Mobility, Data Mining, and Privacy

Yannis Theodoridis

InfoLab, University of Piraeus, Greece
infolab.cs.unipi.gr

Mobile devices and services

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

Wireless networks as mobility data collectors

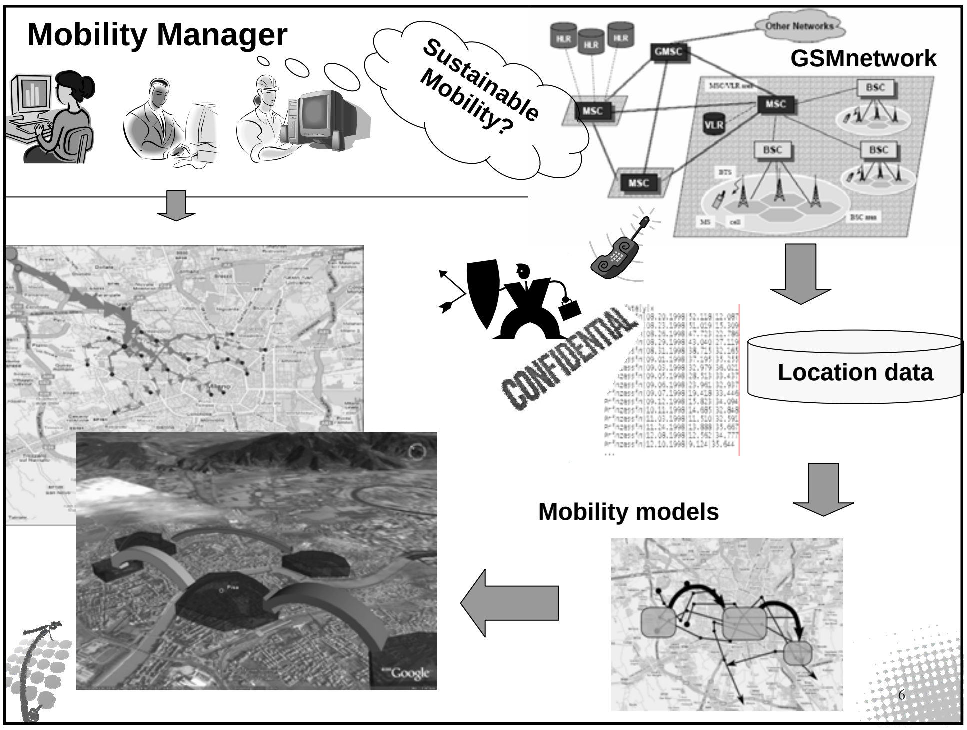
- Wireless networks infrastructures are the nerves of our territory
- besides offering their services, they gather highly informative traces about the human mobile activities
- UbiComp infrastructure will further push this phenomenon
- Miniaturization, wearability, pervasiveness will produce traces of 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
  - Next/current generation of Nokia mobile phones have 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)

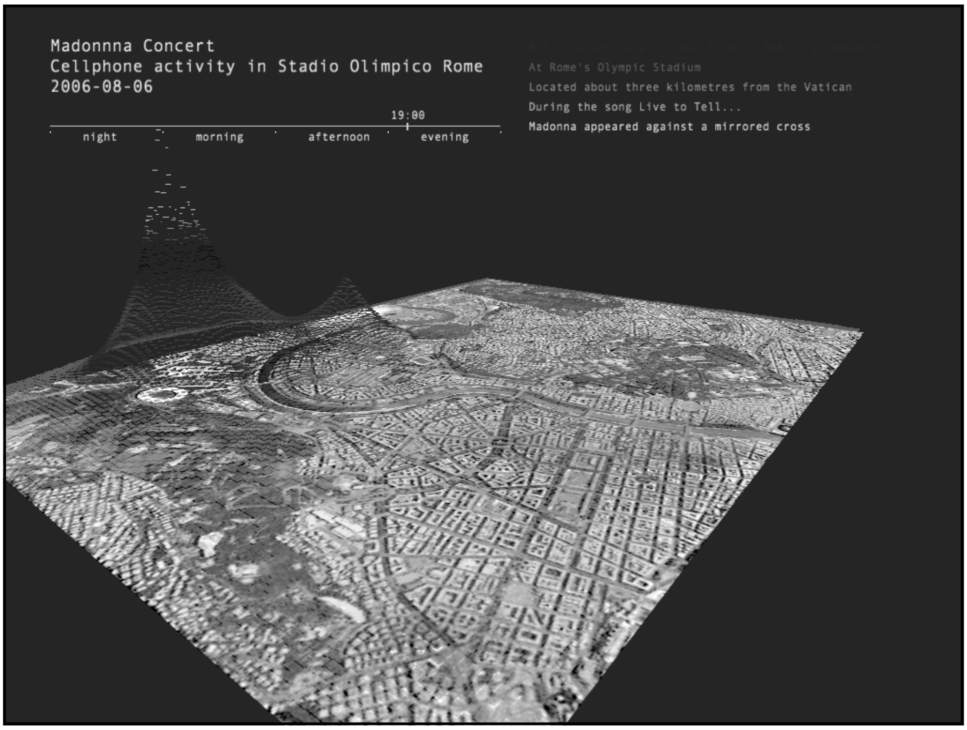
The GeoPKDD scenario

- 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 many 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.

