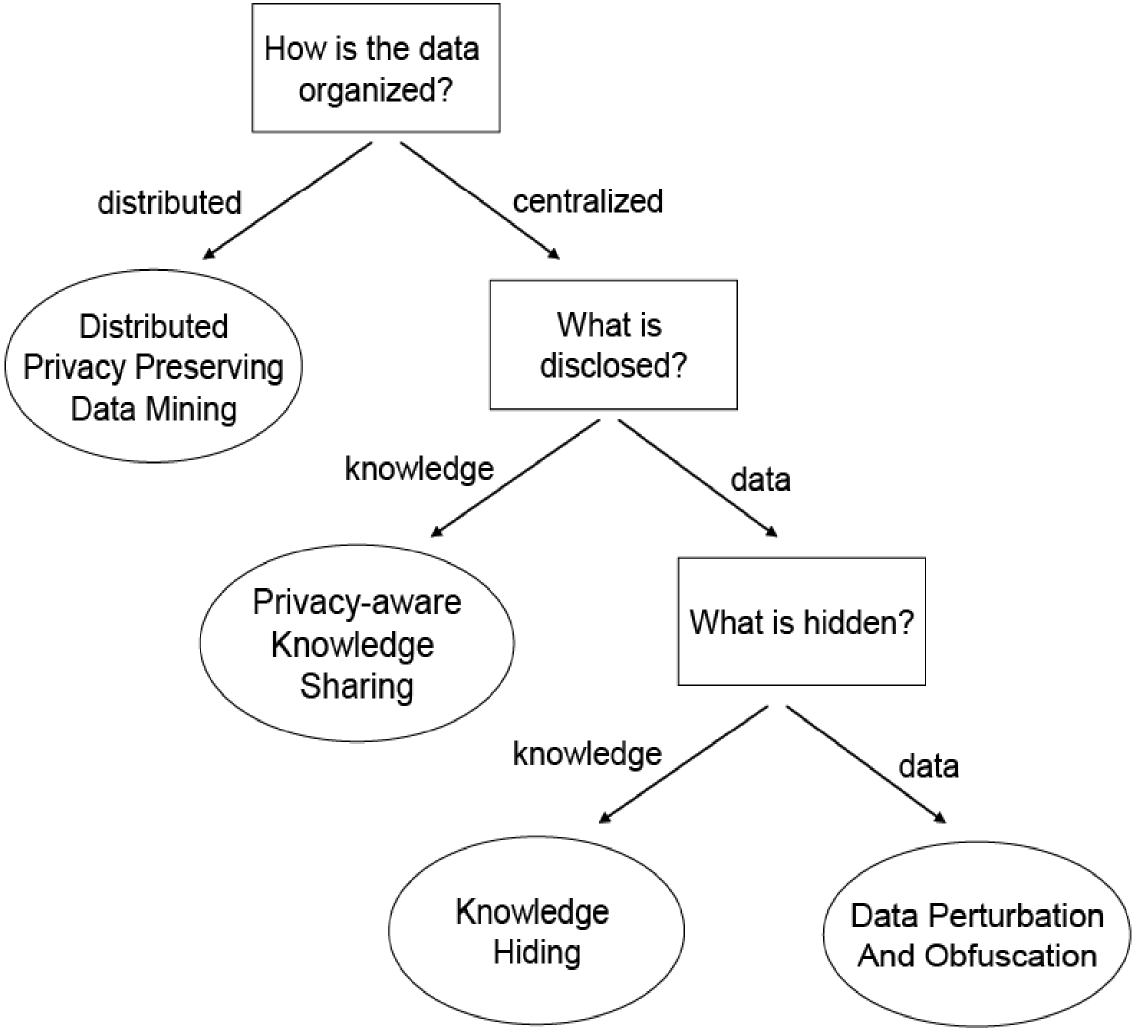
“Individual” Privacy

Distributed Data Hiding

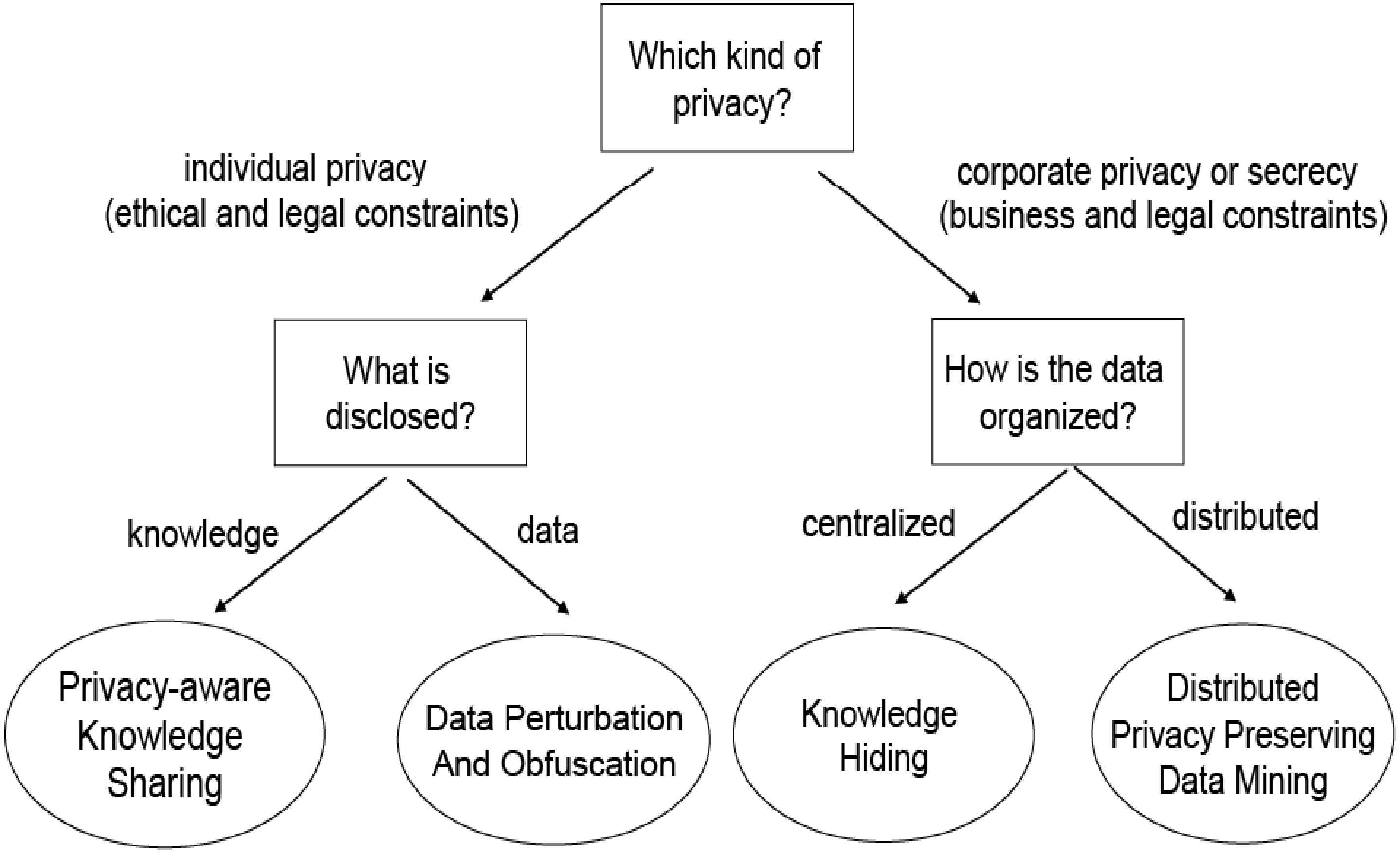
Knowledge Hiding

“Corporate” Privacy (or “Secrecy”)

# A taxonomy tree…


How is the data organized?
distributed
centralized
Distributed Privacy Preserving Data Mining
What is disclosed?
knowledge
data
Privacy-aware Knowledge Sharing
What is hidden?
knowledge
data
Knowledge Hiding
Data Perturbation And Obfuscation

# And another one…



Which kind of privacy?
individual privacy (ethical and legal constraints)
corporate privacy or secrecy (business and legal constraints)
What is disclosed?
How is the data organized?
knowledge
data
centralized
distributed
Privacy-aware Knowledge Sharing
Data Perturbation And Obfuscation
Knowledge Hiding
Distributed Privacy Preserving Data Mining

# 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

- 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.

# Knowledge Hiding

- 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.

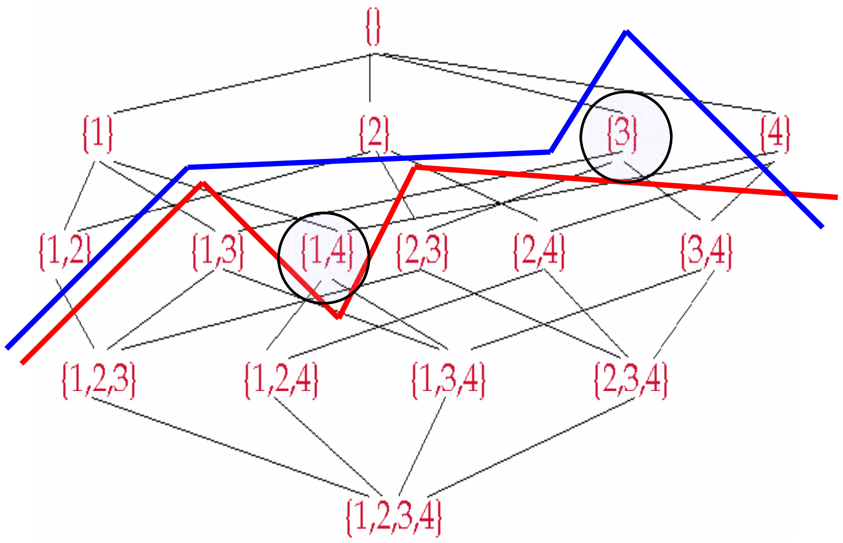
  - Problem: how to transform *D* into *D'* (the database we are going to disclose) in such a way that $R/ R_h$ can be mined from *D'.*

# Knowledge Hiding

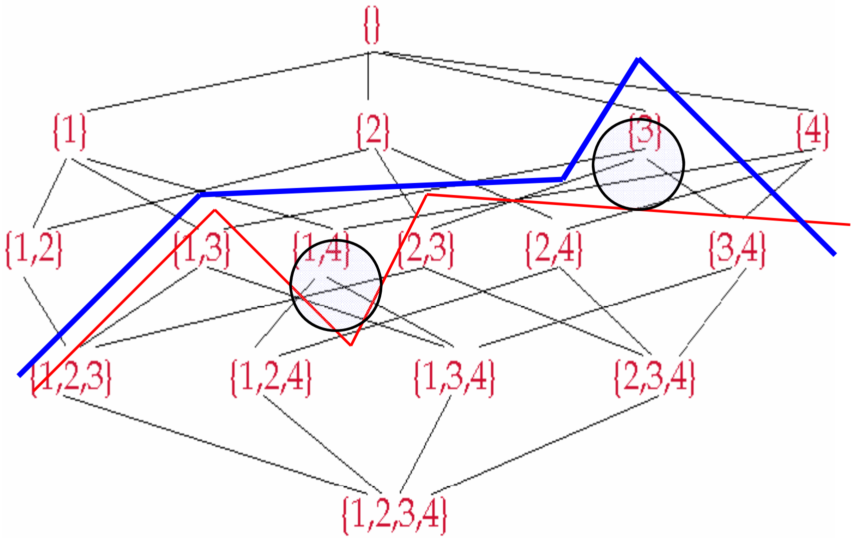
- *Mining frequent itemsets is the fundamental step for mining Association Rules*
- *Suppose min_sup = 2*

| *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 |

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



| *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 |



- [Intermediate table]: itemsets {3} and {1,4} have the ‘1’s turned into ‘?’.
- Some of these ‘?’ will later on be turned into zeros.
- Heuristics:
  - select which of the transactions {T3, T4, T6, T7} will be *sanitized,*
  - to which *extent* (meaning how many items will be affected),
  - and in which relative *order*.

- Heuristics do not guarantee (in any way) the identification of the best possible solution: but they provide overall good solutions efficiently.
- A solution always exists! 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.

# 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** 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.
- K. Liu, H. Kargupta, and J. 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

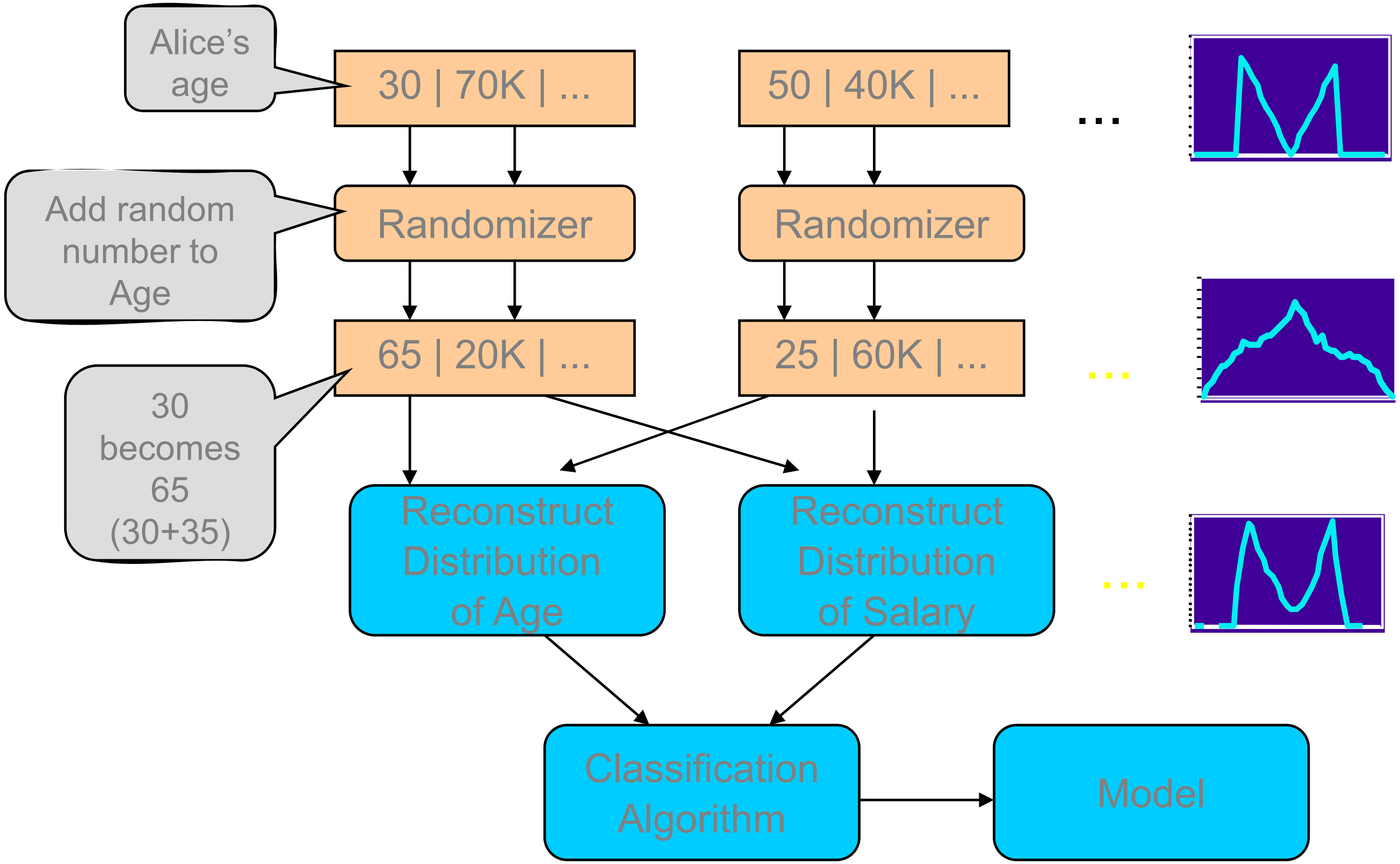
# Data Perturbation and Obfuscation

- 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



Alice's age

Add random number to Age

30 becomes 65 (30+35)

30 | 70K | ...

50 | 40K | ...

...

Randomizer

Randomizer

65 | 20K | ...

25 | 60K | ...

...

Reconstruct Distribution of Age

Reconstruct Distribution of Salary

...

Classification Algorithm

Model

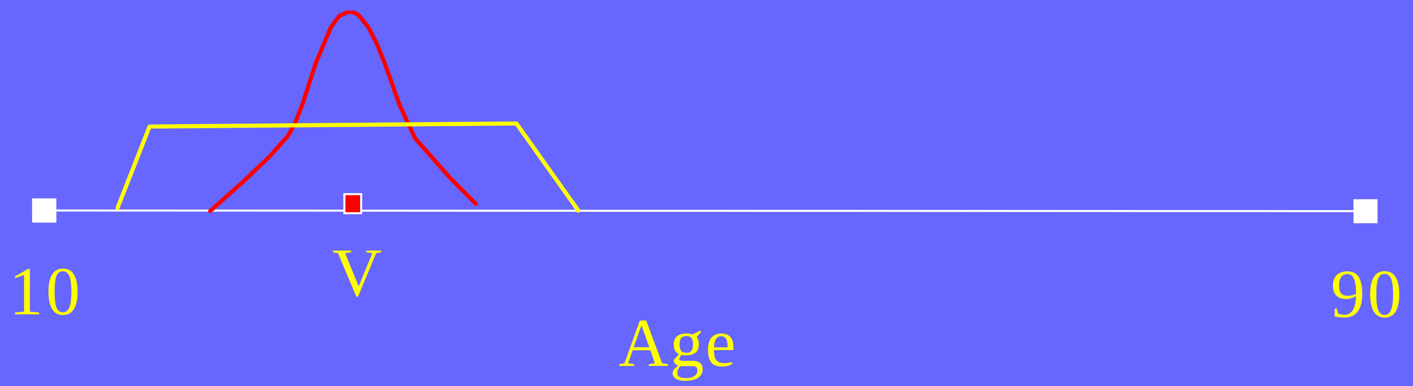
# Reconstruction Problem

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

Estimate the probability distribution of X.

