

# Privacy and anonymity in data publishing and (mobility) data mining

Fosca Giannotti, Dino Pedreschi, Franco Turini

Pisa KDD Laboratory  
Università di Pisa and ISTI-CNR, Pisa, Italy

*Dottorato di Ricerca in Informatica, Scuola Galilei*

*Università di Pisa. Giugno-Luglio 2009*





- We live in a time with unprecedented opportunities of
  - sensing,
  - storing,
  - analyzing
- (micro)-**data**, at mass level, recording human activities at extreme detail
- Paradigm shift: from **statistics** to **data mining**
- Opportunities always come with risks
  - How to protect privacy?





# Wireless networks as mobility data collectors

- Wireless networks infrastructures are the **nerves of our territory**
- Besides offering their services, they gather highly informative **traces** about the human mobile activities
- Miniaturization, wearability, pervasiveness, connectivity, context-awareness will produce traces of ever increasing
  - positioning accuracy
  - semantic richness



# Which mobility data?

- Location data from mobile phones, i.e. cell positions in the GSM/UMTS network.
- Location data from GPS-equipped devices – Galileo in the (near?) future
  - Current generation of smart phones come with on-board GPS receiver, and can transmit GPS tracks by SMS/MMS
- Location data from
  - peer-to-peer mobile networks
  - intelligent transportation environments – VANET
  - ad hoc sensor networks, RFIDs (radio-frequency ids)



# Mobility, Data Mining and Privacy

- Towards an **archaeology of the present**
- A scenario of great opportunities and risks:
  - mining mobility data can yield useful knowledge;
  - but, individual privacy is at risk.
- A new multidisciplinary research area is emerging at this crossroads, with potential for broad social and economic impact
  - F. Giannotti and D. Pedreschi (Eds.)  
*Mobility, Data Mining and Privacy*.  
Springer, 2008.





A paradigmatic project:

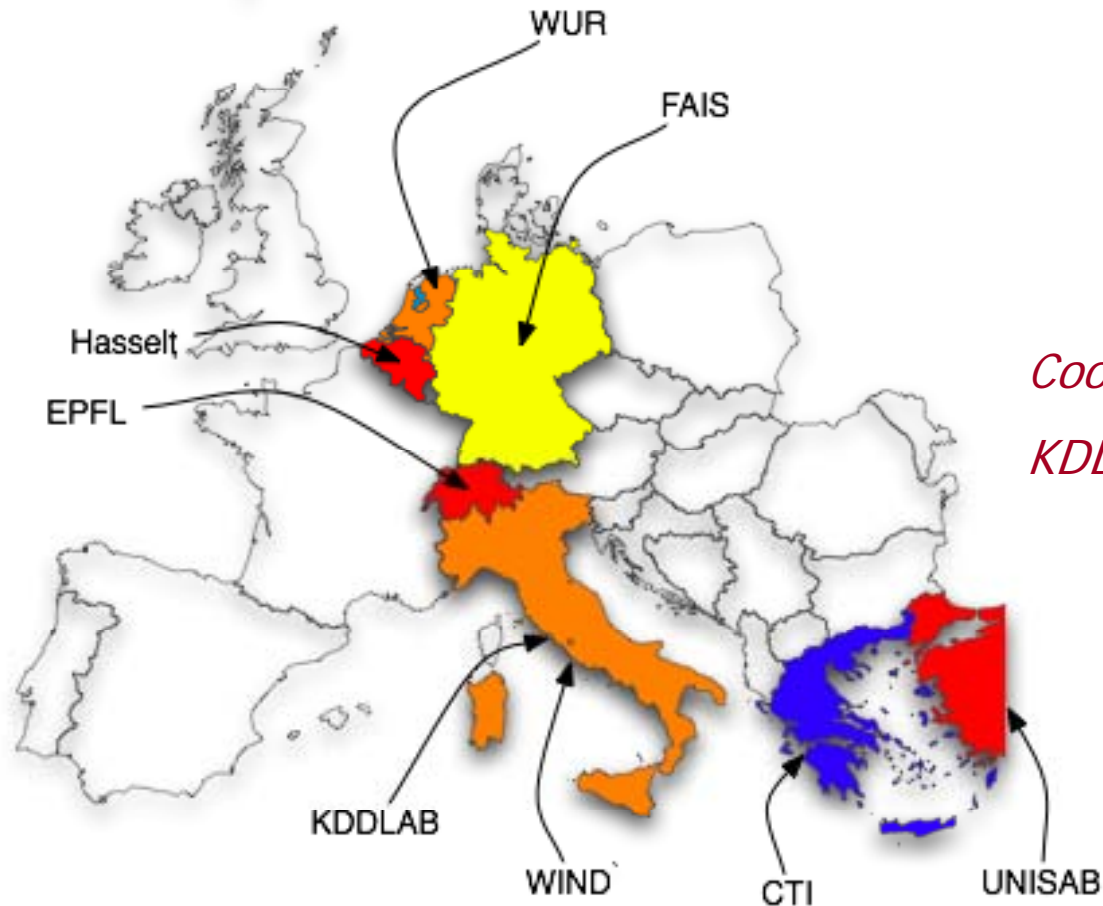
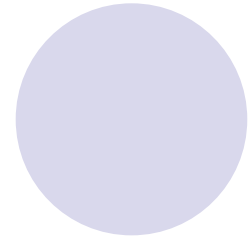
**GeoPKDD**

<http://www.gopkdd.eu>

A European FP6 project

**Geographic Privacy-aware  
Knowledge Discovery and  
Delivery**





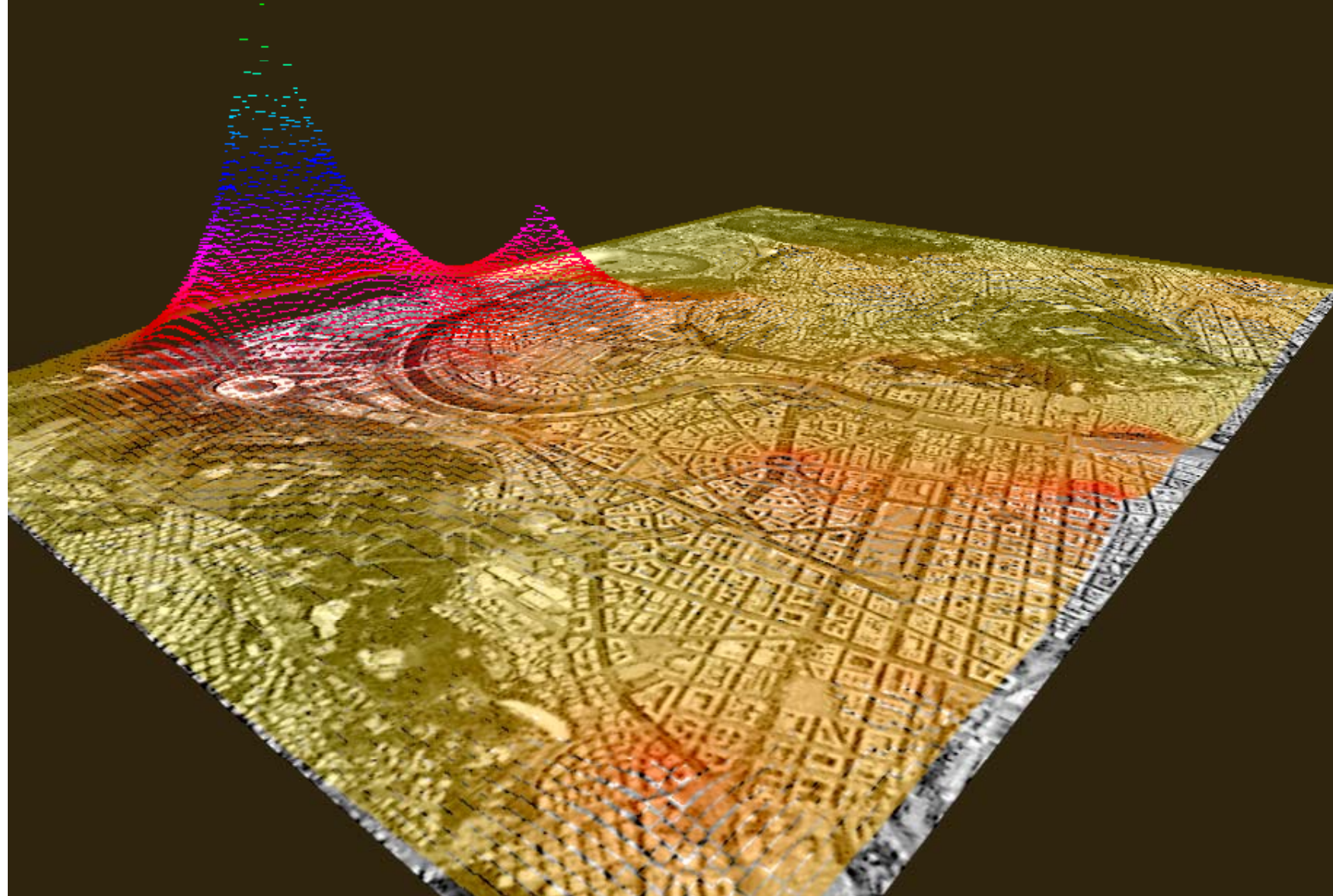
*Coordinator:  
KDD-LAB Pisa, ISTI-CNR*





Madonna Concert  
Cellphone activity in Stadio Olimpico Rome  
2006-08-06

At Rome's Olympic Stadium  
Located about three kilometres from the Vatican  
During the song Live to Tell...  
Madonna appeared against a mirrored cross



# Modelling movement



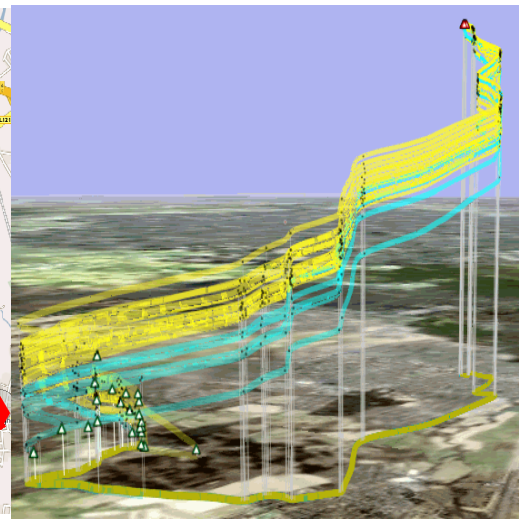
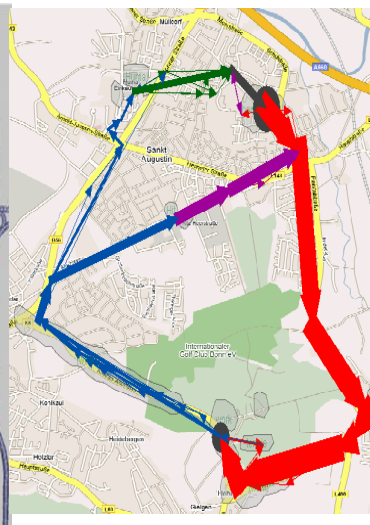
# From mobility data to mobility patterns



# From mobility data to mobility patterns

```
name|date|y|x
Prinzessin|08.20.1998|52.118|12.087
Prinzessin|08.23.1998|51.019|15.309
Prinzessin|08.26.1998|47.723|22.786
Prinzessin|08.29.1998|43.040|27.119
Prinzessin|08.31.1998|38.715|32.165
Prinzessin|09.01.1998|37.195|35.255
Prinzessin|09.03.1998|32.979|36.021
Prinzessin|09.05.1998|28.513|33.437
Prinzessin|09.06.1998|23.961|32.937
Prinzessin|09.07.1998|19.418|33.446
Prinzessin|09.12.1998|15.823|34.094
Prinzessin|10.11.1998|14.685|32.848
Prinzessin|11.03.1998|11.510|32.591
Prinzessin|11.24.1998|13.888|35.667
Prinzessin|12.08.1998|12.562|34.777
Prinzessin|12.10.1998|9.124|35.644
```

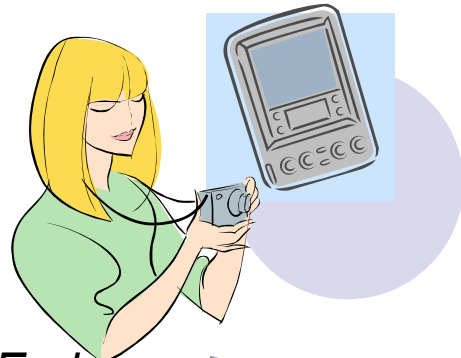
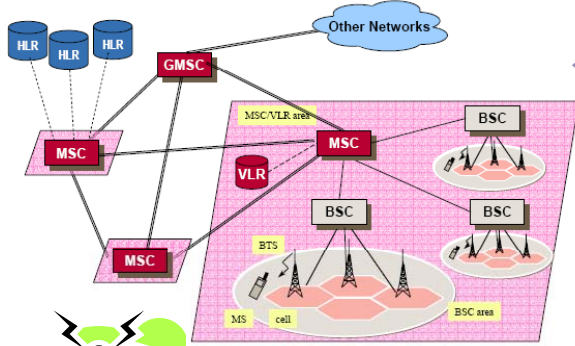
...



# Mobility data mining and the Geographic Knowledge Discovery process



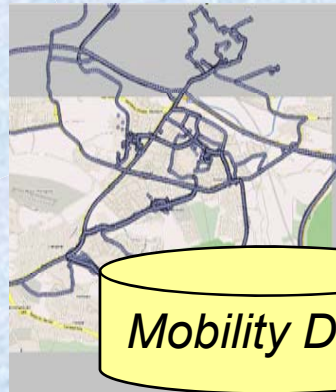
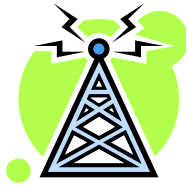
# GSM network, WSN, GPS



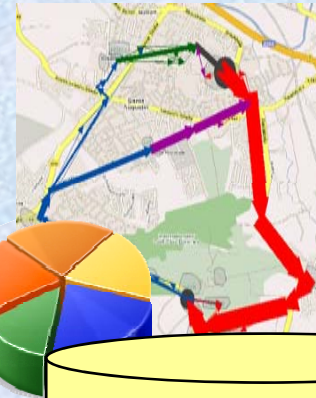
End user



Mobility manager



Mobility Data



Mobility Patterns

name	date	y	x
Prinzessin	08.20.1998	52.118	12.087
Prinzessin	08.23.1998	51.019	15.309
Prinzessin	08.26.1998	47.723	22.786
Prinzessin	08.29.1998	43.040	27.119
Prinzessin	08.31.1998	38.715	32.165
Prinzessin	09.01.1998	37.195	35.255
Prinzessin	09.03.1998	33.070	26.021
Prinzessin	11.08.1998	33.070	26.021
Prinzessin	12.08.1998	12.562	34.777
Prinzessin	12.10.1998	9.124	35.644
...			

Raw data

Privacy and anonymity protection



# GeoPKDD results

- Trajectory DB Management System and data warehouse
  - Theodoridis and colleagues, Athens, Raffaetà and colleagues Venice
- A repertoire of mobility patterns and models
  - Nanni, Pedreschi and colleagues, Pisa
- A visual analytics environment for mobility data
  - Andrienko's, Fraunhofer – Rinzivillo, Pedreschi, Pisa
- A repertoire of privacy-preserving analysis techniques
  - Saygin, Istanbul – Bonchi, Giannotti, Pedreschi, Pisa – Damiani, Milan
- A mobility data mining query language
  - Giannotti, Manco, Renso and colleagues, Pisa + Cosenza
- A reasoning framework for mobility data mining applications
  - Macedo, Spaccapietra, EPFL + Renso, Pisa + Wachowicz, Madrid



# Mobility data analysis in Milano

- WIND Telecomunicazioni spa (major telecom provider, GeoPKDD partner)
    - GSM data (calls & handover data: flows between adjacent cells)
  - Other collaborations:
    - AMA-MI, Comune di Milano, Mobility Agency
    - Infoblu and OctoTelematics (GPS receivers on board of cars with special insurance contract)
- 
- Experience on a massive GPS track dataset
    - 17 K vehicles over one week,
    - 2 M positions,
    - 200 K trajectories

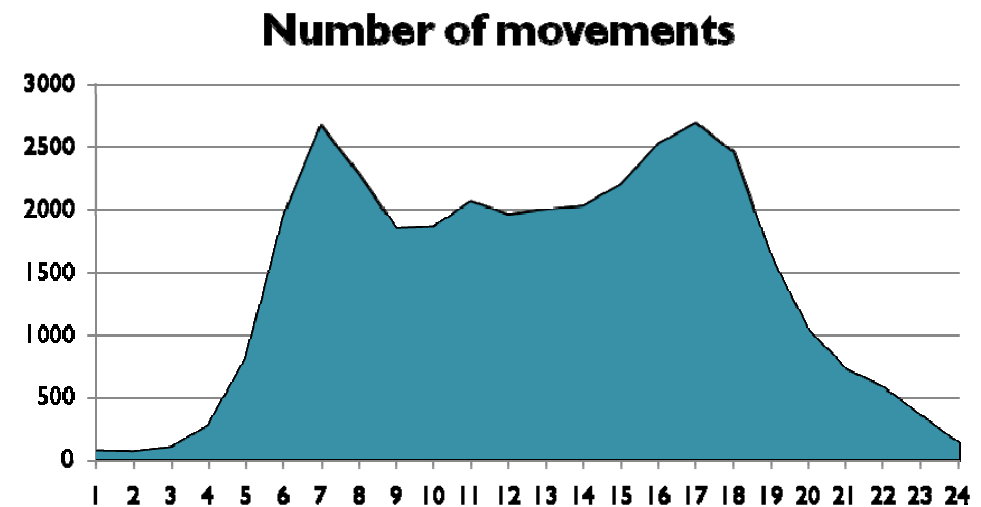
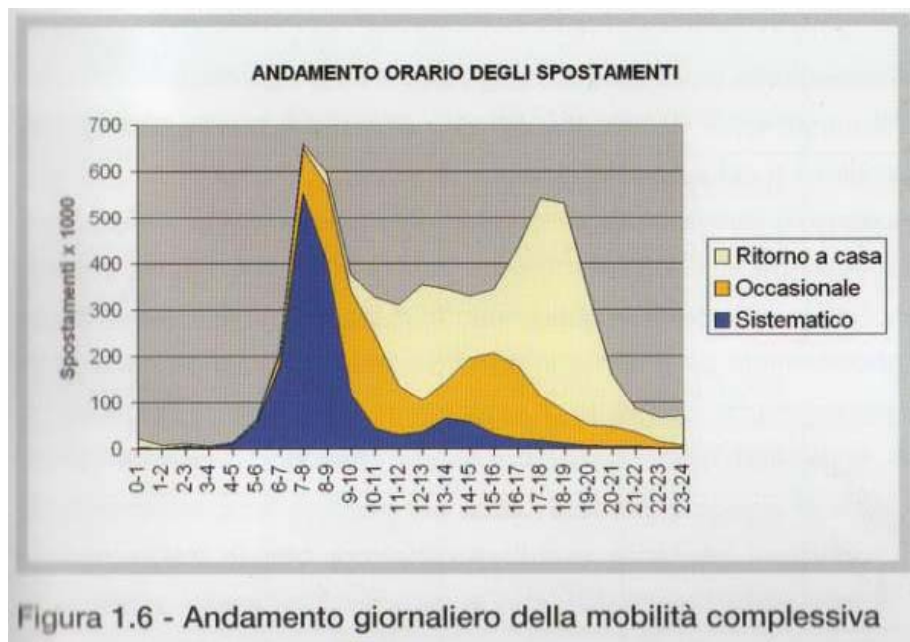




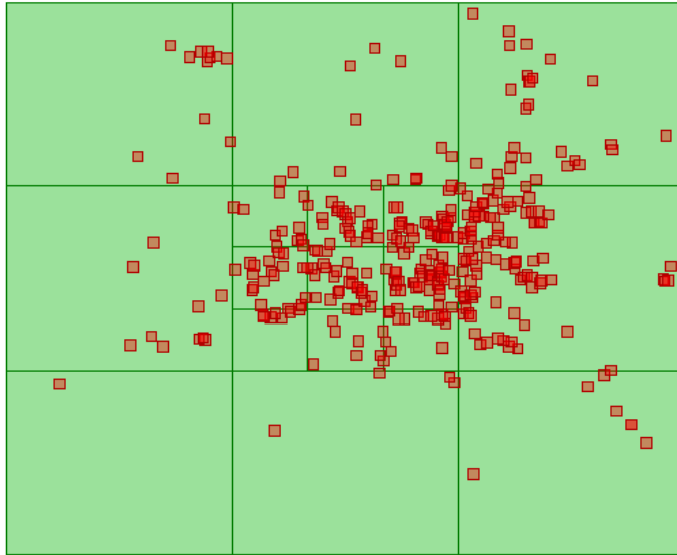


# Consistency with survey data

- Daily mobility trend



# Asymmetric OD- Matrix (Parking Lots)

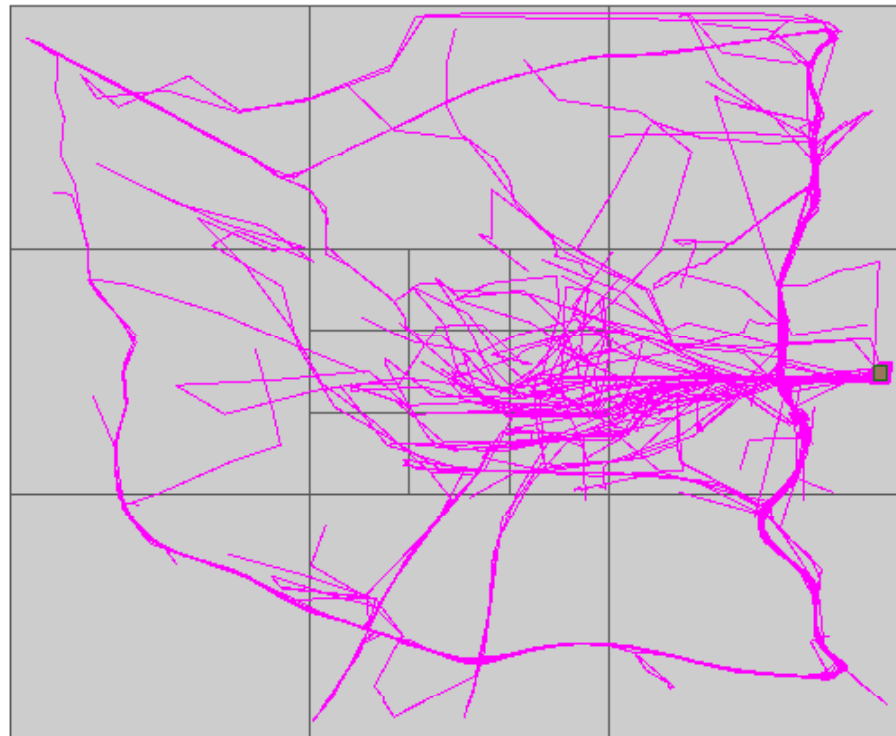


RawData	
Visualize	
TOREGION	SUM(NUM)
317	106
318	83
319	54
343	36
342	36
345	35
349	34
235	33
114	29
214	29
213	28
346	28
269	27
350	27
271	25
341	24
183	24
181	24
119	24
106	24
58	24
121	23
29	23
172	23
238	23
115	22
328	22
19	22
297	22
67	22
234	21
143	21
83	21
107	21
16	21
194	21
279	21
179	21
129	21
45	21
176	20
180	20

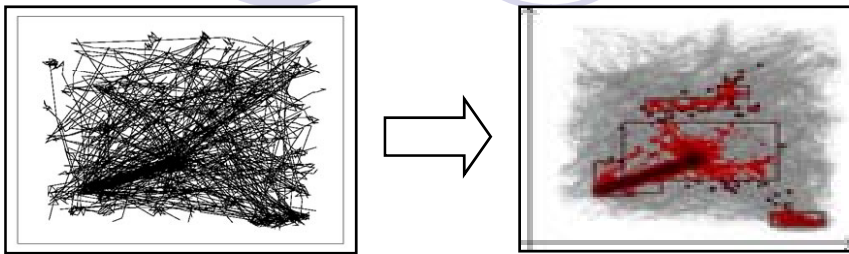


# *Outer Parking Lots (317 - Linate)*

*All the incoming trajectories from each district to the specified outer parking lots*