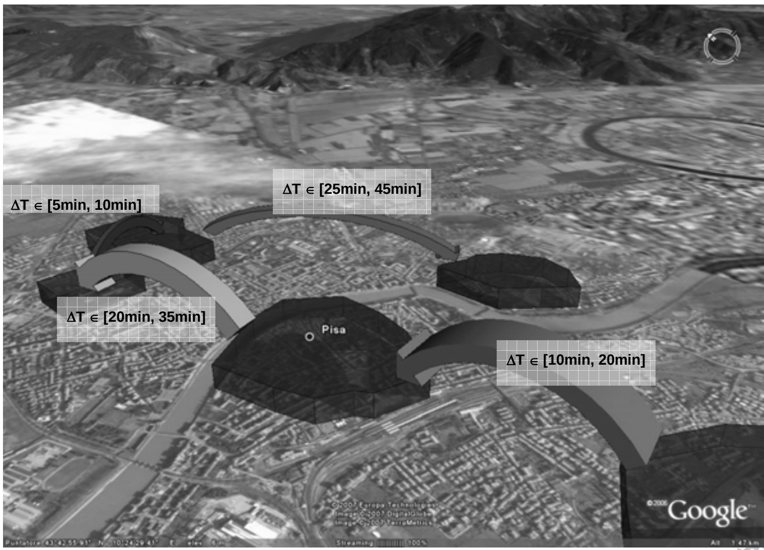


Real-time density estimation in urban areas

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

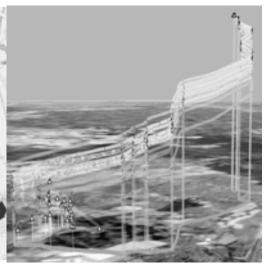


More ambitiously: mobility patterns

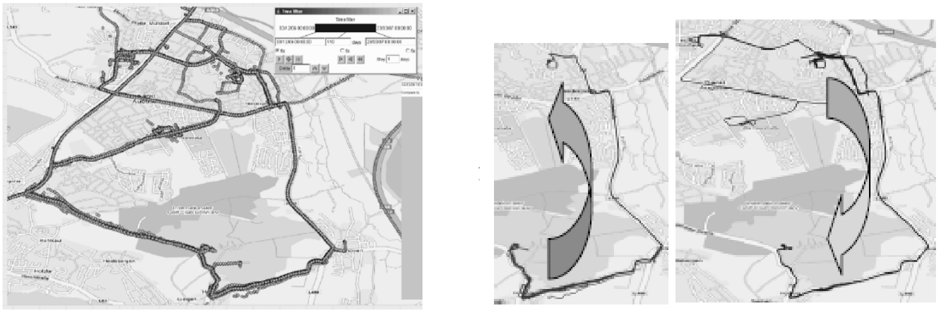


From mobility data to mobility patterns

```
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
...
```



From mobility data to mobility patterns



Key questions

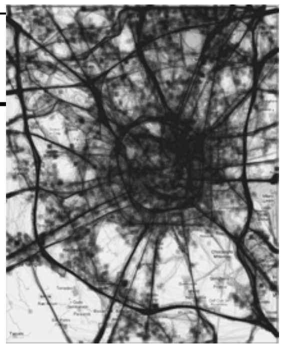
- How to reconstruct a trajectory from raw logs, how to store and query trajectory data?
- How to classify trajectories according to means of transportation (pedestrian, private vehicle, public transportation vehicle, …)?
- Which spatio-temporal patterns and/or models are useful abstractions of mobility data?
  - How to compute such patterns and models efficiently?
- Privacy protection and anonymity – how to make such concepts formally precise and measurable?
  - How to find an optimal trade-off between privacy protection and quality of the analysis?

A guided tour on MODAP technologies

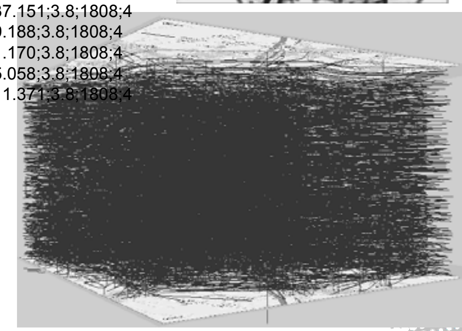
- Trajectory database management
  - Acquiring, storing, indexing, and querying trajectories
  - The Hermes MOD engine
- Trajectory data warehousing and OLAP
- Mobility data mining
  - Frequent pattern mining
  - Trajectory clustering
- Privacy-preserving mobility data querying & mining