# Intuition (Reconstruct single point)

- Use Bayes' rule for density functions

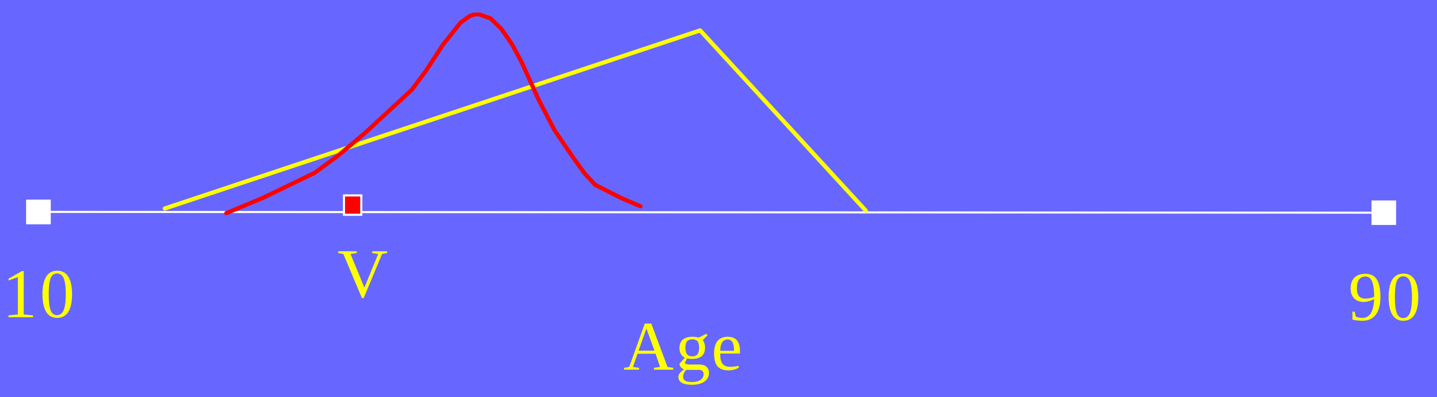


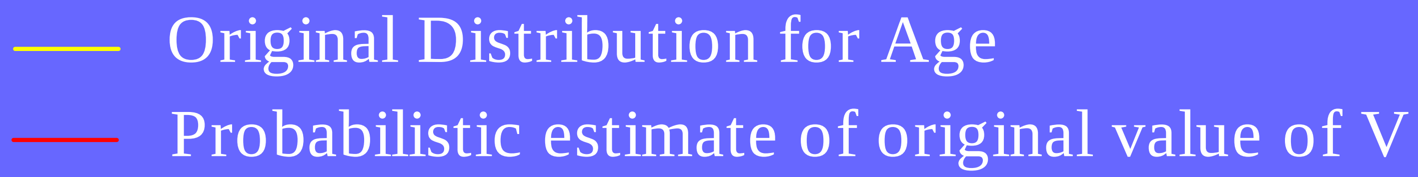
— Original distribution for Age

— Probabilistic estimate of original value of V

# Intuition (Reconstruct single point)

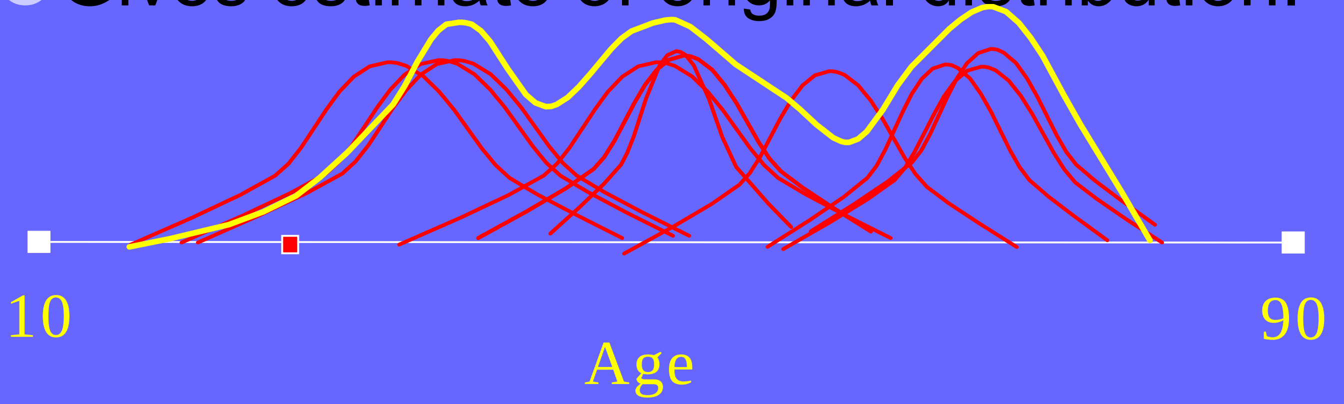
- Use Bayes' rule for density functions





# Reconstructing the Distribution

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



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

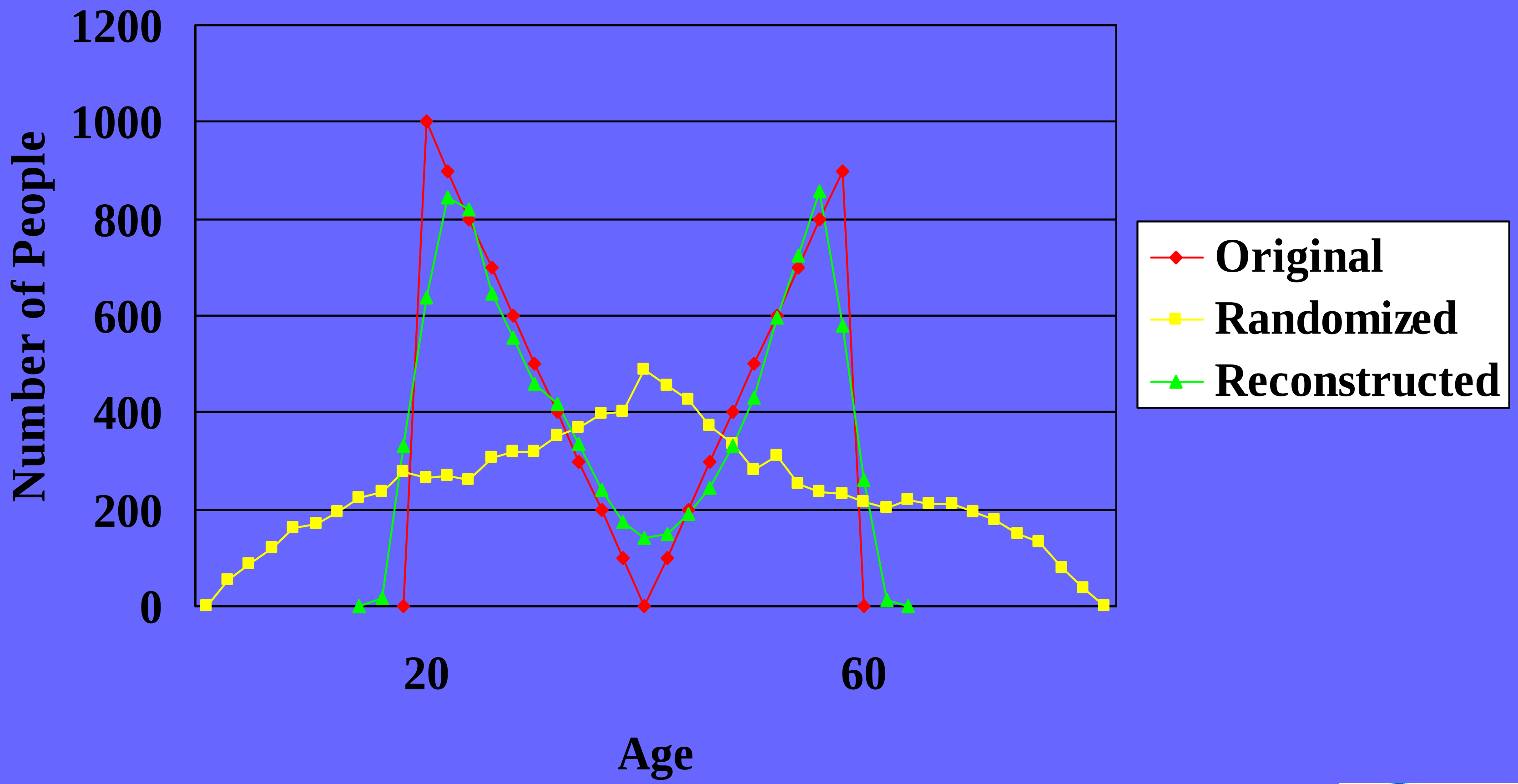
# Reconstruction: Bootstrapping

```
$f_X^0$ := 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.
  - D. Agrawal & C.C. Aggarwal, PODS 2001.

# Works well



Number of People
1200
1000
800
600
400
200
0
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 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.

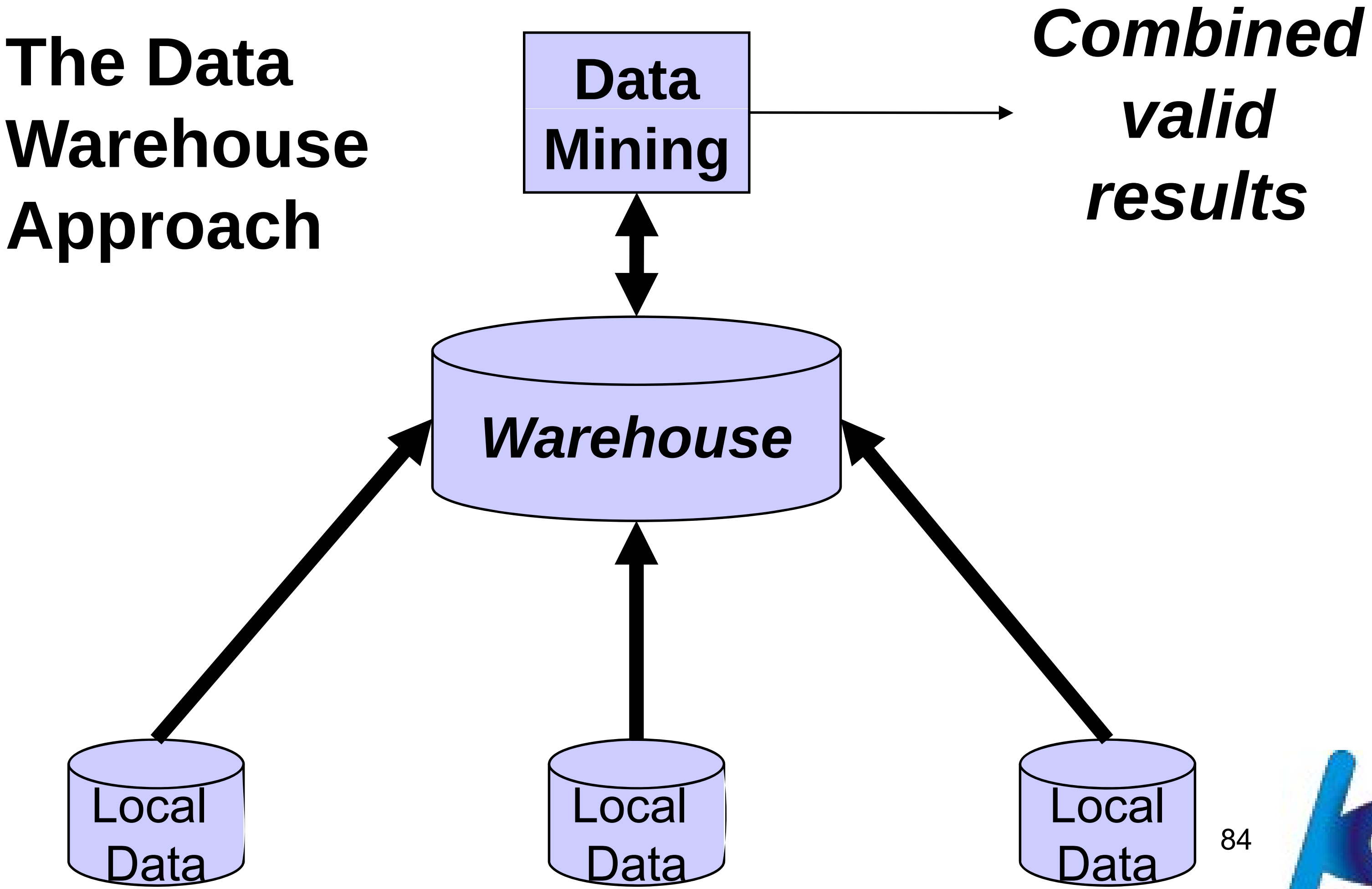
# Distributed Privacy Preserving Data Mining

- 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.

# Data Mining from distributed sources: Standard method



**The Data Warehouse Approach**

**Data Mining**

***Combined valid results***

***Warehouse***

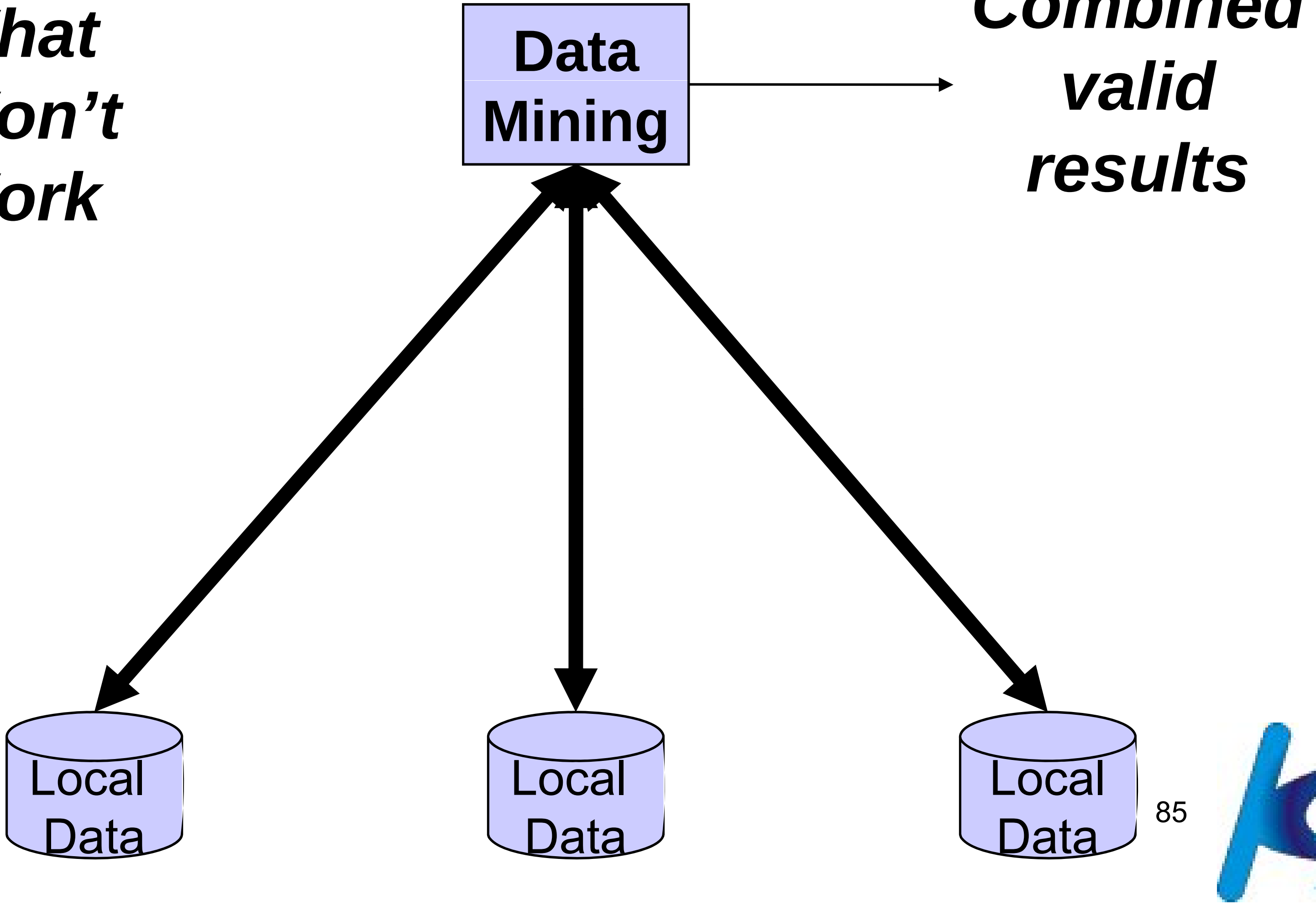
Local Data

Local Data

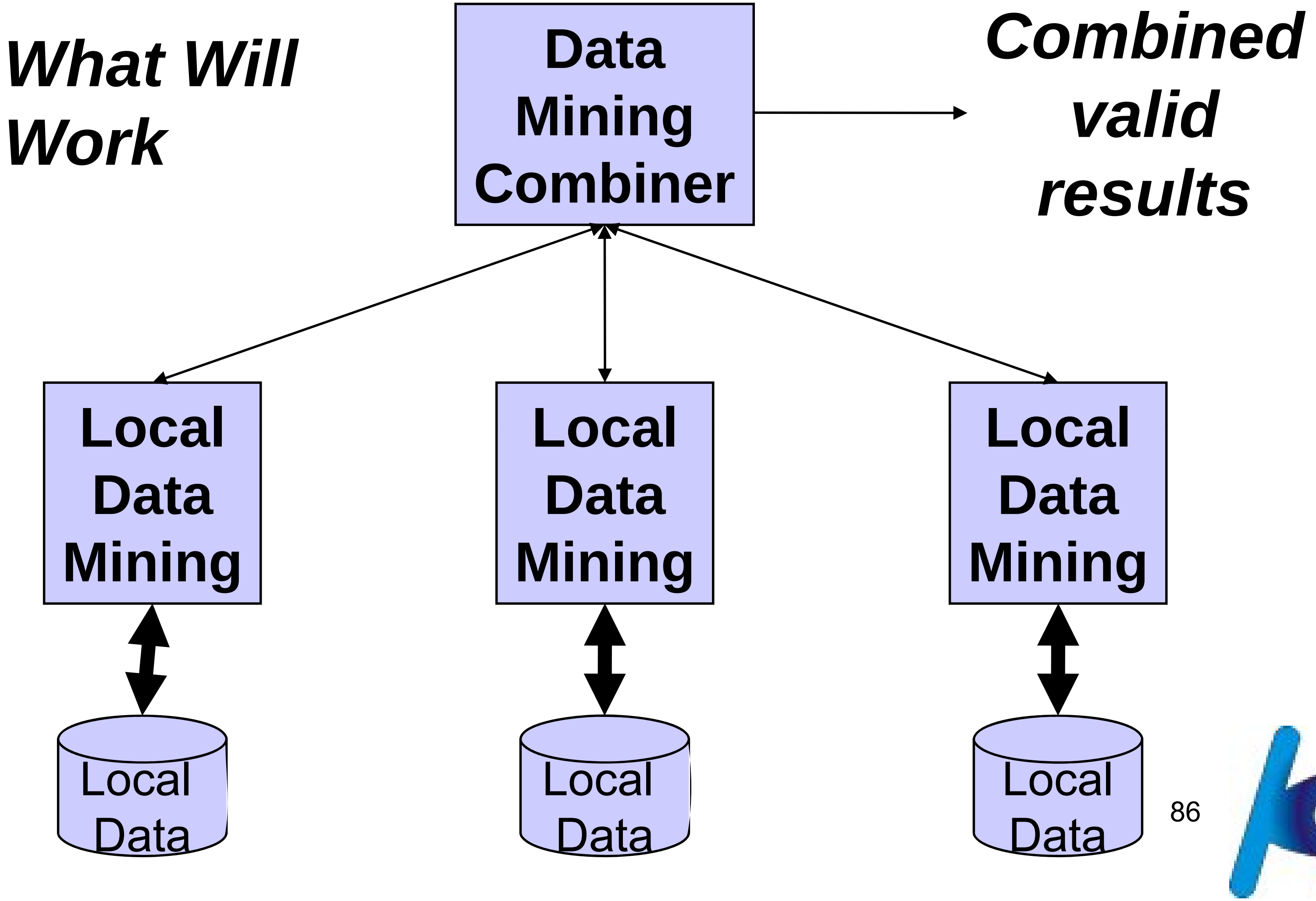
Local Data

# Private Distributed Mining: What is it?

***What Won’t Work***



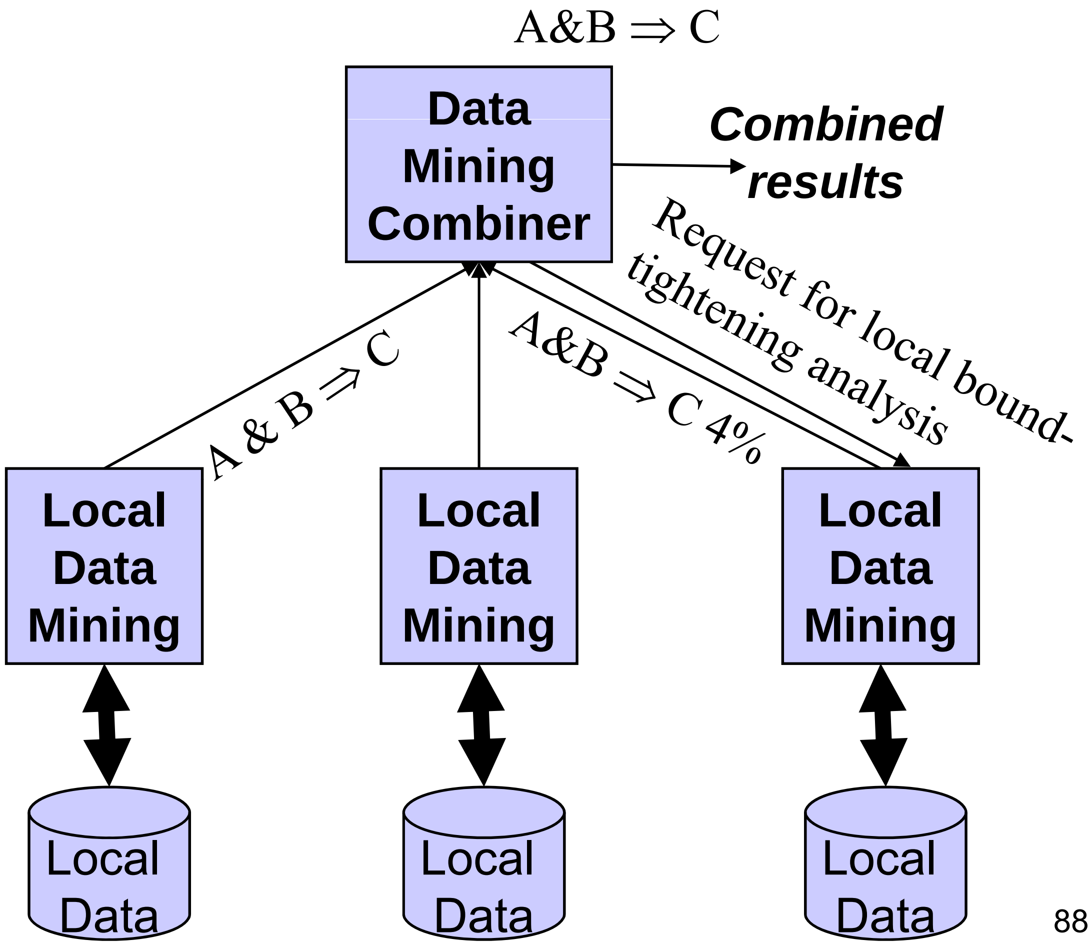
# Private Distributed Mining: What is it?



# 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:

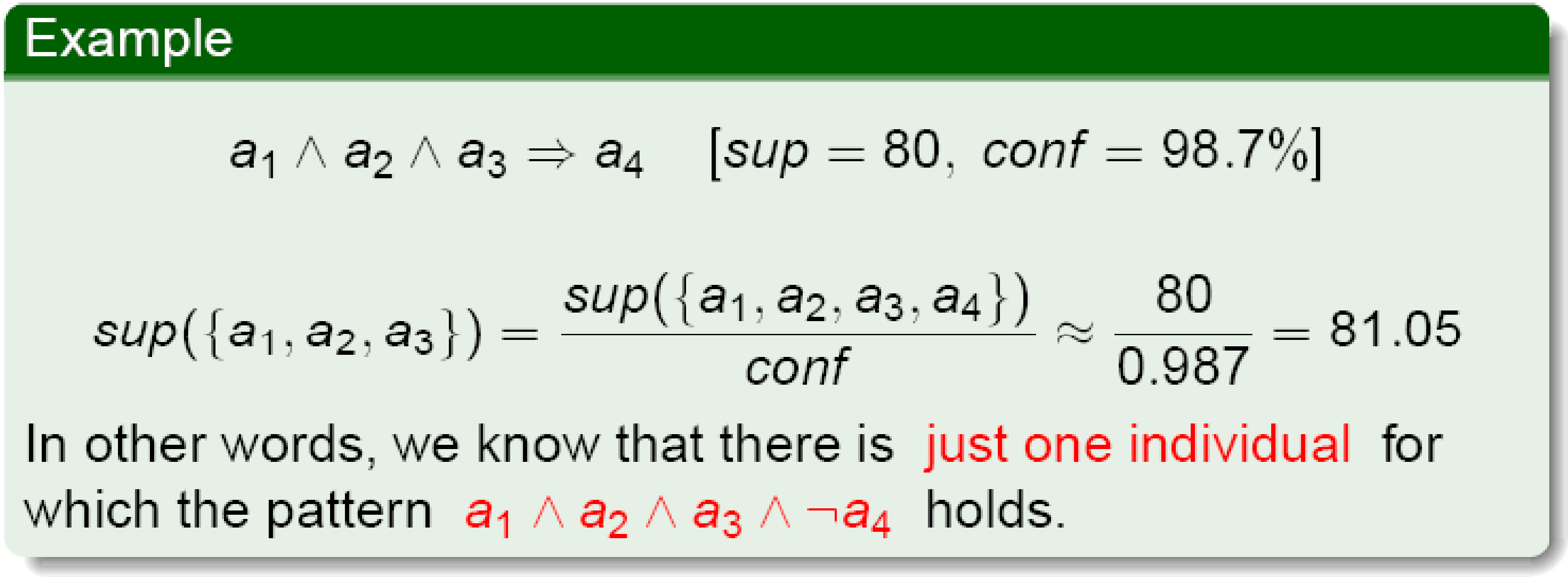
  *“do the data mining results themselves violate privacy”*
- Focus on **individual privacy**: the individuals whose data are stored in the source database being mined.

# 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%)

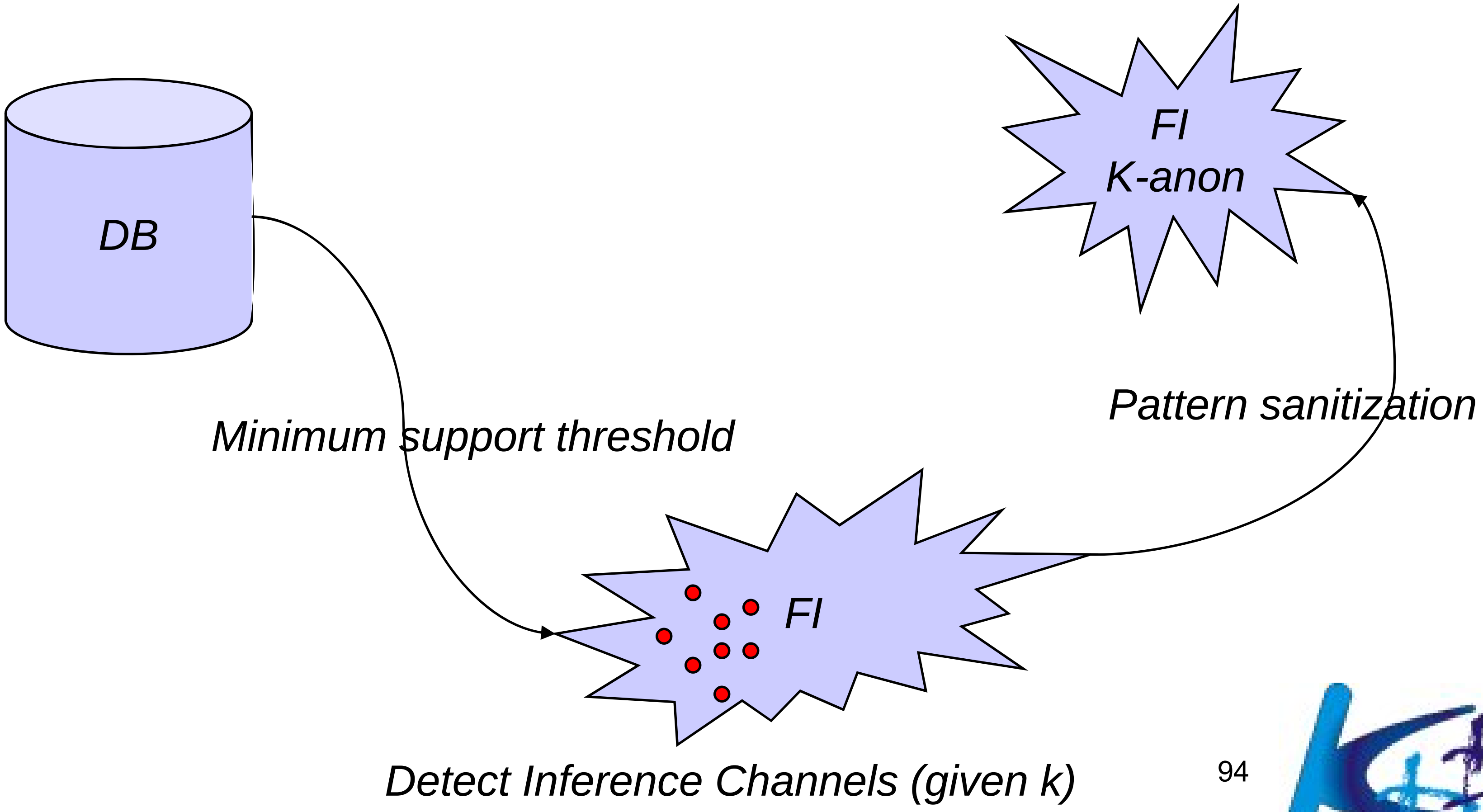
Since *sup(rule) / conf(rule) = sup(head)* we can derive:

**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 <u>not intended to be released</u> as it links public information regarding few people to sensitive data!)

- How to solve this kind of problems?

# The scenario


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

# Detecting Inference Channels

- See Atzori et al. K-anonymous patterns

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

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

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

✓ inclusion-exclusion principle used for support inference

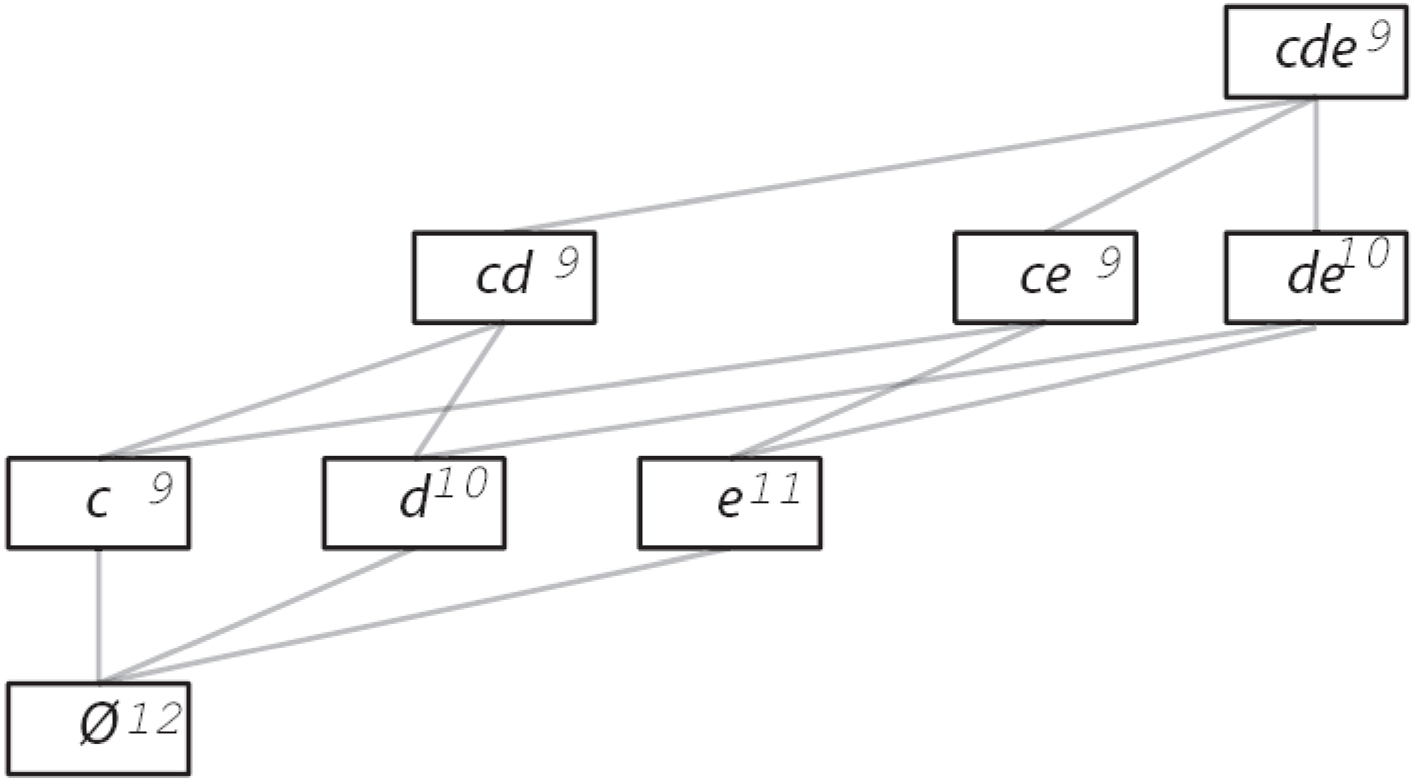
✓ support inference as key attacking technique

✓ 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

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



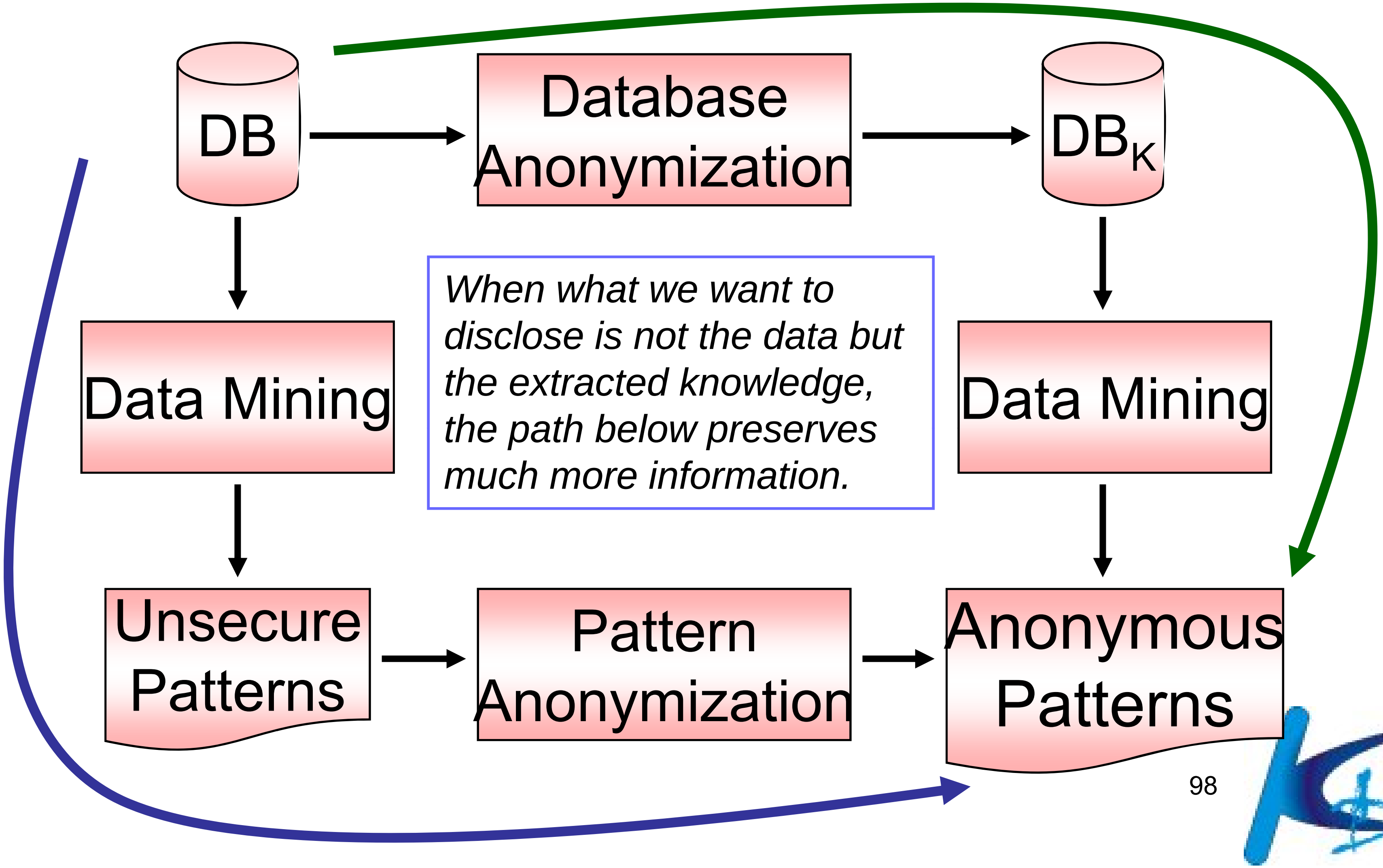
# Blocking Inference Channels

- 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



*When what we want to disclose is not the data but the extracted knowledge, the path below preserves much more information.*

# Plan of the tutorial

- The scenario of ubiquitous computing
  - Analytic opportunities and privacy threats