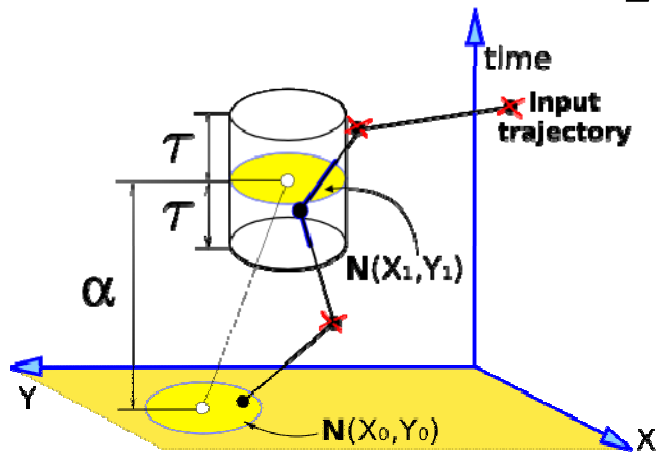


# Trajectory Pattern Mining



*1- Find Regions of Interest*

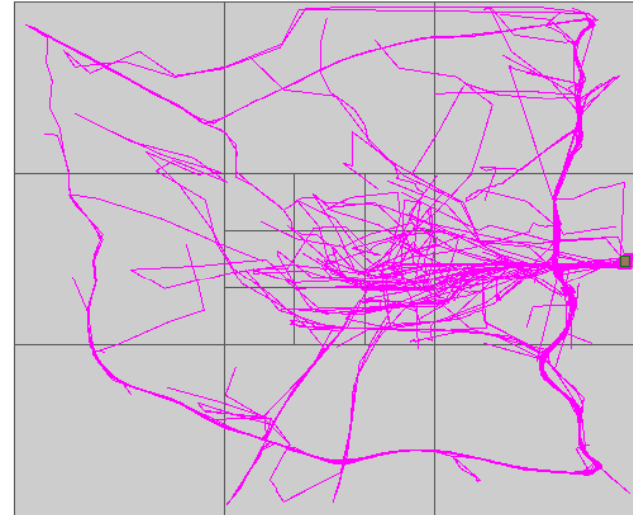
*2- Find similar Trajectory in space and time*



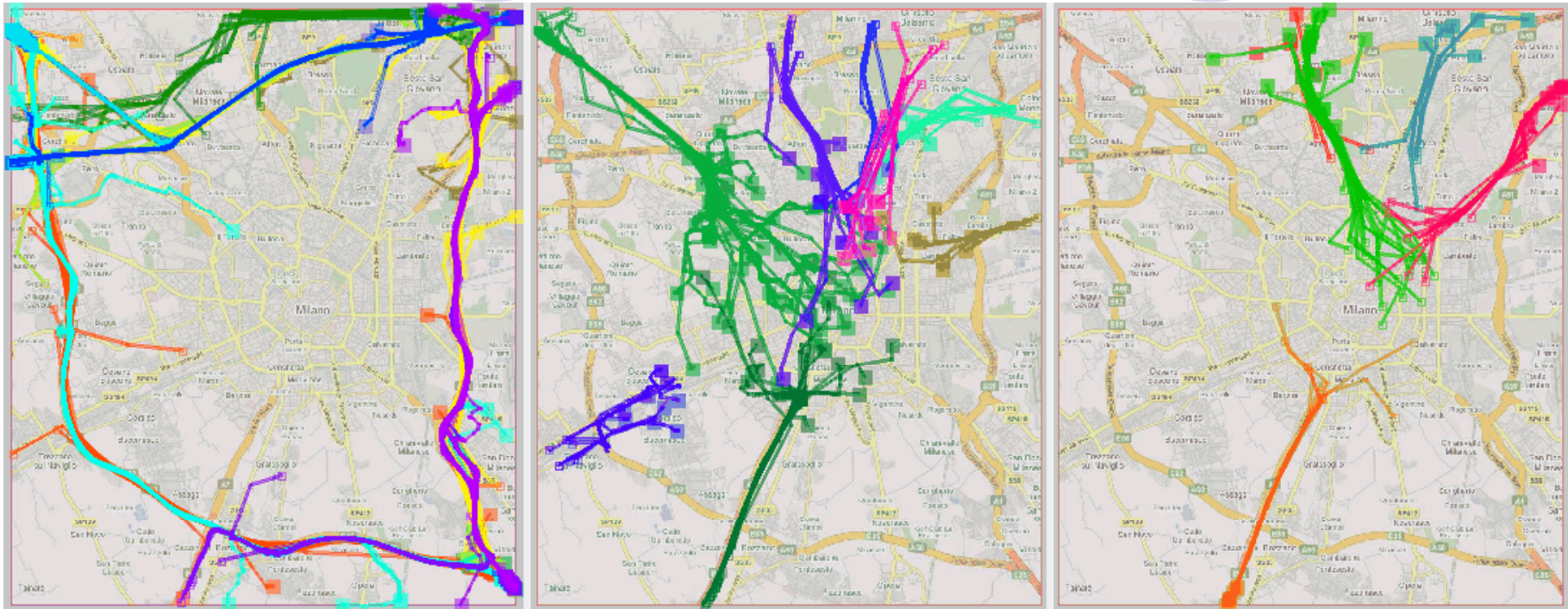
*3- Extract patterns:*



# *T-Patterns for trajectories to Parking lot 317*



# Trajectory Clustering

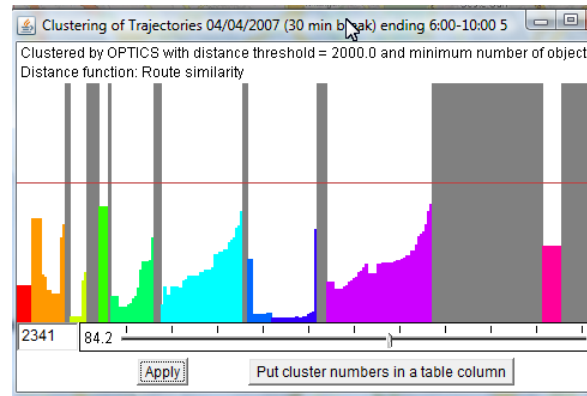
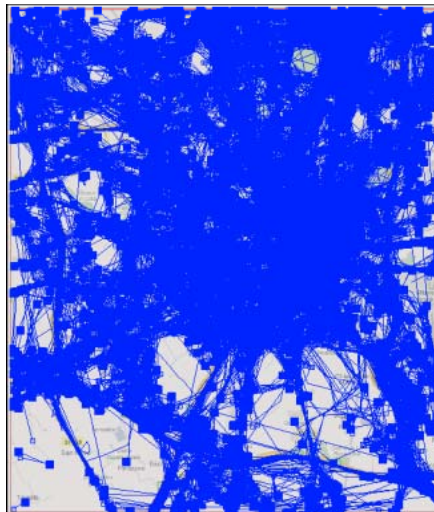
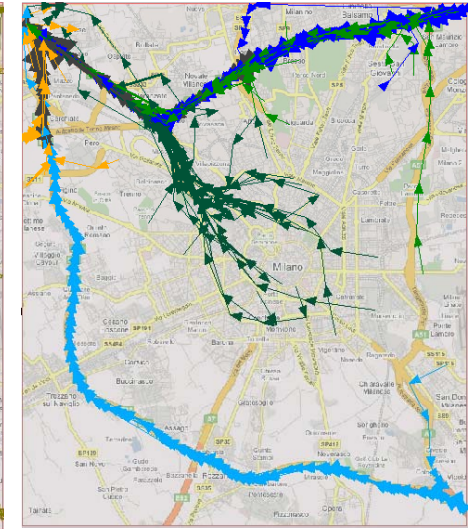
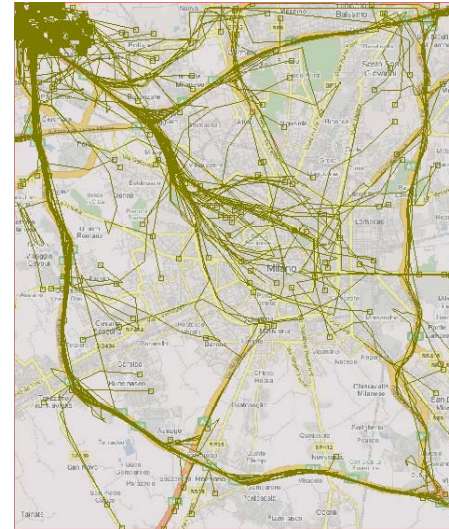
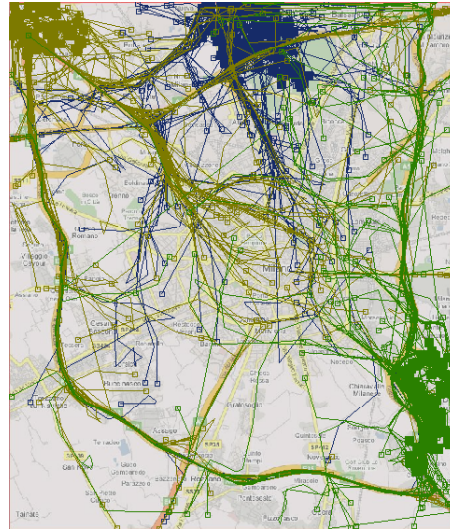


*Left: peripheral routes; middle: inward routes; right: outward routes.*

- *Rinzivillo, Pedreschi, Nanni, Giannotti, Andrienko, Andrienko  
Visually-driven analysis of movement data by progressive  
clustering. J. of Information Visualization, 2008*



# Analytical effect of progressive clustering





# From opportunities to threats

- Personal mobility data, as gathered by the wireless networks, are extremely sensitive
- Their disclosure may represent a brutal violation of the privacy protection rights, i.e., to keep confidential
  - the places we visit
  - the places we live or work at
  - the people we meet
  - ...



# The naive scientist's view

- Knowing the exact identity of individuals is not needed for **analytical** purposes
  - De-identified mobility data are enough to reconstruct aggregate movement behaviour, pertaining to groups of people.
- Reasoning coherent with European data protection laws: personal data, **once made anonymous**, are not subject to privacy law restrictions
- Is this reasoning correct?



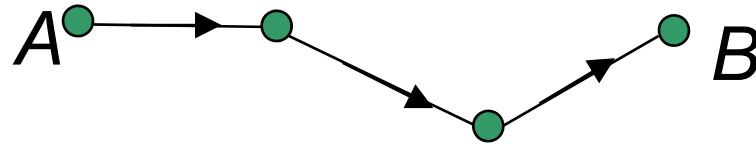
# Unfortunately not!

- Making data (reasonably) anonymous is not easy.
- Sometimes, it is possible to reconstruct the exact identities from the de-identified data.
- Many famous example of re-identification
  - Governor of Massachusetts' clinical records (Sweeney's experiment, 2001)
  - America On Line August 2006 crisis: user re-identified from search logs
- Two main sources of danger:
  - **Many observations** on the same "anonymous" subject
  - **Linking data**, after joining separate datasets

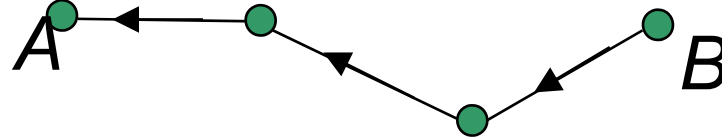


# Spatio-temporal linkage in Mobility Data

Id: 34567



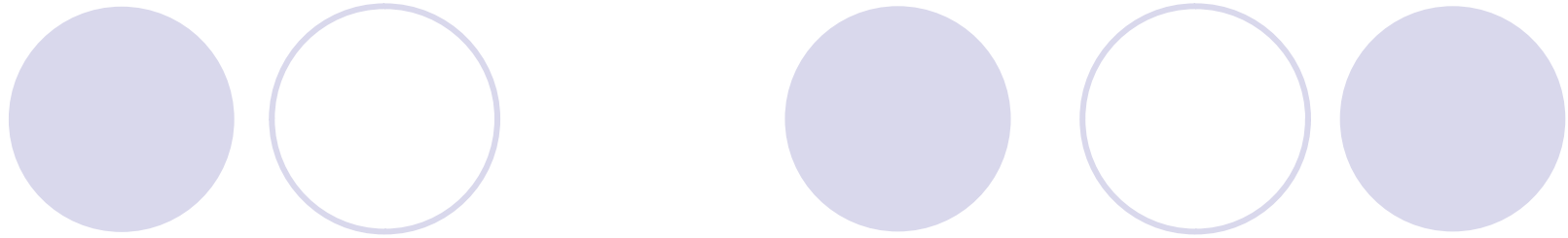
*[almost every day mon-fri  
between 7:45 – 8:15]*



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

- By intersecting the phone directories of locations A and B we 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...





# Privacy- preserving spatio-temporal data mining

**Trajectory randomization is risky!**

**Trajectory anonymization**



# A subtle re-identification attack

- **Disclosure Risks of Distance Preserving Data Transformations**

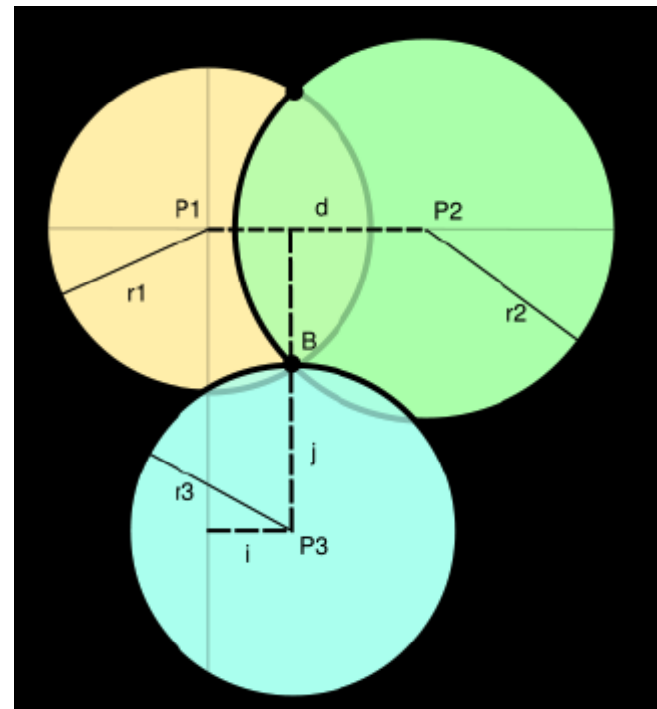
- Erkay Savas, Yucel Saygin, Emre Kaplan, and Thomas B. Pedersen (Sabanci Univ., Istanbul)

- **What if the attacker knows:**

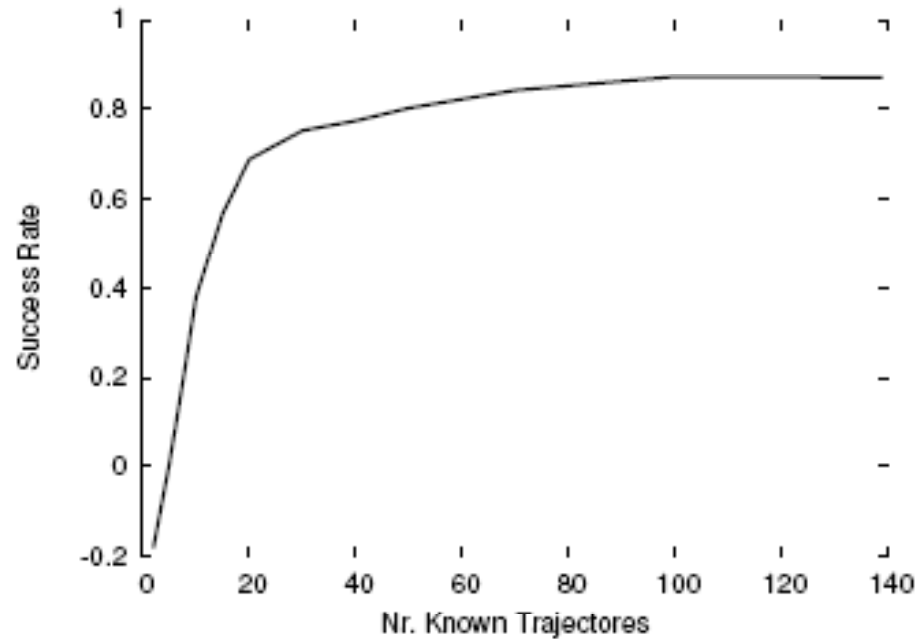
- Some trajectories
  - All mutual distances

- **Hyper-lateralation**

- Works in  $d$  dimensions given  $d + 1$  points
  - If known trajectories are few, then approximate!

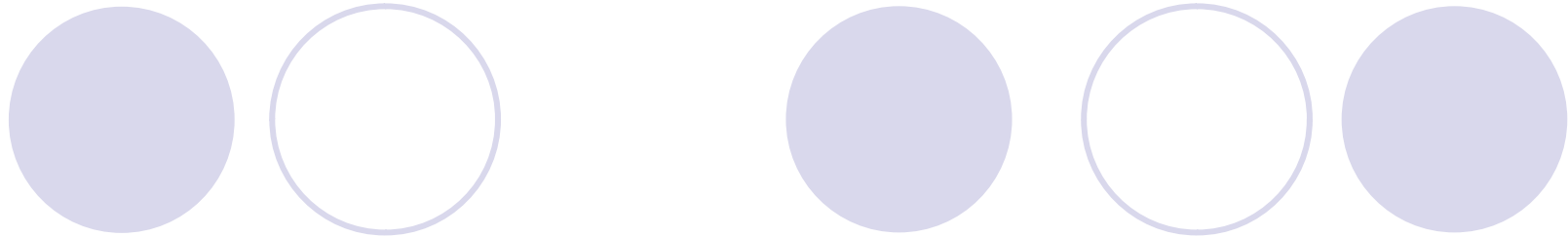


**Red:** true traj **Blue:** approx traj



(b) Success-rate vs. number of known trajectories (Each sample is the average of 60 experiments run for 50.000 iterations).






# Trajectory anonymization: methods and validation

*A. Monreale,  
M. Nanni,  
R. Trasarti,  
R. Vandoni*



*KDDLab ISTI-CNR, Pisa*

*Based on joint work with  
O. Abul  and F. Bonchi*





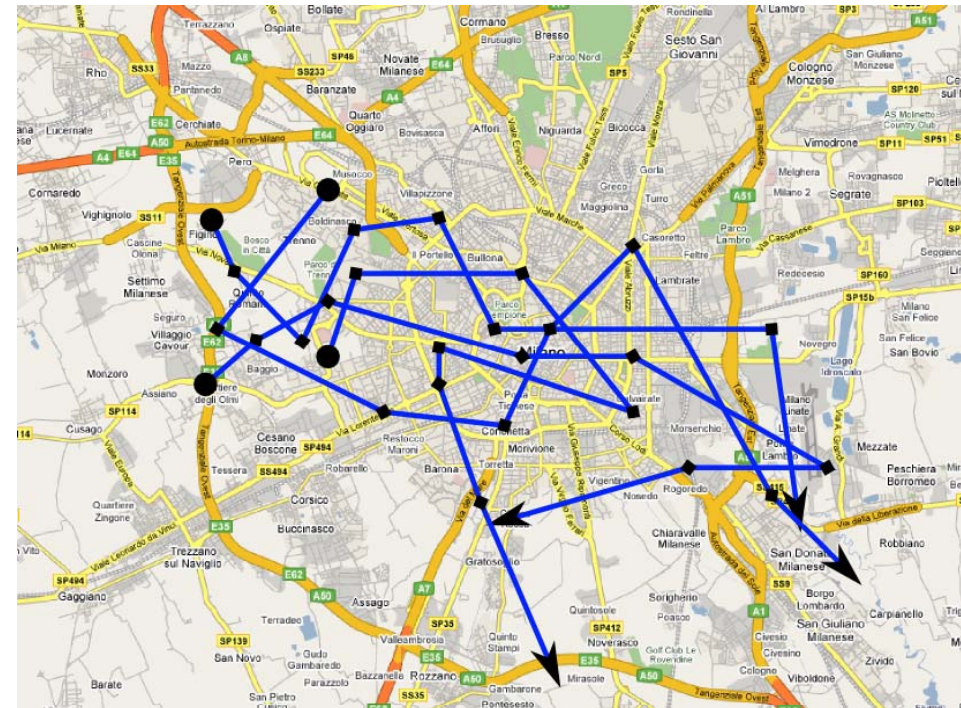
# Context

- *Large diffusion of mobile devices, mobile services and location-based services*



# Context

- *Such devices leave digital traces that can be collected in databases to describe the spatio-temporal trajectories of human companions*





Context

- *Many benefits with sharing moving objects databases (MOD)*
  - *e.g. do-it-yourself computation*
- *The side effect, i.e. risk*
  - *Disclosure of location privacy of individuals*
- *Our solution*
  - *First anonymize, and then publish*



(k,  $\delta$ )-anonymity

- Reasoning

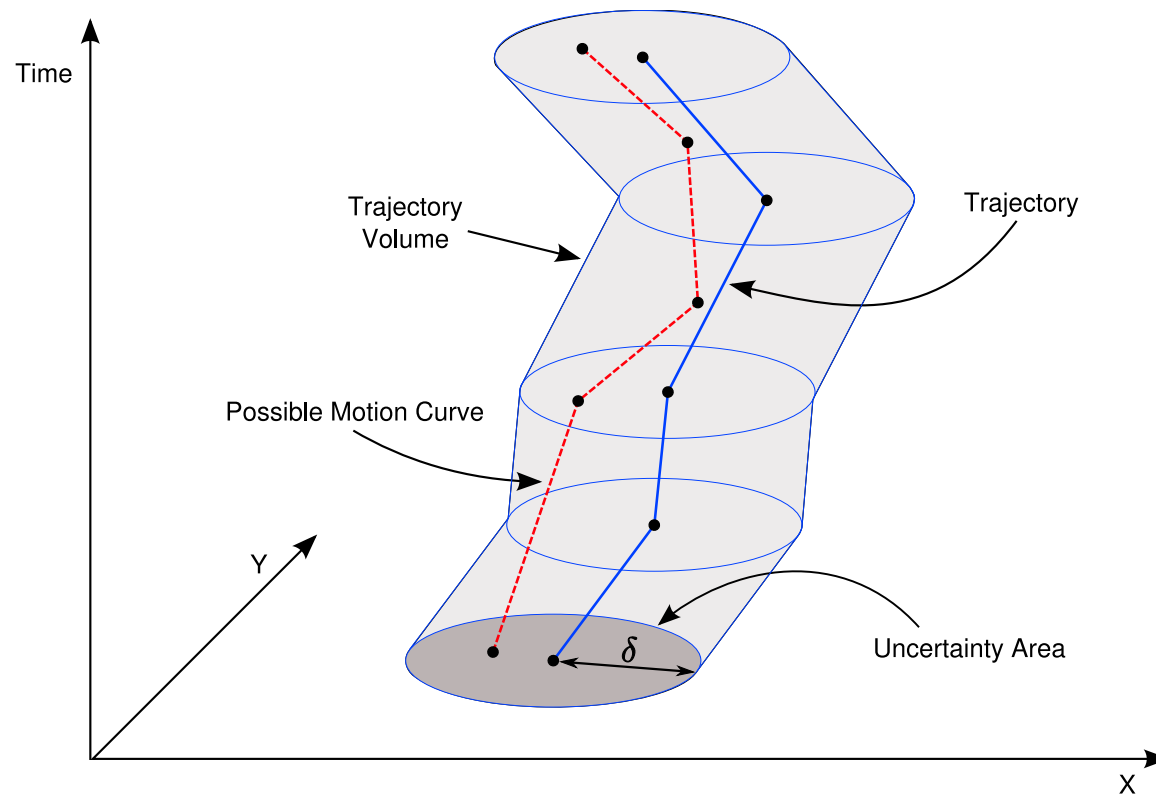
- The parameter k: each trajectory must be indistinguishable from at least k-1 others
  - Originating from the k-anonymity requirement
- The parameter  $\delta$  : inherent uncertainty,  $\delta$ , (impreciseness) in location data measured by mobile devices, e.g. cell-phone, GPS recv.
  - Originating from the particulars of MODs

[Trajcevski et al. TODS 2004]



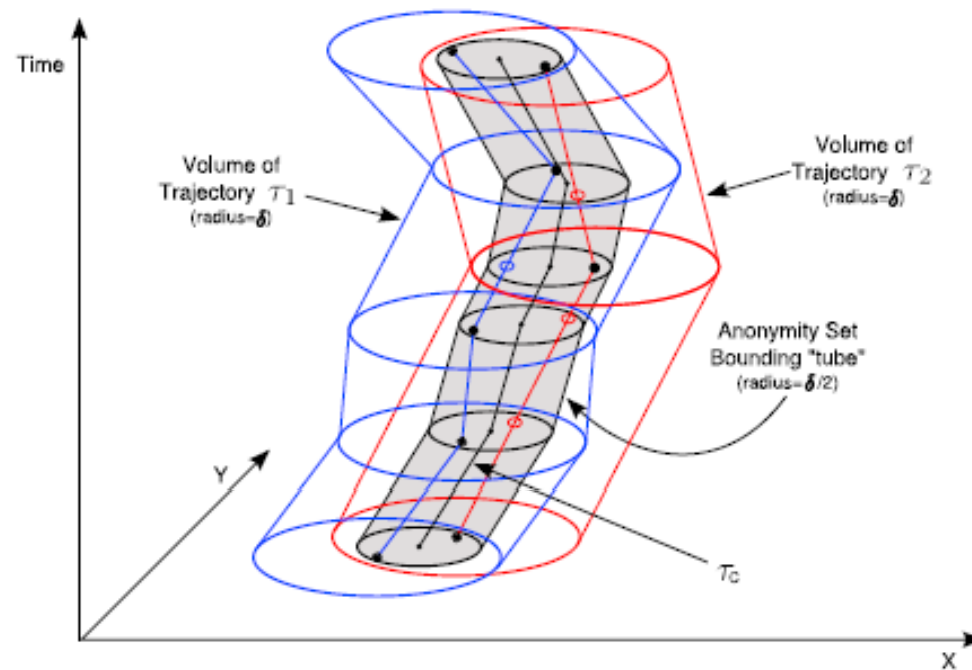
# $(k, \delta)$ -anonymity

- *The location uncertainty:  $\delta$*



# $(k, \delta)$ -anonymity

- A  $(2, \delta)$ -anonymity set





## $(k, \delta)$ -anonymity

- *$(k, \delta)$ -anonymity problem:* Given a dataset of trajectories  $\mathcal{D}$ , an uncertainty threshold  $\delta$  and an anonymity threshold  $k$ , the problem of  $(k, \delta)$ -anonymity requires to transform  $\mathcal{D}$  in a dataset  $\mathcal{D}'$ , such that for each trajectory  $\tau \in \mathcal{D}'$  it exists a  $(k, \delta)$ -anonymity set  $S \subseteq \mathcal{D}'$ ,  $\tau \in S$ ; and the distortion between  $\mathcal{D}$  and  $\mathcal{D}'$  is minimized.