Acquiring, Storing and Querying trajectories

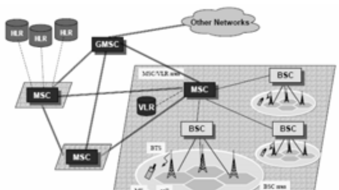
Data: typical structure / size

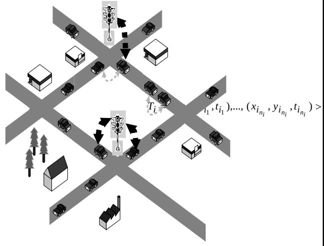


```
N;Time;Lat;Long;Height;Course;Speed;PDOP;State;NSat
…
8;22/03/07 08:51:52;50.777132;7.205580; 67.6;345.4;21.817;3.8;1808;4
9;22/03/07 08:51:56;50.777352;7.205435; 68.4;35.6;14.223;3.8;1808;4
10;22/03/07 08:51:59;50.777415;7.205543; 68.3;112.7;25.298;3.8;1808;4
11;22/03/07 08:52:03;50.777317;7.205877; 68.8;119.8;32.447;3.8;1808;4
12;22/03/07 08:52:06;50.777185;7.206202; 68.1;124.1;30.058;3.8;1808;4
13;22/03/07 08:52:09;50.777057;7.206522; 67.9;117.7;34.003;3.8;1808;4
14;22/03/07 08:52:12;50.776925;7.206858; 66.9;117.5;37.151;3.8;1808;4
15;22/03/07 08:52:15;50.776813;7.207263; 67.0;99.2;39.188;3.8;1808;4
16;22/03/07 08:52:18;50.776780;7.207745; 68.8;90.6;41.170;3.8;1808;4
17;22/03/07 08:52:21;50.776803;7.208262; 71.1;82.0;35.058;3.8;1808;4
18;22/03/07 08:52:24;50.776832;7.208682; 68.6;117.1;11.371;3.8;1808;4
…
```



Location data producers: GSM, GPS, WiFi





Location data (id, x, y, t) are collected

Trajectory stream manager + Trajectory reconstruction

trajectory data (obj-id, traj-id, (x, y, t)*) are reconstructed

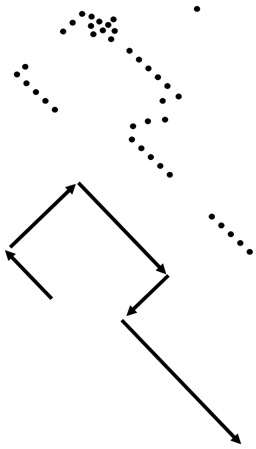
$$T_i = <(x_{i_1}, y_{i_1}, t_{i_1}), \ldots, (x_{i_{n_i}}, y_{i_{n_i}}, t_{i_{n_i}})>$$

Moving Object Database

The trajectory reconstruction problem

- From raw location data (obj-id, x, y, t)

a sample of a user's movement (GPS recordings)

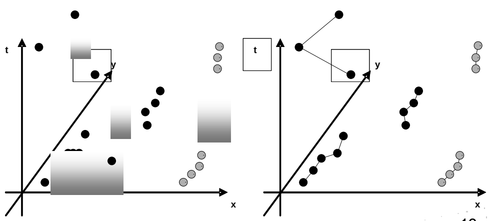


- To trajectory data (obj-id, traj-id, (x, y, t)+)

a sample of reconstructed trajectories

Reconstructing trajectories

- Collected raw data represent time-stamped geographical locations
  - Raw points arrive in bulk sets
  - We need a filter that decides if the new series of data is to be appended to an existing trajectory or not:
    - Tolerance distance
    - Temporal gap
    - Spatial gap
    - Maximum speed
    - Maximum noise duration



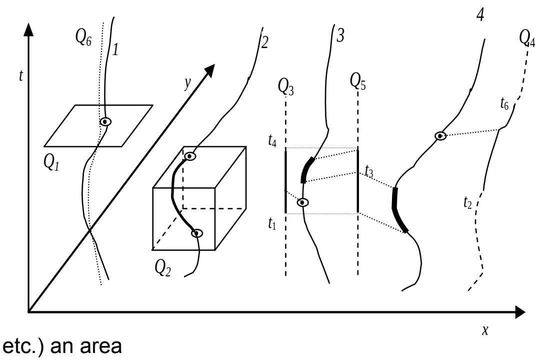
18

Moving Objects Databases

- The traditional database technology has been extended into Moving Object Databases (MODs) that handle modeling, indexing and query processing issues for trajectories
- Spatial and temporal dimensions are considered as first-class citizens.
- Both past and current (as well as anticipated future) positions of moving objects are of interest.
  - SECONDO (Guting et. al.) ICDE'05.
  - PLACE (Mokbel et al.) VLDB'04.