- Privacy and anonymity: prognosis and therapy
  - In data publishing: attack models and privacy-preserving techniques
  - In data mining: attack models and privacy-preserving data mining techniques
- **Privacy and anonymity in Location Based Services**
- Preliminary results on privacy and anonymity techniques in mobility data analysis
- Conclusion

# Privacy and Anonymity in Location- and Movement-Aware Data Analysis

# Service-Privacy Trade-off

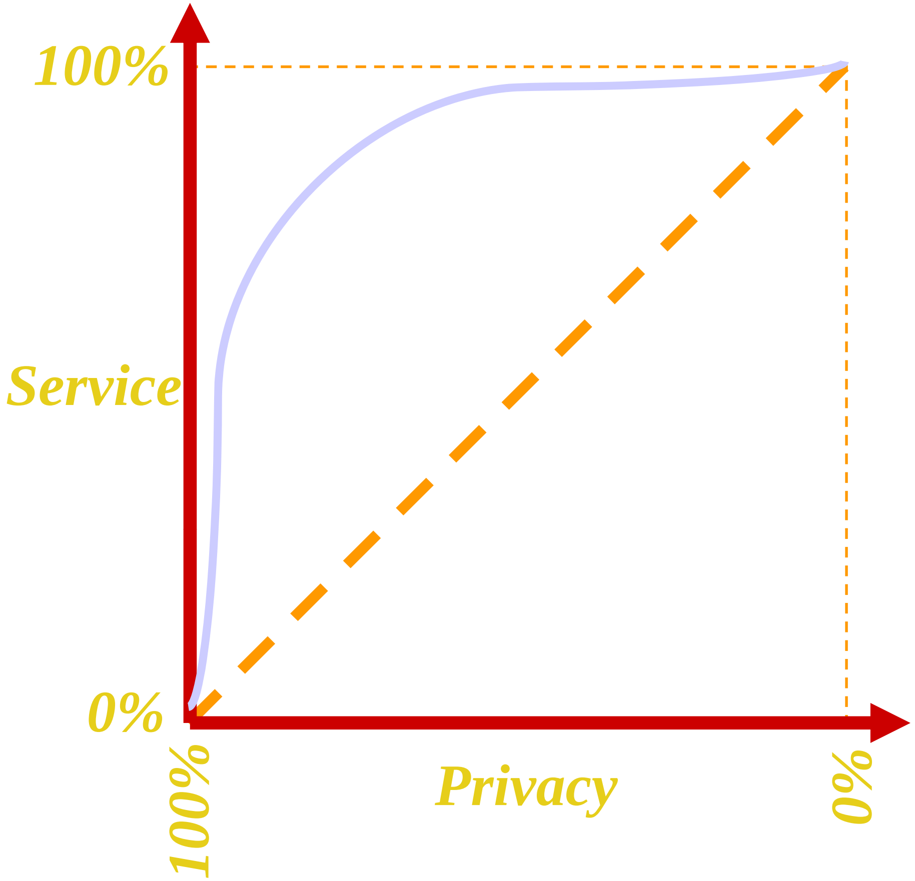
- **First extreme:**
  - A user reports her exact location ➔ 100% service

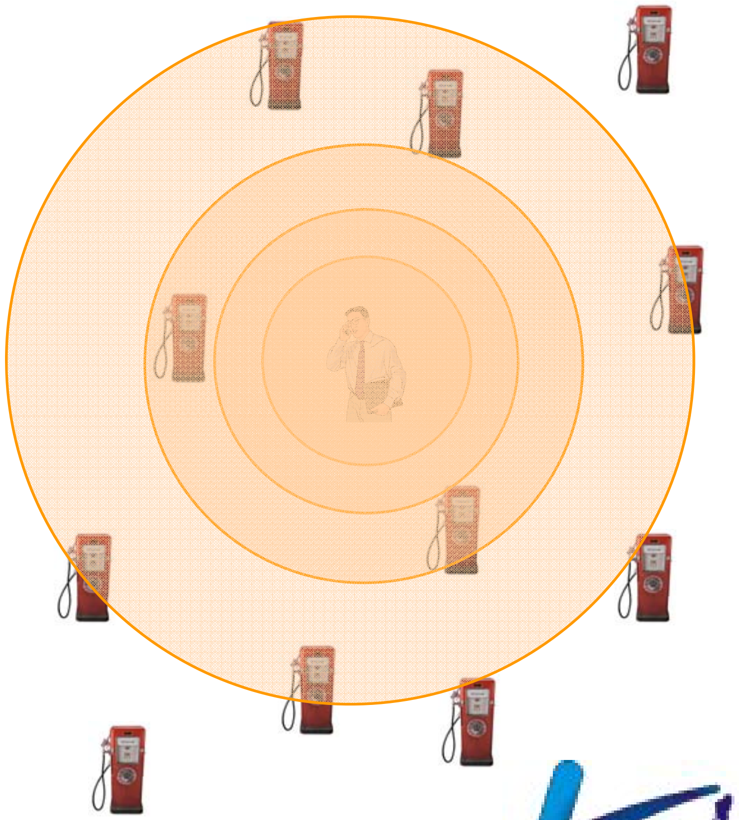
- **Second extreme:**
  - A user does NOT report her location ➔ 0% service

Desired Trade-off: A user reports a perturbed version of her location ➔ *x*% service

# Service-Privacy Trade-off

- Example: *What is my nearest gas station*

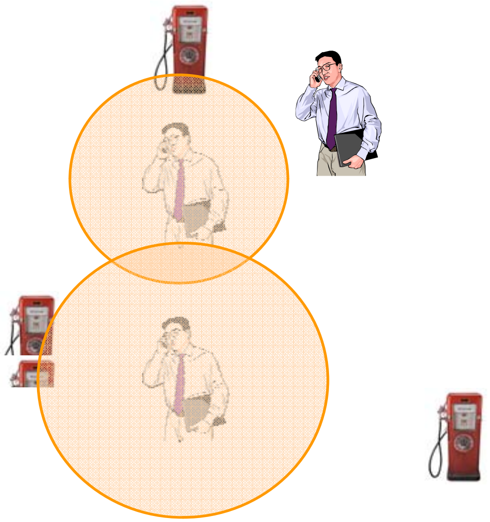




# Concepts for Location Privacy

## Location Perturbation

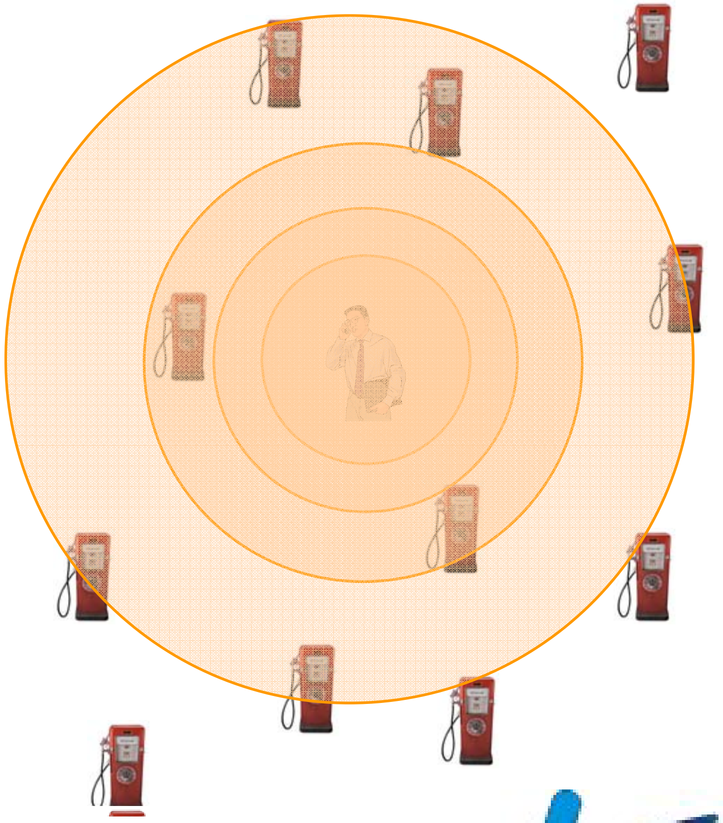
- The user location is represented with a wrong value
- The privacy is achieved from the fact that the reported location is false
- The accuracy and the amount of privacy mainly depends on how far the reported location form the exact location



# Concepts for Location Privacy

## Spatial Cloaking

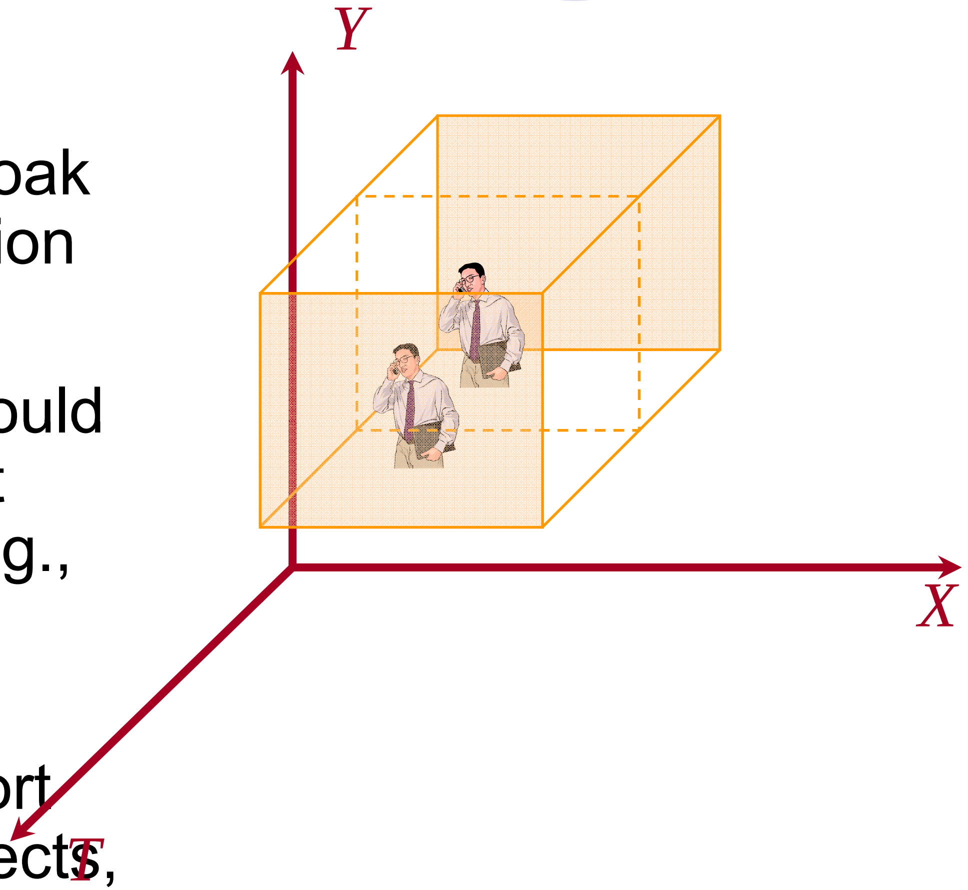
- Location *cloaking*, location *blurring*, location *obfuscation*
  - The user exact location is represented as a region that includes the exact user location
  - An adversary does know that the user is located in the *cloaked* region, but has no clue where the user is exactly located
  - The area of the *cloaked* region achieves a trade-off between the user privacy and the service



# Concepts for Location Privacy

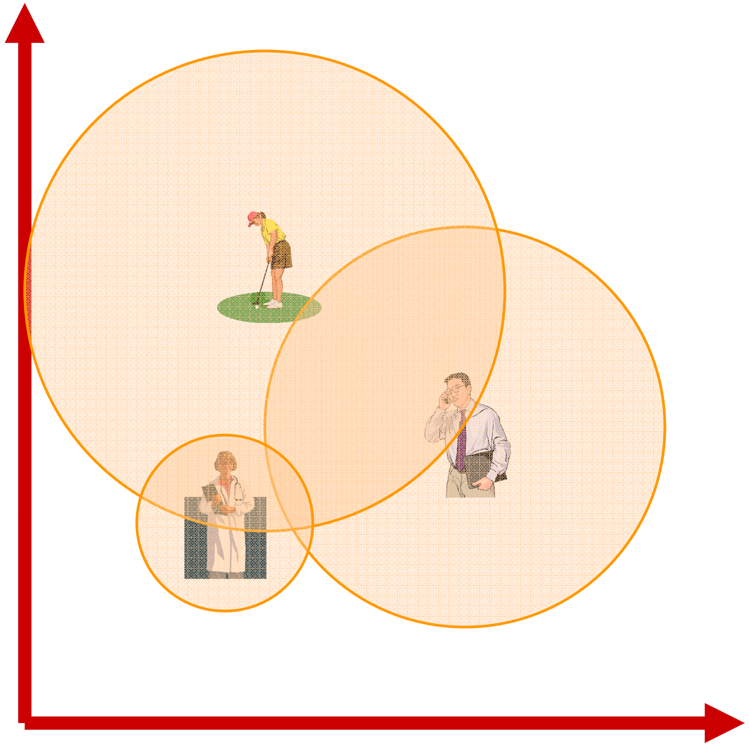
## Spatio-temporal Cloaking

- 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

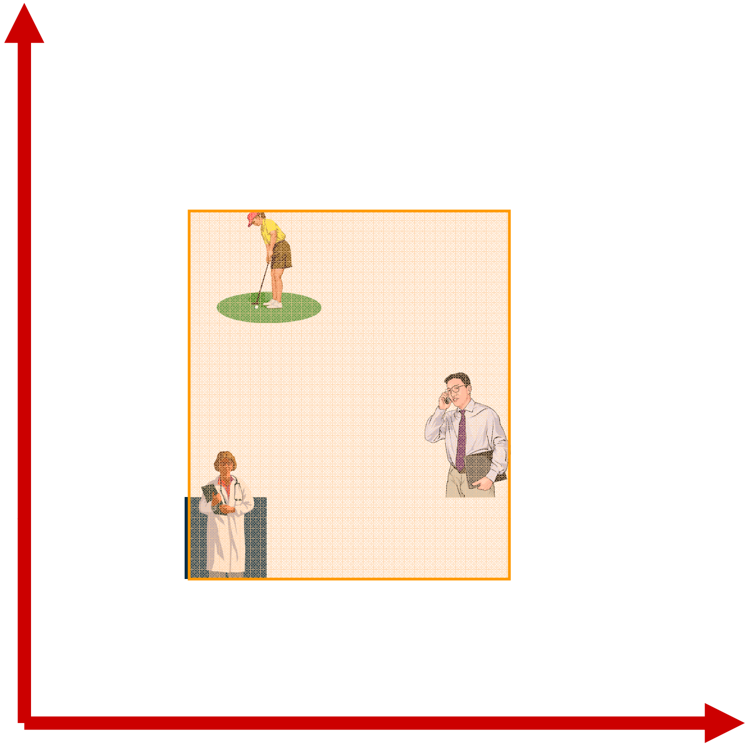


# Concepts for Location Privacy

## Data-Dependent Cloaking



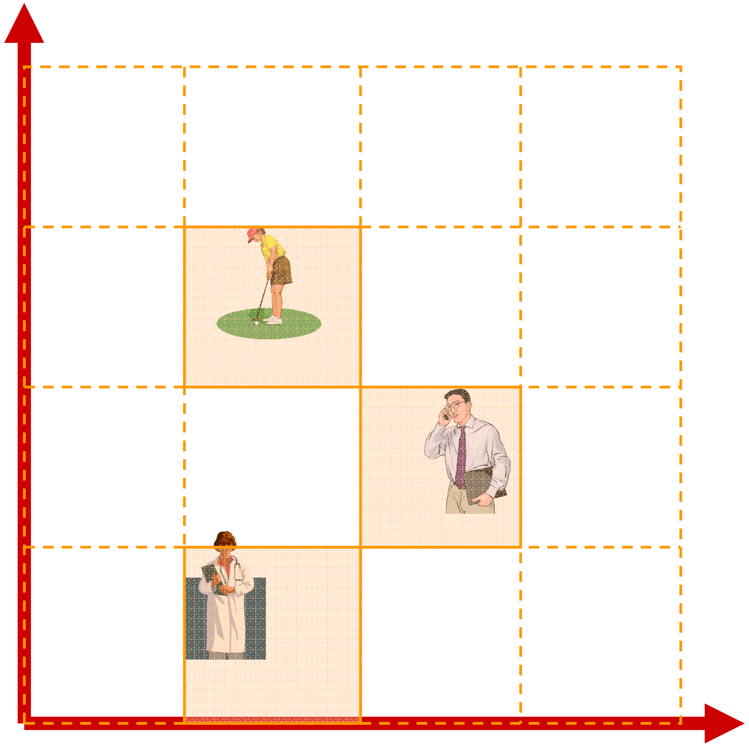
*Naïve cloaking*



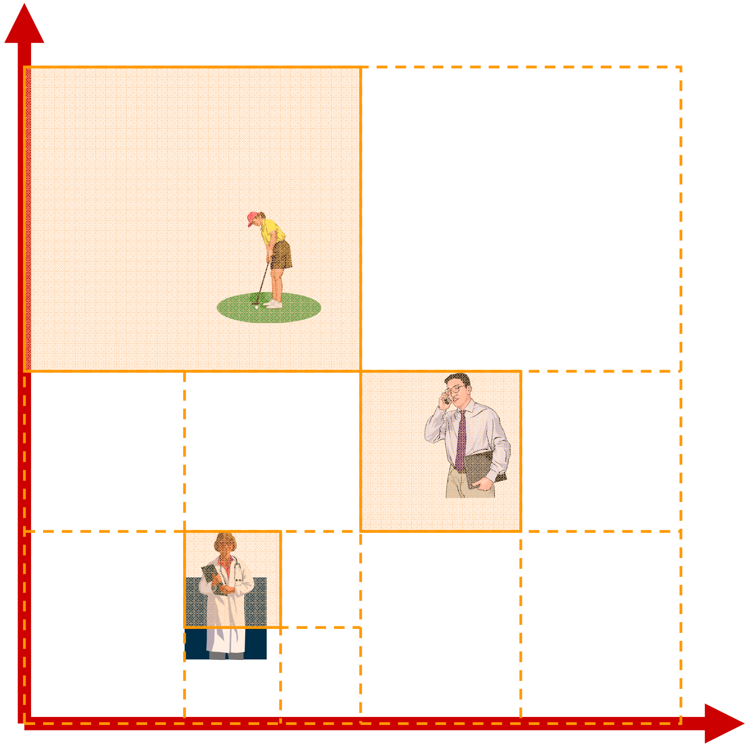
*MBR cloaking*

# Concepts for Location Privacy

## Space-Dependent Cloaking



*Fixed grid cloaking*

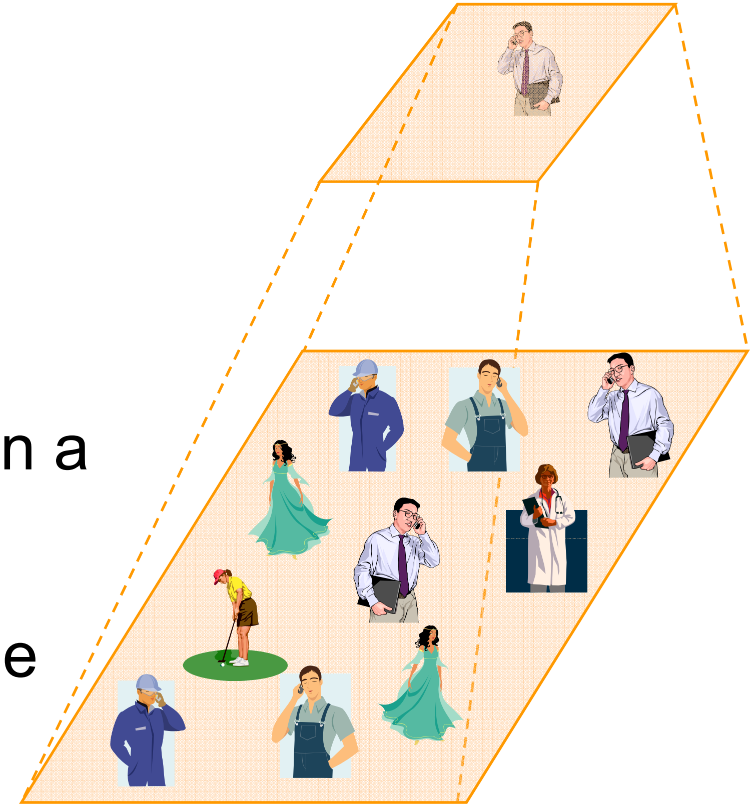


*Adaptive grid cloaking*

# 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***

# Concepts for Location Privacy

## Privacy Profile

- **Each mobile user will have her own *privacy-profile* that includes:**
  - *K*. A user wants to be *k*-anonymous
  - $A_{min}$. The minimum required area of the blurred area
  - $A_{max}$. The maximum required area of the blurred area
  - Multiple instances of the above parameters to indicate different privacy profiles at different times

| ***Time*** | ***k*** | $A_{min}$ | $A_{max}$ |
|---|---|---|---|
| 8:00 AM - | 1 | —— | —— |
| 5:00 PM - | 100 | 1 mile | 3 miles |
| 10:00 PM - | 1000 | 5 miles | —— |

# Summary

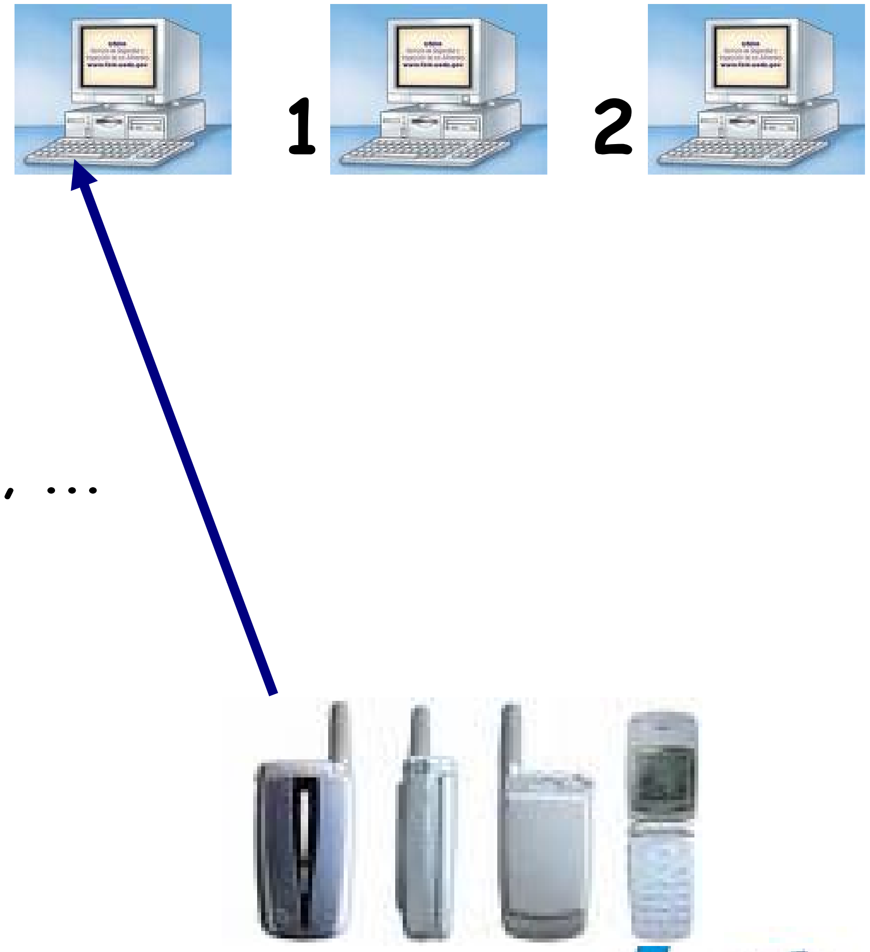
- Location-based services scenario and privacy issues