*!Note that the problem is generic since the distortion metric is not concrete yet*



# Space Translation

- First stage (Candidate Optimal Clustering)
  - Clusters all trajectories such that every cluster has between  $k$  and  $2k-1$  trajectories,
    - an optimality condition for any  $k$ -anonymization
- Second stage (Space Translation)
  - Within each cluster, move trajectories towards the cluster center such that they form a  $(k, \delta)$ -anonymity set.
    - Optimally solved when  $\delta=0$ , and approximated otherwise
- First stage is more involved than the second





# Greedy clustering algorithm (GC)

1. Choose a pivotal trajectory:
  - At the first step: random trajectory
  - In the following steps: choose the trajectory that is farthest from the previous pivot
2. Find its  $(k-1)$ -Nearest Neighbors to form a cluster, and remove them from the dataset
3. If at least  $k$  objects are left, continue from 1.





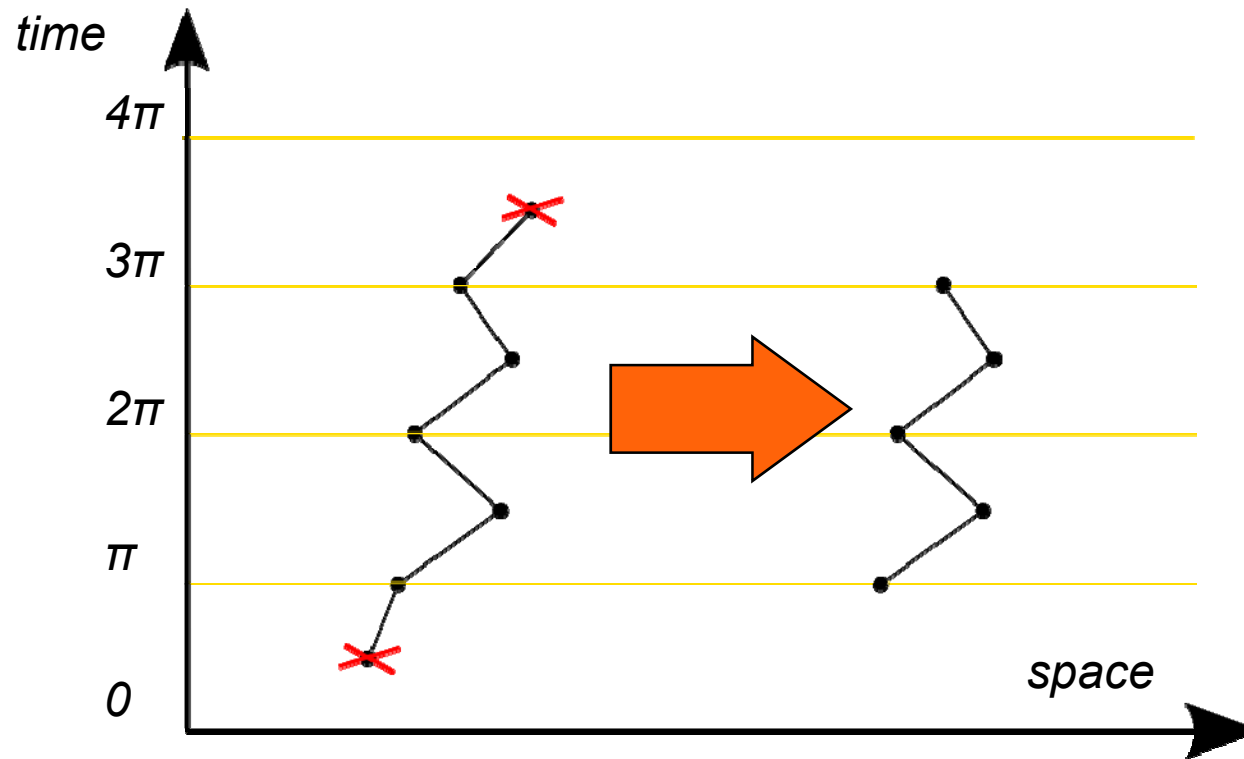
# Never Walk Alone (NWA)

- GC algorithm with ad-hoc preprocessing and outlier detection and removal
- Clustering constrained to yield
  - compact clusters → to limit distortion
  - only little trash → to limit deletions
- Two conflicting goals
- A trade-off is sought:
  - Start with tight constraints, then relax iteratively, till a satisfactory solution can be found



# Same Time Span Grouping

- *Cut heads and tails w.r.t. time period  $\pi$*

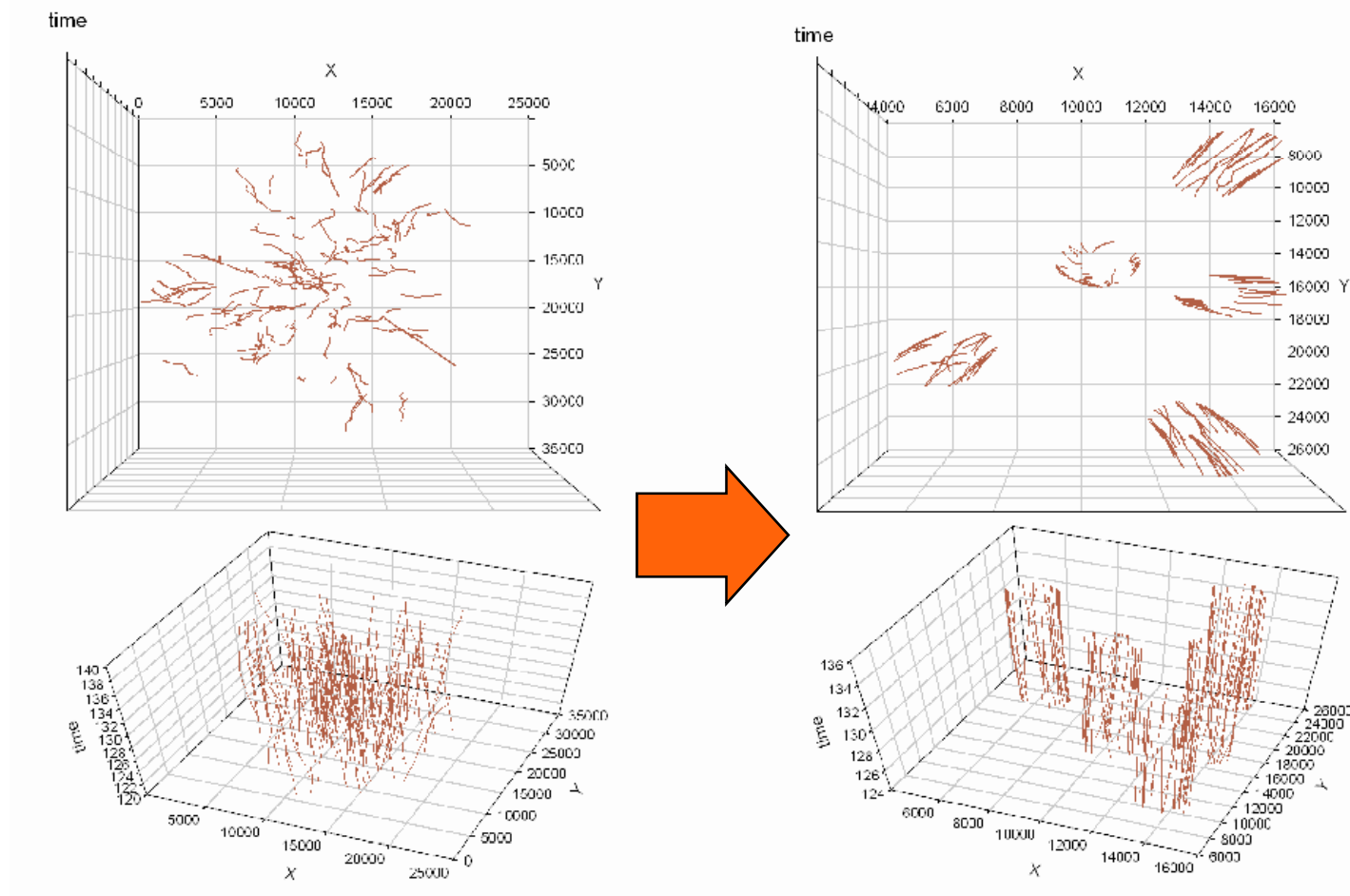


- *Group trajectories with same start & end times*

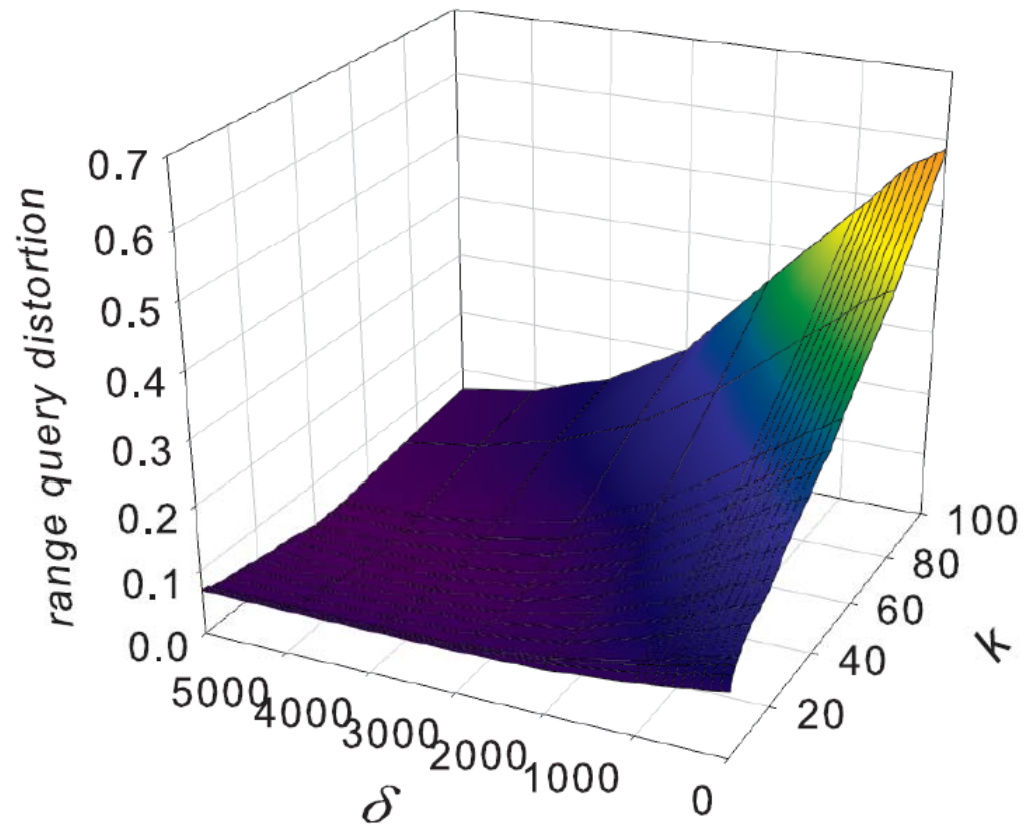


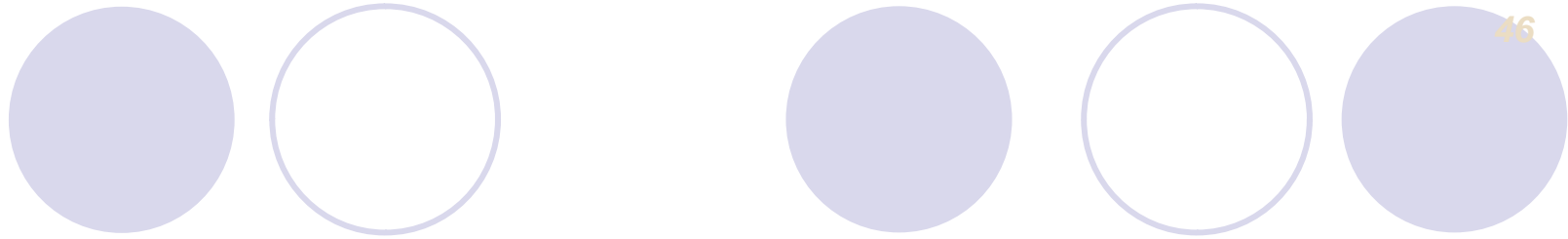
# Sample result

(NWA on Oldenburg synthetic dataset)



# NWA on Oldenburg (PSI)

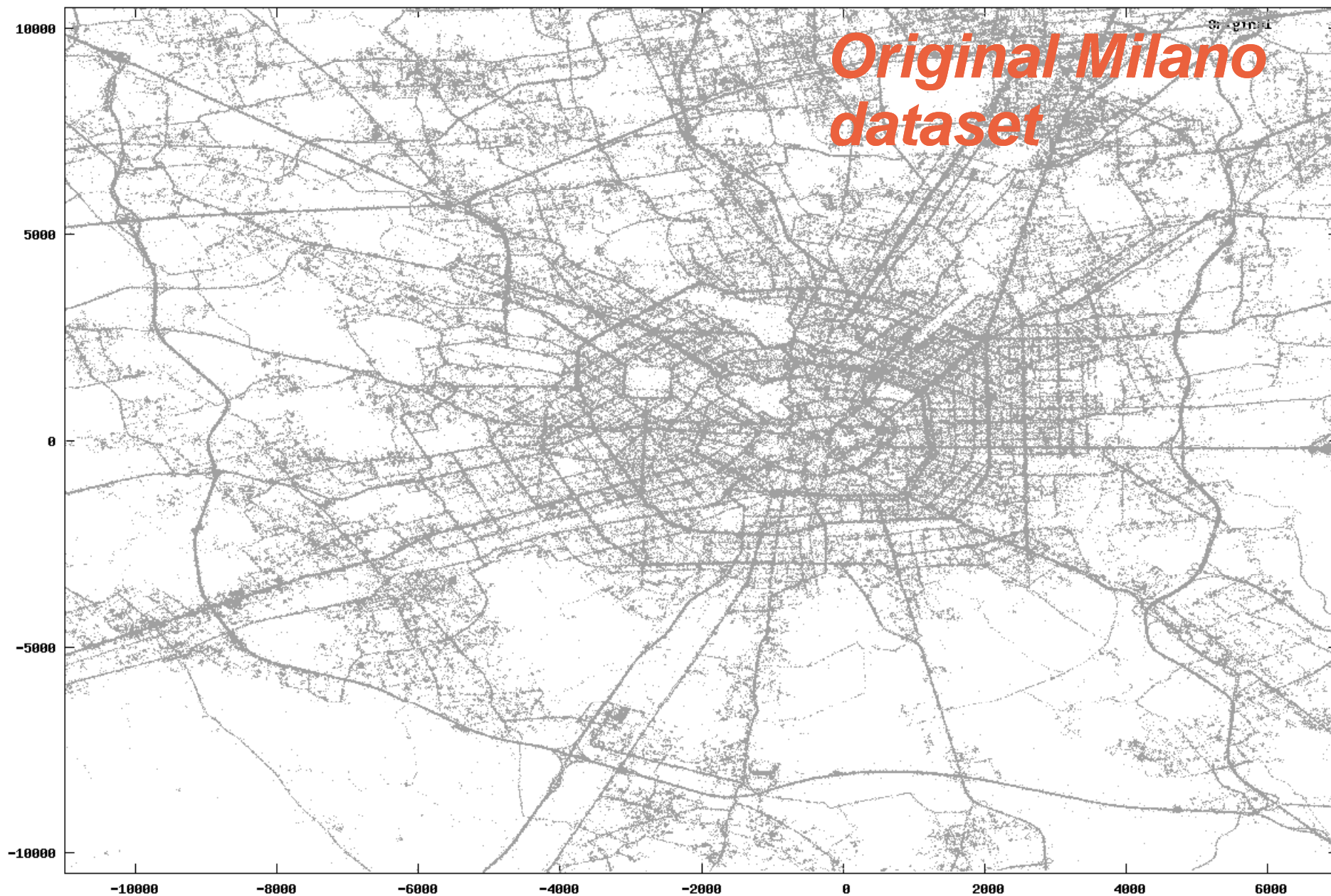


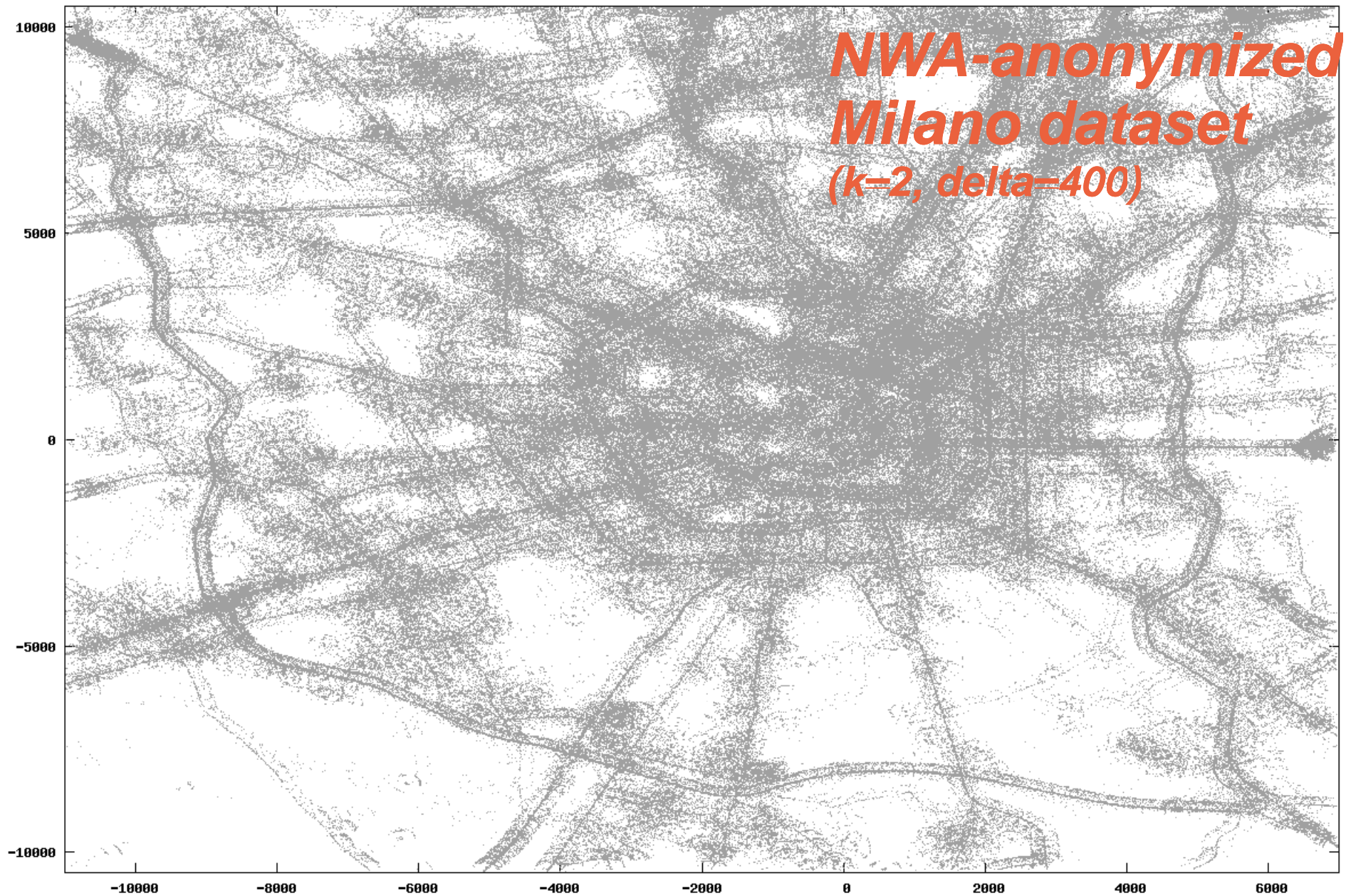


# EVALUATION OF NWA



# Original Milano dataset

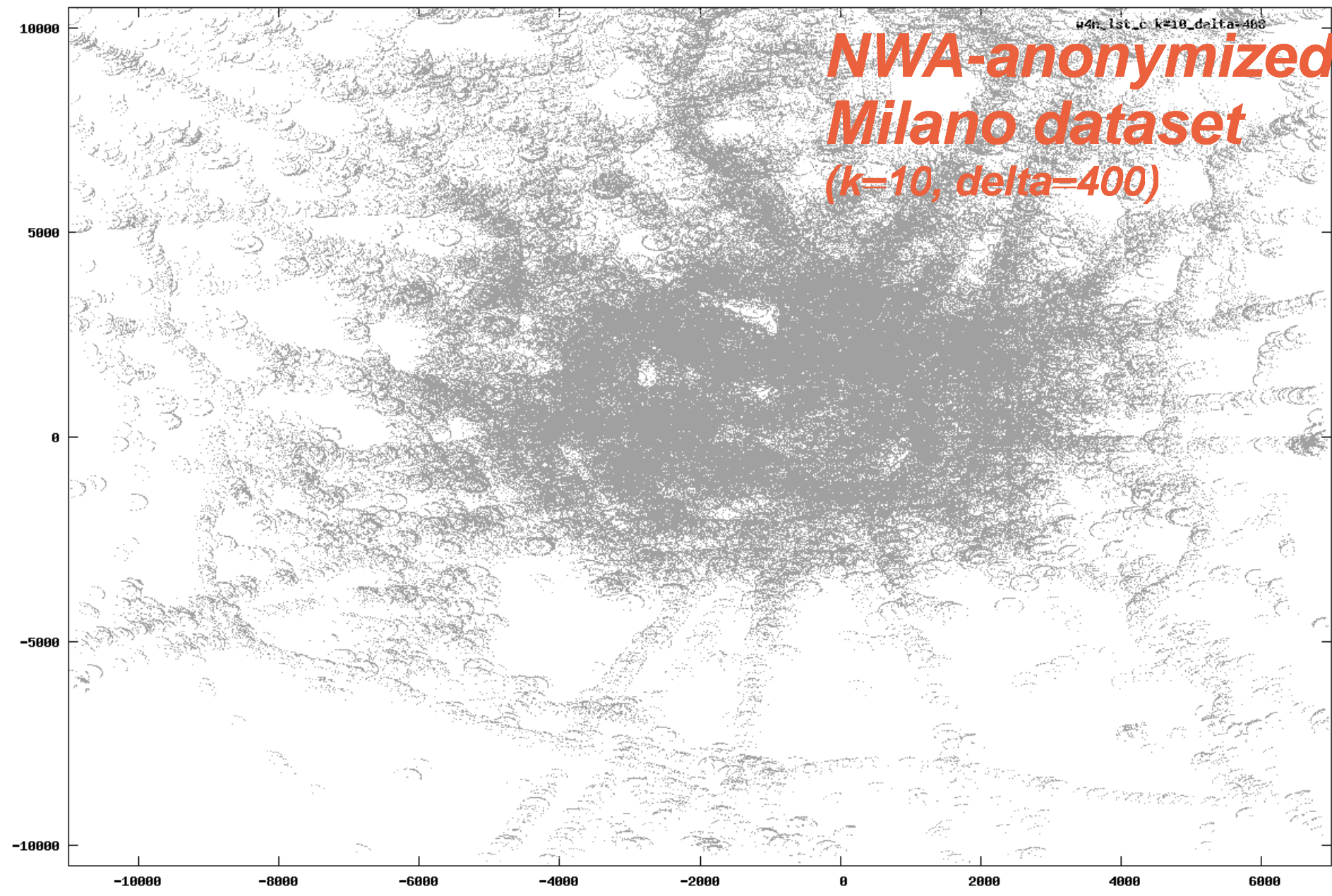






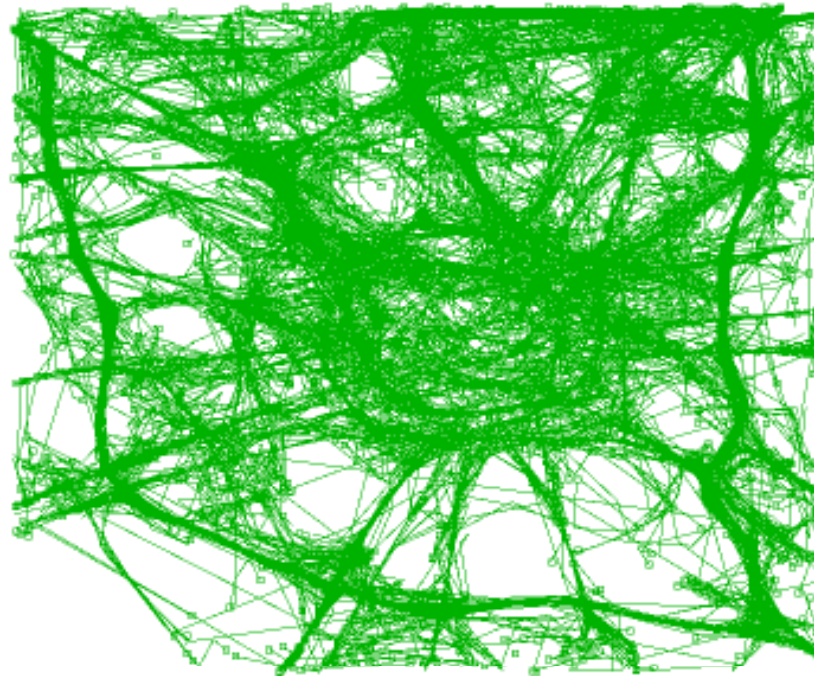
#40 lst\_c k=10\_delta=400

***NWA-anonymized  
Milano dataset  
(k=10, delta=400)***

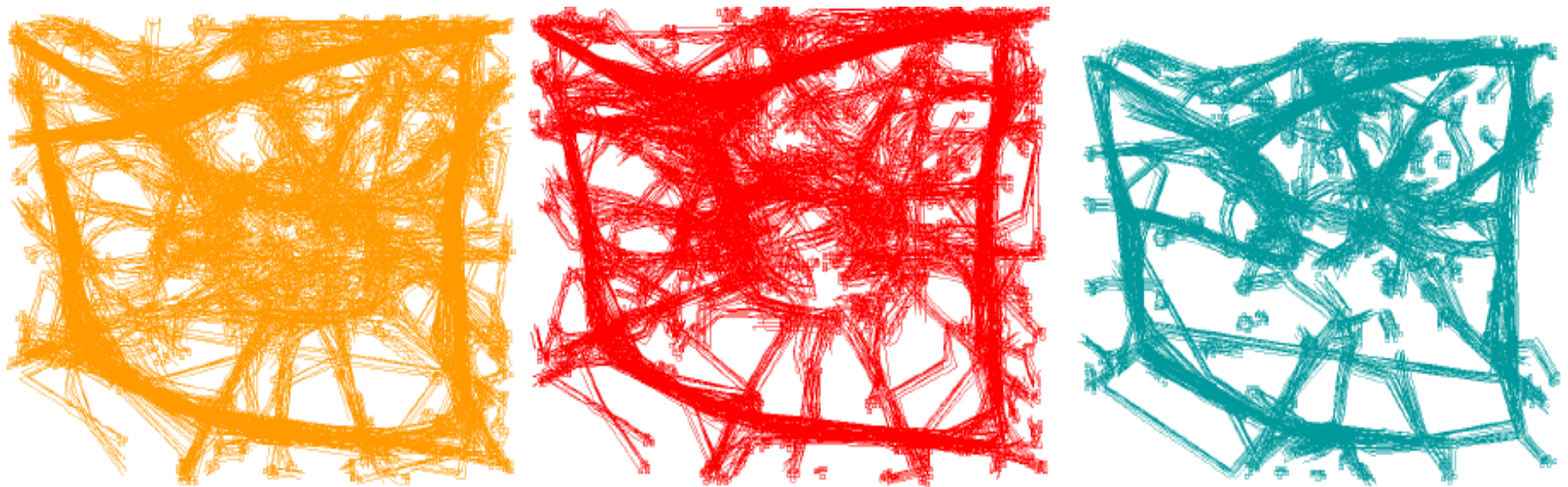


# Dataset

- *Dataset of GPS trajectories (project GeoPKDD)*
  - *3200 trajectories Sunday morning from 6:00 -11:00 am*