Querying the Moving Object Database

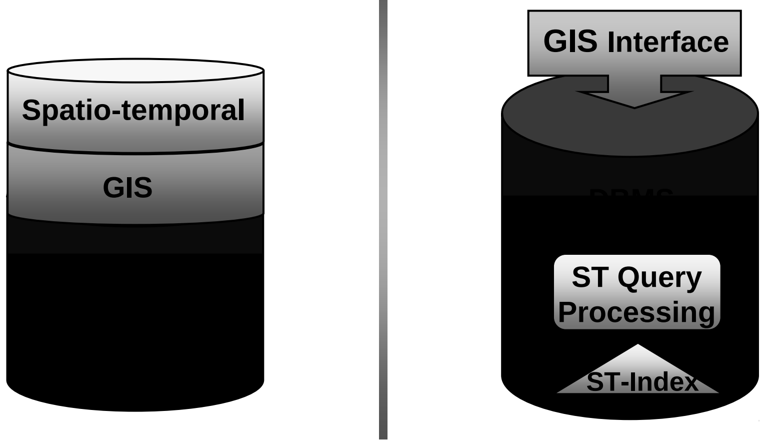
- Traditional spatial search
  - Range / distance-based / NN queries
- Trajectory-sub-sequence search
  - Spatial / temporal intersections of trajectories
- Topological / directional search
  - enter (cross, leave, bypass, etc.) an area
  - located west (south, etc.) of a (static) area
  - located left of (right of, in front of, etc.) a (moving) object



Location-based Database Servers

Layered Approach

Built-in Approach

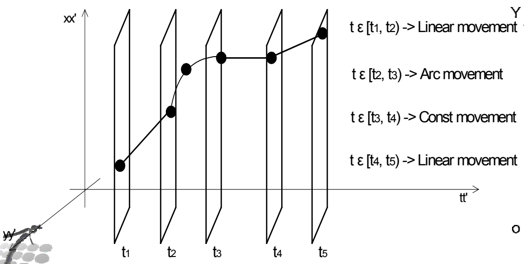


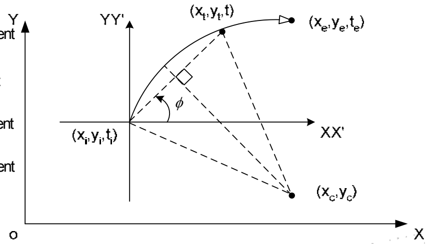
HERMES: An Engine for MODs

- Built on top of ORACLE 10
- Data model: absolute vs. relative location coordinates
  - Current location as a function in time over the starting location
  - linear and arc movement functions
- Trajectory management
  - Insert/Update/Delete a moving object or a segment of its trajectory
  - Functions over trajectories or sets of trajectories
- Data management
  - Supported indices: R-tree (for stationary data)
  - Development of a specialized index (TB-tree)

Hermes: trajectory data type

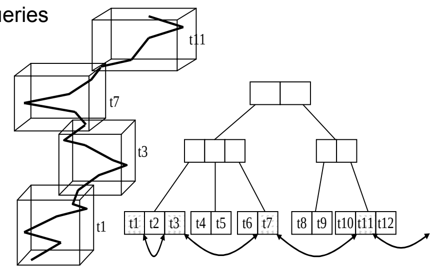
- Primitive definition:
  - Unit_Function = d 〈 xi:double, yi:double, xe:double, ye:double, xc:double, yc:double, v:double, a:double, flag:TypeOfFunction 〉, where
  - TypeOfFunction={ CONST, PLNML_1, ARC_<1..8> }
  - Unit_Moving_Point = d 〈 p: Period〈SEC〉, m: Unit_Function〉
  - Moving_Point = d { tab: set〈Unit_Moving_Point〉 | …constraints…}





TB-Tree support in Hermes MOD engine

- TB-Tree Index
  - Maintains the 'trajectory' concept
    - Each node consists of segments of a single trajectory
    - Nodes are linked together in a chain
  - Effective for trajectory-oriented queries
  - Implemented in Hermes using Oracle's indexing extensibility



HERMES includes

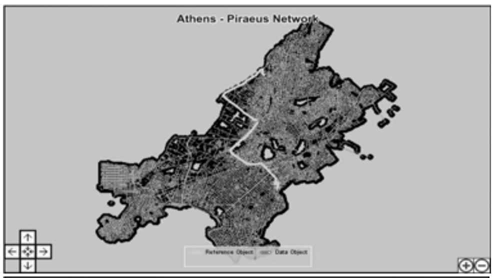
- Spatial entities:
  - Road Network Data (Nodes, Links)
  - Landmarks (ID, geometry, address, area, type)
  - Regions (ID, name, geometry)
- “Moving” entities:
  - Vehicles (object_id, traj_id, route)