- Real-time Anonymity of point-based services
- Real-time Anonymity of trajectory-based services
- Enhancing privacy in trajectory data
  - By confusing paths
  - By introducing dummy trajectories
  - By reducing frequency of user requests
- Introducing Dummy trajectories for enhancing privacy
- Privacy-aware location query systems

# Location-Based Services and Privacy Issues

# Location-Based Services and Privacy Issues

- Context: communication for Location-based services (LBS)

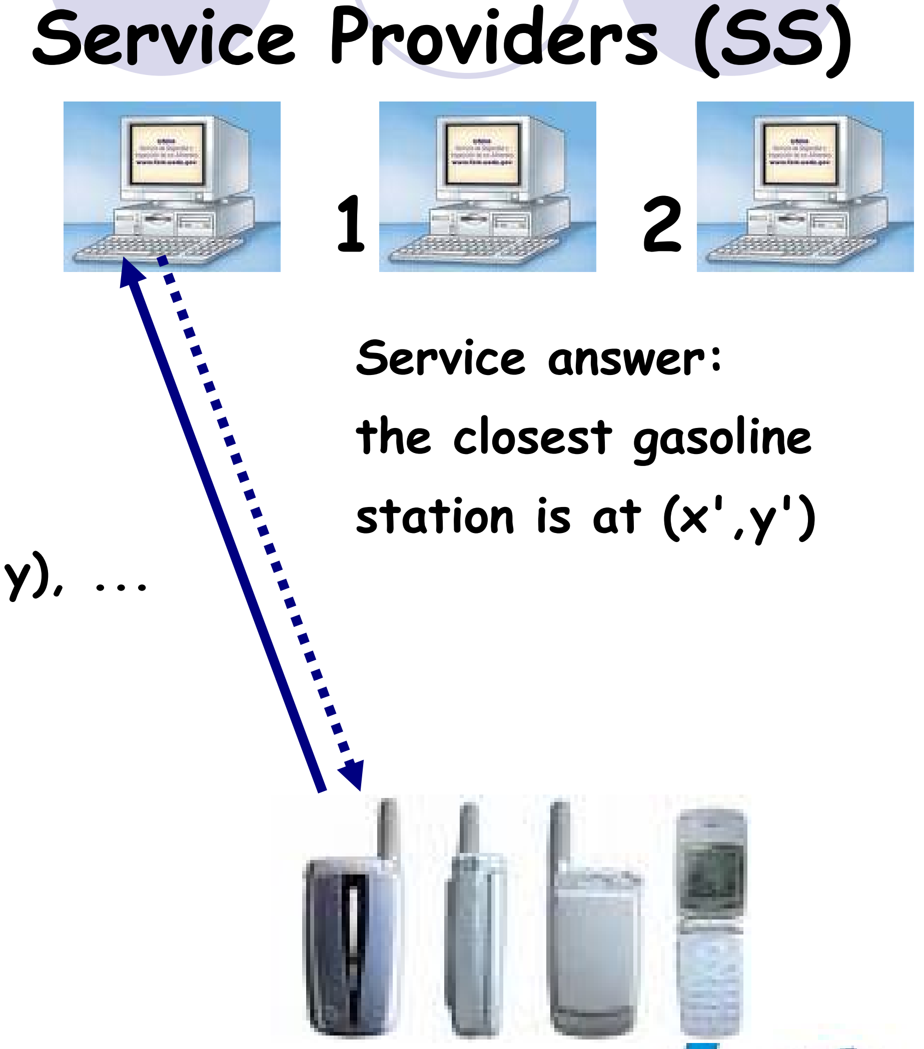
Service Providers (SS)



User Request: Jack, (x,y), ...

# Location-Based Services and Privacy Issues

- Context: communication for Location-based services (LBS)



Service Providers (SS)

Service answer:

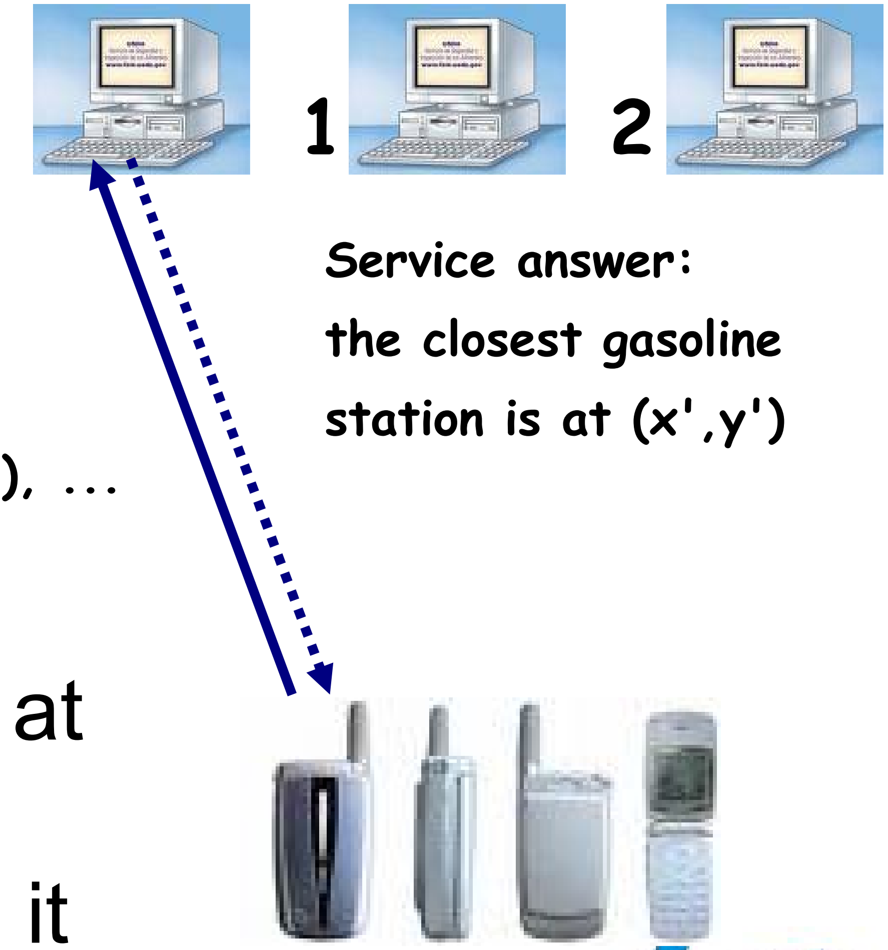
the closest gasoline station is at (x',y')

User Request: Jack, (x,y), ...

# Location-Based Services and Privacy Issues

- Context: communication for Location-based services (LBS)

Service Providers (SS)



Service answer: the closest gasoline station is at (x',y')

User Request: Jack, (x,y), ...

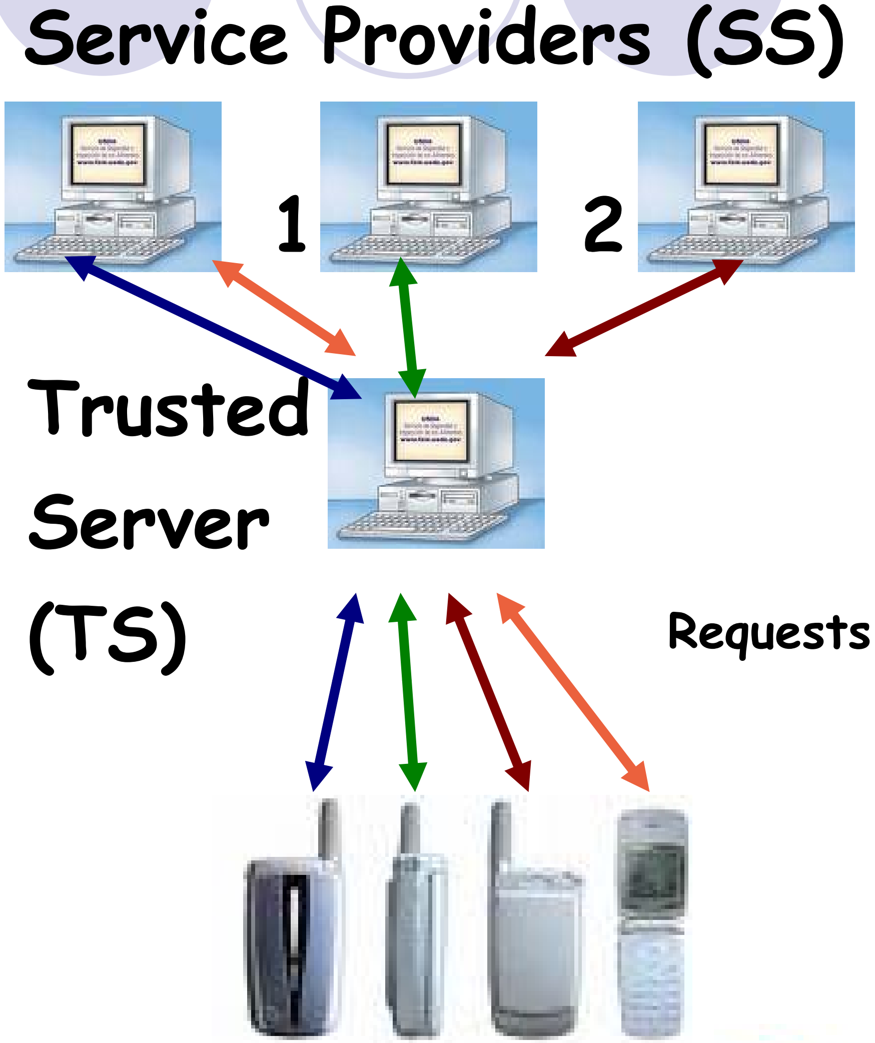
Privacy Issues:

SS knows that Jack is at x,y at time of request

With several requests, it is possible to trace Jack

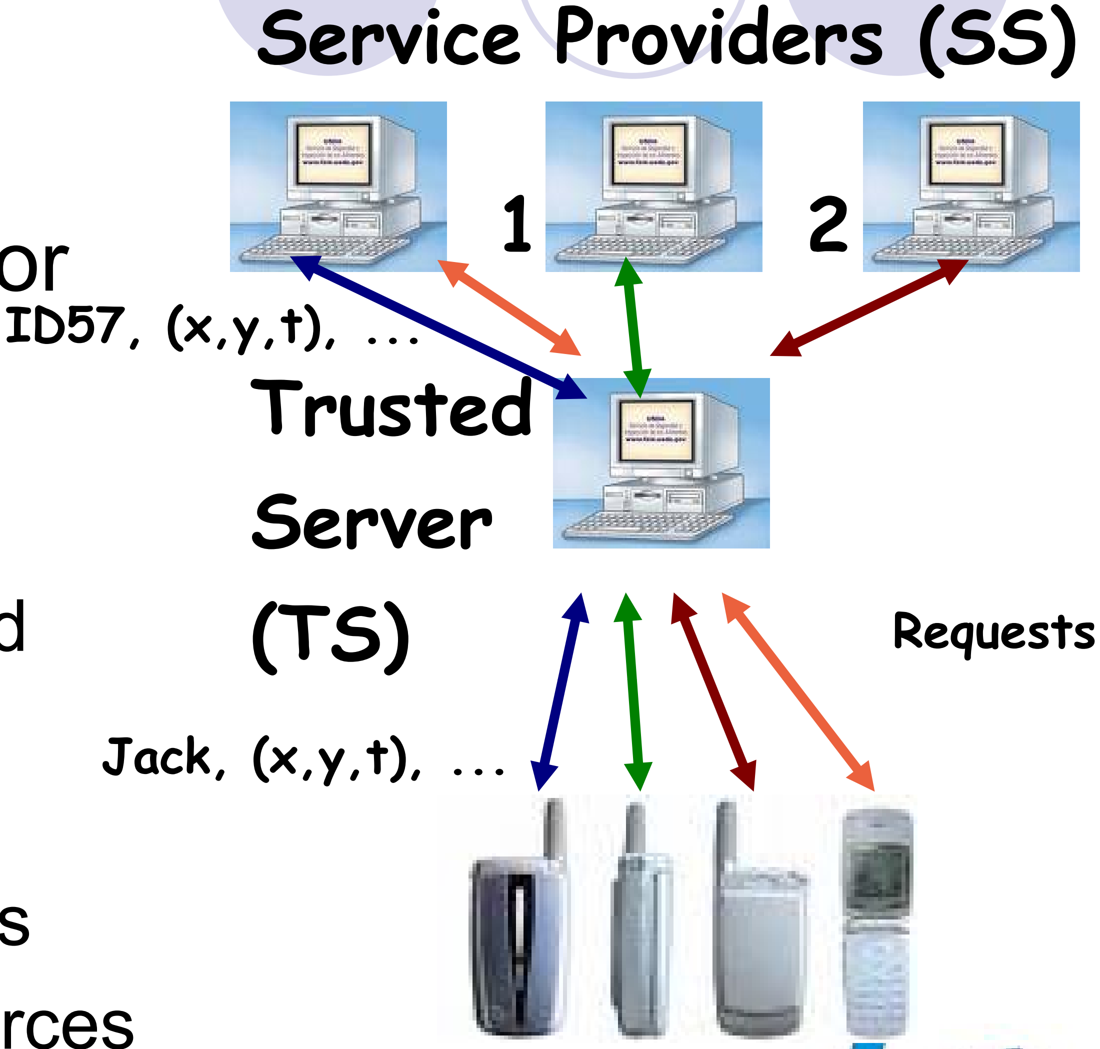
# Personalized Anonymization for Location Privacy

- Context: communication for Location-based services (LBS)
  - Trusted Server between user and LBS



# Trusted Server

- Context: communication for Location-based services (LBS)
  - Trusted Server between user and LBS
- Privacy:
  - TS masks Names
  - Optionally it enforces other privacy policies

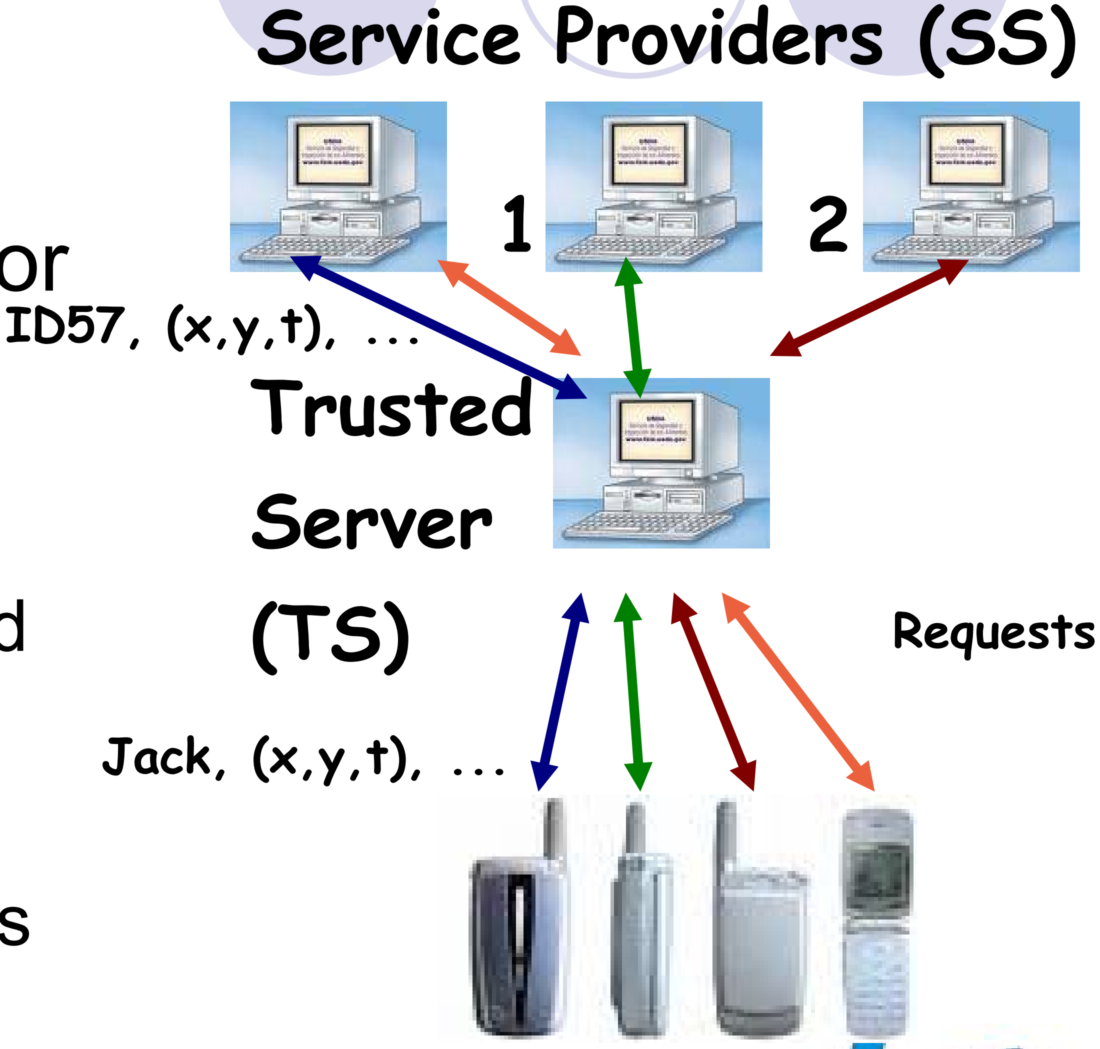


# Real-time Anonymity of point-based services

## Location-Privacy in Mobile Systems:
## A Personalized Anonymization Model
## [*Gedik & Liu, ICDCS05*]

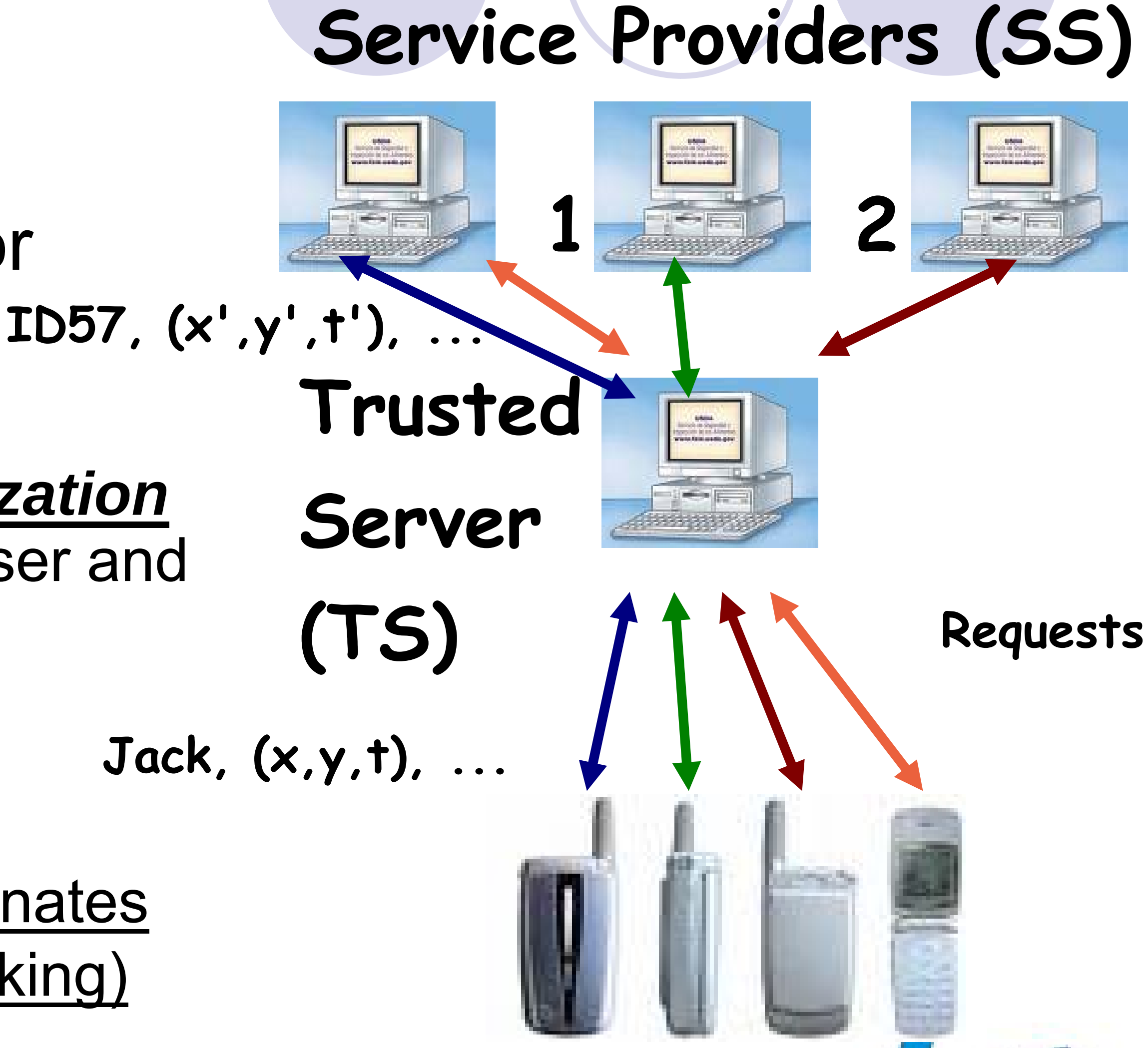
# Personalized Anonymization for Location Privacy

- Context: communication for Location-based services (LBS)
  - Trusted Server between user and LBS
- Privacy:
  - TS masks Names



# Personalized Anonymization for Location Privacy

- Context: communication for Location-based services (LBS)
  - Trusted ***Anonymization*** Server between user and LBS
- Privacy:
  - TS masks Names
  - Space-time coordinates are distorted (cloaking)



# Personalized Anonymization for Location Privacy

- CliqueCloak Algorithm
  - Mask location and temporal data by perturbation
  - Based on delaying messages and lowering the spatio/temporal resolution

- Each user can specify her own parameters
  - K, QoS (Space Resolution, Time Precision)
- It relies on K-Anonymity
  - A privacy framework developed in the context of relational tables

# K-Anonymization

- Anonymity: "*a state of being not identifiable within a set of subjects, the Anonymity Set*"
- K-Anonymity: |Anonymity Set| ≥ k
- Subjects of the data cannot be re-identified while the data remain practically useful
  - By attribute generalization and tuple suppression

# An example on tables: Original Database