# Visual effects of anonymity



(b) Anonymous Dataset  $k=5$   $\delta=600m$

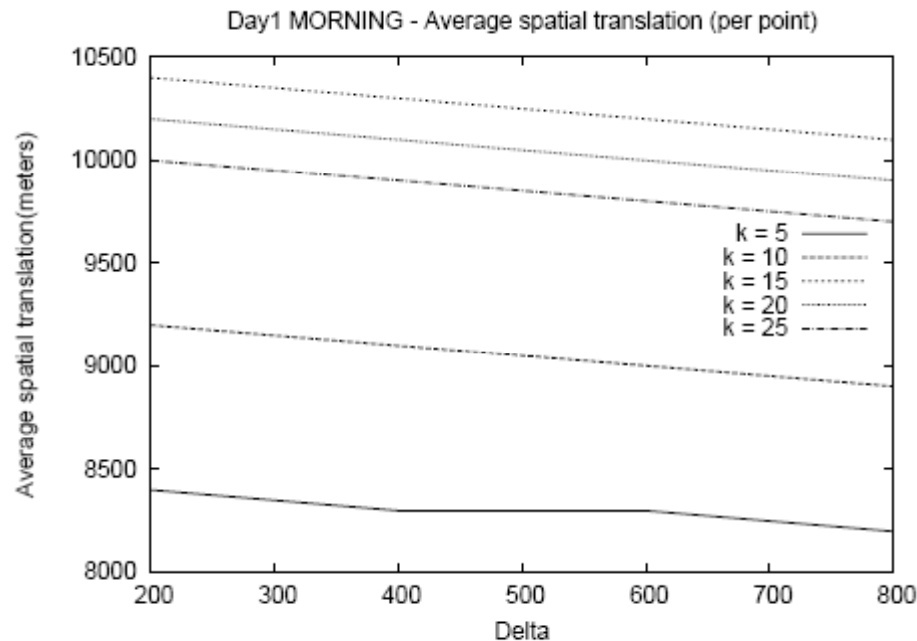
(c) Anonymous Dataset  $k=10$   $\delta=600m$

(d) Anonymous Dataset  $k=20$   $\delta=600m$

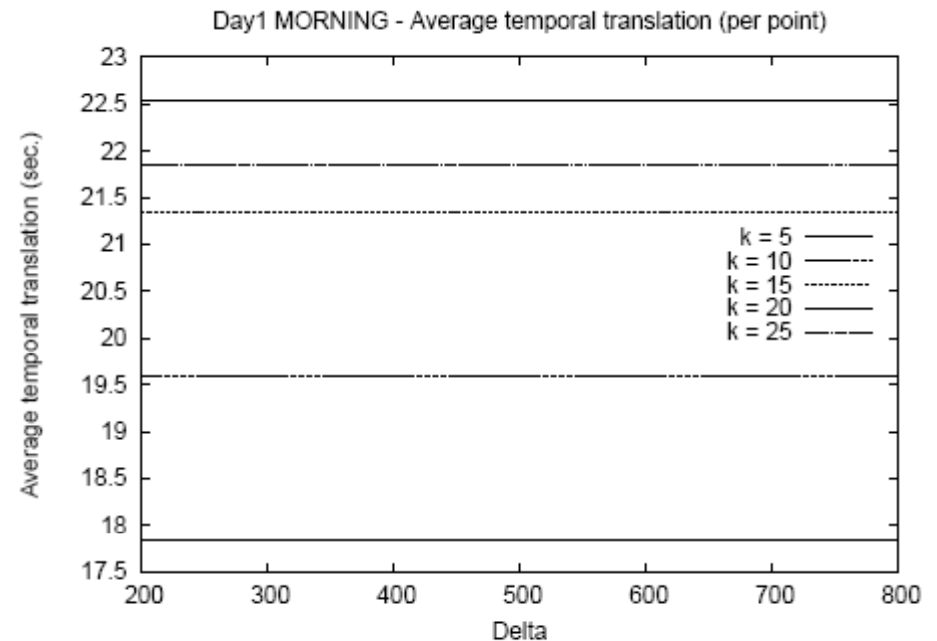


# Spatial and Temporal distortion

- *Distorsion for different  $k$  and  $d$  values*



(a) Average spatial translation

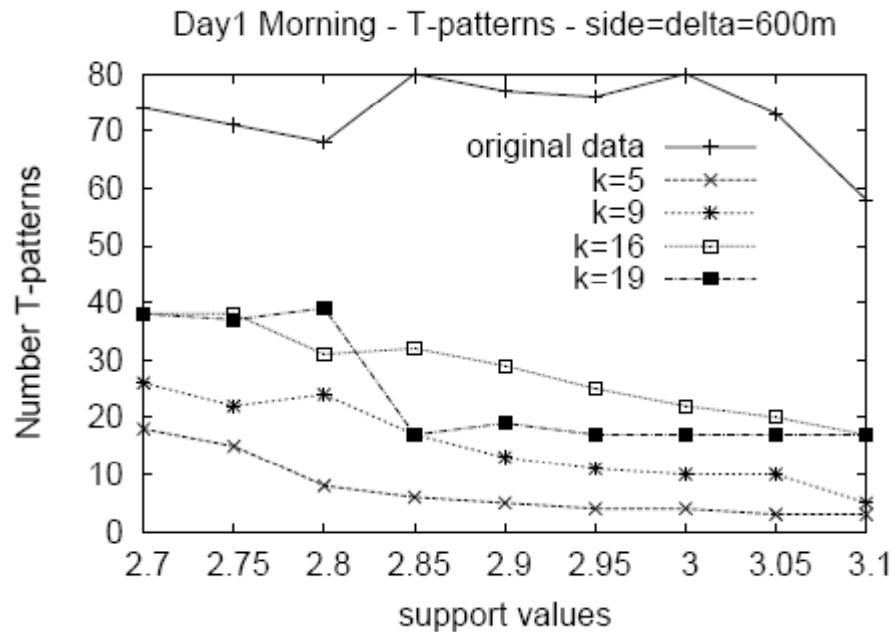


(b) Average temporal translation

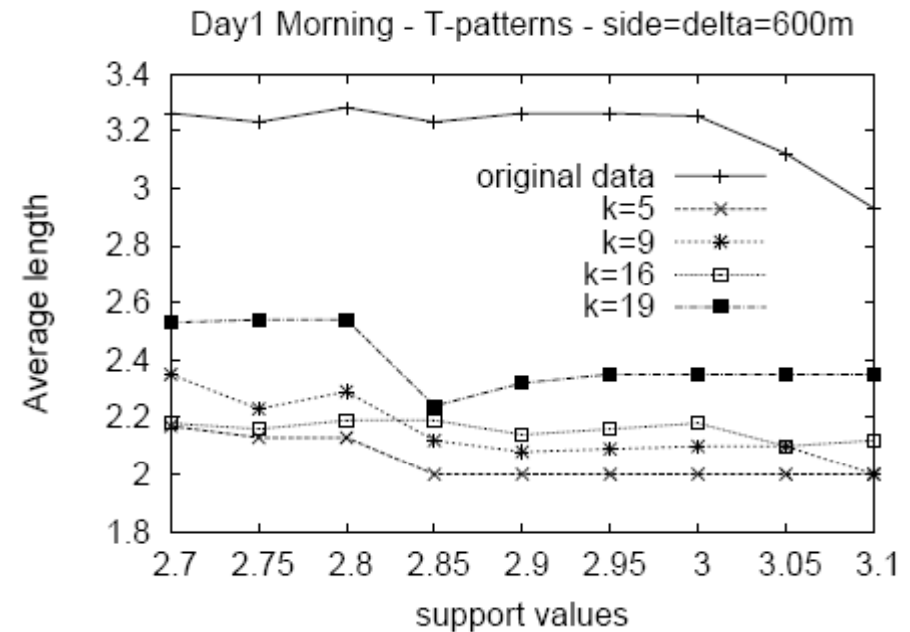


# Trajectory patterns preservation

- *For different  $k$  values:*
  - *Analyze the number of  $t$ -patterns*
  - *Analyze the average length of  $t$ -patterns*



(a) Number of T-patterns



(b) Average Length of T-patterns



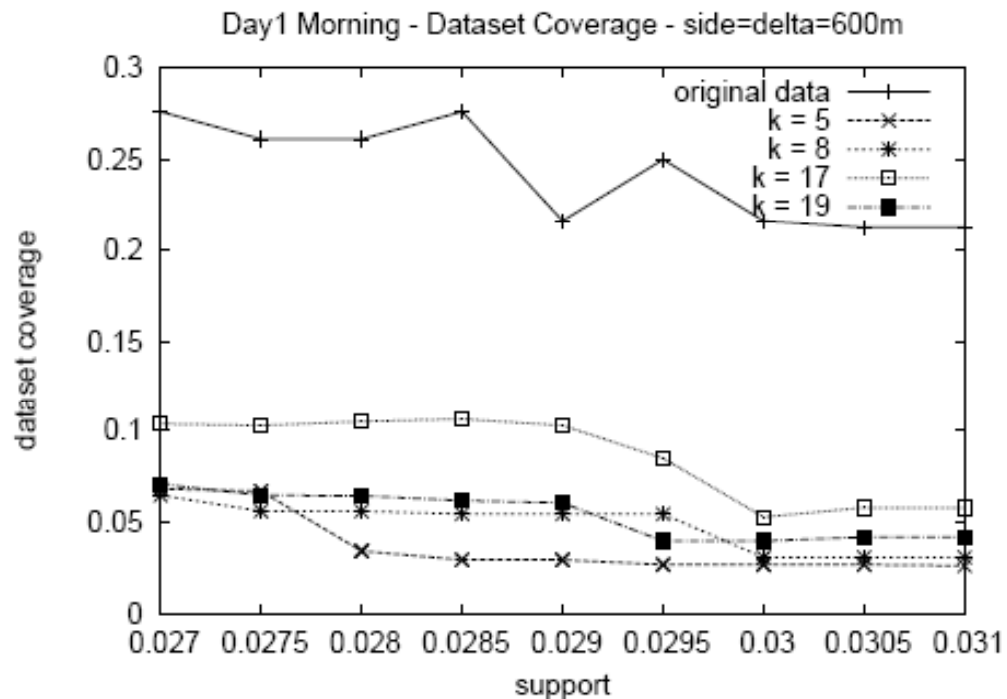
# Predictive power of anonymous T-patterns

- Use evaluation functions to estimate the predictive power of a T-pattern set
  - **Dataset Coverage:** measures the ratio of trajectories covered by the T-pattern set
  - **Spatial Coverage:** measures the ratio of space covered by the T-pattern set
  - **Region Separation:** measures ratio of regions crossed by the analyzed T-patterns



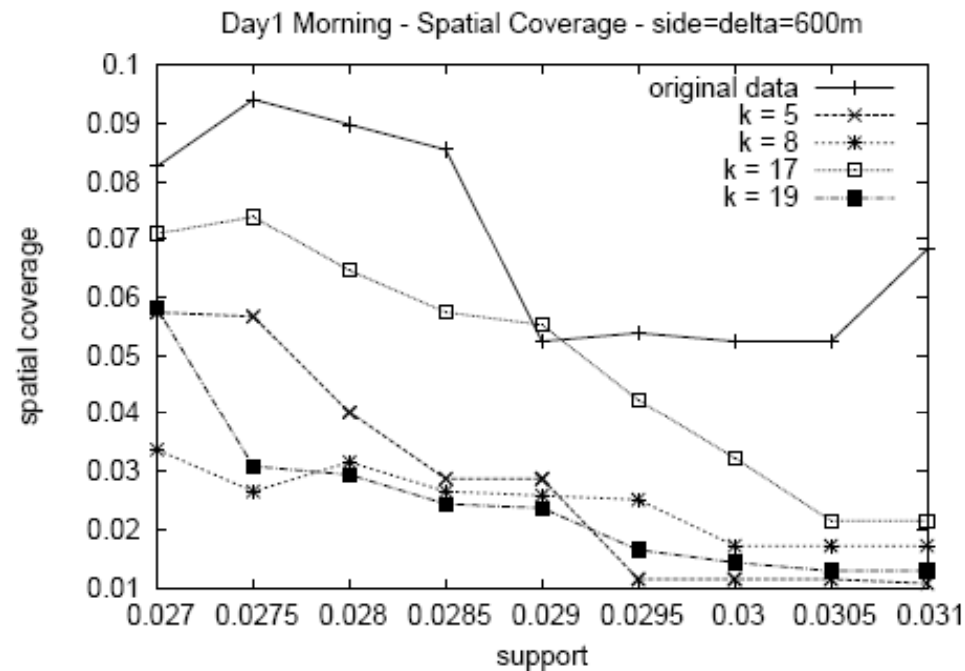
# Dataset Coverage

- For anonymous t-patterns this function is computed w.r.t. the original dataset
- Measures how the anonymous t-patterns cover the original dataset
- No certainty that a trajectory that supports an original t-pattern also supports an anonymous pattern.



# Spatial Coverage

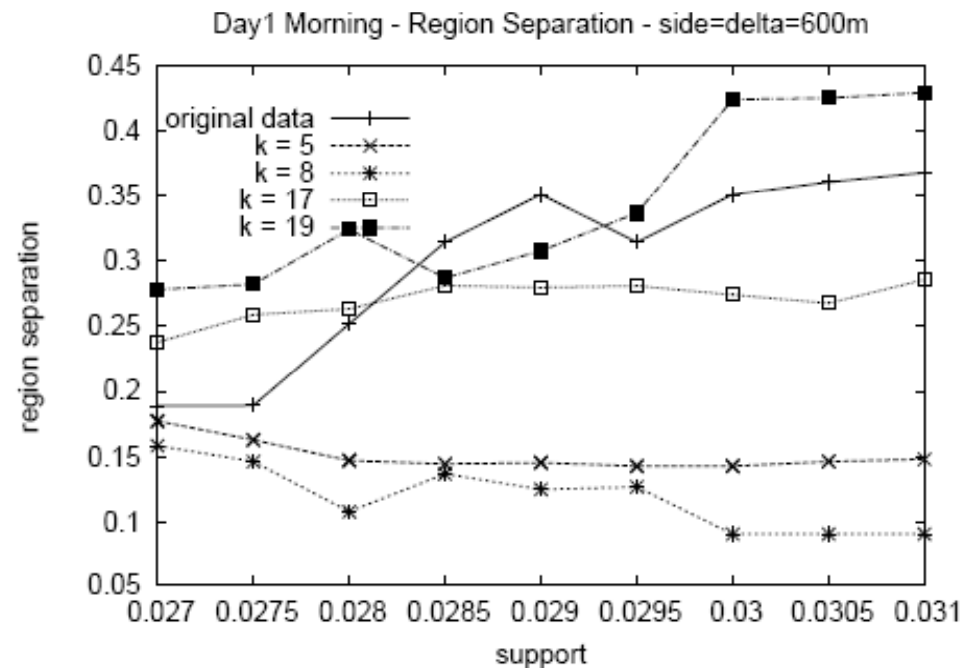
- After the anonymity the value of this function is lower for any value of  $k$
- This effect is due to both the reduction of the number of t-patterns and the clustering and translation of the original trajectories applied to guarantee the  $(k, \delta)$ -anonymity





# Region Separation

- After the anonymization the value of this function can be:
  - **Lower:** NWA approach with the perturbations has generated many near dense regions and this leads to larger regions
  - **Higher:** if the anonymization by the perturbations distributes some trajectories in different groups then, it is possible that the *t*-patterns contain small regions





# Future work

- Ongoing improvements:
  - Try to enforce  $(k, \delta)$ -anonymity specific for maintaining high quality in subsequent data mining analysis
  - Keep in consideration background knowledge (e.g., road network)
    - avoid making people walk through walls
    - avoid making people walk on the water
- Open question – beyond plain  $k$ -anonymity:
  - How to provide **diversity** while providing  $k$ -anonymity?



# Always Walk with Others ! (AWO)

- Nergiz, Atzori, Saygin (Sabanci Univ. + Pisa KDD LAB). ACMGIS 2008



# Trajectory Anonymization (AWO)

- We protect privacy of the individuals by using the following techniques:
  - K-Anonymity: anonymize the dataset so that every trajectory is indistinguishable from  $k - 1$  other trajectories
  - Reconstruction: release atomic trajectories sampled randomly from the area covered by anonymized trajectories



# Trajectory Anonymization (AWO)

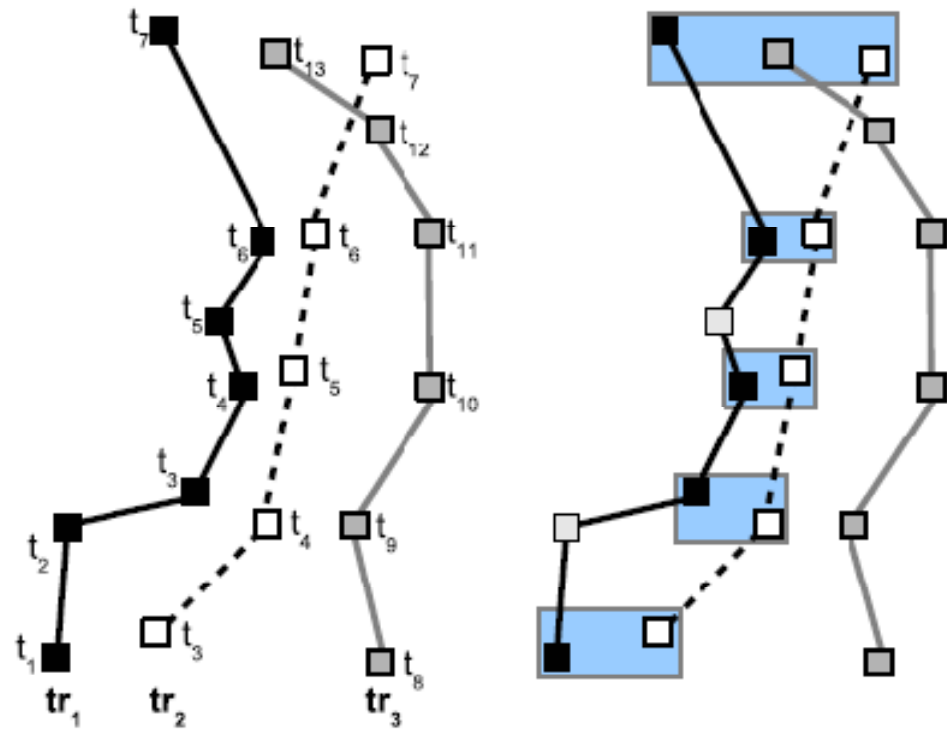
- Clustering for efficient anonymization
  - Define cluster size to be greater than  $k$
- Any clustering algorithm where you can constrain the cluster sizes to be greater than or equal to  $k$
- But the devil is the detail of
  - How to align trajectories
  - Define a cost metric
  - Condense the trajectories
  - Prevent further privacy leaks



# Trajectory Anonymization (AWO)

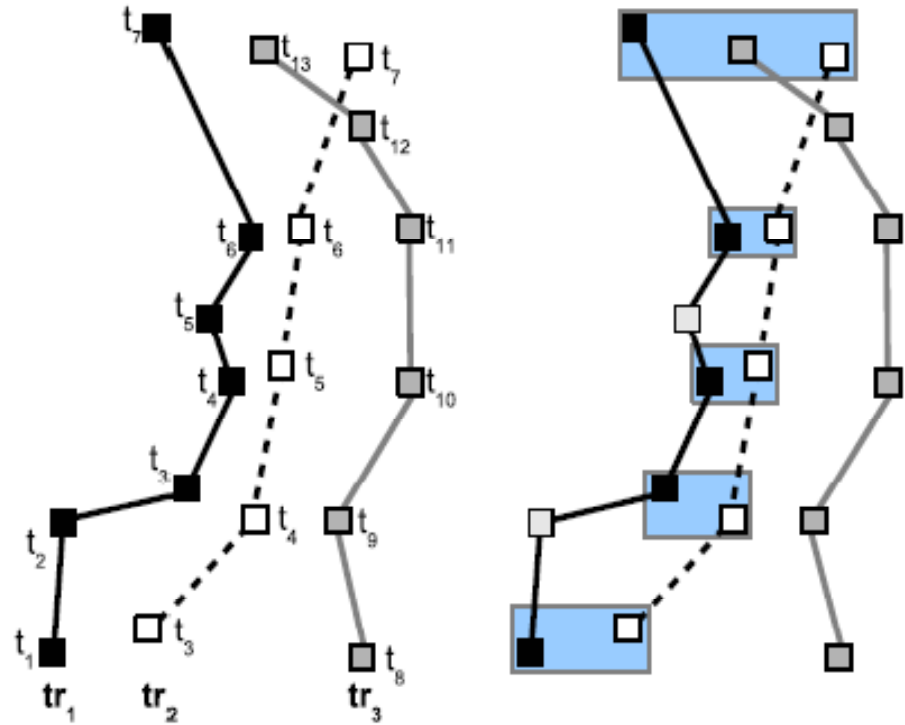
$$LCM(tr^*) = \sum_{p_i \in tr^*} [w_s(\log |x_i| + \log |y_i|) + w_t \log |t_i|] \\ + (|tr| - |tr^*|) \cdot (w_s \log S + w_t \log T)$$

- Trajectories within each cluster need to be condensed into an anonymous trajectory
- Need a cost metric to incorporate space and time



# Trajectory Anonymization (AWO)

- Distance between two trajectories is defined as the cost of their optimal anonymization.
- The problem is reduced to finding the cost-optimal anonymization of two trajectories