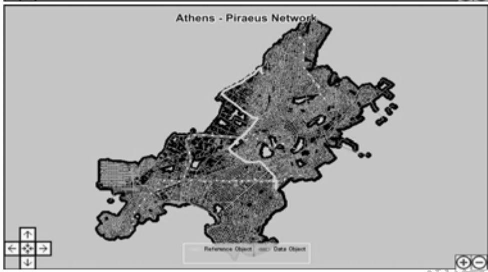
Query Operations

- Entities involved in a query
  - Reference Object: the type (trajectory or spatial entity) of the object based on which query answers are retrieved
  - Data Object: the type (trajectory or spatial entity) of the objects participating in the posed query answer
- Query classification
  - Moving Point – Moving Point
  - Moving Point – Static Spatial
  - Static Spatial – Moving Point

Moving Point – Moving Point

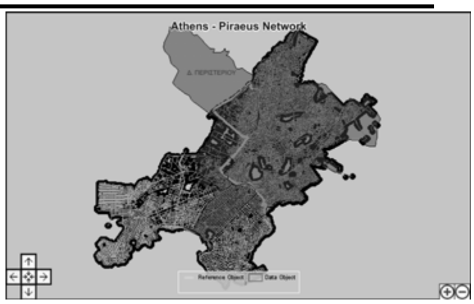
- Nearest Neighbor queries
  - Given a trajectory T, find the K nearest (during T's lifetime) parts of other trajectories
- Similarity queries
  - Spatial similarity
  - Spatiotemporal similarity
  - Speed-pattern similarity
  - Direction-pattern similarity

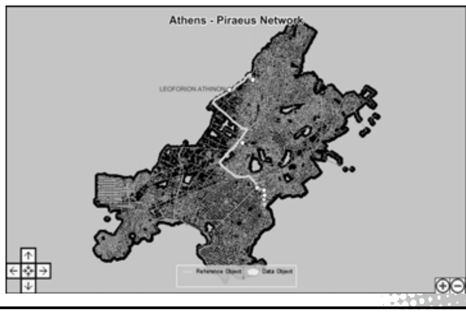




Moving Point – Static Spatial

- Point query
  - Find the regions that intersect with a given trajectory
- Topological query
  - Find the regions that contain, overlap by intersect, overlap by disjoint etc with a given trajectory
- Nearest-Neighbor query
  - Find the K nearest landmarks (POIs) to a given trajectory

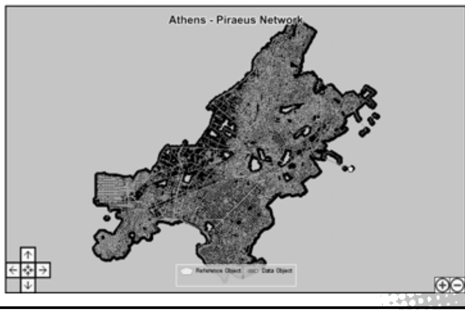




Static Spatial– Moving Point (1/2)

- Range query
  - Find trajectory parts fully contained in a given spatiotemporal window

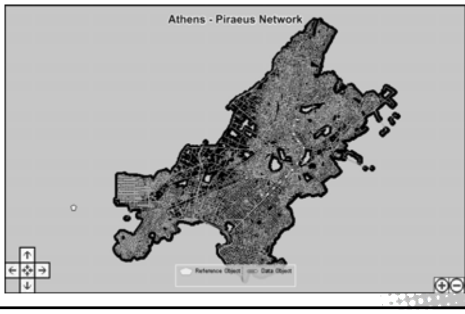
- Nearest Neighbor query
  - Find the K nearest trajectory parts to a POI, within a given time period



Static Spatial– Moving Point (2/2)

- Topological query
  - Find the trajectories that enter/leave an area within a given time period

- Directional query
  - Find trajectories whose location is east, west, north, south, left, right, front, behind of a POI