```
Race  DOB        Sex ZIP   Problem
----- ---------- --- ----- ---------------
black 05/20/1965 M   02141 short of breath
black 08/31/1965 M   02141 chest pain
black 10/28/1965 F   02138 painful eye
black 09/30/1965 F   02138 wheezing
black 07/07/1964 F   02138 obesity
black 11/05/1964 F   02138 chest pain
white 11/28/1964 M   02138 short of breath
white 07/22/1965 F   02139 hypertension
white 08/24/1964 M   02139 obesity
white 05/30/1964 M   02139 fever
white 02/16/1967 M   02138 vomiting
white 10/10/1967 M   02138 back pain
```

# An example on tables:
# A 2-anonymized database

```
Race  DOB        Sex ZIP   Problem
----- ---------- --- ----- ---------------
black 1965       M   02141 short of breath
black 1965       M   02141 chest pain
black 1965       F   02138 painful eye
black 1965       F   02138 wheezing
black 1964       F   02138 obesity
black 1964       F   02138 chest pain
white 196*       *   021** short of breath
white 196*       *   021** hypertension
white 1964       M   02139 obesity
white 1964       M   02139 fever
white 1967       M   02138 vomiting
white 1967       M   02138 back pain
```

# Messages

- ms = <uid, rno, {t,x,y}, k, {dt, dx, dy}, C>

- Where
  - (uid, rno) = user-id and message number
  - {t,x,y} = L(ms) = spatio-temporal location
  - K = anonymity threshold
  - dt, dx, dy = quality of service constraints
  - C = the actual message

  - Bcn(ms) = [t-dt, t+dt] [x-dx, x+dx] [y-dy, y+dy]
  - Bcl(ms) = spatio-temporal **cloaking box** of ms, contained in Bcn(ms)

# Definition of Location k-anonymity

- For a message ms in S and its perturbed format mt in T, the following condition must hold:

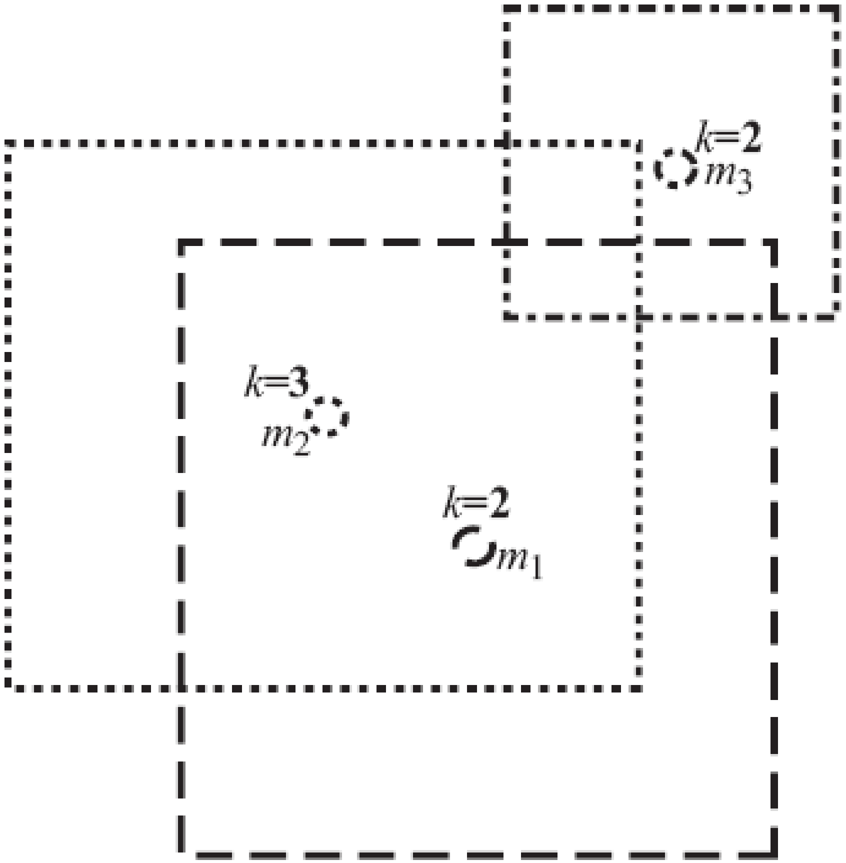
$\forall$ T' $\subset$ T, s.t. mt $\in$ T', |T'| $\geq$ ms.k,

$\forall$ {$mt_i$, $mt_j$} $\subset$ T', $mt_i$.uid $\neq$ $mt_j$.uid and

$\forall$ mti $\in$ T', Bcl($mt_i$) = Bcl(mt)

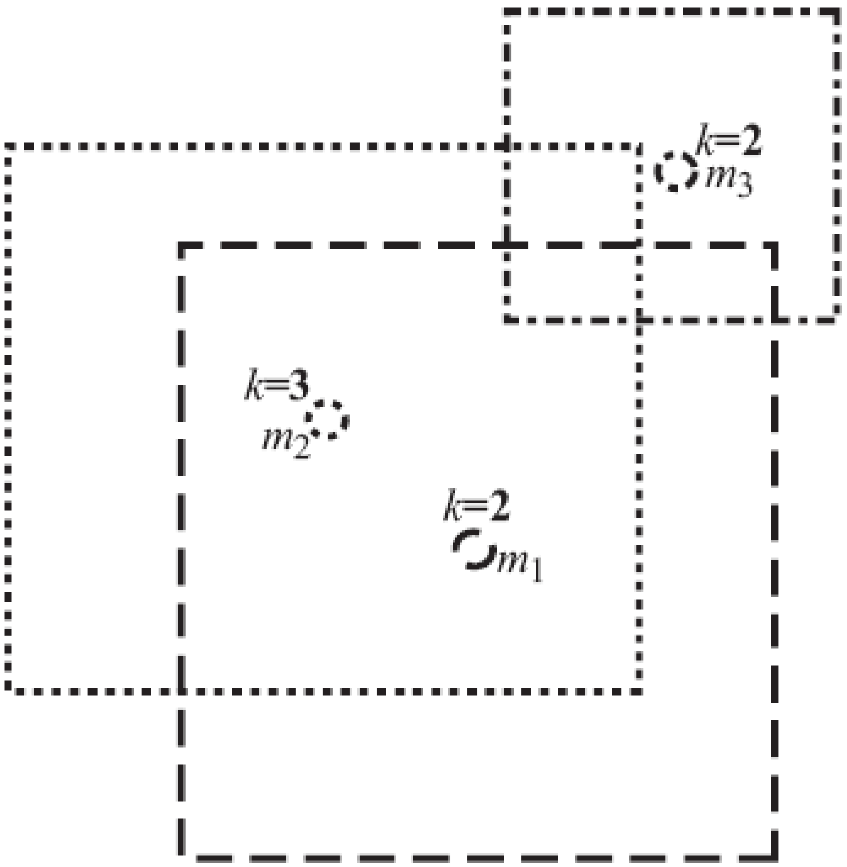
- ms.C = mt.C , mt.uid = hash(ms.uid)

# Clique-Cloak Algorithm: Spatial Layouts

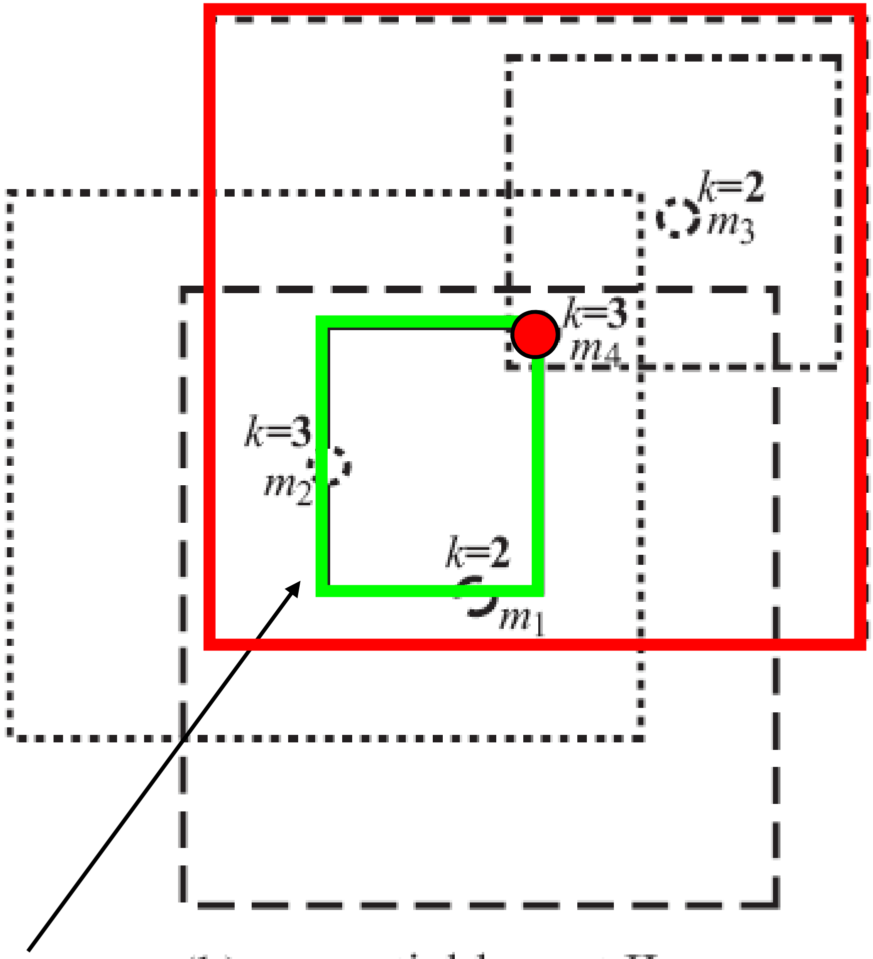


(a) spatial layout I

# Clique-Cloak Algorithm: Spatial Layouts



(a) spatial layout I



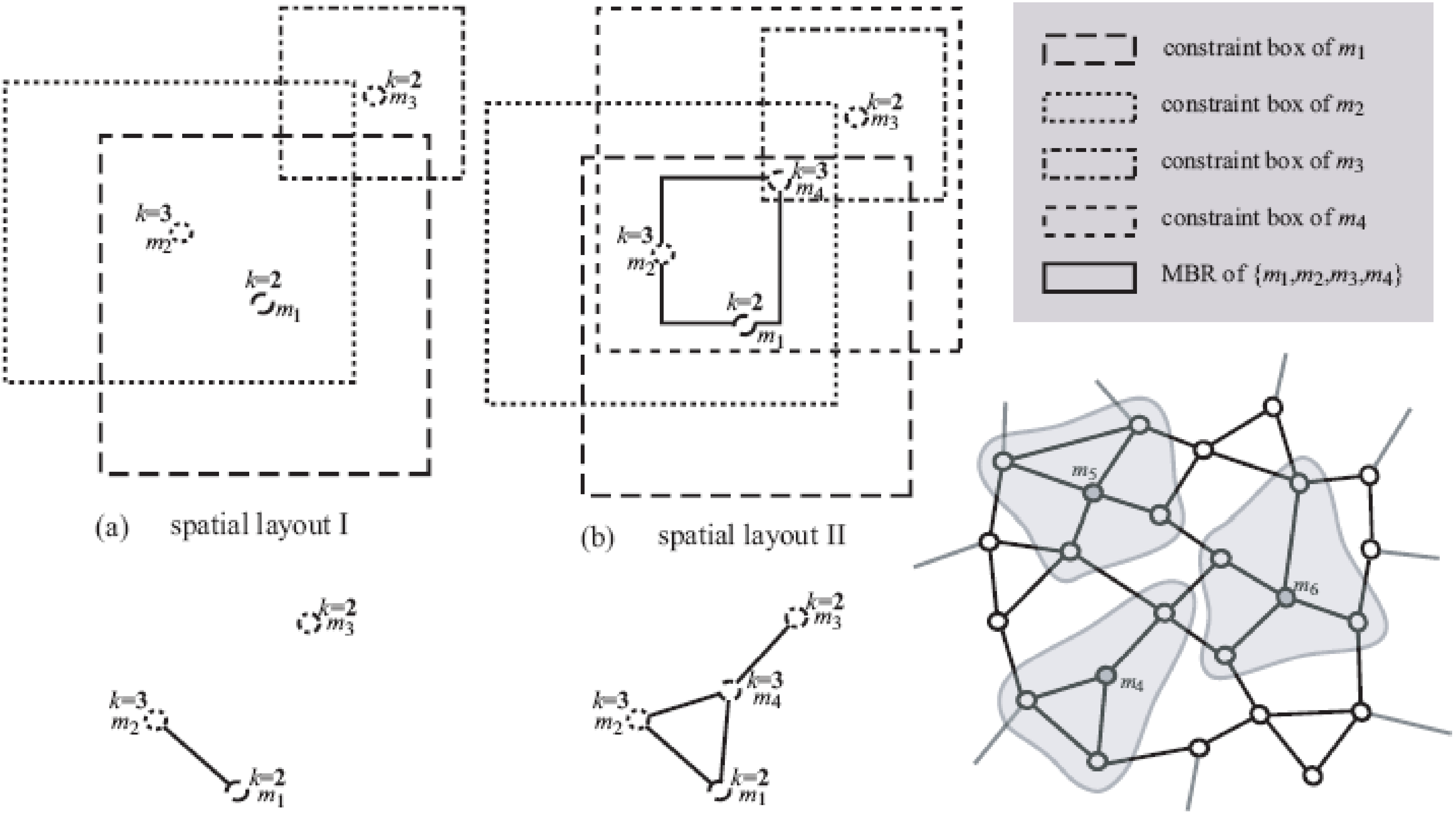
(b) spatial layout II

**minimum bounding rectangle**

# Constraint Graphs

- G(S,E) is an undirected graph
- S is the set of vertices
  - Each representing a message received at the message perturbation engine
- E is the set of edges, $(ms_i, ms_j) \in E$ iff
  1. $L(ms_i) \in Bcn(ms_j)$
  2. $L(ms_j) \in Bcn(ms_i)$
  3. $ms_i.uid \neq ms_j.uid$
- $ms_i$ is anonymizable iff $\exists$ an l-clique M s.t. $\forall\ ms_i \in M$ we have $ms_i.k \leq l$

# Clique-Cloak Algorithm: Constraint Graphs



(a) spatial layout I

(b) spatial layout II

(c) constraint graph I

(d) constraint graph II

(e) constraint graph

# Clique-Cloak Algorithm: Four Steps

- Data structures: Message Queue, Multidimensional Index, Constraint Graph, Expiration Heap
- Steps:
  1. Zoom-in, **i.e. Locate neighbors messages of popped message m, update data structures (Index and Graph)**
  2. Detection, **(local k-search sub-algorithm) find a m.k-clique in the subgraph {m} U {$m_j$ $\in$ neighbor of m | $m_j$.k $\leq$ m.k }**
  3. Perturbation, **use the MBR of the clique as cloaking box of the messages in the clique**
  4. Expiration, **through an expiration heap**

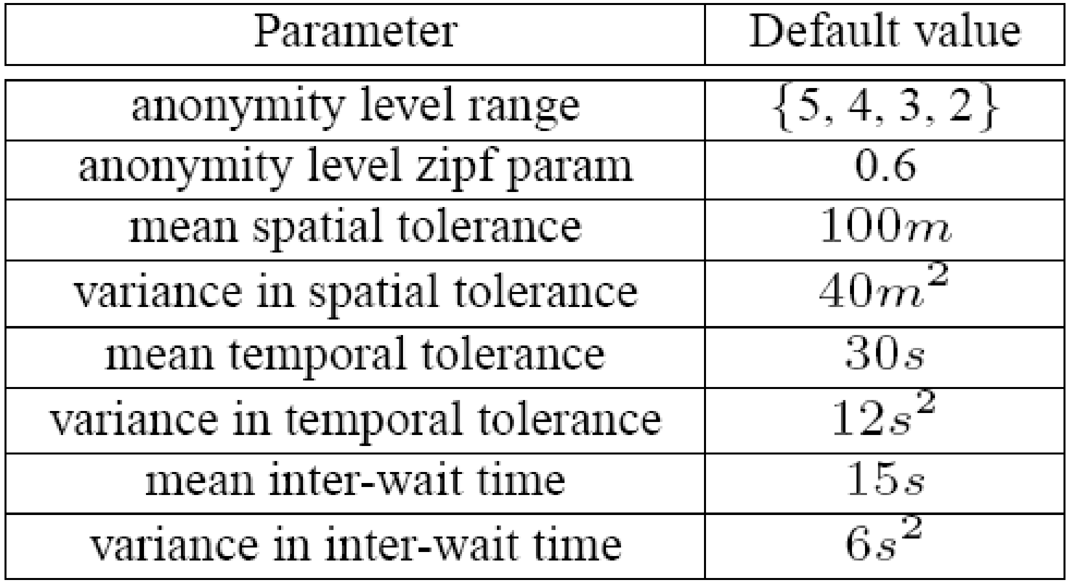
# An Optimization: nbr-k Search Algorithm

Detection, (local k-search) find a m.k-clique in the subgraph of the message and its neighbors $m_j$ s.t. $m_j.k \leq m.k$

Detection, (nbr k-search) find the *largest* clique M in the subgraph of the message and its neighbors $m_j$ s.t. $m_j.k \leq |M|$

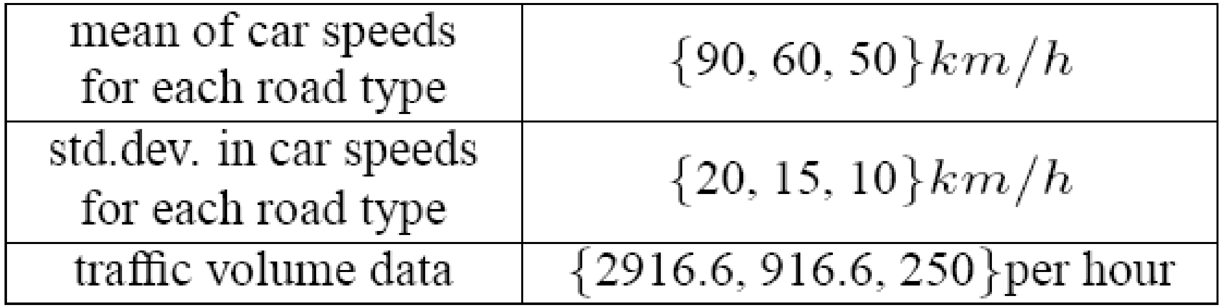
The suggested implementation makes use of local k-search varying k in a decreasing order

# Synthetic Data Generator



| Parameter | Default value |
|---|---|
| anonymity level range | $\{5, 4, 3, 2\}$ |
| anonymity level zipf param | 0.6 |
| mean spatial tolerance | $100m$ |
| variance in spatial tolerance | $40m^2$ |
| mean temporal tolerance | $30s$ |
| variance in temporal tolerance | $12s^2$ |
| mean inter-wait time | $15s$ |
| variance in inter-wait time | $6s^2$ |

Table 1: Message generation parameters



| | |
|---|---|
| mean of car speeds for each road type | $\{90, 60, 50\}km/h$ |
| std.dev. in car speeds for each road type | $\{20, 15, 10\}km/h$ |
| traffic volume data | $\{2916.6, 916.6, 250\}$per hour |

Table 2: Car movement parameters

Chamblee region of state of Georgia in USA (160km$^2$)

10,000 cars

# Experiments: Success rate and anonymity level

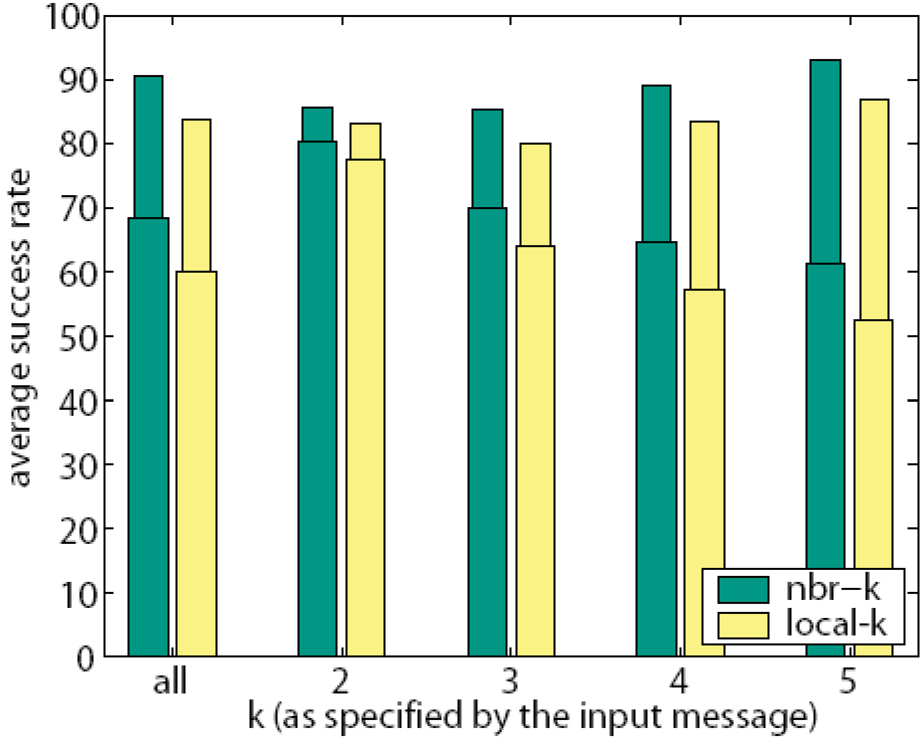


Figure 2: Success rates for different $k$ values

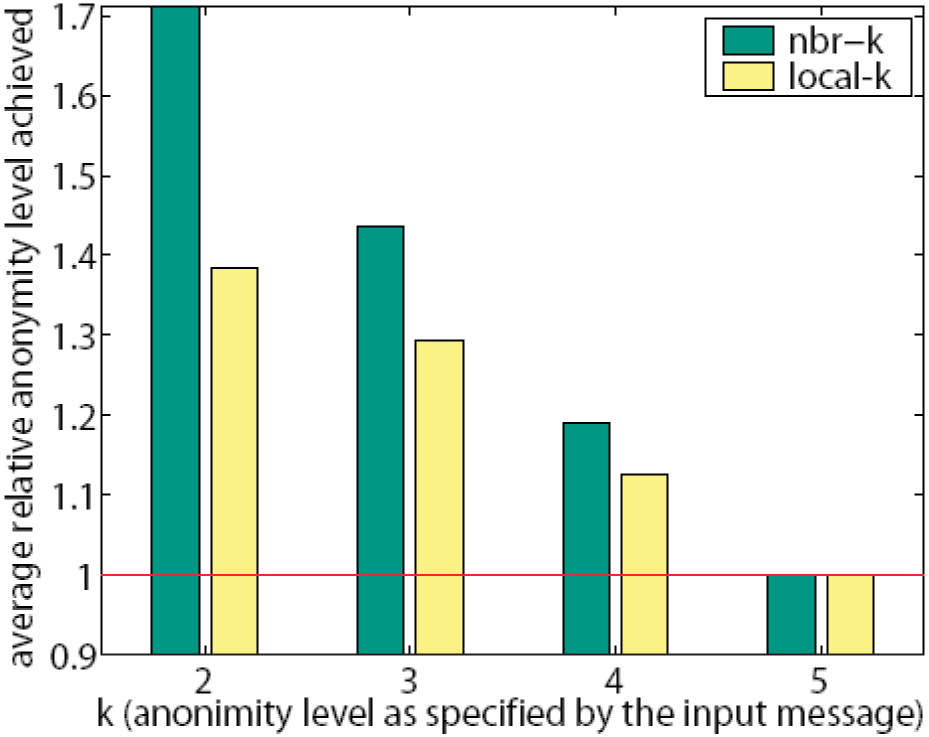


Figure 3: Relative anonymity levels for different $k$ values

**Accuracy < 18m in 75% of the cases!**

# Other Approaches to privacy-preserving point-based services

- Other different privacy-preserving algorithms have been presented
  - Most of them rely on the concept of k-anonymity
- Noise Addiction / Uncertainty
  - Other authors proposed a framewok to augment uncertainty to location data in a controlled way

**Privacy-preserving location-dependent query processing**

**[Atallah and Frikken, *ICPS04*]**

**Preserving User Location Privacy in Mobile Data Management Infrastructures**

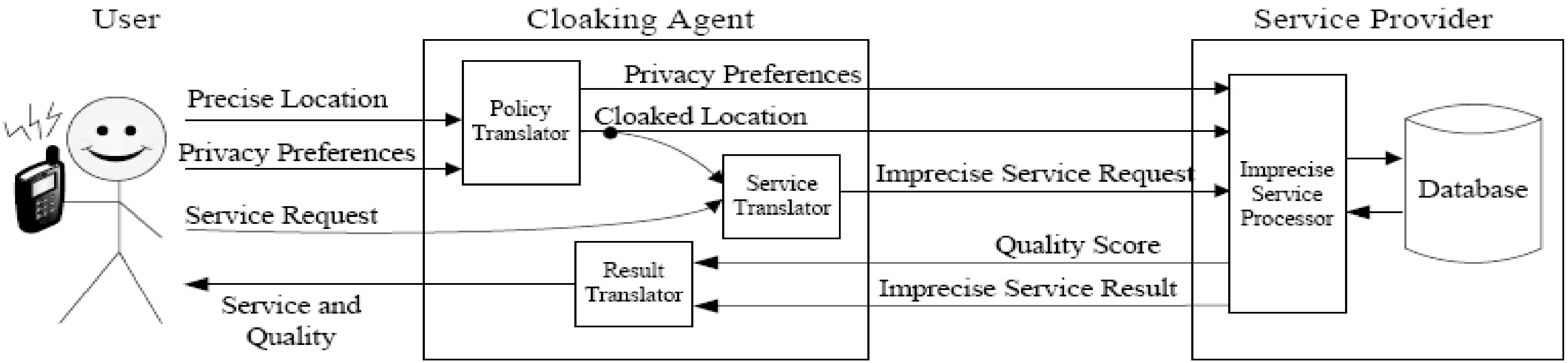
**[Cheng *et al.*, *PET06*]**

# Approach 1: Perturbation

- Random perturbation of client's location
  - Chosen by client
  - Variable, and not known to server
- Large enough to "hide" exact location (privacy)
- Small enough to avoid "too much damage" to quality of answer
- Issue: Quantifying the damage to answer
- Requests are ST regions

# Approach 2: Grid Method

- The plane is covered with squares tiles
- Client sends as “query” the tile that contains the true query point
  - Hence tile size known to both client and server
- Large tiles imply better privacy, but also a cost
  - Cost in efficiency (if exact answer)
  - Cost in quality of answer (if most efficient)



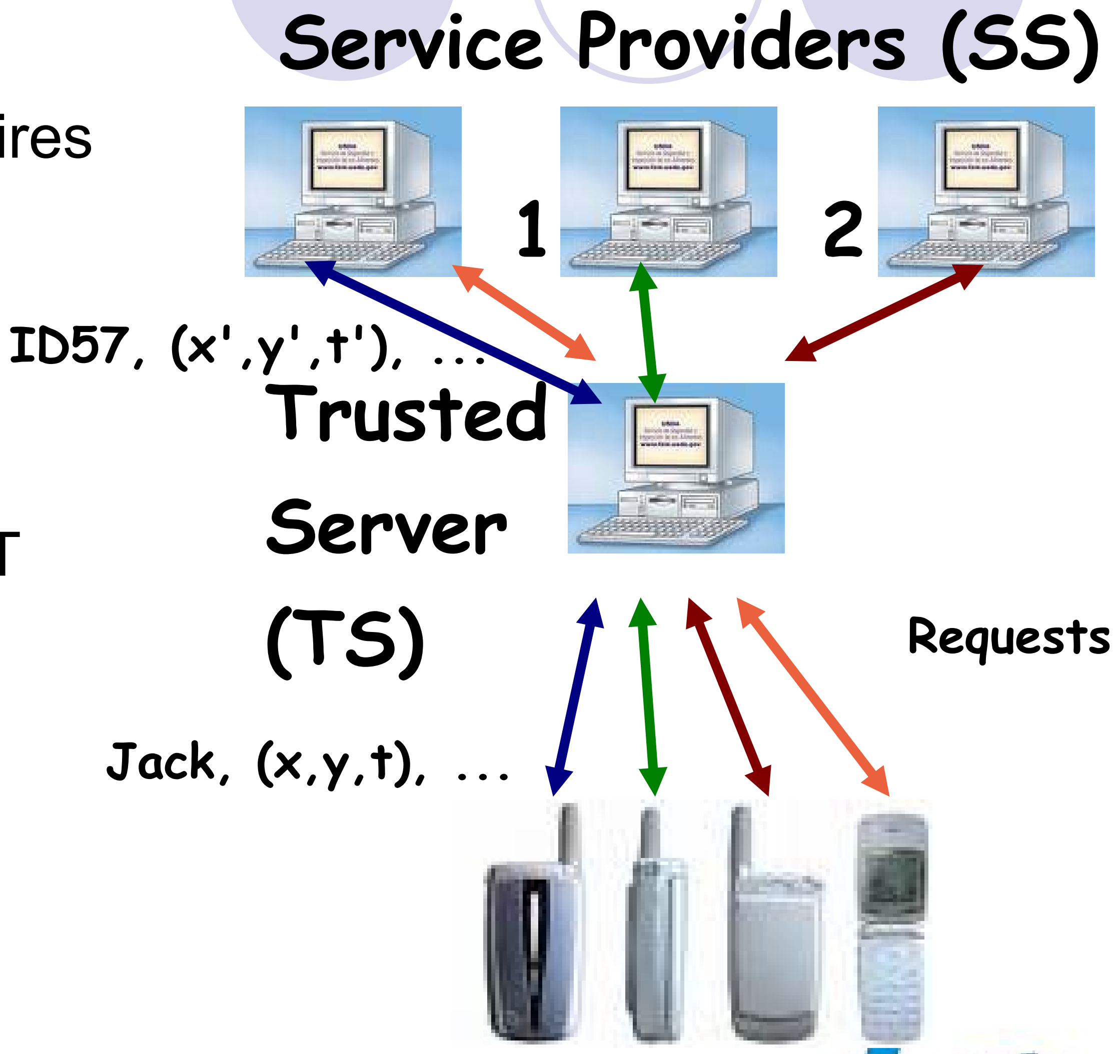
# Real-time Anonymity of trajectory-based services

## Location Privacy of ST Sequences

[*Bettini et al., SDM workshop, VLDB05*]

# Location Privacy of ST Sequences

- **Problem**:
  - What if the service requires authentication and the same user makes a number of requests?
  - Threat: sequences of ST points can be used to breach anonymity (e.g., tracing users from their own homes)



# Location Privacy of ST Sequences: LBQID

- Location-Based Quasi-Identifiers are Spatio-temporal patterns
  - <AreaCondominium [7am,8am], AreaOfficeBldg [8am,9am], AreaOfficeBldg [4pm,5pm], AreaCondominium [5pm,6pm]> Recurrence: 3.Weekdays * 2.Weeks
- If the pattern(s) matches the sequence of requests of a user, then enforcing k-anonymity is required (over trajectories)

# Historical k-Anonymity

- Personal History of Locations (PHL)
  - sequence of ST points associated to a given user (its trajectory, not necessarily requests)
  - e.g. <x1,y1,t1> , <x2,y2,t2> , ...<xn,yn,tn>
- Historical *k*-Anonymity (H*k*A)
  - A set of requests issued by the same user statisfies H*k*A if there exist k-1 PHLs P1,...P(k-1) for k-1 different users s.t. The set of requests “match” P1... P(k-1)
    - Requests are ST regions

# ST generalization algorithm

- A simple algorithm is presented
  - O(k*n) where n is the number of location point in the TS
  - Very naïve and unpractical for a number of reasons
    - Mainly, too many suppressions
- Another unlinking technique suggested
  - Changing ID or disabling requests for a period of time to confuse the SP (necessary since as the sequence length grows, HkA become impossible to reach)

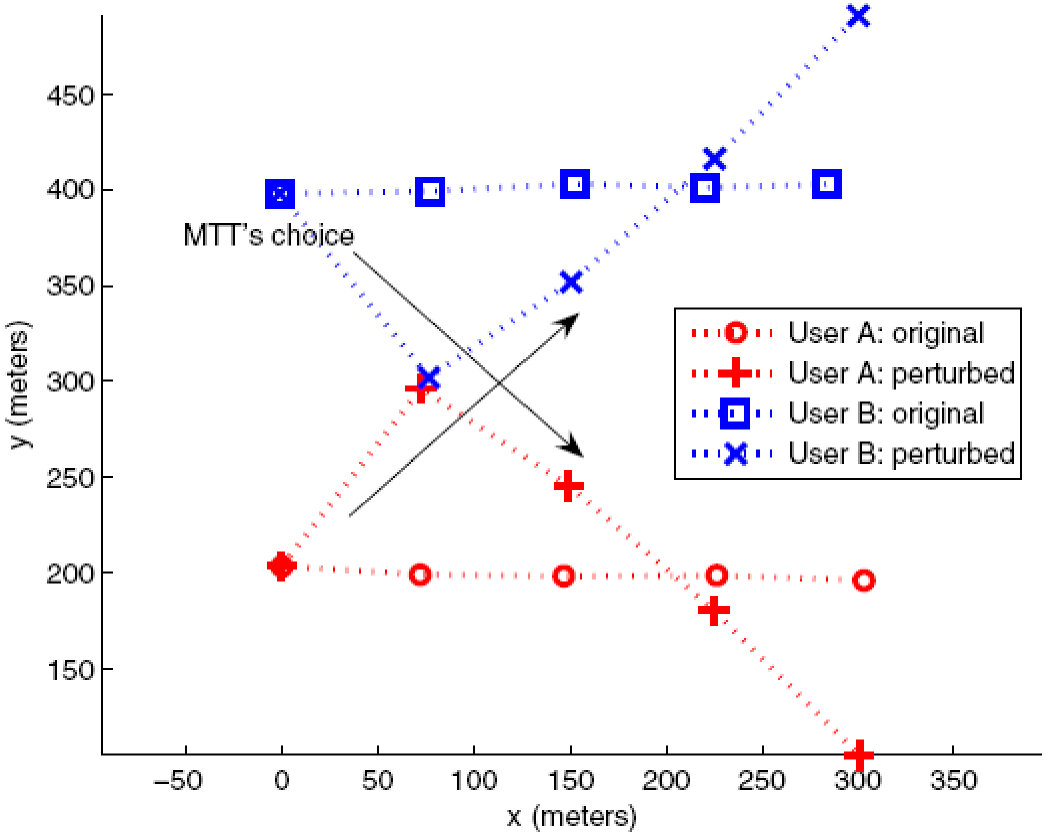
# Enhancing privacy in trajectory data:
# by path confusion
# by introducing dummies
# by reducing frequency of user requests

**Protecting location privacy through Path Confusion**
**[*Baik Hoh and Marco Gruteser, SECURECOMM05*]**

**Anonymous Communication Technique using Dummies for Location-Based Services**
**[*Hidetoshi Kido, Yutaka Yanagisawa,*
*Tetsuji Satoh, ICPS05*]**

**Protecting Privacy in Continuous Location-Tracking Applications**
**[*Marco Gruteser and Xuan Liu*,**
**IEEE Security and Privacy March/April 2004]**

# Path confusion forces paths to cross each other reducing traceability of users



- blue and red users move in parallel.
- Path-Perturbation algorithm perturbs the parallel segments into a crossing path

# Dummies: Possible Threat / Motivations

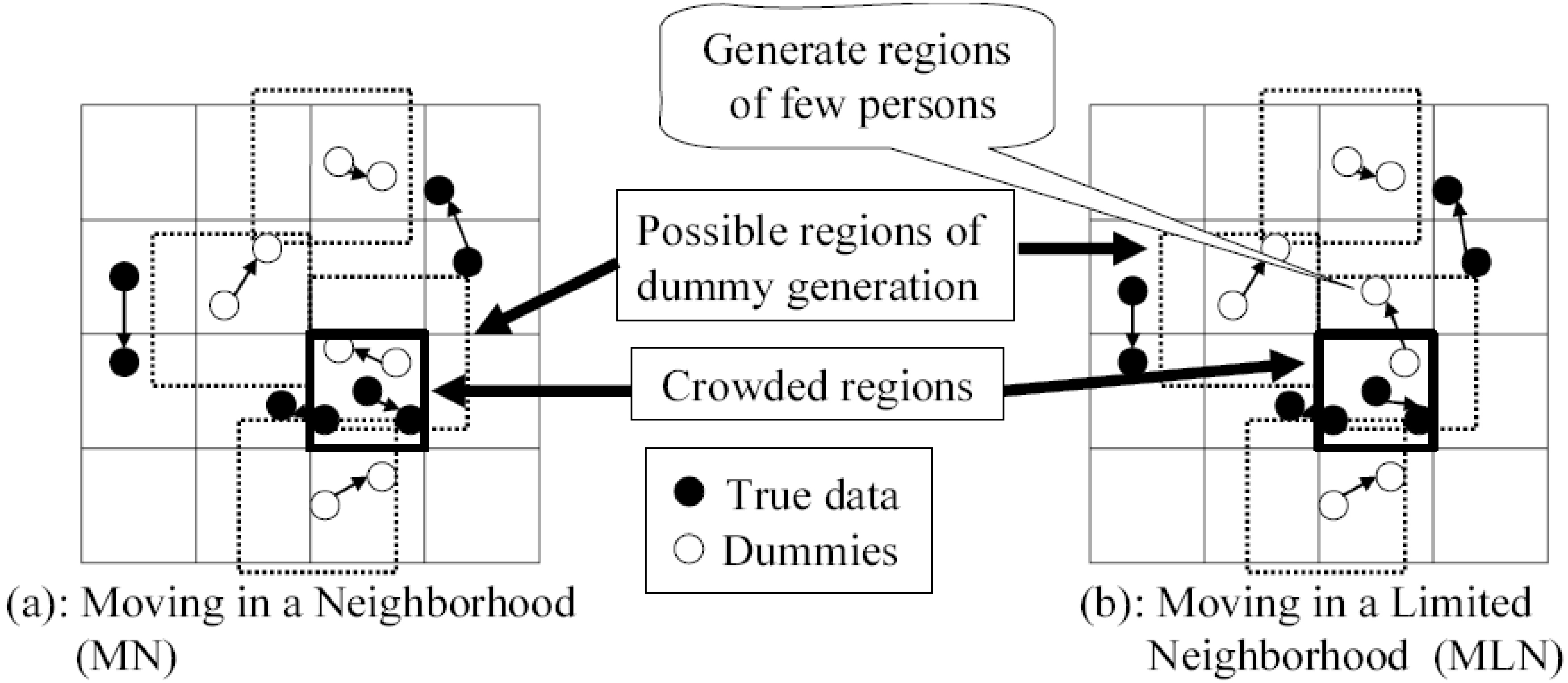
- An LBS gives a user information about when buses will arrive at the nearest stop in a particular vicinity. For example, a person goes to a clinic every week and uses this service at his house and the clinic each time. If such position data are accumulated and analyzed, a staff member or a patient of the clinic may learn the person's address.
- Based on position data, location privacy can be invaded. To protect it, service providers must be prevented from learning the true position of users
  - It is necessary to anonymize the position data
  - Try to solve problems in Path-Confusion ***when users are traced for long times***

# Introducing Dummy Data

- To preserve location privacy, users send dummy data together with real data, and the server cannot distinguish, replying both kind of request
- IDs are assumed not known by the server
- Problems described in the paper:
  - Generation of Realistic dummy movements: MN and MLN (Moving in a Limited Neighborhood alg)
  - Reduction of communication costs
  - Experiments using GeoLink Kyoto Map Applet

***http://www.digitalcity.gr.jp/openlab/kyoto/map_guide.html***

# MN and MLN algorithms
# Moving in a (Limited) Neighborhood



Generate regions of few persons
Possible regions of dummy generation
Crowded regions
● True data
○ Dummies
(a): Moving in a Neighborhood (MN)