# Trajectory Anonymization (AWO)

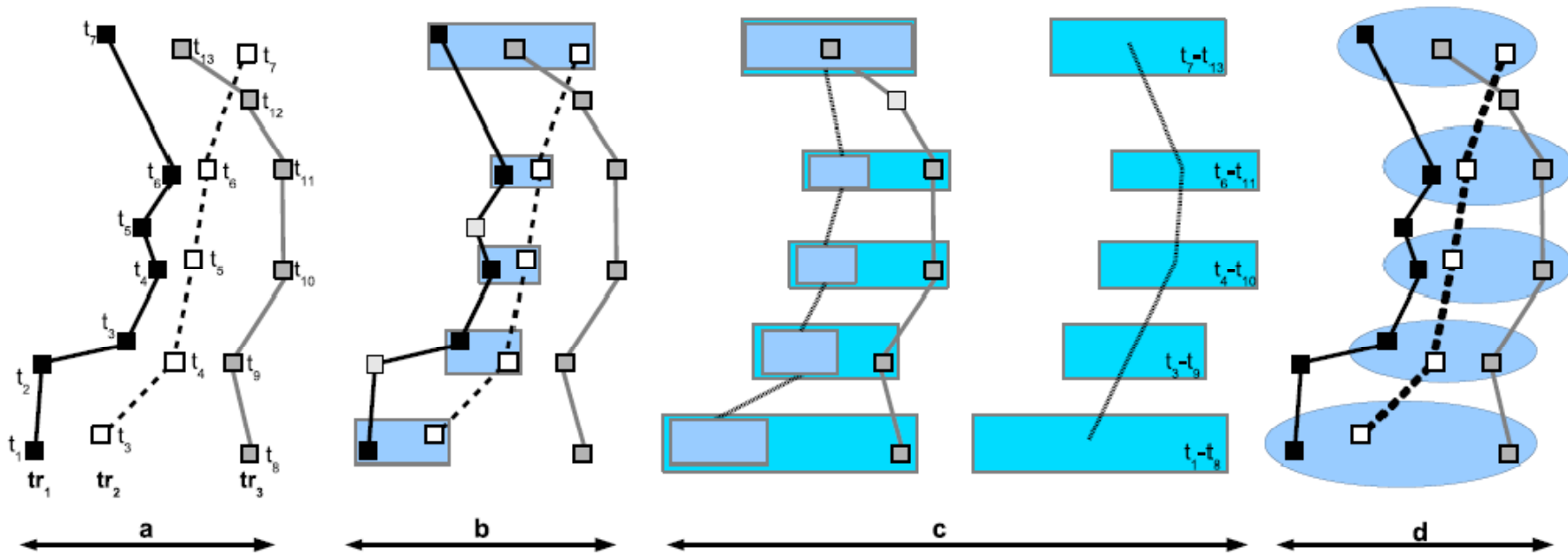


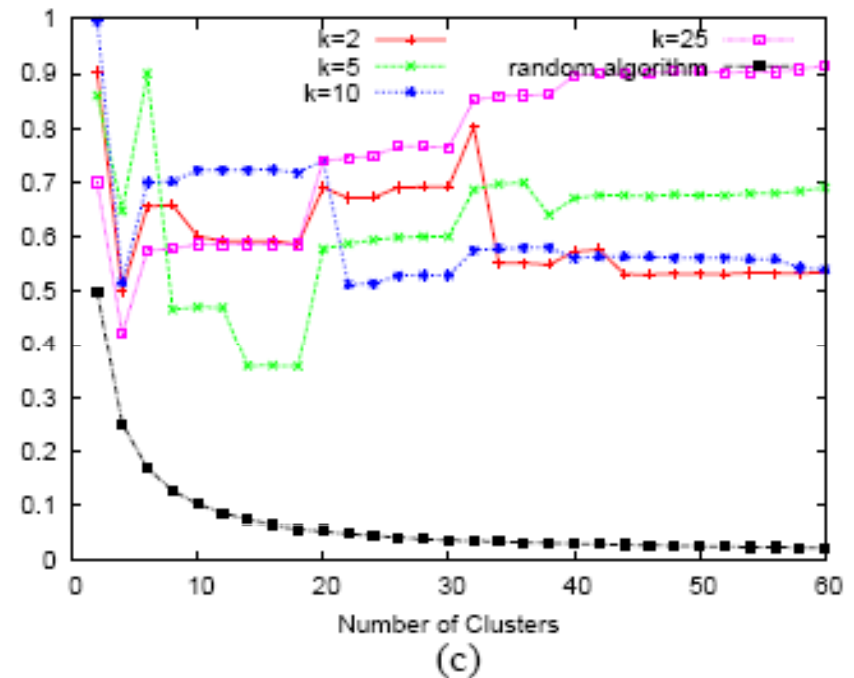
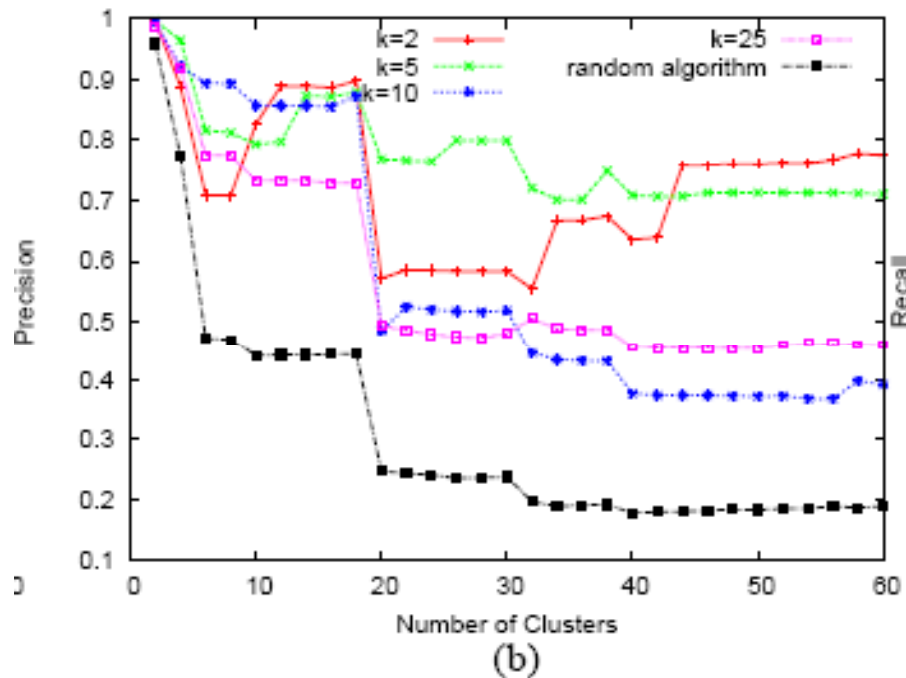
Figure 1. Anonymization Process

a. trajectories  $tr_1, tr_2$ , and  $tr_3$ ; b. anonymization  $tr^*$  of  $tr_1$  and  $tr_2$ ; c. anonymization of  $tr^*$  and  $tr_3$ ; d. point matching used in the anonymization of  $tr_1, tr_2$ , and  $tr_3$ . Matching contains five point links



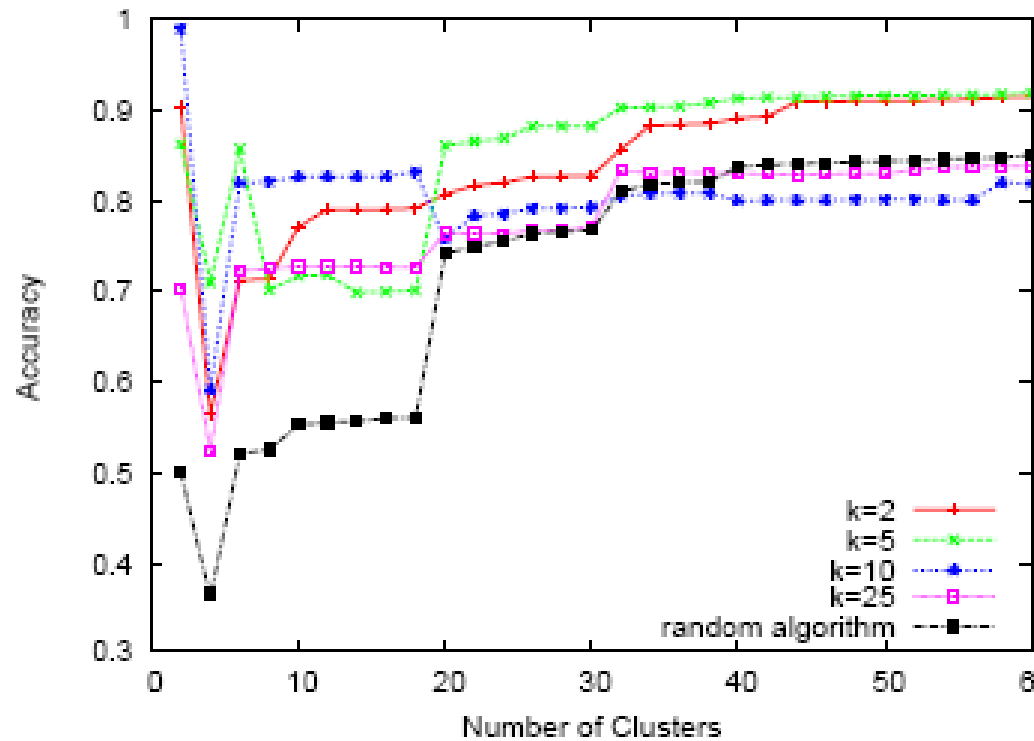
# Distortion evaluation on clustering results

Compare results of a “random algorithm” and AWO with different k values



# Accuracy on clustering results

Compare results of a “random algorithm” and AWO with different k values



# Hiding Sensitive Trajectory Patterns

Fosca Giannotti, ISTI-CNR, Italy.

joint work with **Osman Abul, Maurizio Atzori and Francesco Bonchi**



# Trajectory Pattern Hiding

- In our framework
  - The domain of data ( $D$ ) is spatio-temporal, i.e. spatio-temporal traces of moving entities.
  - The sensitive information is in the form of **spatio-temporal patterns** (see Giannotti *et al.* SIAM DM'06, KDD'07)
  - Trajectories are network based, i.e. background road network is available and publicly known.
  - Road network is represented as a directed labeled graph, and trajectories as temporarily annotated sequences of vertexes.
- Our approach
  - **Hiding sensitive patterns by trajectory coarsening**

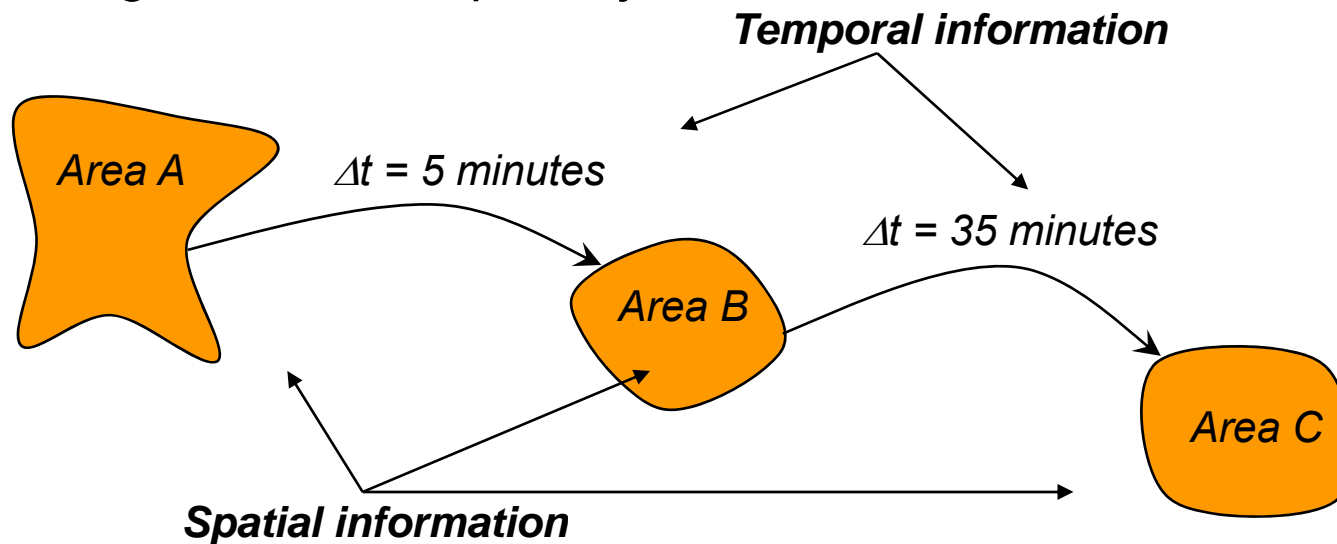


# Pattern hiding –

Please refer to the paper for technical details  
(i.e., definitions of containment, support etc.)

## *Preliminaries: what kind of 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;
- the transition between two consecutive regions in such a sequence is annotated with a **typical travel time** that, again, emerges from the input trajectories.



# Pattern hiding: Problem Statement

**Problem 1 (Trajectory Pattern Hiding Problem)** *It is given a set of sensitive trajectory patterns that must be hidden from  $\mathcal{D}$ :  $\mathcal{P}_h = \{P_1, \dots, P_n\}$ . Given a disclosure threshold  $\psi$ , the Trajectory Pattern Hiding Problem requires to transform  $\mathcal{D}$  in a database  $\mathcal{D}'$  such that:*

1.  $\mathcal{D}'$  is still consistent with  $\mathcal{BN}$ ;
2.  $\forall P_i \in \mathcal{P}_h, \sup_{[\mathcal{D}', \tau]}(P_i) \leq \psi$ ;
3. the difference between  $\mathcal{D}$  and  $\mathcal{D}'$  is minimized.

- How to achieve this?
- Our approach: by trajectory *coarsening*



# Pattern hiding by coarsening

- Coarsening a trajectory means to just suppress some points of the trajectory.
- Two nice properties:
  - Coarsening a trajectory does not make it violate time/space consistency with the background road network.
  - If a coarsened trajectory supports a pattern, so it does the original trajectory (directly consequence of the anti-monotonicity of frequency).



# Pattern hiding by coarsening: The Problem

## Problem 2 (Hiding by Coarsening)

Given a trajectory  $T$ , a set of sensitive trajectory patterns  $\mathcal{P}_h = \{P_1, \dots, P_n\}$  and time tolerance  $\tau$ . The Pattern Hiding Problem by Coarsening requires to transform  $T$  in a trajectory  $T'$  using only coarsening operations, and such that:

1.  $\forall P_i \in \mathcal{P}_h. T \not\subseteq_{\tau} P_i$ ;
2. number of coarsening operations is minimized.

- The problem is **NP-Hard**
- We must devise heuristics







# Hiding Sequential Patterns

- Key notion is the *matching set* (number of different ways a trajectory supports sensitive patterns)
- We define **matching set** as a notion of identifying all instances of sensitive patterns in a sequence
- Es:  $S = \langle c, a, b, \rangle$ ,  $T = \langle a, a, b, c, c, b, a, \rangle$
- $M(ts) = \{(1, 2, 3), (1, 3, 5), (2, 3, 4), (2, 3, 5)\}$



# Pattern hiding by coarsening- *The Sanitization Algorithm*

- Our proposal exploits two heuristics for points removal from trajectories.
- Local heuristic:
  - Which points to be removed (suppressed) from a selected trajectory:  
*suppress the point that is involved in the largest number of matches:*
  - *This operation is iterated till the matching set of T is empty*
- Global heuristic:
  - The concern is which trajectories to be selected for sanitization:  
*sanitize the trajectories that have a larger matching set size first*
  - *The trajectories are sorted in a descending order of matching set size*



# Attack by background network

- The problem

- Suppose the point  $(v_i, t_i)$  is suppressed from the trajectory segment  $(v_{i-1}, t_{i-1}), (v_i, t_i), (v_{i+1}, t_{i+1})$
- Also assume that there is only one path in the background road network from  $(v_{i-1}, t_{i-1})$  to  $(v_{i+1}, t_{i+1})$  which is of course passing through  $(v_i, t_i)$
- So, given the publicly available background road network, attacker can reconstruct the segment
  - pseudo-hiding case



# Attack by background network

- We secure our algorithm to such kind of attacks by requiring every consecutive points in published trajectories *k-secure*, borrowing the ideas from *k-anonymity*

**Definition 8 (*k-secure path*)** Given two spatio-temporal points  $p_1 = (x_1, y_1, t_1)$  and  $p_2 = (x_2, y_2, t_2)$  over  $\mathcal{BN}$ , a path from  $p_1$  to  $p_2$  time consistent with  $\mathcal{BN}$ , is said to be *k-secure* if there exist at least other  $k - 1$  paths from  $p_1$  to  $p_2$  time consistent with  $\mathcal{BN}$ .



# Attack by background network

- The new algorithm under background attack assumption
  - For every trajectory  $T'$  in  $D'$  (the sanitized data), *k-secure path property* of every consecutive point pair is checked (the efficient  *$k^{\text{th}}$  shortest path* algorithm from the literature is used to do so)
  - A point violating the *k-secure path property* is suppressed and the process is repeated until  $T'$  is *k-secure*
- Note that the distortion with attack by background network is always at least as the original sanitization algorithm



# Experimental evaluation

- Dataset

- The dataset is generated with *Brinkhoff trajectory generator* [Brinkhoff 2003] on the road network of city of Oldenburg, Germany
  - 4000 trajectories with varying lengths

- Sensitive patterns

- 4 patterns, randomly chosen from all frequent spatio-temporal patterns mined with *TAS miner* [Giannotti *et al.* 2006] with parameters, support at %5 and time-tolerance 15



# Experimental evaluation

- Performance metrics, **M1** and **M2**

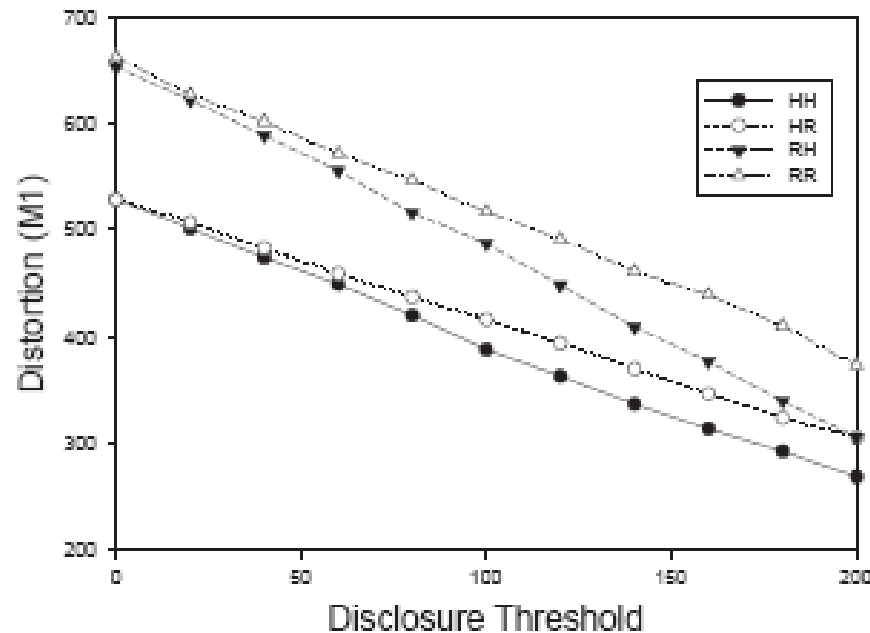
- M1 (Data distortion): total number of points removed in  $\mathcal{D}'$
- M2 (Frequent Pattern Distortion):

$$\frac{|\mathcal{F}(\mathcal{D}, \sigma)| - |\mathcal{F}(\mathcal{D}', \sigma)|}{|\mathcal{F}(\mathcal{D}, \sigma)|}$$



# Experimental evaluation

- Exploring effect of global/local heuristics on M1



*Legend:*

**HH:** Both local/global heuristic

**HR:** only local heuristic

**RH:** only global heuristic

**RR:** random suppression

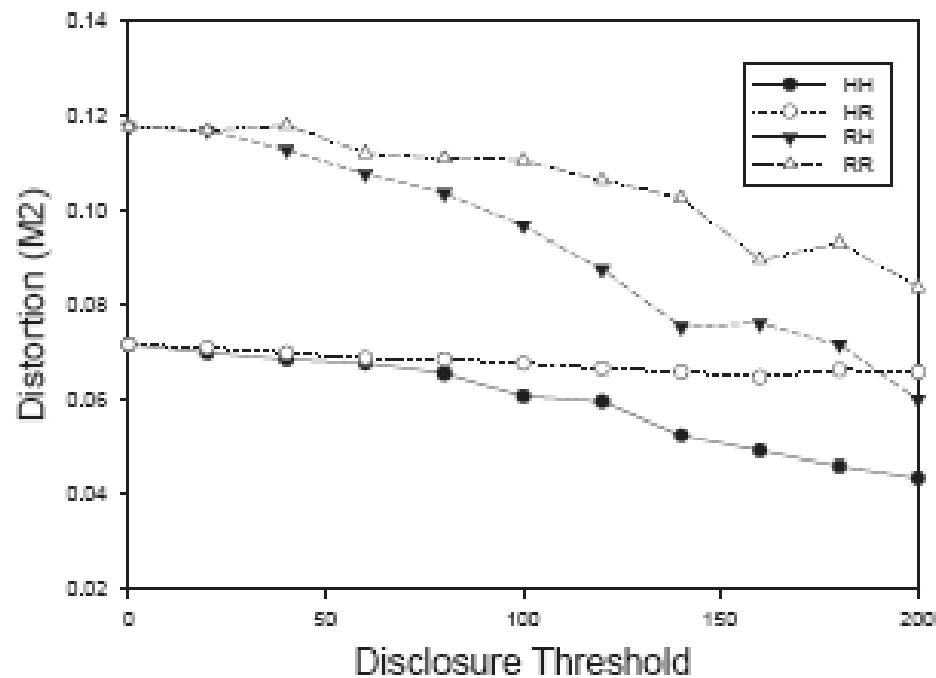
*The heuristics are effective, i.e. causes less distortion at all disclosure thresholds*





# Experimental evaluation

- Exploring effect of global/local heuristics on M2

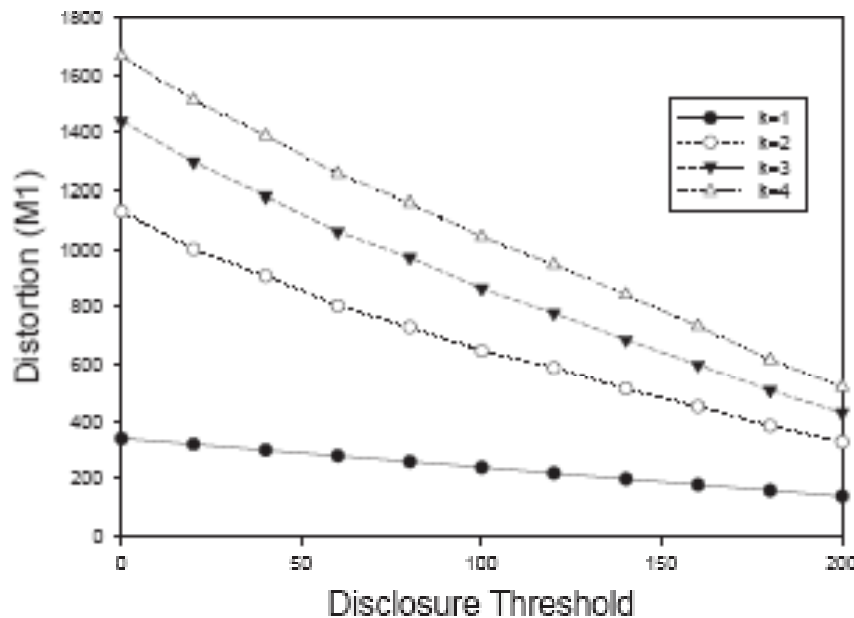


*The heuristics are effective, i.e. causes less distortion at all disclosure thresholds*



# Experimental evaluation

- Exploring effect of anonymity parameter  $k$ , under attack by background network, on distortion



*Higher the  $k$ , higher the distortion as expected,*

*The distortion gaps at different  $k$  levels indicate the potential danger for the attack by background network*



# Conclusion

- Novel problem of hiding sensitive trajectory patterns from spatio-temporal data is studied
- The optimization problem is proven to be NP-Hard
- A polynomial sanitization algorithm is developed
- The algorithm is further enhanced to be secure against attack by background knowledge (road network)
- The effectiveness of the heuristics is assessed experimentally