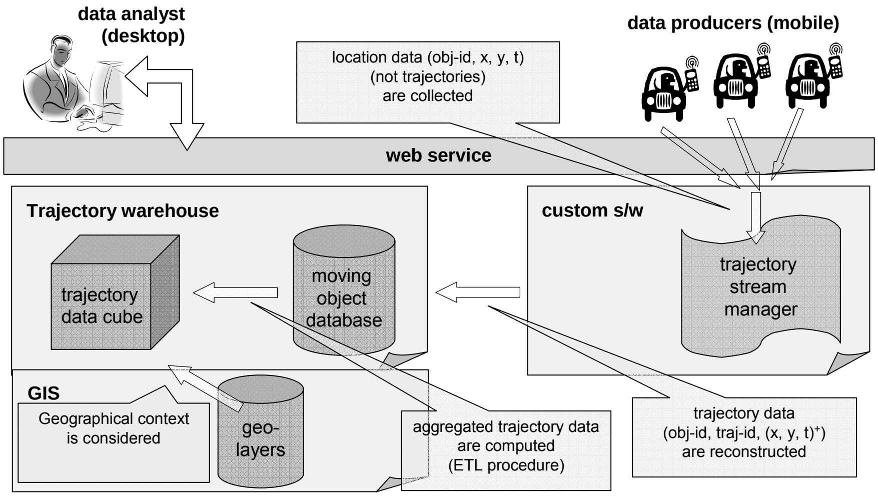


HERMES References

- Nikos Pelekis, Elias Frentzos, Nikos Giatrakos, Yannis Theodoridis: HERMES: aggregative LBS via a trajectory DB engine. SIGMOD Conference 2008: 1255-1258
- Elias Frentzos, Kostas Gratsias, Yannis Theodoridis: Towards the Next Generation of Location-Based Services. W2GIS 2007: 202-215
- Elias Frentzos, Kostas Gratsias, Nikos Pelekis, Yannis Theodoridis: Algorithms for Nearest Neighbor Search on Moving Object Trajectories. GeoInformatica 11(2): 159-193 (2007)
- Nikos Pelekis, Yannis Theodoridis: Boosting location-based services with a moving object database engine. MobiDE 2006: 3-10
- Nikos Pelekis, Yannis Theodoridis, Spyros Vosinakis, Themis Panayiotopoulos: Hermes - A Framework for Location-Based Data Management. EDBT 2006: 1130-1134
- Yannis Theodoridis: Ten Benchmark Database Queries for Location-based Services. Comput. J. 46(6): 713-725 (2003)

Trajectory Data Warehouses and OLAP

A trajectory warehouse system architecture



Data warehouses (DW)

- Widely investigated for conventional, non-spatial data.
- Some research on spatial DW, pioneering work by Han et al. in 1998.
  - Spatial and non-spatial dimensions and measures.
  - OLAP operations in a spatial data cube.
- Recent research direction: developing spatio-temporal DW and supporting spatio-temporal OLAP operations in order to extract summarized spatio-temporal information.
  - Useful for: traffic supervision systems, transportation and supply chain managements, mobile e-commerce.
  - Focus on methods for an efficient implementation of spatio-temporal aggregate queries.

Trajectory data warehousing

- Trajectory data warehousing should
  - extract aggregate information from MOD
  - support a variety of dimensions (temporal, spatial, thematic, …) and measures (about space, time and their derivatives)
  - Storing measures associated with facts, concerning the set of trajs crossing the cell
    - ⇒ aggregate information in base cells
- Challenges
  - high volume and complex nature of data; special query processing requirements
- Results so far:
  - design of a trajectory-oriented data cube
  - extensions of traditional aggregation techniques to produce summary information for OLAP analysis

Basic definitions & schemas

- Moving Object Database
  - A collection of trajectories

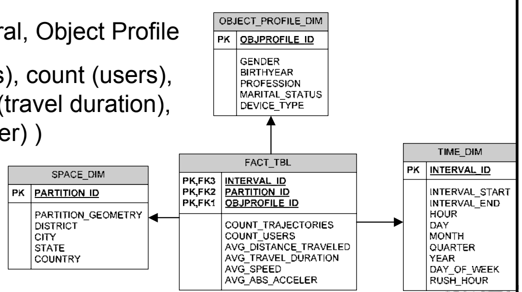
$$T_i = < (x_{i_1}, y_{i_1}, t_{i_1}), \ldots, (x_{i_{n_i}}, y_{i_{n_i}}, t_{i_{n_i}}) >$$

OBJECTS (object-id: *identifier*, description: *text*, gender: {M | F}, birth-date: *date*, profession: *text*, device-type: *text*)

RAW_LOCATIONS (object-id: *identifier*, timestamp: *datetime*, eastings-x: *numeric*, northings-y: *numeric*, altitude-z: *numeric*)

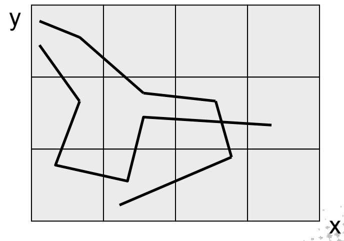
MOD_TRAJECTORIES (trajectory-id: *identifier*, object-id: *identifier*, trajectory: *3D geometry*)

- Trajectory Data Warehouse
  - Dimensions: Spatial, Temporal, Object Profile
  - Measures: count (trajectories), count (users), avg (distance traveled), avg (travel duration), avg (speed), avg (abs (acceler))



ETL processing: loading

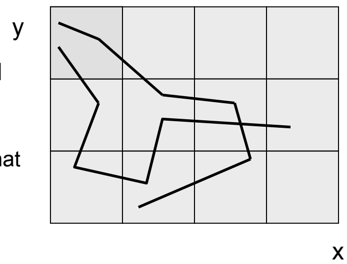
- Loading data into the dimension tables ➔ straightforward
- Loading data into the fact table ➔ complex
  - Fill in the measures with the appropriate numeric values
  - In order to calculate the measures, we have to extract the portions of the trajectories that fit into the base cells of the cube
    - Alternative solutions
      - cell-oriented
      - trajectory-oriented



ETL processing: algorithms

Cell-oriented approach **(COA)**

- Search for the portions of trajectories that they reside inside a spatiotemporal cell
  - Perform a **spatiotemporal range query** that returns the portions of trajectories that satisfy the range constraints
  - This is efficiently supported by the **TB-tree** [VLDB'00]
- Decompose the trajectory portions with respect to the user profiles they belong to
- Compute measures for this cell
- Repeat for the next cells