(b): Moving in a Limited Neighborhood (MLN)

- MN: generate a random point in the neighborhood of the previous dummy positions
- MLN: like MN, but using also requests distributions of other users

# Reducing frequency of user requests in Continuous Location-Tracking

- Hiding single locations of each users may be not enough:
  - Services require authentication (history of requests is mandatory to provide the service)
  - Frequent requests can be linked to the same user
- The architecture:
  - User inform a Location broker about his exact location
  - Location broker uses a privacy manager (policy matching + path sensitivity analysis)
  - After anonymization, the request is forwarded to the service provider (an ID is used instead of real name)

# Privacy Manager Algorithms

- It may cancel the forwarding of user requests (location updates) to service providers
  - User privacy policies (which the privacy manager has access to) specify sensitive zones (e.g., buildings)
    - Base algorithm
  - Also non-sensitive requests can be cancelled to reduce frequency of requests (weakening the attacker knowledge)
    - Bounded-rate algorithm
- Forwards only when they do not give away which of at least k sensitive areas the user visited
    - k-Area algorithm

# Privacy-aware location query systems

**The New Casper: Query Processing for Location Services without Compromising Privacy**

**[*Mohamed F. Mokbel, Chi-Yin Chow, Walid G. Aref*, VLDB*06*]**

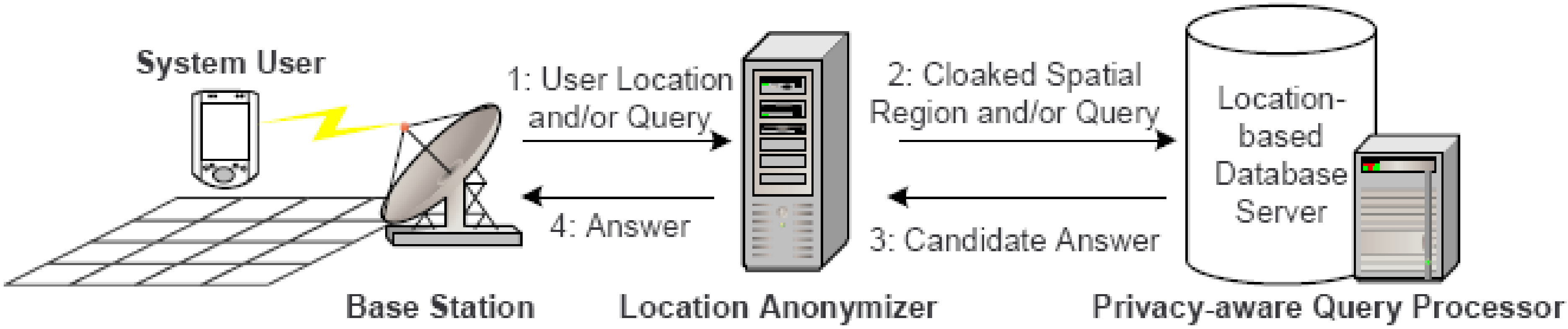
# New Casper

- There are major privacy concern in current LBS when users have to continuously report their locations to the DB server in order to be queried later
- Casper is a sophisticated query processing system which allow to maintain an updated location dbserver and allow different kind of queries
- Named after the friendly ghost that can hide its location and help people :-)

# Kind of Queries

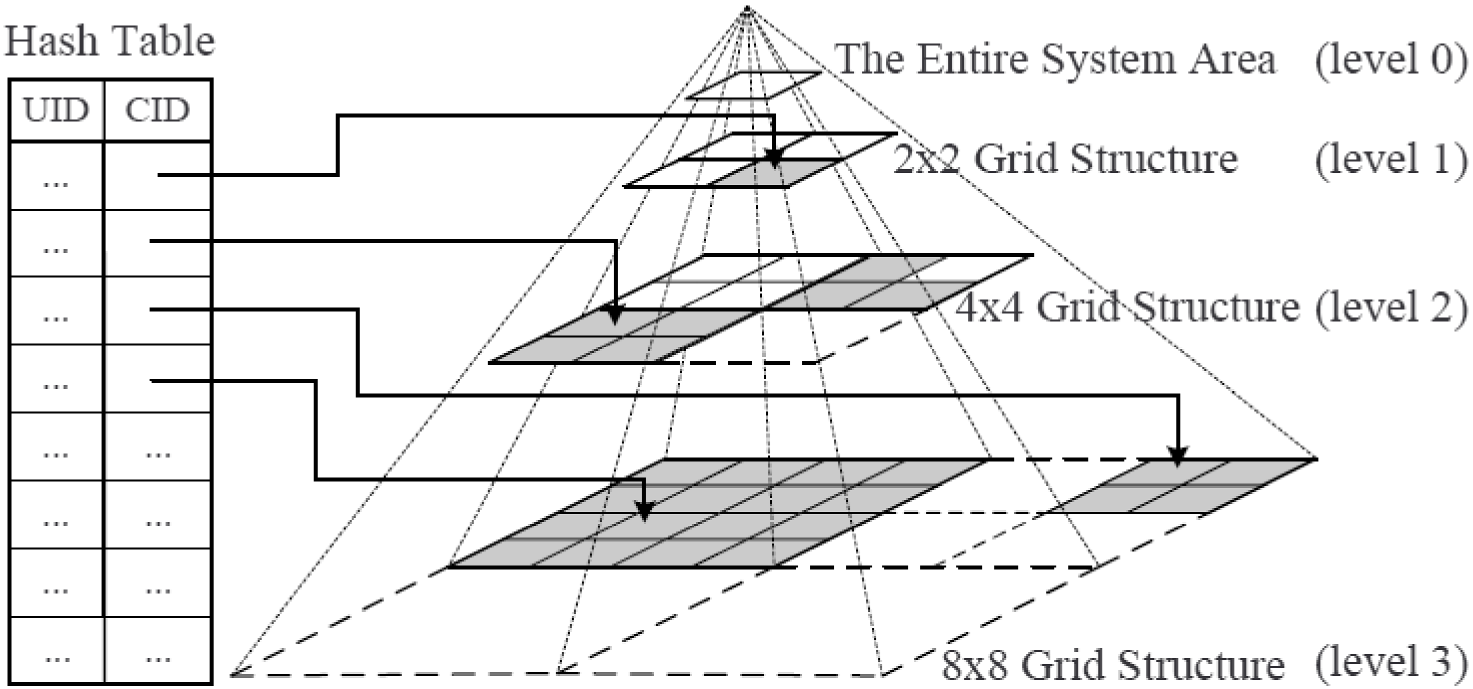
- Private queries over public data
  - “Where is my nearest gas station”, in which the person who issues the query is a private entity while the data (i.e., gas stations) are public
- Public queries over private data
  - “How many cars in a certain area”, in which a public entity asks about personal private location
- Private queries over private data
  - “Where is my nearest buddy” in which both the person who issues the query and the requested data are private

# Architecture


System User
1: User Location and/or Query
2: Cloaked Spatial Region and/or Query
Location-based Database Server
4: Answer
3: Candidate Answer
Base Station
Location Anonymizer
Privacy-aware Query Processor

- Casper framework mainly consists of two components:
  - location anonymizer (client cloaking algorithm)
  - privacy-aware query processor (server side reconstruction algorithm)

# Adaptive Location Anonymizer


Hash Table
The Entire System Area (level 0)
2x2 Grid Structure (level 1)
4x4 Grid Structure (level 2)
8x8 Grid Structure (level 3)

| UID | CID |
|---|---|
| ... | |
| ... | |
| ... | |
| ... | |
| ... | ... |
| ... | ... |
| ... | ... |
| ... | ... |

**According to user preferences, location updates are de-identified and stored at different details**

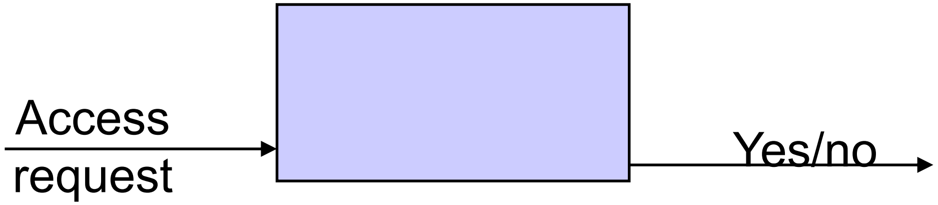
# Features wrt Previous Approaches

- Location anonymizer distinguishes itself from previous proposals:
  - Provides a customizable privacy profile for each mobile user that contains the k-anonymity and minimum cloaked area requirements
  - Scales well to a large number of mobile users with arbitrary privacy profiles
  - Cannot be reverse engineered to give any information about the exact user location

# Policies for location-base access control (hints)

# Location based access control

- Security mechanisms are often transparent or nearly transparent to end users
- A basic mechanism  is access control



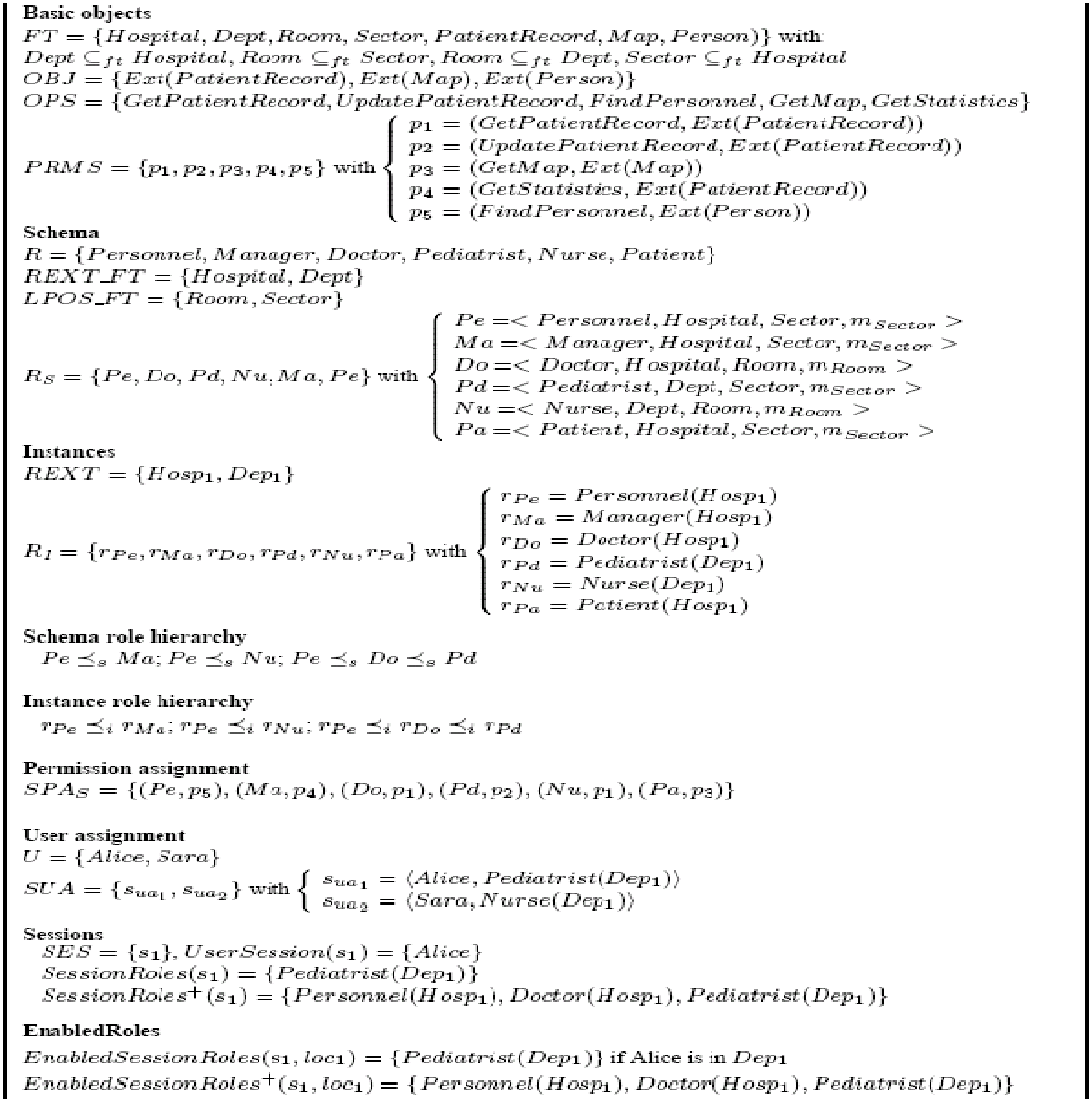
- Focus is on the geographical dimension of access control
- M. L. Damiani, E. Bertino, B. Catania, P. Perlasca: *GEO-RBAC: A spatially aware RBAC*. ACM Trans. Inf. Syst. Secur. 10(1): (2007)

# Example

- A (mobile) doctor cannot disclose patients’ records outside the hospital in which the doctor works
- A doctor, however, cannot be also a patient in the same hospital at the same time

# Specifying policies for LB access control

- The Geo-RBAC model
  - Spatially-constrained disclosure of information
  - Dynamic computation of user’s position at different granularities
  - First attempt to integrate access control  and location privacy



**Basic objects**

$FT = \{Hospital, Dept, Room, Sector, PatientRecord, Map, Person)\}$ with

$Dept \subseteq_{ft} Hospital, Room \subseteq_{ft} Sector, Room \subseteq_{ft} Dept, Sector \subseteq_{ft} Hospital$

$OBJ = \{Ext(PatientRecord), Ext(Map), Ext(Person)\}$

$OPS = \{GetPatientRecord, UpdatePatientRecord, FindPersonnel, GetMap, GetStatistics\}$

$$PRMS = \{p_1, p_2, p_3, p_4, p_5\} \text{ with } \begin{cases} p_1 = (GetPatientRecord, Ext(PatientRecord)) \\ p_2 = (UpdatePatientRecord, Ext(PatientRecord)) \\ p_3 = (GetMap, Ext(Map)) \\ p_4 = (GetStatistics, Ext(PatientRecord)) \\ p_5 = (FindPersonnel, Ext(Person)) \end{cases}$$

**Schema**

$R = \{Personnel, Manager, Doctor, Pediatrist, Nurse, Patient\}$

$REXT\_FT = \{Hospital, Dept\}$

$LPOS\_FT = \{Room, Sector\}$

$$R_S = \{Pe, Do, Pd, Nu, Ma, Pe\} \text{ with } \begin{cases} Pe = < Personnel, Hospital, Sector, m_{Sector} > \\ Ma = < Manager, Hospital, Sector, m_{Sector} > \\ Do = < Doctor, Hospital, Room, m_{Room} > \\ Pd = < Pediatrist, Dept, Sector, m_{Sector} > \\ Nu = < Nurse, Dept, Room, m_{Room} > \\ Pa = < Patient, Hospital, Sector, m_{Sector} > \end{cases}$$

**Instances**

$REXT = \{Hosp_1, Dep_1\}$

$$R_I = \{r_{Pe}, r_{Ma}, r_{Do}, r_{Pd}, r_{Nu}, r_{Pa}\} \text{ with } \begin{cases} r_{Pe} = Personnel(Hosp_1) \\ r_{Ma} = Manager(Hosp_1) \\ r_{Do} = Doctor(Hosp_1) \\ r_{Pd} = Pediatrist(Dep_1) \\ r_{Nu} = Nurse(Dep_1) \\ r_{Pa} = Patient(Hosp_1) \end{cases}$$

**Schema role hierarchy**

$Pe \preceq_s Ma; Pe \preceq_s Nu; Pe \preceq_s Do \preceq_s Pd$

**Instance role hierarchy**

$r_{Pe} \preceq_i r_{Ma}; r_{Pe} \preceq_i r_{Nu}; r_{Pe} \preceq_i r_{Do} \preceq_i r_{Pd}$

**Permission assignment**

$SPA_S = \{(Pe, p_5), (Ma, p_4), (Do, p_1), (Pd, p_2), (Nu, p_1), (Pa, p_3)\}$

**User assignment**

$U = \{Alice, Sara\}$

$$SUA = \{s_{ua_1}, s_{ua_2}\} \text{ with } \begin{cases} s_{ua_1} = \langle Alice, Pediatrist(Dep_1)\rangle \\ s_{ua_2} = \langle Sara, Nurse(Dep_1)\rangle \end{cases}$$

**Sessions**

$SES = \{s_1\}, UserSession(s_1) = \{Alice\}$

$SessionRoles(s_1) = \{Pediatrist(Dep_1)\}$

$SessionRoles^+(s_1) = \{Personnel(Hosp_1), Doctor(Hosp_1), Pediatrist(Dep_1)\}$

**EnabledRoles**

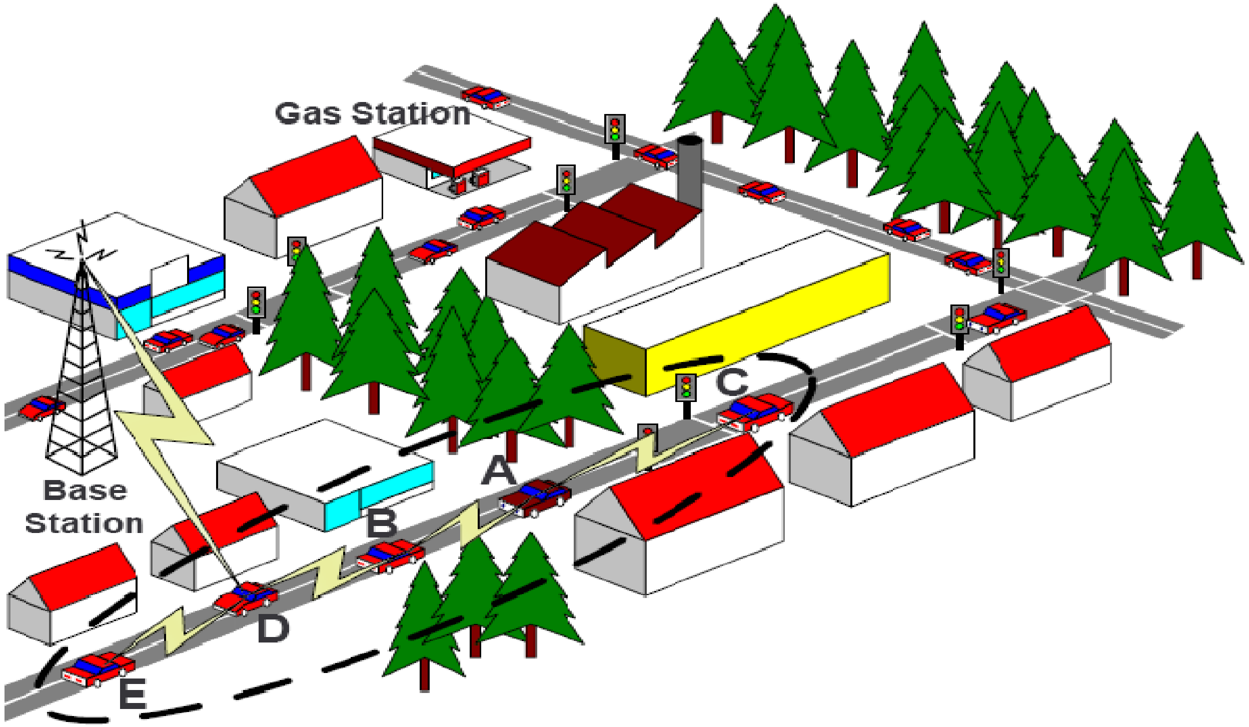
$EnabledSessionRoles(s_1, loc_1) = \{Pediatrist(Dep_1)\}$ if Alice is in $Dep_1$

$EnabledSessionRoles^+(s_1, loc_1) = \{Personnel(Hosp_1), Doctor(Hosp_1), Pediatrist(Dep_1)\}$

# Peer-to-peer architectures for LBS

# Peer-to-Peer Cooperative Architecture

## Group Formation



- **Main idea: whenever a user want to issue a location-based query, the user broadcasts a request to its neighbors to form a group. Then, a random user of the group will act as the query sender.**

# Trading privacy for trust

[Bhargava and colleagues,
Purdue Univ.]

# Problem motivation

- Privacy and trust form an adversarial relationship
  - Users have to provide digital credentials that contain private information in order to build trust in open environments like Internet or peer-to-peer (LBS) systems.
- Research is needed to quantify the tradeoff between privacy and trust

# Subproblems

- How much privacy is lost by disclosing a piece of credential?
- How much does a user benefit from having a higher level of trust?
- How much privacy a user is willing to sacrifice for a certain amount of trust gain?

# Bhargava’s approach

- Formulate the privacy-trust tradeoff problem
- Design metrics and algorithms to evaluate the privacy loss. We consider:
  - Information receiver
  - Information usage
  - Information disclosed in the past
- Estimate trust gain due to disclosing a set of credentials
- Develop mechanisms empowering users to trade trust for privacy

# Formulation of tradeoff problem (1)

- Set of private attributes that user wants to conceal
- Set of credentials
  - R(i): subset of credentials *revealed* to receiver *i*
  - U(i): credentials *unrevealed* to receiver *i*
- Credential set with minimal privacy loss
  - A subset of credentials *NC* from *U* (i)
  - *NC* satisfies the requirements for trust building
  - PrivacyLoss(*NC*∪*R(i)*) – PrivacyLoss(*R(i)*)) is minimized

# Plan of the tutorial

- The scenario of ubiquitous computing
  - Analytic opportunities and privacy threats
- Privacy and anonymity: prognosis and therapy
  - In data publishing: attack models and privacy-preserving techniques
  - In data mining: attack models and privacy-preserving data mining techniques
- Privacy and anonymity in Location Based Services
- **Preliminary results on privacy and anonymity techniques in mobility data analysis**
- Conclusion

# Preliminary results and research trends in privacy-preserving mobility data publishing

# Mobility data publishing

- Very little work on mobility data publishing
- Main reasons
  - Data is not yet available due to privacy issues
  - Work focused on “online” LBS where a location data warehouse does not need to be maintained/released
- Privacy-preserving techniques for data publishing exist for relational tables
  - They can be easily extended to ST data, but privacy concerns are not well-studied for ST data
  - Ad-hoc (offline) solutions would enable more accuracy while preserving anonymity of data donors
- Stay tuned … new results on k-anonymous trajectories arriving from GeoPKDD