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



# Location-Based 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

# Service-Privacy Trade-off

- **First extreme:**

- A user reports her exact location → 100% service

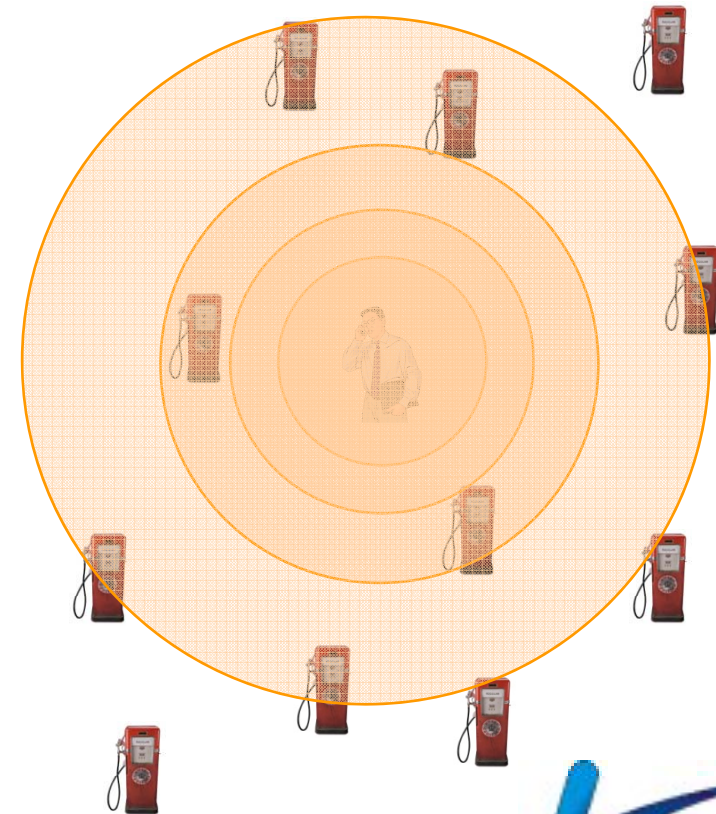
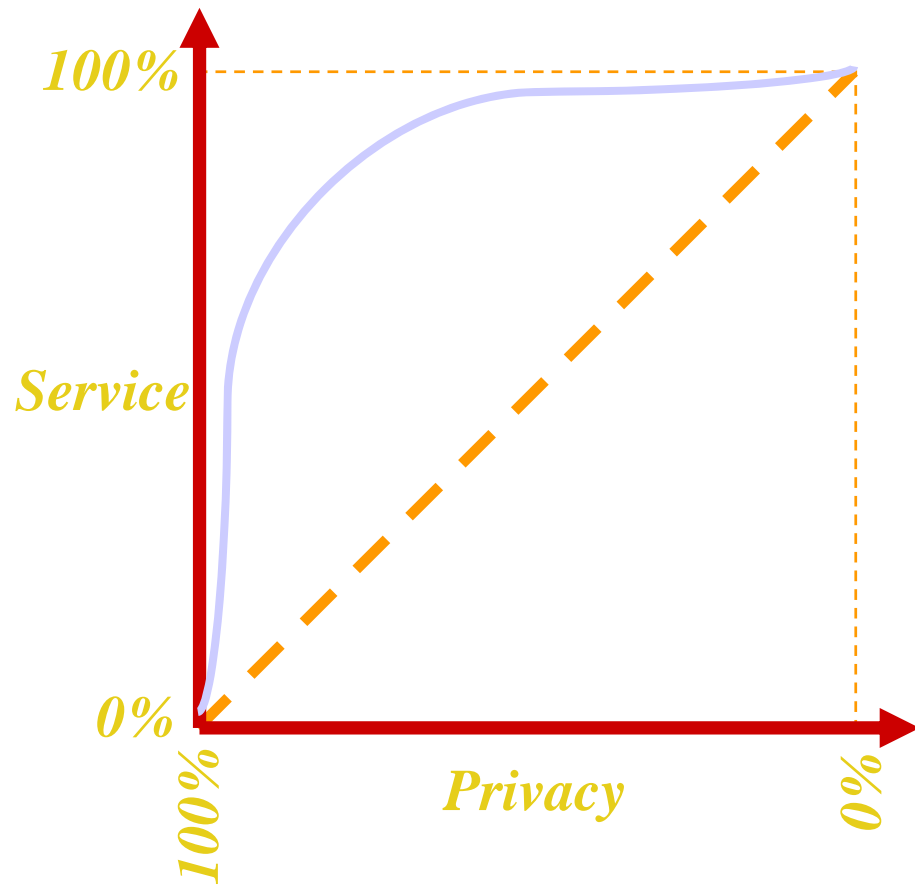
- **Second extreme:**

- 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

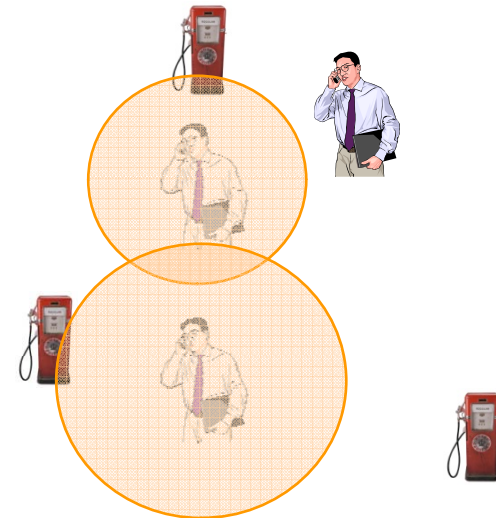
- Example: *What is my nearest gas station*



# 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

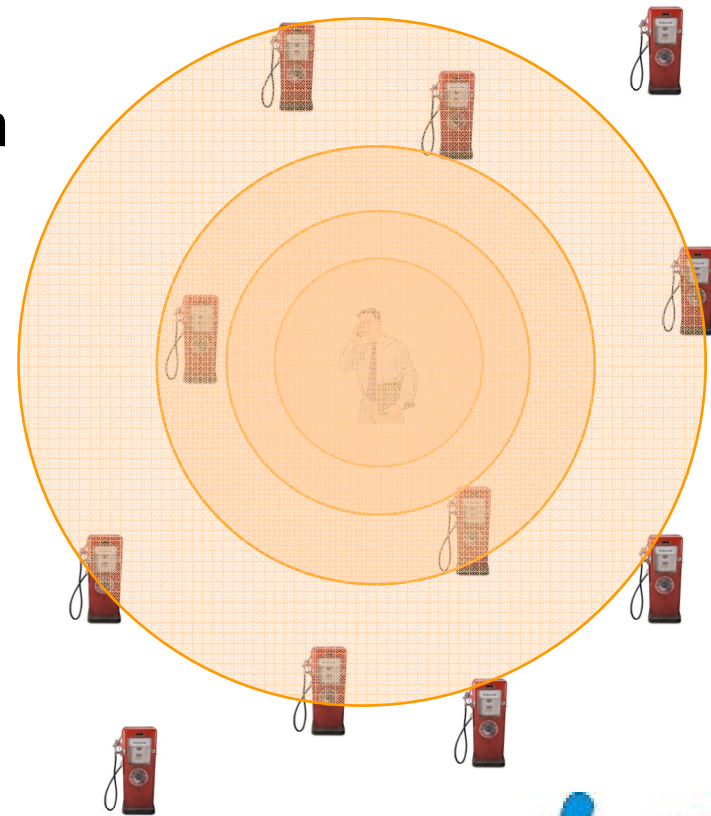




# Concepts for Location Privacy

## Spatial Cloaking

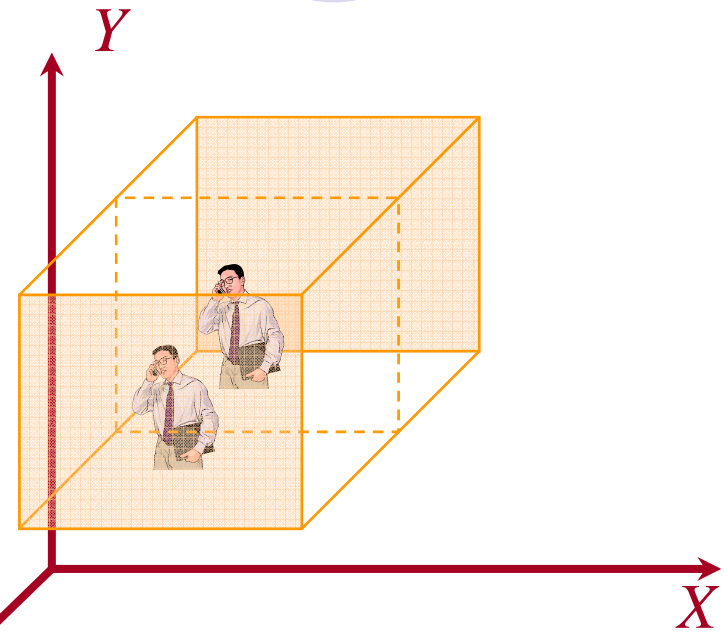
- Location *cloaking*, location *blurring*, location *obfuscation*
- The user exact location is represented as a region that includes the exact user location
- An adversary does know that the user is located in the *cloaked* region, but has no clue where the user is exactly located
- The area of the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



# Concepts for Location Privacy

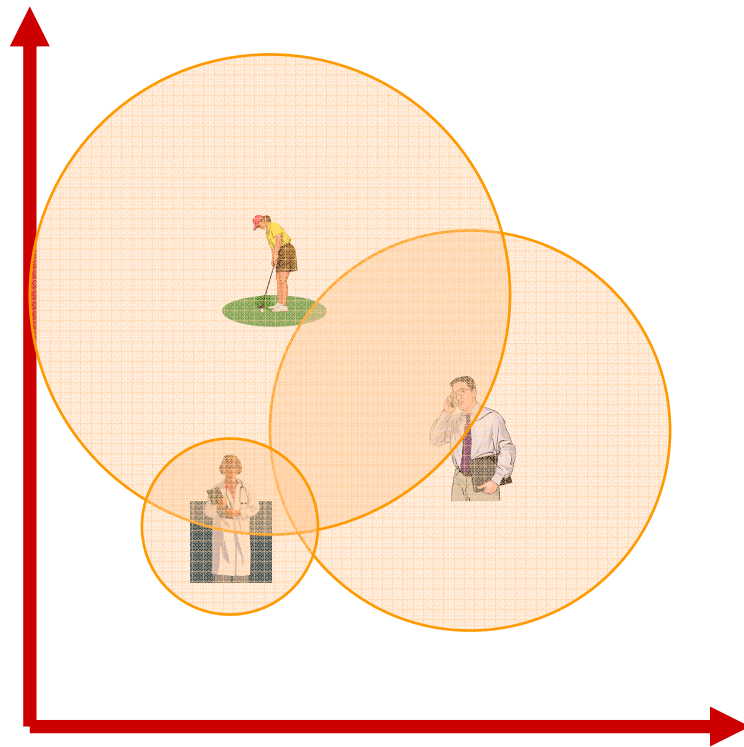
## Spatio-temporal Cloaking

- 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

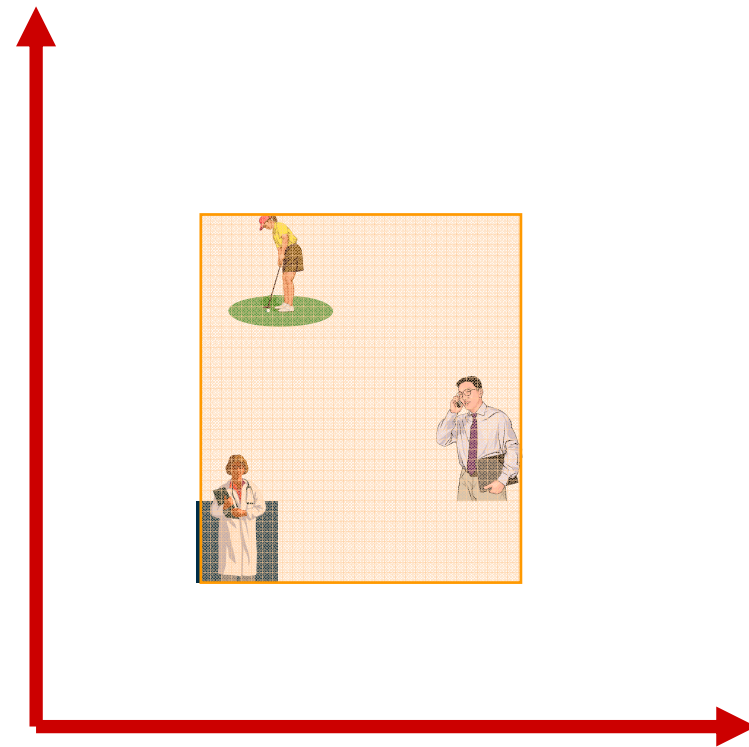


# Concepts for Location Privacy

## Data-Dependent Cloaking



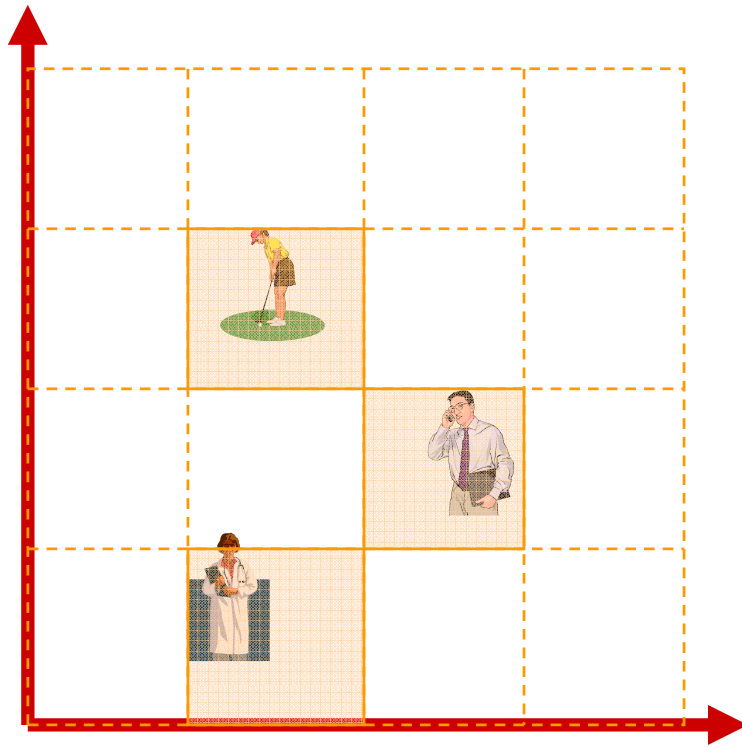
*Naïve cloaking*



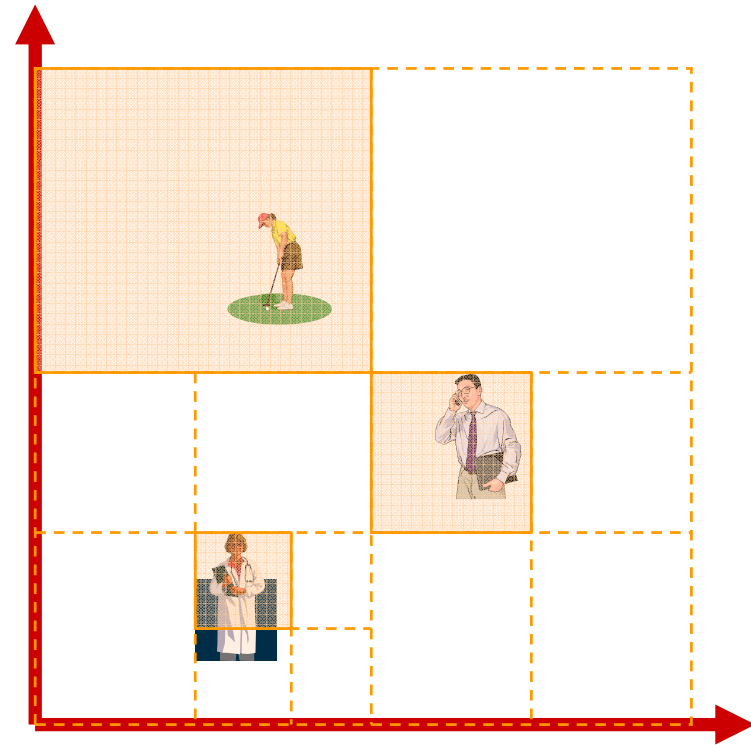
*MBR cloaking*

# 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

<i>Time</i>	<i>k</i>	$A_{min}$	$A_{max}$
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

## Service Providers (SS)

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

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



1



2



# Location-Based Services and Privacy Issues

## Service Providers (SS)

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

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



1



2



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



# Location-Based Services and Privacy Issues

## Service Providers (SS)

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

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



1



2



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

## 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

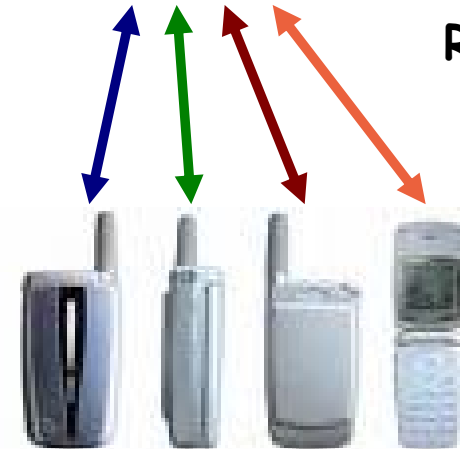
## Service Providers (SS)



## Trusted Server (TS)

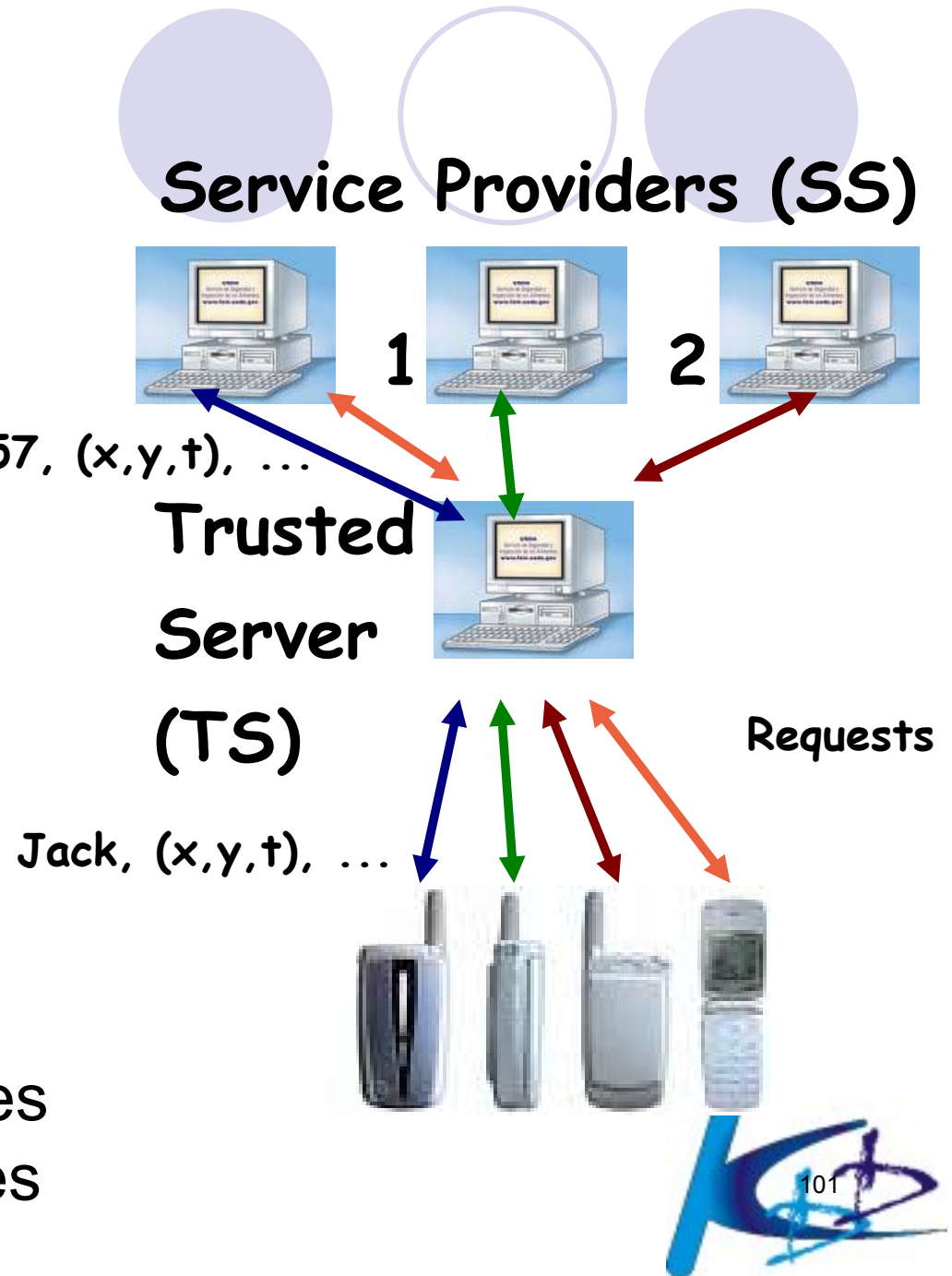


## Requests



# 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]*

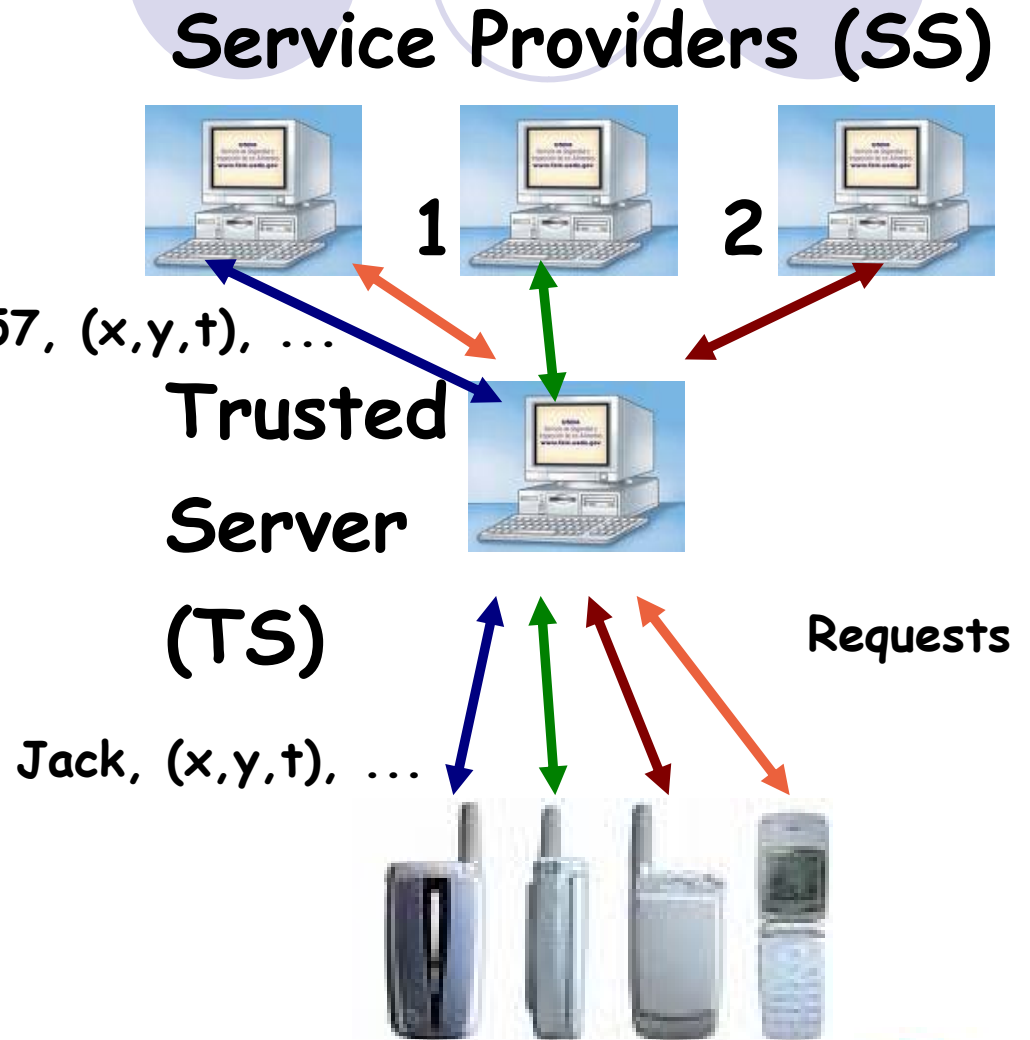
# Personalized Anonymization for Location Privacy

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

- Trusted Server between user and LBS

- Privacy:

- TS masks Names

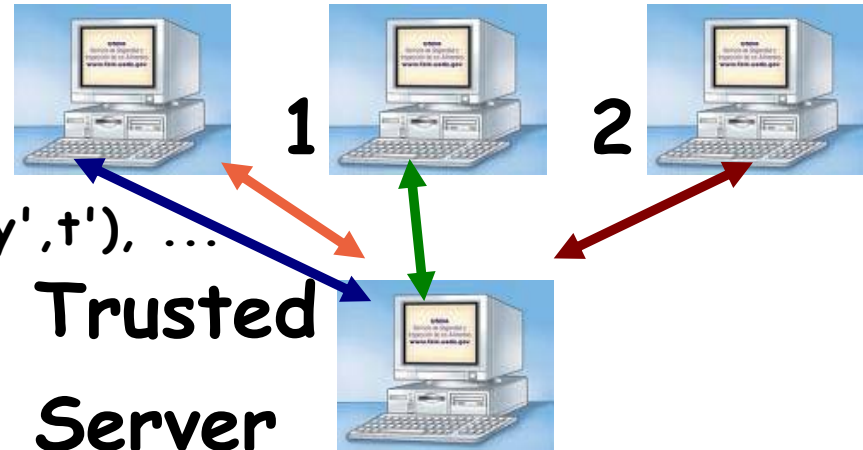


# Personalized Anonymization for Location Privacy

## Service Providers (SS)

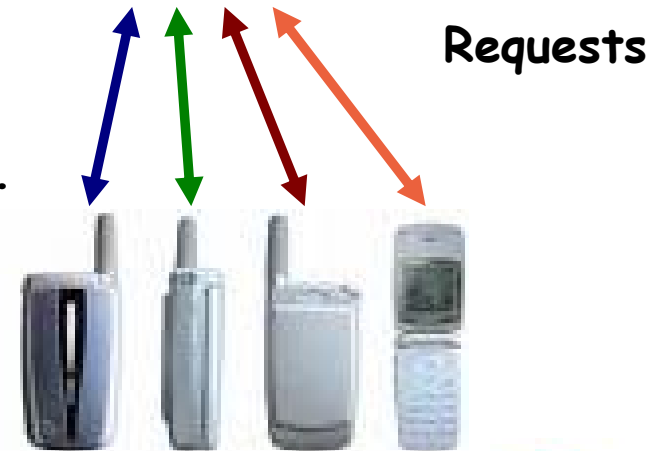
- Context: communication for Location-based services (LBS)
  - Trusted Anonymization Server between user and LBS
- Privacy:
  - TS masks Names
  - Space-time coordinates are distorted (cloaking)

ID57, (x', y', t'), ...