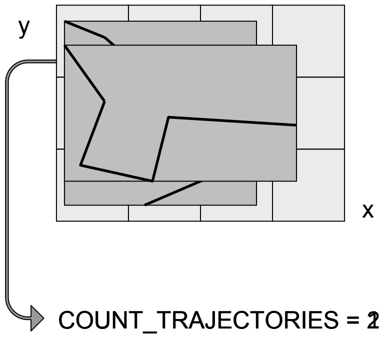


COUNT_TRAJECTORIES = 2
COUNT_USERS = 2
…

ETL processing: algorithms

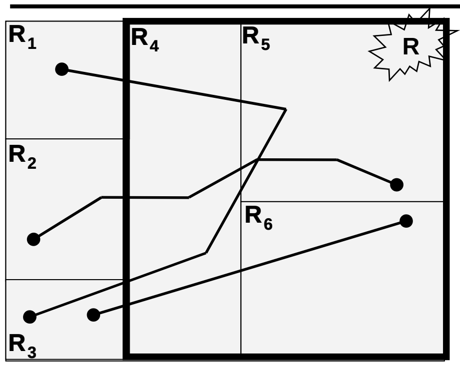
Trajectory-oriented approach **(TOA)**

- Discover the spatiotemporal cells where each trajectory resides in
  - In order to avoid checking all cells, use the trajectory MBR
- Identify the cells that overlap with the MBR and contain portions of the trajectory
- Compute measures for each cell
- …
- Repeat for the next trajectories
- …



COUNT_TRAJECTORIES = 2
COUNT_USERS = 2
…

Aggregating measures in the cube



At the lowest hierarchy level:

count of trajectories in R_4 = 3

count of trajectories in R_5 = 2

count of trajectories in R_6 = 1

Roll up in R

count of trajectories in R = 6
(according to traditional roll up)

Correct answer: 3 (!!) due to the fact that the contents (trajectories) of the partitions are overlapping

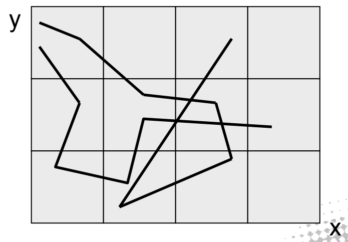
How to compute the correct answer?

•A naïve solution is to query back the raw data.

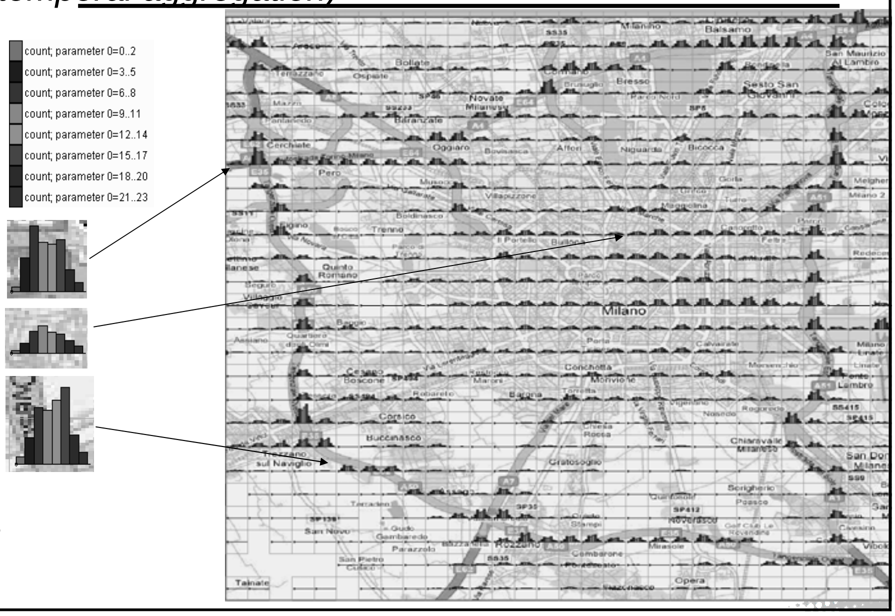
•Can we do something better?