# **Strong K-Anonymity for ST data** [Bettini & Mascetti, PRISE 2006 + TR]

- Anonymity: "*a state of being not identifiable within a set of subjects, the Anonymity Set*"
- K-Anonymity: |Anonymity Set| ≥ k

- Strong k-anonymity allows multiple presence of same user in the anonymity set
- Also fine-grained anonymization are suggested for time intervals

# Protecting users’ privacy from LBQIDs [Verykios et al., 2007]

- Use of frequent spatio-temporal patterns to serve as LBQIDs (Location-Based Quasi Identifiers)
- Knowledge of movement patterns makes easy the identification of the user
- Use of a spatio-temporal K-anonymization system to protect users’ privacy whenever the user exhibits a behavior that partially matches with a frequent movement pattern

# Generalization & Unlinking Strategies

- Generalization technique
  - When a user’s behavior matches with one or more of his/her LBQIDs, the location and time of request are expanded to cover an area that contains k-1 other subjects who may have sent a similar request

- Unlinking technique
  - When the generalization algorithm fails, the system dynamically creates a mix-zone where the user is dissociated from his/her previous system identity and is provided with a new one

# Never Walk Alone

[**Bonchi, Abul, Nanni**, March 2007]

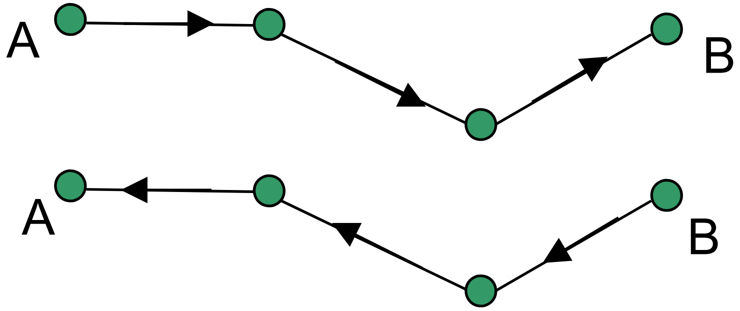
ISTI-CNR Technical Report, Submitted

# Re-identification Example

- Your city municipality traffic management office has collected a database of trajectories of vehicles (equipped with a GPS device) driving on the city road network.
- The traffic management office wants to mine this dataset to find behavioural patterns, but it has not the *know-how* needed to mine the data.
- Data mining is outsourced to your research lab.
- Due to privacy laws, the dataset is “anonymized” before the release: in a naive tentative of preserving anonymity, the car identifiers are not disclosed but instead replaced with pseudonyms.

# Re-identification Example

Id:
34567



[almost every day mon-fri
between 7:45 – 8:15]

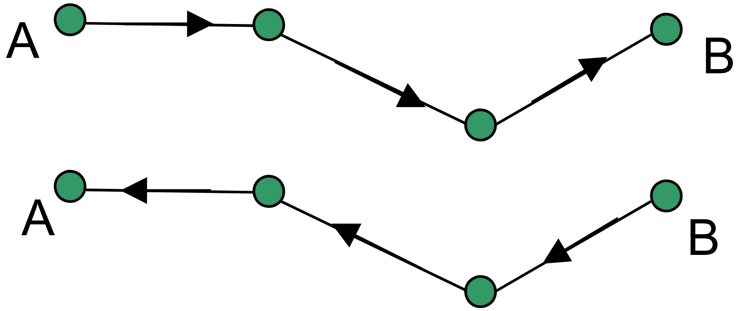
[almost every day mon-fri
between 17:45 – 18:15]

- By intersecting the phone directories of A and B you find that only one individual lives in A and works in B.
- Id:34567 = Prof. Smith
- Then you discover that on Saturday night Id:34567 usually drives to the city red lights district…

# k-anonymity principle

- *k*-anonymity principle: each release of data must be such that each individual is indistinguishable from at least *k* - 1 other individuals.
- Is this a **local pattern**?

Support = 3



[almost every day mon-fri between 7:45 – 8:15]

[almost every day mon-fri between 17:45 – 18:15]

- Any **local pattern** describing (supported by) less than *k* individuals is a possible threat to *k*-anonymity…
- **Local patterns** can be dangerous!!!

# "Exploiting Uncertainty for Anonymity in Moving Objects Databases" [Abul, Bonchi, Nanni] (submitted)

- Motivation:

  location data enables intrusive inferences, which may reveal habits, social customs, religious and sexual preferences of individuals, and can be used for unauthorized advertisement and user profiling.

- Problem:

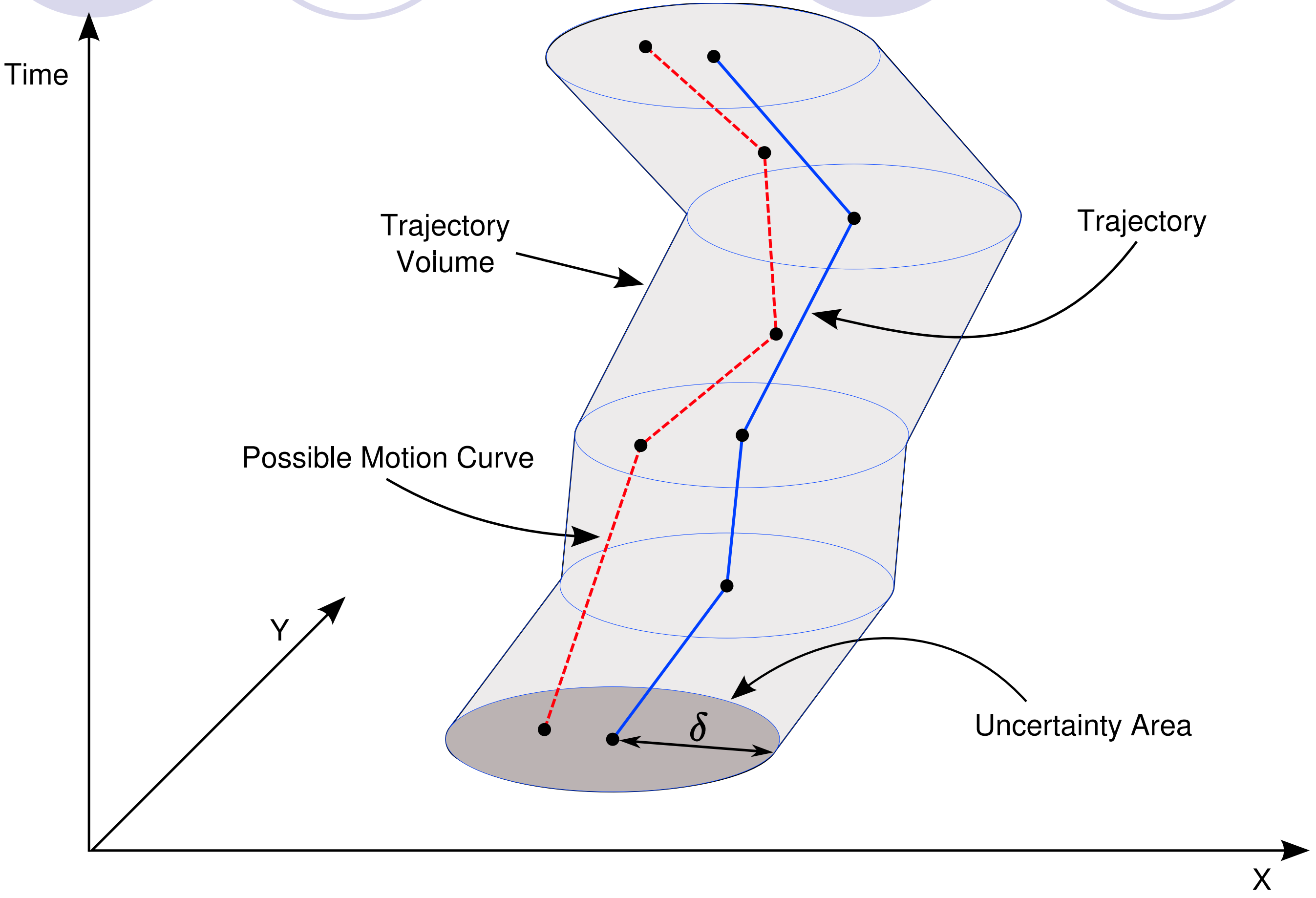
  Anonymity preserving data publishing from MOD

- Basic idea:

  to exploit the inherent uncertainty of moving objects position for enforcing anonymity with less information distortion

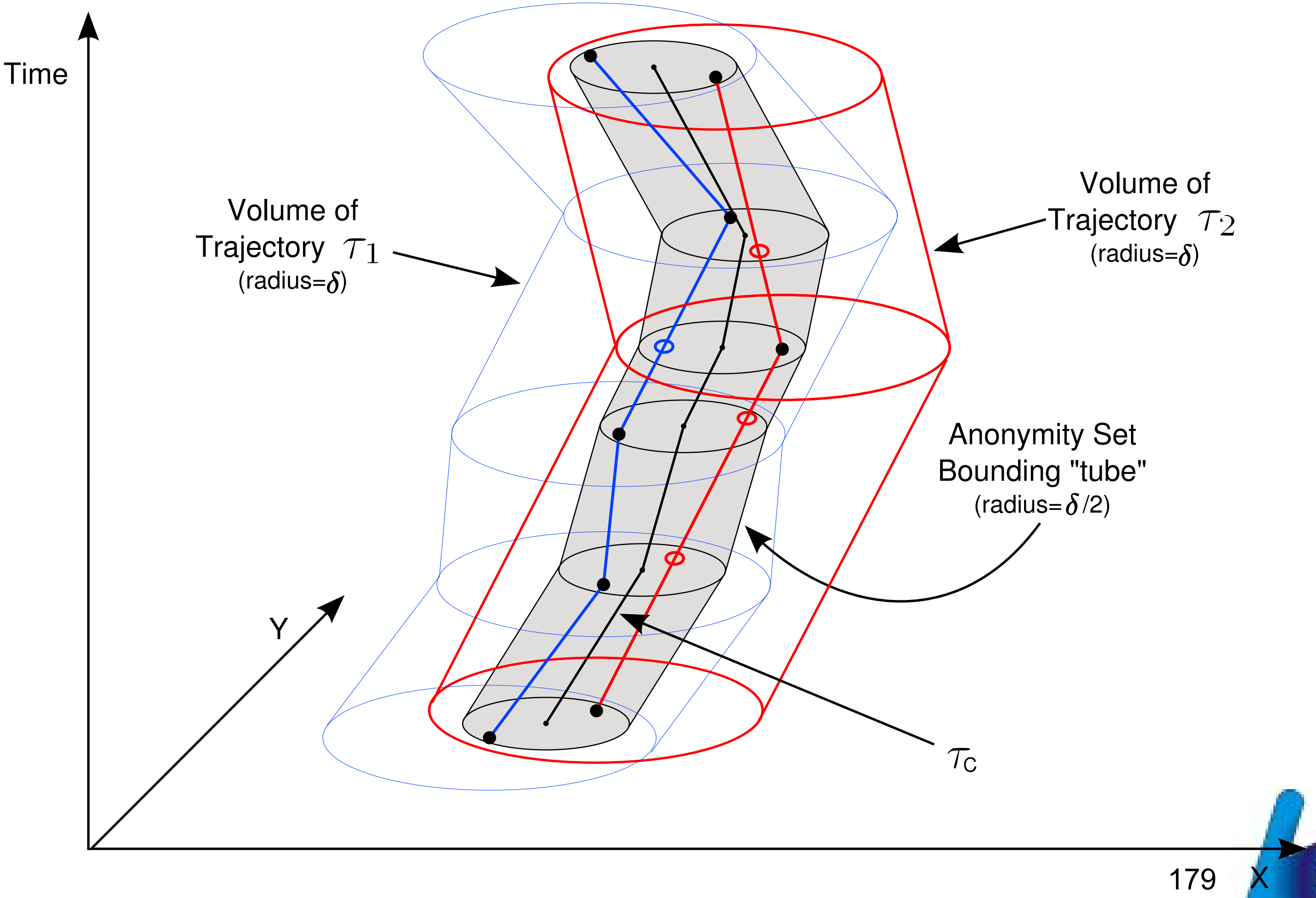
- Main contributions:
  - concept of $(k,\delta)$-anonymity
  - deep characterization of the problem
  - *NWA* ("*N*ever *W*alk *A*lone") method for enforcing $(k,\delta)$-anonymity

# Uncertainty and Trajectories



G. Trajcevski, O. Wolfson, K. Hinrichs, and S. Chamberlain. *“Managing uncertainty in moving objects databases.” ACM Trans. Database Syst.*, 29(3):463–507, 2004.

# Anonymity and Trajectories


Time
Y
X
Volume of Trajectory $\mathcal{T}_1$ (radius=$\delta$)
Volume of Trajectory $\mathcal{T}_2$ (radius=$\delta$)
Anonymity Set Bounding "tube" (radius=$\delta/2$)
$\mathcal{T}_C$

# Never Walk Alone

- Based on clustering and space translation:
  1. Create clusters under the constraint population $\geq k$
  2. Transform each cluster in a (k,$\delta$)-anonymity set by space translation
- Distance measure adopted: simple Euclidean
- Limitation: only trajectories starting and ending at the same time can be clustered together.
- *NWA* tries to overcome the limitation by means of a pre-processing step.
- *NWA* is also equipped with outliers identification and removal.

# Preliminary results and research trends in privacy-preserving mobility data analysis

# Mobility data analysis

- Data analysis of ST data is fundamental for emerging applications like, e.g.:
  - Traffic analysis
  - Sustainable mobility management
  - Studies on animal behaviours
  - to improve/save resources, e.g. GPRS antennas location
  - make reliable models that describe (application-depending) moving objects for decision-making purpose
    - E.g., in urban traffic applications, what if I change the driving direction of a one-way street?

# Hiding Sequences

[Abul , Atzori, Bonchi, Giannotti,
PDM07 @ ICDE07]

# Hiding Sequences

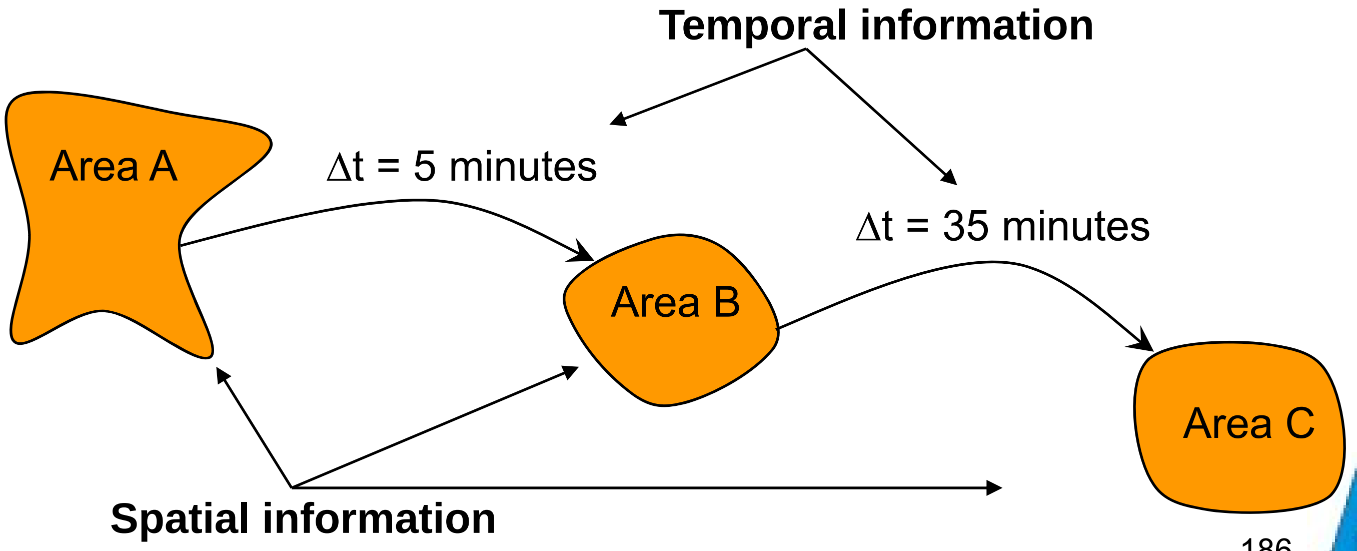
- Trajectory data can be represented as sequences of ST points
- Global privacy policies (specified as sensitive subsequences, i.e., private paths) can be enforced through slightly distorting the original trajectory database
  - Optimal problem has been shown to be NP-Hard
  - Efficient Heuristics are provided
  - the algorithms handle time constraints like *max/min gap* and *max window*
  - Empirical studies also on side-effects on the pattern mined from the distorted dataset

# Motivation

- Knowledge hiding is explored in depth in the context of *frequent itemsets* from tabular data,
- However, many real-world applications demand sequential data,
  - e.g. web usage data, biomedical patient data, spatio-temporal trajectory data of moving entities, etc.
  - Clearly, in all of these applications privacy is a concern. For instance, linking a trajectory to its owner may reveal individual's sensitive habits and social preferences.
- Here we address knowledge hiding in the context where both data and patterns have sequential structure

# What is a trajectory pattern?

- A trajectory pattern is a **sequence of spatial regions** that, on the basis of the source trajectory data, emerge as frequently visited in the order specified by the sequence;
- Possibly  with a **typical travel time**



# Hiding Sequential Patterns

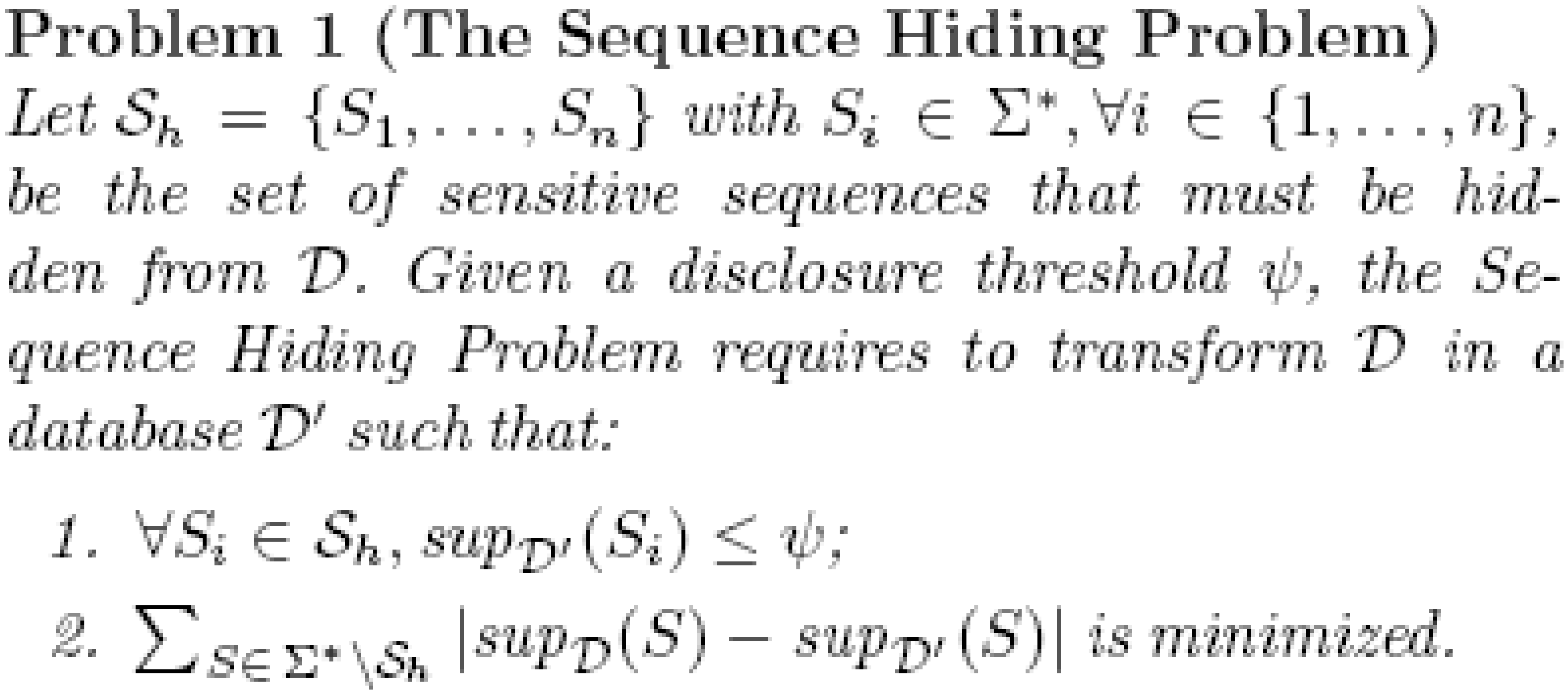
- Definitions
  - Let $S$ be a simple sequence[§] defined over an alphabet $\Sigma$, i.e. $S \in \Sigma^*$, and $D$ be a database of simple sequences.
  - $S \in \Sigma^*$ is a subsequence of $T \in \Sigma^*$, denoted $S \sqsubseteq T$, iff $S$ can be obtained by deleting some elements (not necessarily contiguous) from $T$
  - Support of sequence of $S$ on $D$ is defined as

$$sup_{\mathcal{D}}(S) = |\{T \in \mathcal{D} \,|\, S \sqsubseteq T\}|$$

[§] This is not a restriction but preferred for the sake of simplicity. Later it will be generalized s.t. each element of $S$ is a subset of $\Sigma$.

# Hiding Sequential Patterns

- The Sequence Hiding Problem



**Problem 1 (The Sequence Hiding Problem)** *Let $\mathcal{S}_h = \{S_1, \ldots, S_n\}$ with $S_i \in \Sigma^*, \forall i \in \{1, \ldots, n\}$, be the set of sensitive sequences that must be hidden from $\mathcal{D}$. Given a disclosure threshold $\psi$, the Sequence Hiding Problem requires to transform $\mathcal{D}$ in a database $\mathcal{D}'$ such that:*

1. $\forall S_i \in \mathcal{S}_h, sup_{\mathcal{D}'}(S_i) \leq \psi$;
2. $\sum_{S \in \Sigma^* \setminus \mathcal{S}_h} |sup_{\mathcal{D}}(S) - sup_{\mathcal{D}'}(S)|$ *is minimized.*

Note that a special case occurs when $\psi=0$, where every instance needs to be hidden.

# A Sanitization Algorithm

- A 2-stage greedy algorithm
  - First stage: Select a subset of *D* for sanitization
  - Second stage: For each sequence chosen to be sanitized (the output from the first stage), select marking positions
- The heuristic
  - Recalling the objective is introducing minimum number of $\Delta$s,
    - For the first stage: Sort the sequences in ascending order of matching set size, and select top |*D*|- $\psi$ for sanitization
    - For the second stage: Choose the marking position that is involved in most matches

# **Privacy Preserving Spatio-Temporal Clustering on Horizontally Partitioned Data** [A. Inan and Y. Saygin, DaWaK 2006]

- Introduction of secure multiparty solution to privacy problems in spatio-temporal data without loss of accuracy
  - clusters can be computed when trajectories are stored in different data repositories
  - Data doesn't have to be shared,; only the mining model is eventually shared
- Different distance metrics can be used
  - Euclidean distance
  - Longest Common Subsequence
  - Dynamic Time Warping
  - Edit Distance

# Conclusions

# 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.

# European Union Data Protection Directives

- Directive 95/46/EC
  - Passed European Parliament 24 October 1995
  - Goal is to ensure free flow of information
    - *Must preserve privacy needs of member states*
  - Effective October 1998
- Effect
  - Provides guidelines for member state legislation
    - Not directly enforceable
  - Forbids sharing data with states that don't protect privacy
    - Non-member state must provide adequate protection,
    - Sharing must be for "allowed use", or
    - Contracts ensure adequate protection

# 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;

# US Healthcare Information Portability and Accountability Act (HIPAA)

- Govern’s use of patient information
  - Goal is to protect the patient
  - Basic idea: Disclosure okay if anonymity preserved
- Regulations focus on outcome