## Trusted Server (TS)

Jack, (x, y, t), ...





# 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:  $|\text{Anonymity Set}| \geq k$
- 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

- $ms = \langle uid, rno, \{t,x,y\}, k, \{dt, dx, dy\}, C \rangle$
- 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.  $mt \in T'$ ,  $|T'| \geq ms.k$ ,

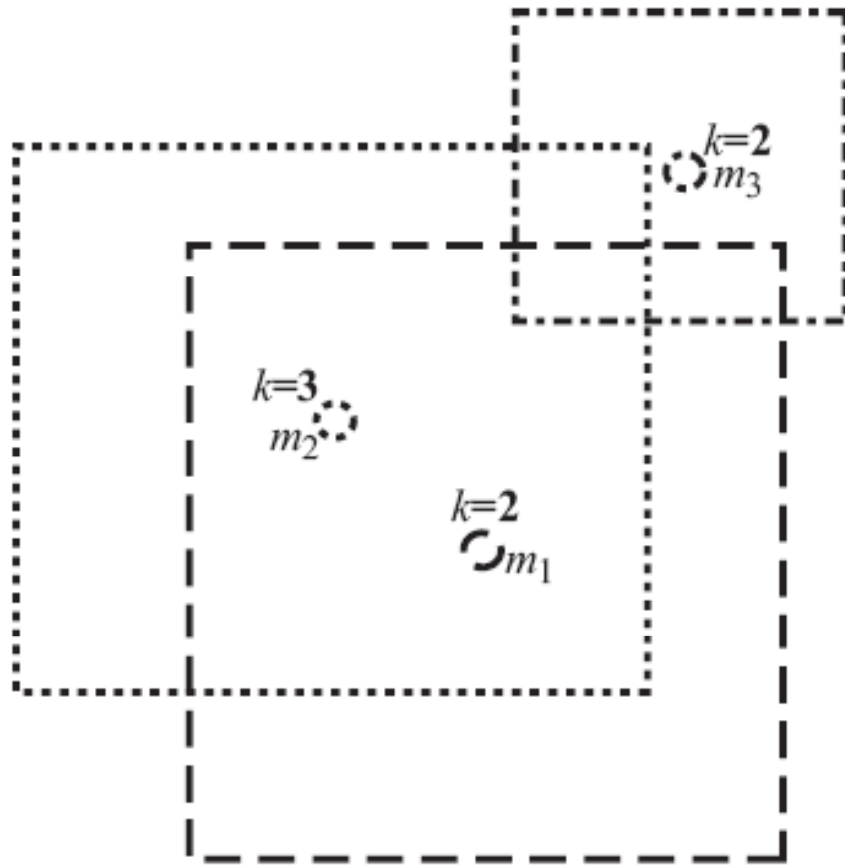
$\forall \{mt_i, mt_j\} \subset T'$ ,  $mt_i.uid \neq mt_j.uid$  and

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

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

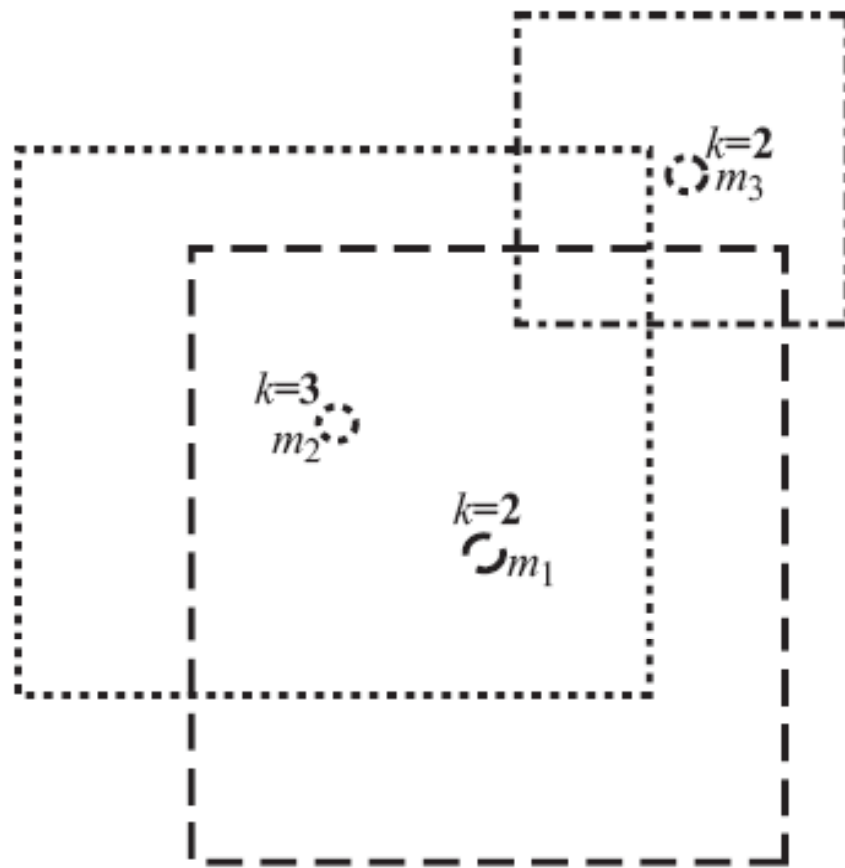


# Clique-Cloak Algorithm: Spatial Layouts

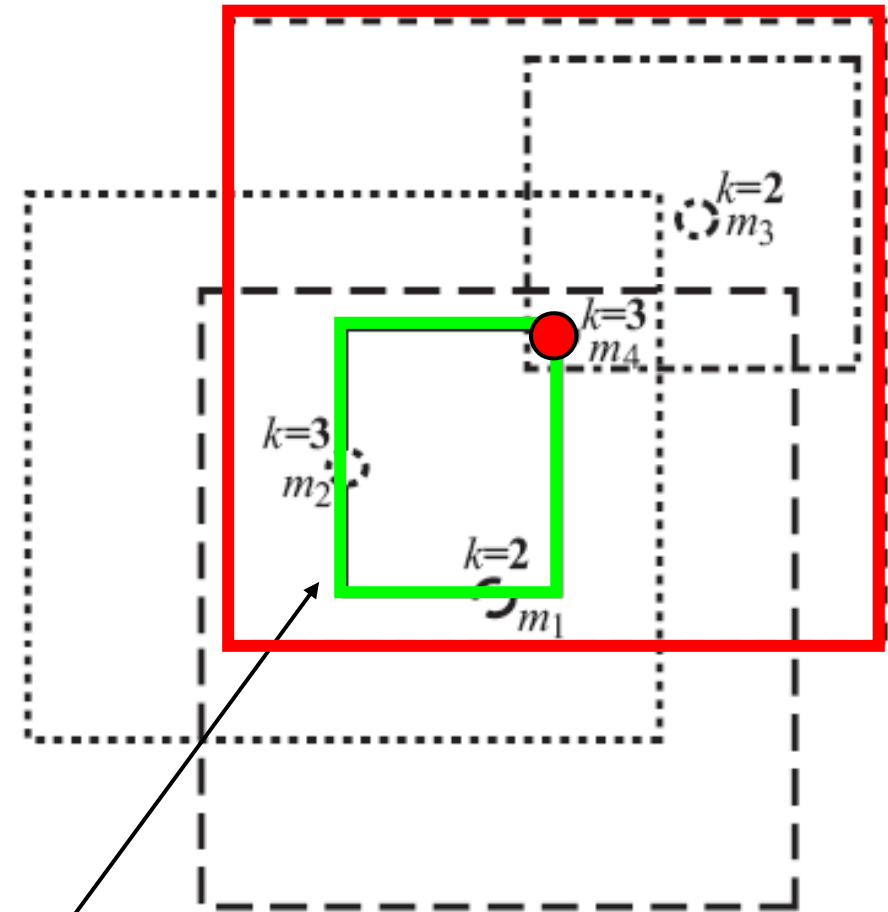


(a) spatial layout I

# Clique-Cloak Algorithm: Spatial Layouts



(a) spatial layout I

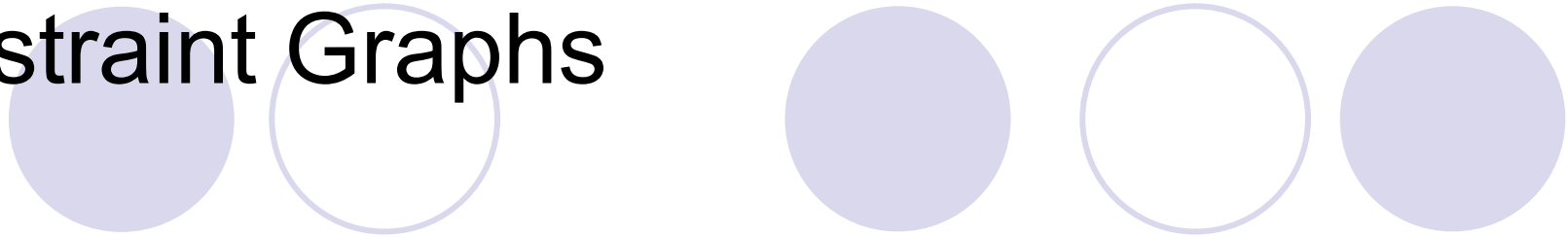


(b) spatial layout II

**minimum bounding rectangle**



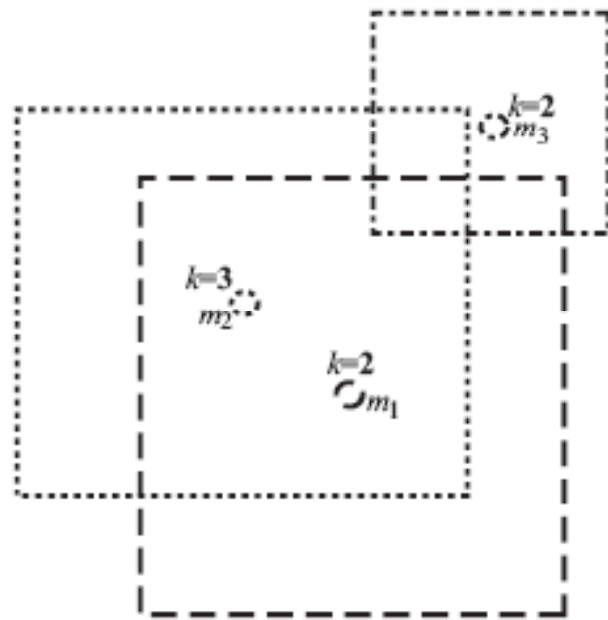
# Constraint Graphs



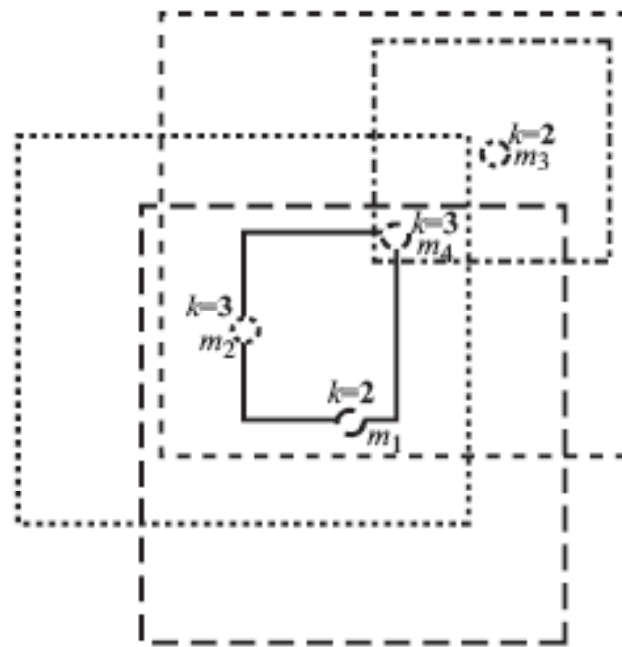
- $G(S,E)$  is an undirected graph
- $S$  is the set of vertices
  - Each representing a message received at the message perturbation engine
- $E$  is the set of edges,  $(ms_i, ms_j) \in E$  iff
  1.  $L(ms_i) \in Bcn(ms_j)$
  2.  $L(ms_j) \in Bcn(ms_i)$
  3.  $ms_i.uid \neq ms_j.uid$
- $ms_i$  is anonymizable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



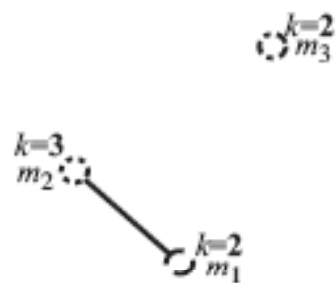
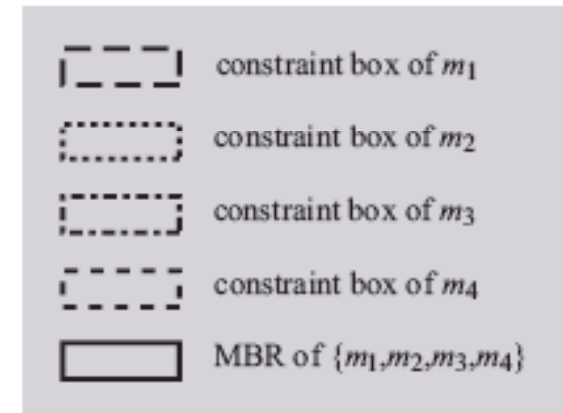
# Clique-Cloak Algorithm: Constraint Graphs



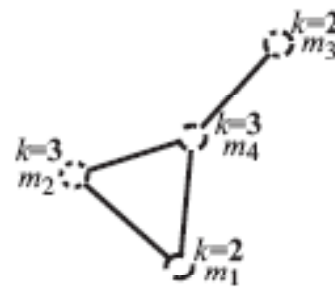
(a) spatial layout I



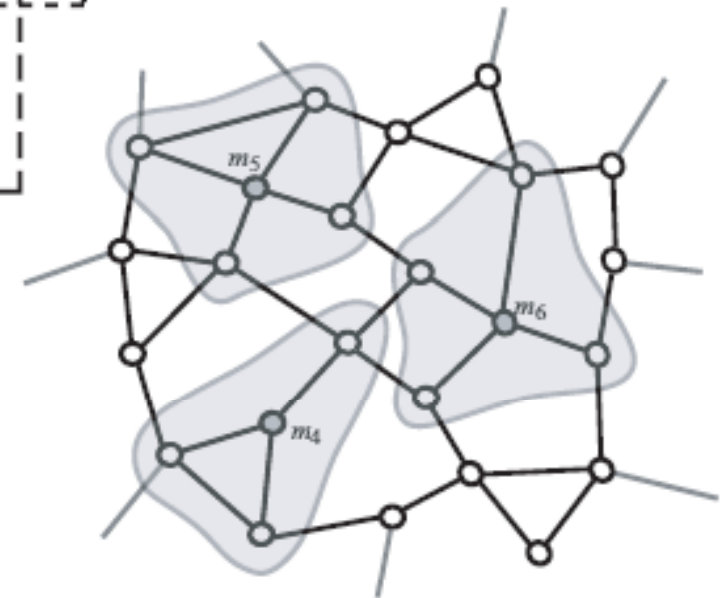
(b) spatial layout II



(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  $\{m\} \cup \{m_j \in \text{neighbor of } m \mid m_j.k \leq m.k\}$**
  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



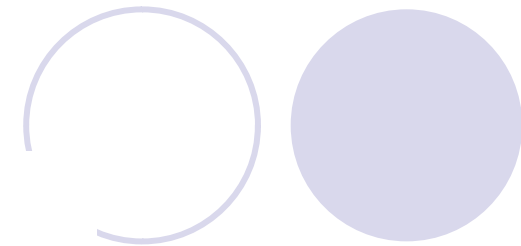
# Synthetic Data Generator

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 <sup>2</sup>
mean temporal tolerance	30s
variance in temporal tolerance	12s <sup>2</sup>
mean inter-wait time	15s
variance in inter-wait time	6s <sup>2</sup>

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<sup>2</sup>)

10,000 cars



# Experiments: Success rate and anonymity level

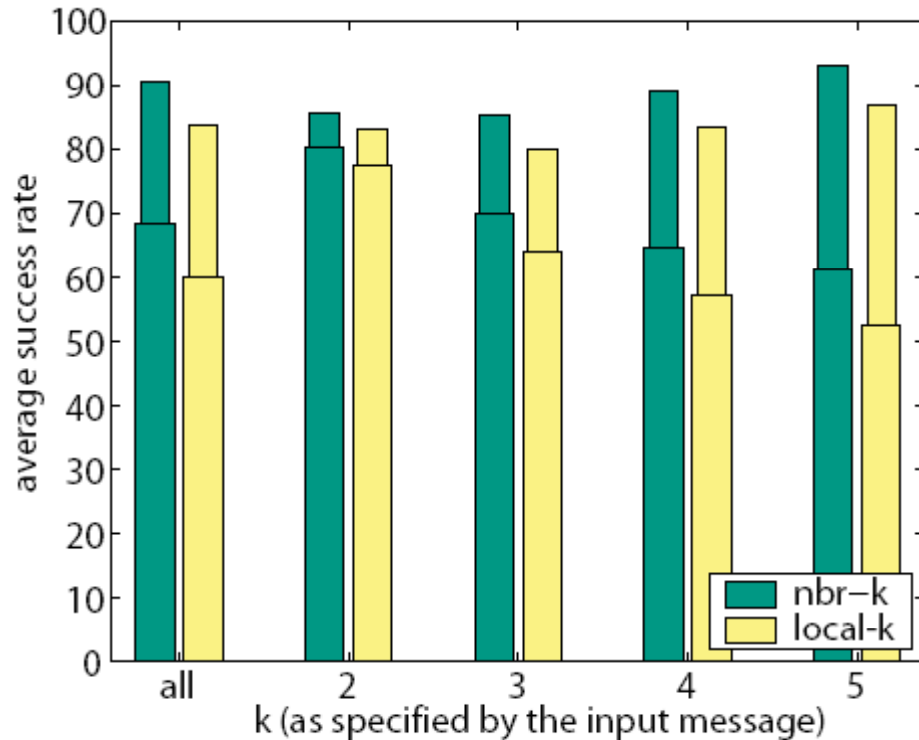


Figure 2: Success rates for different  $k$  values

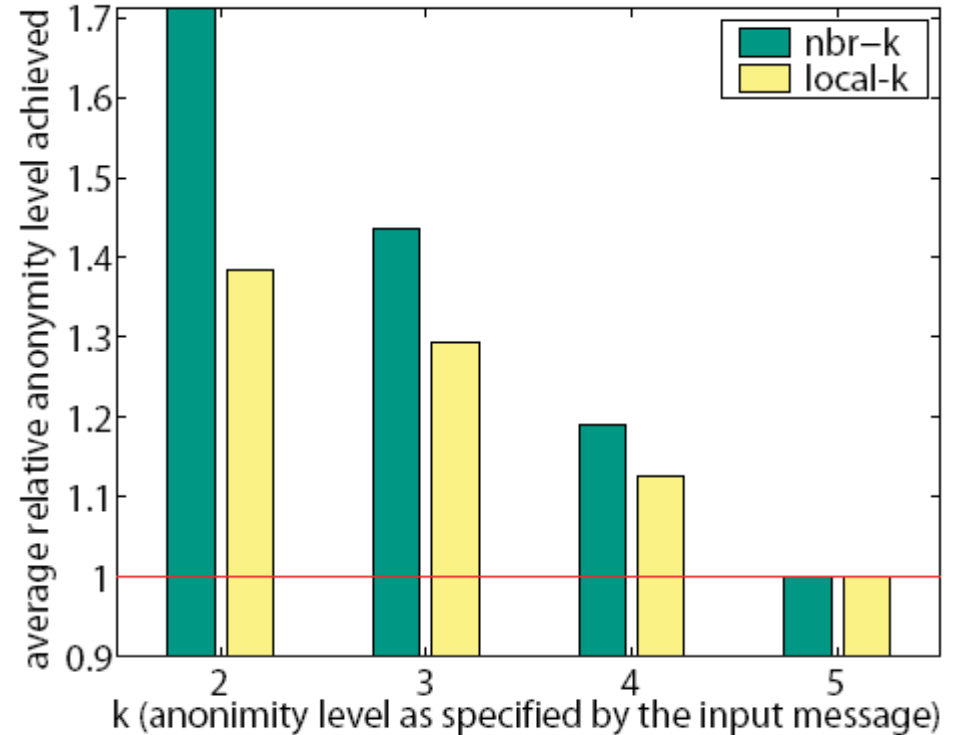


Figure 3: Relative anonymity levels for different  $k$  values

**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 framework 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

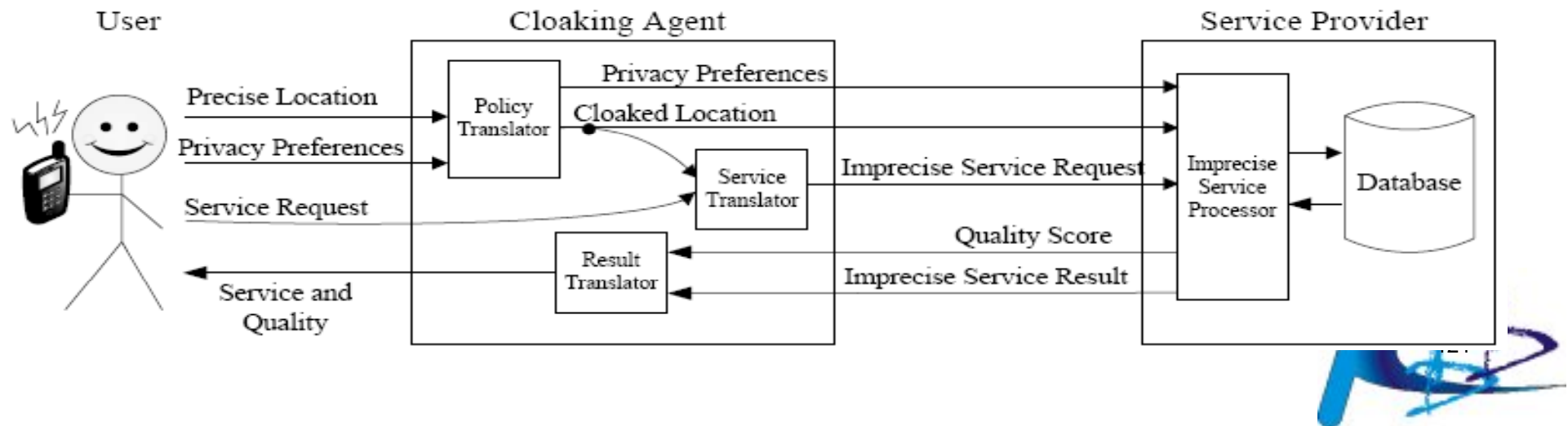
- Random perturbation of client's location
  - Chosen by client
  - Variable, and not known to server
- Large enough to “hide” exact location (privacy)
- Small enough to avoid “too much damage” to quality of answer
- Issue: Quantifying the damage to answer
- Requests are ST regions





## 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)





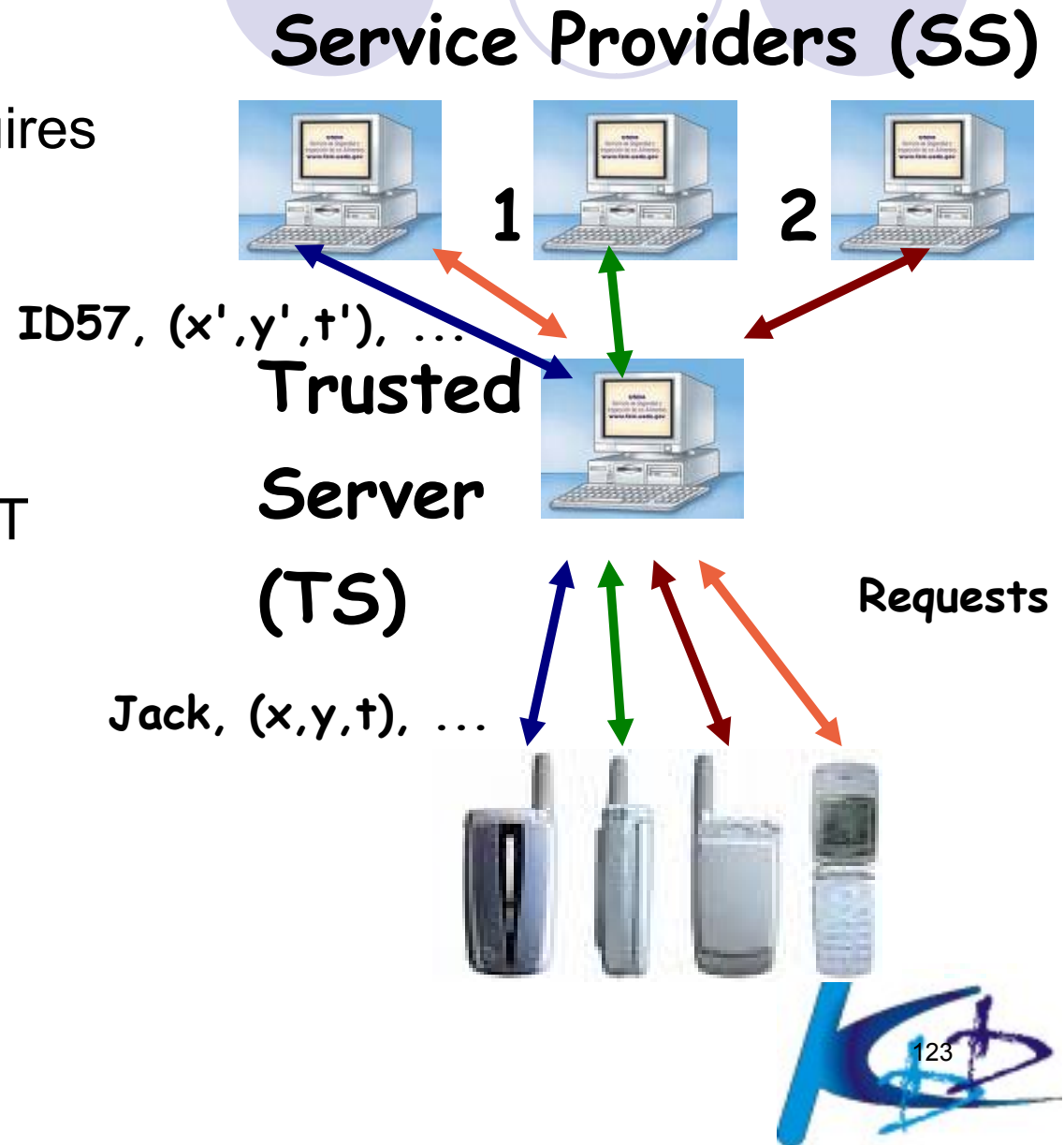
# Real-time Anonymity of trajectory-based services

**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.  $\langle x_1, y_1, t_1 \rangle$  ,  $\langle x_2, y_2, t_2 \rangle$  , ...  $\langle x_n, y_n, t_n \rangle$
- Historical  $k$ -Anonymity (HkA)
  - A set of requests issued by the same user satisfies HkA if there exist  $k-1$  PHLs  $P_1, \dots, P_{(k-1)}$  for  $k-1$  different users s.t. The set of requests “match”  $P_1 \dots P_{(k-1)}$ 
    - 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]**

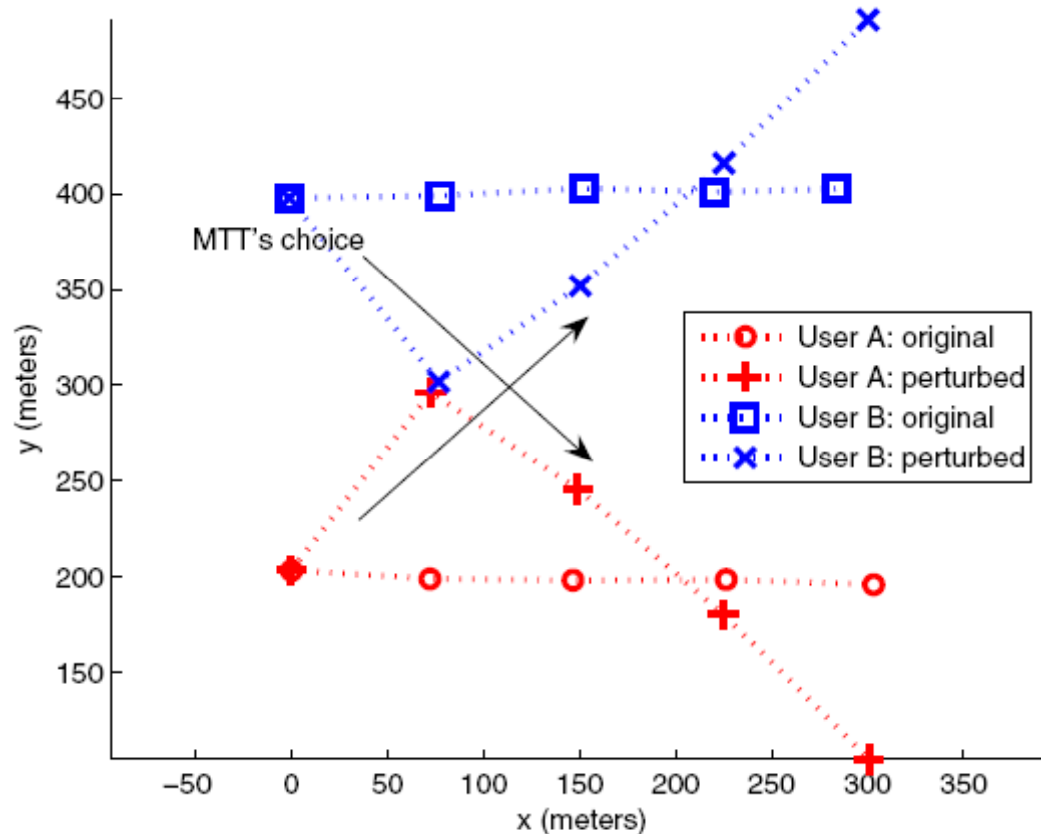
**Anonymous Communication Technique using Dummies for Location-Based  
Services**

**[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***



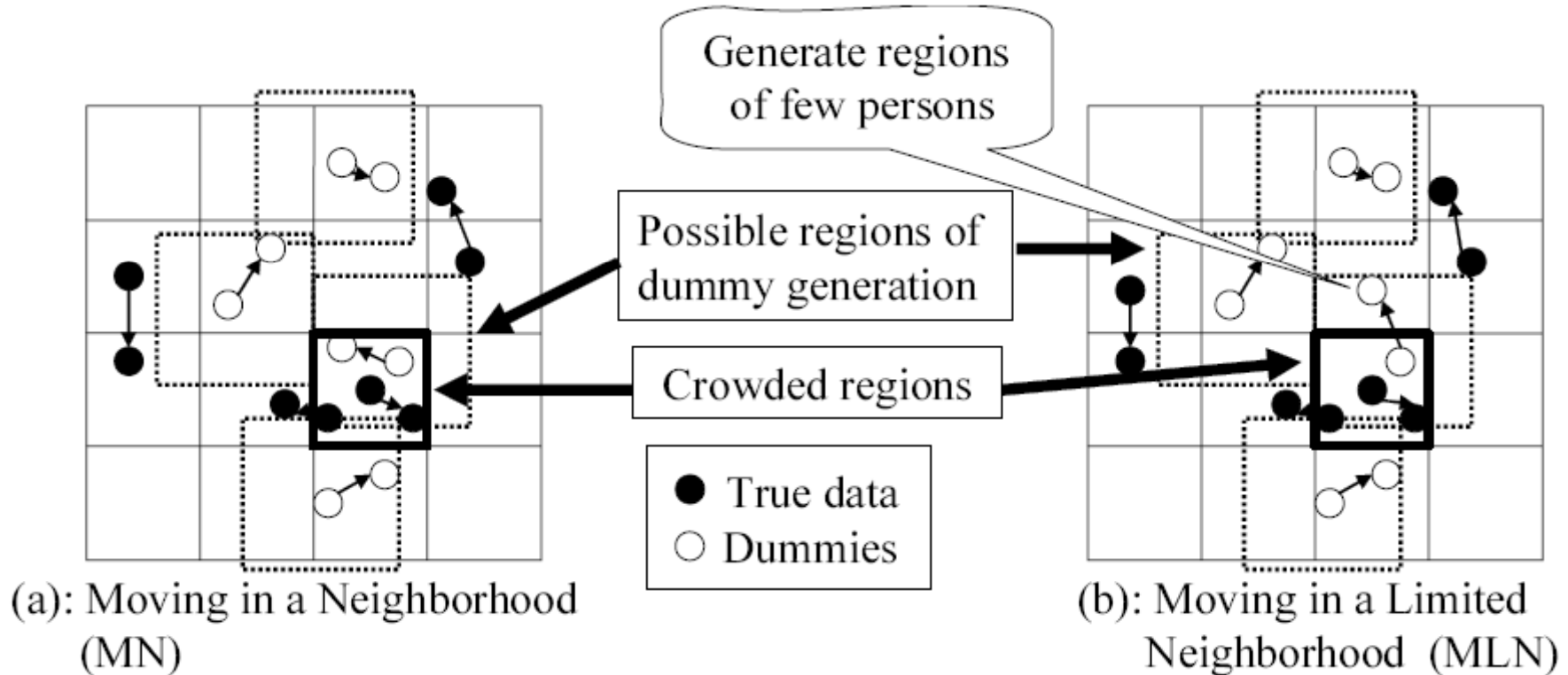
# Introducing Dummy Data

- 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](http://www.digitalcity.gr.jp/openlab/kyoto/map_guide.html)*



# MN and MLN algorithms

## Moving in a (Limited) Neighborhood



- 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, VLDB06]*



# 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 :-)

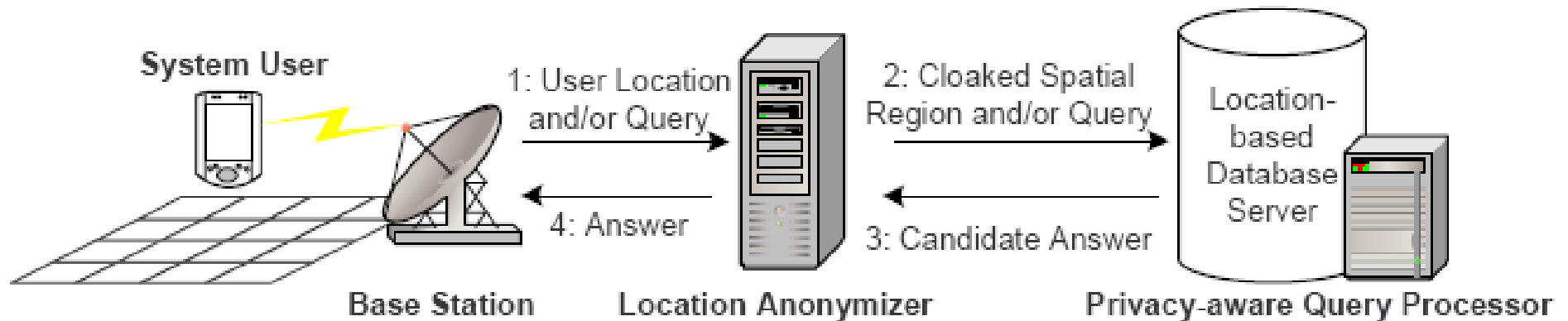


# Kind of Queries

- 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

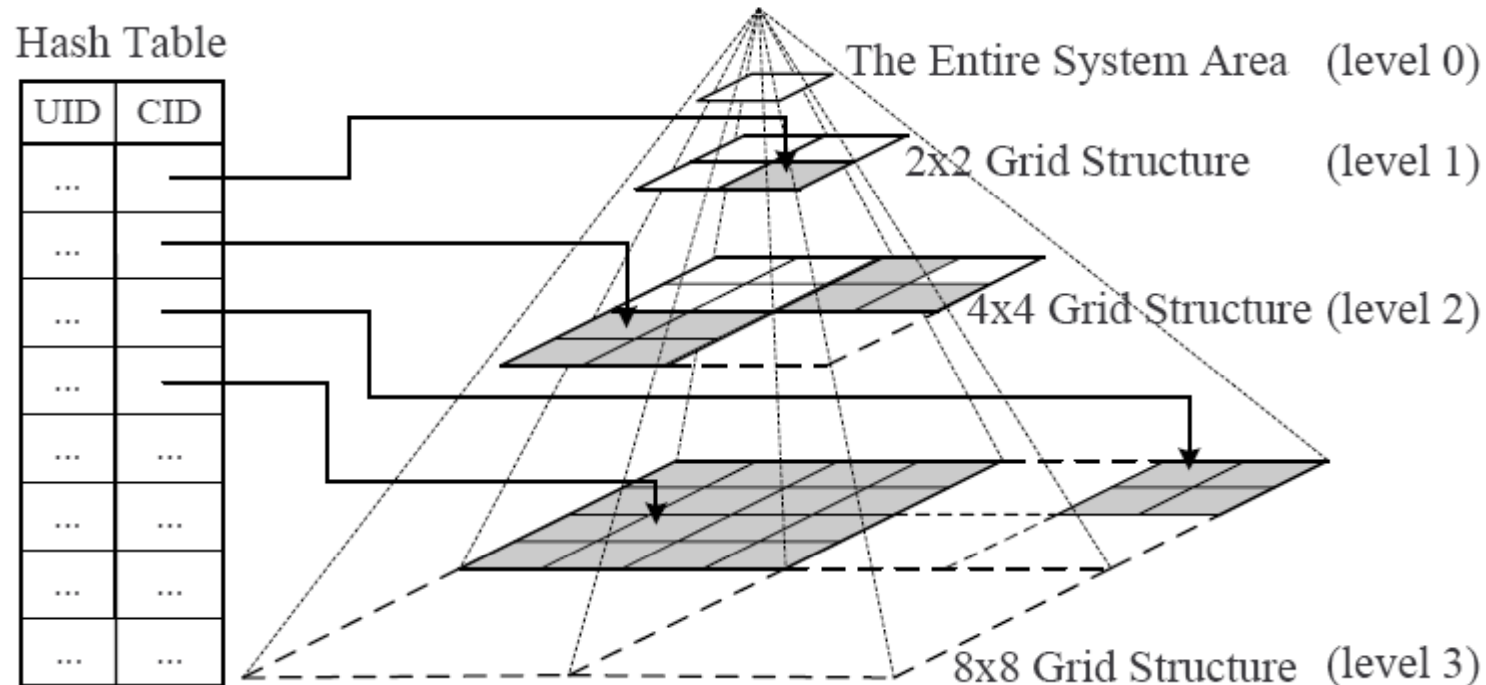


# Architecture



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

# Adaptive Location Anonymizer



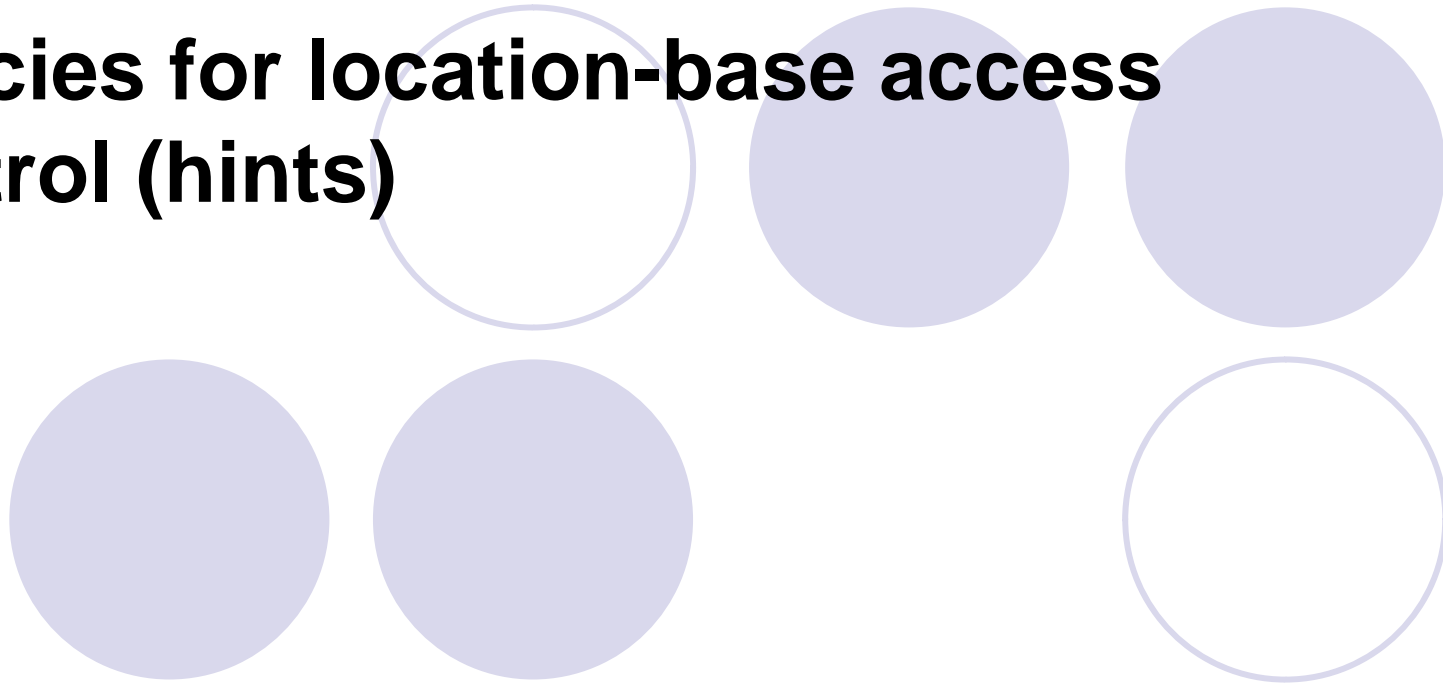
According to user preferences, location updates are de-identified and stored at different details

# 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  
 $Dept \subseteq_{ft} Hospital, Room \subseteq_{ft} Sector, Room \subseteq_{ft} Dept, Sector \subseteq_{ft} Hospital$   
 $OBJ = \{Ext(PatientRecord), Ext(Map), Ext(Person)\}$   
 $OPS = \{GetPatientRecord, UpdatePatientRecord, FindPersonnel, GetMap, GetStatistics\}$

$PRMS = \{p_1, p_2, p_3, p_4, p_5\}$  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

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

$R_{EXT\_FT} = \{Hospital, Dept\}$

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

$R_S = \{Pe, Do, Pd, Nu, Ma, Pa\}$  with  $\begin{cases} Pe = \langle Personnel, Hospital, Sector, m_{Sector} \rangle \\ Ma = \langle Manager, Hospital, Sector, m_{Sector} \rangle \\ Do = \langle Doctor, Hospital, Room, m_{Room} \rangle \\ Pd = \langle Pediatricist, Dept, Sector, m_{Sector} \rangle \\ Nu = \langle Nurse, Dept, Room, m_{Room} \rangle \\ Pa = \langle Patient, Hospital, Sector, m_{Sector} \rangle \end{cases}$

### Instances

$R_{EXT} = \{Hosp_1, Dep_1\}$

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

### 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

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

### User assignment

$U = \{Alice, Sara\}$

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

### Sessions

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

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

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

### EnabledRoles

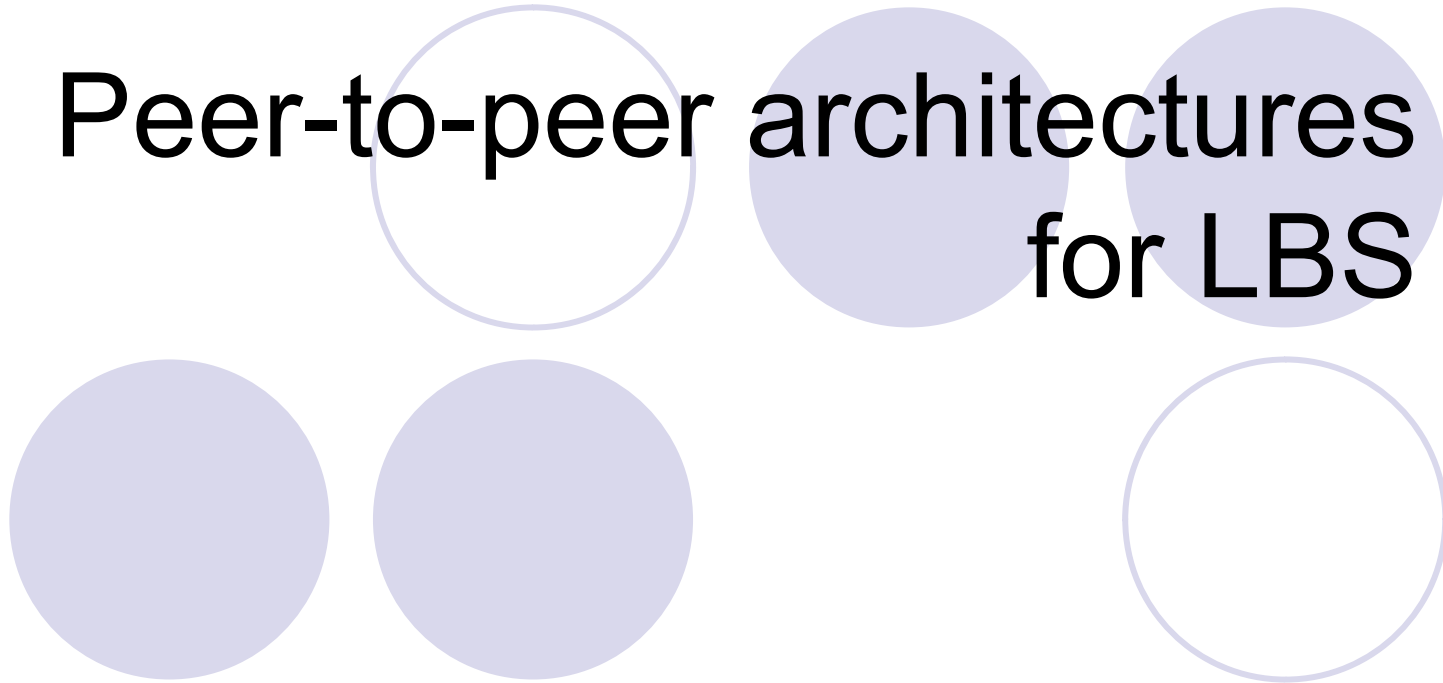
$EnabledSessionRoles(s_1, loc_1) = \{Pediatricist(Dep_1)\}$  if Alice is in  $Dep_1$

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



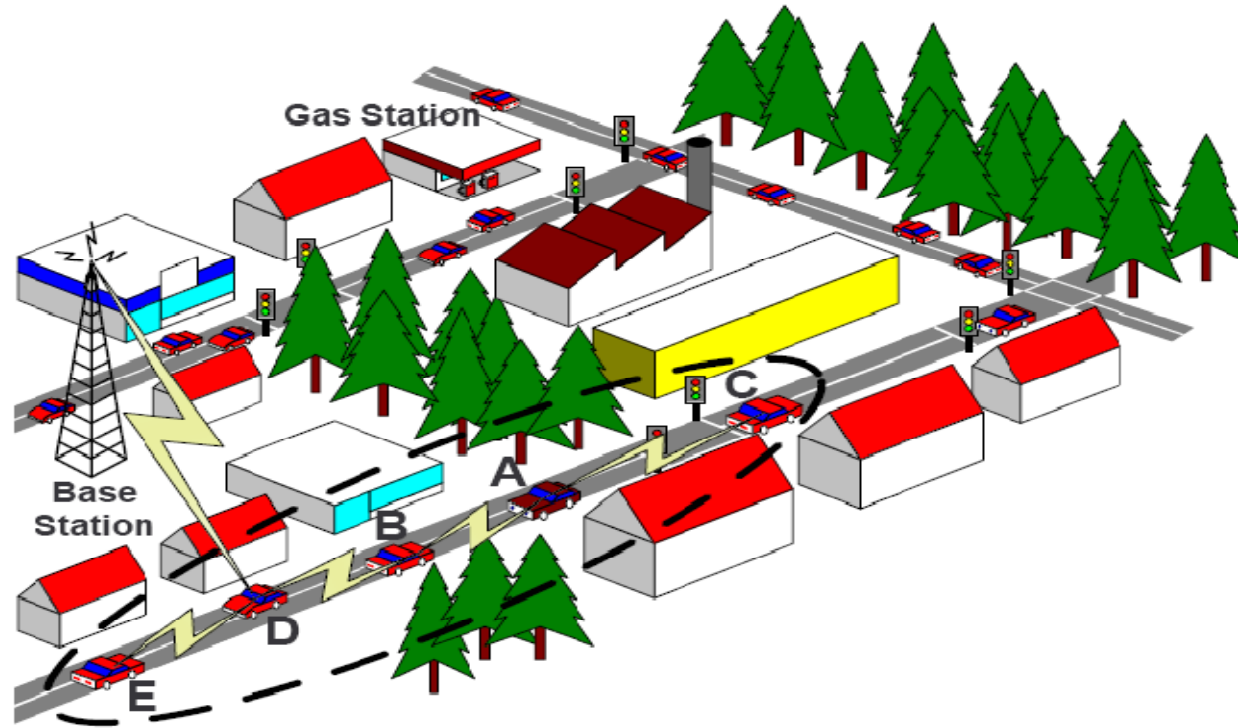


# Peer-to-peer architectures for LBS



# 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.

The slide features a decorative arrangement of six circles. Three circles are solid light purple, and three are hollow with a light purple outline. They are arranged in two rows of three, with the text centered between them.

# 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 <sup>(1)</sup>

- 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
  - $\text{PrivacyLoss}(NC \cup R(i)) - \text{PrivacyLoss}(R(i))$  is minimized