  - A covered entity may not use or disclose protected health information, except as permitted or required…
    - To individual
    - For treatment (generally requires consent)
    - To public health / legal authorities
  - Use permitted where “there is no reasonable basis to believe that the information can be used to *identify an individual*”

# The Safe Harbor “atlantic bridge”

- In order to bridge EU and US (different) privacy approaches and provide a streamlined means for U.S. organizations to comply with the European Directive, the U.S. Department of Commerce in consultation with the European Commission developed a "Safe Harbor" framework.
- Certifying to the Safe Harbor will assure that EU organizations know that US companies provides “adequate” privacy protection, as defined by the Directive.

# The Safe Harbor “atlantic bridge”

- Data presumed not identifiable if 19 identifiers removed (§ 164.514(b)(2)), e.g.:
    - Name,
    - location smaller than 3 digit postal code,
    - dates finer than year,
    - identifying numbers

- Shown not to be sufficient (Sweeney)

# Pointers to Resources

# Web Links on Privacy Laws

English

- europa.eu.int/comm/justice_home/fsj/privacy/law/index_en.htm
- www.privacyinternational.org/
- www.export.gov/safeharbor/

Italian

- www.garanteprivacy.it
- www.interlex.it/
- www.iusreporter.it/
- www.privacy.it/

# Web Resources on PPDM

- **Privacy Preserving Data Mining Bibliography (maintained by Kun Liu)**
  http://www.cs.umbc.edu/~kunliu1/research/privacy_review.html

- **Privacy Preserving Data Mining Blog**
  http://www.umbc.edu/ddm/wiki/index.php/PPDM_Blog

- **Privacy Preserving Data Mining Bibliography (maintained by Helger Lipmaa)**
  http://www.cs.ut.ee/~lipmaa/crypto/link/data_mining/

- **The Privacy Preserving Data Mining Site (maintained by Stanley Oliveira)**
  http://www.cs.ualberta.ca/%7Eoliveira/psdm/psdm_index.html [no longer updated]

- **IEEE International Workshop on Privacy Aspects of Data Mining (every year in conjunction with IEEE ICDM conference)**

  **PADM'06 webpage:** http://www-kdd.isti.cnr.it/padm06/

# Bibliography on Spatio-Temporal Anonymity and Privacy in Location Based Services (1/3)

- **Mohamed F. Mokbel, Chi-Yin Chow and Walid G. Aref. The New Casper: Query Processing for Location Services without Compromising Privacy, In Proceedings of VLDB 2006**
- **Chi-Yin Chow, Mohamed F. Mokbel and Xuan Liu. A Peer-to-Peer Spatial Cloaking Algorithm for Anonymous Location-based Services. In Proceedings of the ACM International Symposium on Advances in Geographic Information Systems, ACM GIS06**
- **R. Cheng, Y. Zhang, E. Bertino and S. Prabhakar. Preserving User Location Privacy in Mobile Data Management Infrastructures. In Proceedings of Privacy Enhancing Technology Workshop (PET) 2006**
- **M. Duckham and L. Kulik. Location privacy and location-aware computing. Book chapter in Dynamic & Mobile GIS: Investigating Change in Space and Time, CRC Press, Boca Rator, FL, pp 35-51**
- **L. Kazatzopoulos C. Delakouridis G. F. Marias, and P. Georgiadis, iHIDE: Hiding Sources of Information in WSNs. In Proceedings of 2nd International Workshop on Security, Privacy and Trust in Pervasive and Ubiquitous Computing (IEEE SecPerU2006)**
- **Bugra Gedik and Ling Liu. Location-Privacy in Mobile Systems: A Personalized Anonymization Model, In Proceedings of ICDCS 2005**
- **P. Kamat, Y. Zhang, W. Trappe and C. Ozturk. Enhancing Source-Location Privacy in Sensor Network Routing. In Proceedings of ICDCS 2005**
- **M. Gruteser, Baik Hoh. On the Anonymity of Periodic Location Samples. 2nd Int Conf. On Security in Pervasive Computing, Boppard, Germany 2005.**

# Bibliography on Spatio-Temporal Anonymity and Privacy in Location Based Services (2/3)

- **B. Hoh and M. Gruteser. *Protecting* location privacy through path confusion, In First International Conference on Security and Privacy for Emerging Areas in Communications Networks, SecureComm 2005**
- **Claudio Bettini, X. SeanWang and Sushil Jajodia. Protecting Privacy Against Location-Based Personal Identification, In Proceedings of 2nd VLDB Workshop on Secure Data Management (SDM) 2005**
- **M. Youssef, V. Atluri and N. R. Adam . Preserving Mobile Customer Privacy: An Access Control System for Moving Objects and Customer Profiles, In Proceedings of the 6th International Conference on Mobile Data Management (MDM) 2005**
- **Marco Gruteser and Xuan Liu. Protecting Privacy in Continuous Location Tracking Applications, IEEE Security and Privacy Magazine, 2(2), pp 28-34, 2004**
- **A. R. Beresford and F. Stajano. Mix zones: User privacy in location-aware services, In Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, March 2004**
- **Marco Gruteser and Dirk Grunwald, Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking, In Proceedings of MobySys 2003**
- **A. R. Beresford and F. Stajano. Location privacy in pervasive computing, IEEE Pervasive Computing, 2(1):46–55, 2003**
- **J. Al-Muhtadi, R. Campbell, A. Kapadia, M. D. Mickunas and S. Yi. Routing Through the Mist: Privacy Preserving Communication in Ubiquitous Computing Environments. In Proceedings of ICDCS 2002**

# Bibliography on Spatio-Temporal Anonymity and Privacy in Location Based Services (3/3)

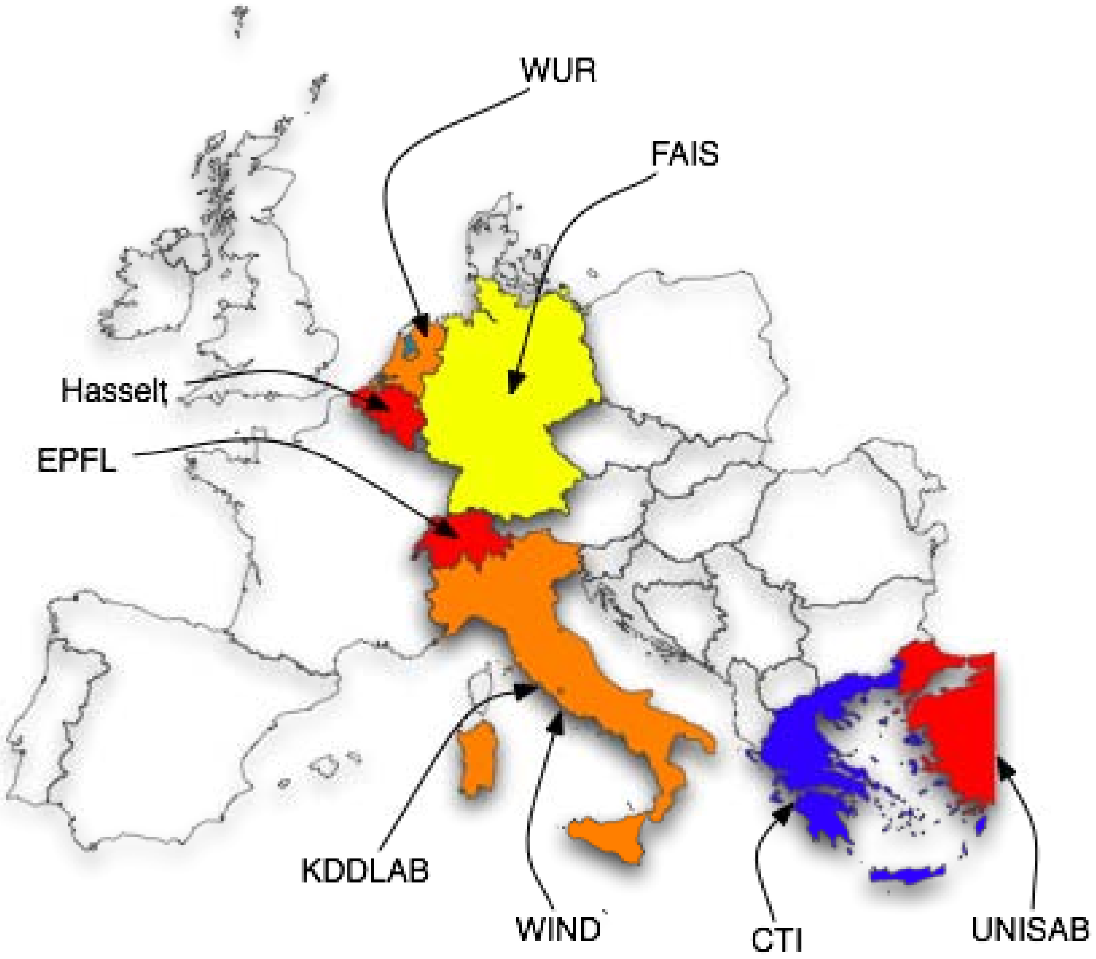
- ***Hidetoshi Kido, Yutaka Yanagisawa, Tetsuji Satoh.*** **Anonymous Communication Technique using Dummies for Location-Based Services. Proceedings of the Intern. Conf. On Pervasive Services, 2005. ICPS05.**
- **Panos Kalnis, Gabriel Ghinita. Query Privacy and Spatial Anonymity in Location Based Services. http://anonym.comp.nus.edu.sg/**
- **Panos Kalnis, Gabriel Ghinita, Kyriakos Mouratidis and Dimitris Papadias. Preserving Anonymity in Location Based Services. Technical Report TRB6/06, Department of Computer Science, National University of Singapore**
- **Gabriel Ghinita, Panos Kalnis and Spiros Skiadopoulos. PRIVÉ: Anonymous Location-Based Queries in Distributed Mobile Systems. Technical Report TRB7/06, Department of Computer Science, National University of Singapore**

# Announcements

www.geopkdd.eu

GeoPKDD

Geographic Privacy-aware Knowledge Discovery and Delivery



# GeoPKDD Privacy Observatory

- Privacy cannot be achieved by technology alone
  - it’s a social, ethical, legal and technological matter.
- The GeoPKDD Observatory interacts with stakeholders in privacy issues. Activities:
  - create and maintain relationships with European and national authorities for data protection and other privacy related organizations,
  - implement regulations into KDD methods and tools,
  - provide ideas for revisions of regulations themselves by means of novel privacy preserving technologies.
- http://www.geopkdd.eu/pro

# GeoPKDD book (forthcoming)

Fosca Giannotti and Dino Pedreschi (Eds.)

***Mobility, Privacy, and Data Mining*.**

Lecture Notes in Computer Science Series, Springer, Berlin, 2007.

# Part I: Setting the stage

- Chapter 1 **Basic Concepts of Movement Data**
  - Natalia Andrienko, Gennady Andrienko, Nikos Pelekis, Stefano Spaccapietra (FHG, CTI, EPFL)
- Chapter 2 **Characterising mobile applications through a privacy-aware geographic knowledge discovery process**
  - Monica Wachowicz, Arend Ligtenberg, Chiara Renso, Seda Gürses (WUR, KDDLAB)
- Chapter 3 **Wireless Network Data Sources: Tracking and Synthesizing Trajectories**
  - Chiara Renso, Simone Puntoni, Elias Frentzos, Andrea Mazzoni, Bart Moelans, Nikos Pelekis, Fabrizio Pini (KDDLAB, CTI, HASSELT,WIND)
- Chapter 4 **Privacy Protection: Regulation, Threats and Opportunities**
  - Francesco Bonchi, Franco Turini, Bart Moelans, Dino Pedreschi, Yücel Saygin, Vassilios Verykios (KDDLAB, HASSELT, UNISAB,CTI)

# Part II: Managing moving object and trajectory data

- Chapter 5 **Trajectory Data Models**
  - Jose Macedo, Christelle Vangenot, Walied Othman, Nikos Pelekis, Elias Frentzos, Bart Kuijpers, Irene Ntoutsi, Stefano Spaccapietra, Yannis Theodoridis, (EPFL,CTI,HASSELT)
- Chapter 6 **Trajectory Database Systems**
  - Elias Frentzos, Nikos Pelekis, Irene Ntoutsi, Yannis Theodoridis (CTI)
- Chapter 7 **Towards Trajectory Data Warehouses**
  - Nikos Pelekis, Alessandra Raffaetà, Maria-Luisa Damiani, Christelle Vangenot, Gerasimos Marketos, Elias Frentzos, Irene Ntoutsi, Yannis Theodoridis (CTI, KDDLAB, EPFL)
- Chapter 8 **Privacy and Security in Spatio-temporal Data and Trajectories**
  - Vassilios S. Verykios, Maria Luisa Damiani, Aris Gkoulalas-Divanis (CTI, EPFL)

# Part III: Mining spatial and temporal data

- Chapter 9 **Knowledge Discovery from Geographical Data**
  - Salvatore Rinzivillo, Franco Turini, Vania Bogorny, Christine Körner, Bart Kuijpers, Michael May (KDDLAB, HASSELT, FAIS)
- Chapter 10 **Spatio-temporal Data Mining**
  - Bart Kuijpers, Mirco Nanni, Christine Körner, Michael May, Dino Pedreschi (KDDLAB, HASSELT, FAIS)
- Chapter 11 **Privacy in Spatio-temporal Data Mining**
  - Francesco Bonchi, Yücel Saygin, Vassilios S. Verykios, Maurizio Atzori, Aris Gkoulalas-Divanis, Selim Volkan Kaya, Erkay Savaş (KDDLAB, UNISAB, CTI)
- Chapter 12 **Querying and Reasoning for Spatio-Temporal Data Mining**
  - Giuseppe Manco, Chiara Renso, Miriam Baglioni, Fosca Giannotti, Bart Kujpers, Alessandra Raffaetà (KDDLAB, HASSELT)
- Chapter 13 **Visual Analytics Methods for Movement Data**
  - Gennady Andrienko, Natalia Andrienko, Ioannis Kopanakis, Arend Ligtenberg (FAIS, KDDLAB, WUR)

# **KDubiq**: European coordination action on Ubiquitous KD

- Working Group: *Privacy and Security issues in Ubiquitous Knowledge Discovery*
- Chair: Fosca Giannotti,
- Co-Chair: Yucel Saygin, Sabanci Univ., Istanbul, Turkey
- http://www.kdubiq.org

# PinKDD 2007

- First ACM SIGKDD Int. Workshop on **Privacy, Security, and Trust in KDD**
  - August 12th, 2007, San Jose, CA, USA
  - Paper submission deadline: May 31st, 2007
- Full-day workshop at ACM SIGKDD ’07

***http://www-kdd.isti.cnr.it/pinKDD07***