The distinct count problem

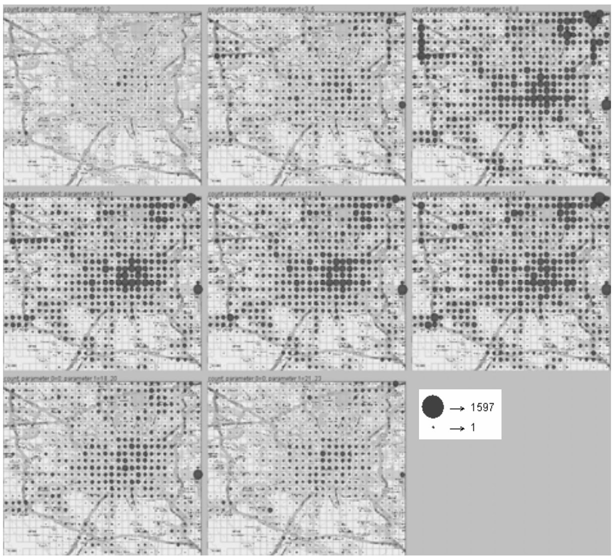
- During the ETL process, measures can be computed in an accurate way by executing MOD queries
- Once the fact table has been fed, aggregate-only information is stored inside the TDW (no trajectory / user ids)
- When rolling up, COUNT_USERS, COUNT_TRAJECTORIES and, hence, all other measures defined over COUNT_TRAJECTORIES are subject to the distinct count problem [ICDE'04]:
  - if an object remains in the query region for several timestamps during the query interval, instead of counting this object once, it is counted multiple times in the result



OLAP example: traffic density patterns (spatio-temporal aggregation)



OLAP example: low- vs. high- speed movement (counts, 3h intervals)



TDW References

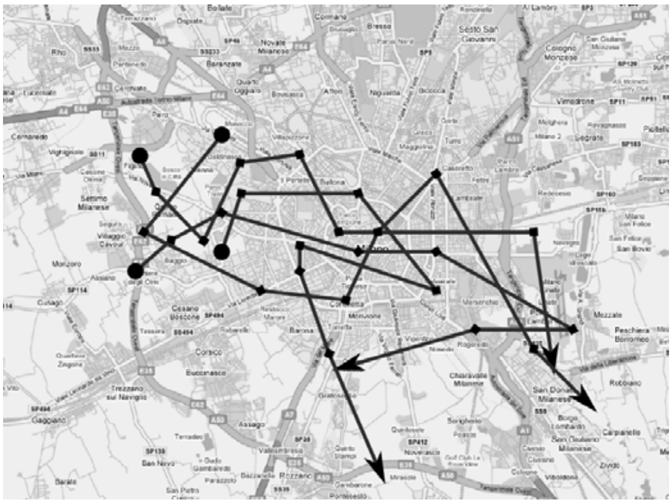
- eCourier.co.uk dataset, http://api.ecourier.co.uk/.
- Han, J., Stefanovic, N., and Koperski, K. Selective Materialization: An Efficient Method for Spatial Data Cube Construction. Proc. PAKDD, 1998.
- Orlando, S., Orsini, R., Raffaetà, A., Roncato, A., and Silvestri, C. Spatio-Temporal Aggregations in Trajectory Data Warehouses. Proc. DaWaK, 2007.
- Gerasimos Marketos, Elias Frentzos, Irene Ntoutsi, Nikos Pelekis, Alessandra Raffaetà, and Yannis Theodoridis. Building Real World Trajectory Warehouses. Proc. MobiDE'08, Vancouver, Canada
- Pelekis, N., Raffaetà, A., Damiani, M.-L., Vangenot, C., Marketos, G., Frentzos, E., Ntoutsi, I., and Theodoridis, Y. Towards Trajectory Data Warehouses. Chapter in Mobility, Data Mining and Privacy: Geographic Knowledge Discovery. Springer-Verlag. 2008.
- Pfoser, D., Jensen, C.S., and Theodoridis, Y. Novel Approaches to the Indexing of Moving Object Trajectories, Proc. VLDB, 2000.
- Tao, Y., Kollios, G., Considine, J., Li, F., and Papadias, D. Spatio-Temporal Aggregation Using Sketches. Proc. ICDE, 2004.

Mobility data mining

Frequent Pattern Mining

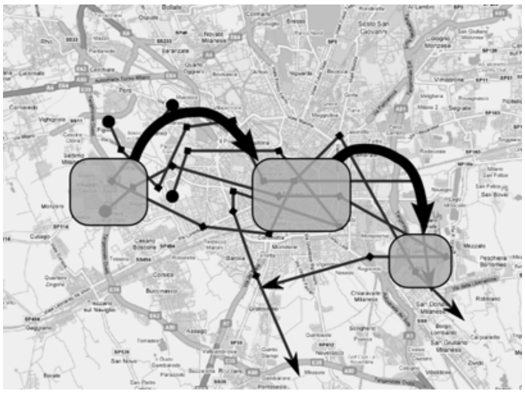
Trajectory Clustering

Q: What is a trajectory pattern?



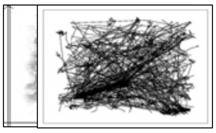
A: A spatio-temporal sequential pattern

- A sequence of visited regions, frequently visited in the specified order with similar transition times



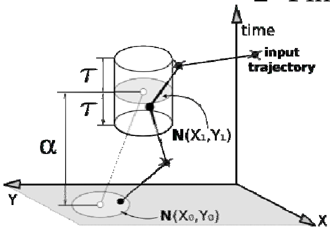
- Giannotti, Nanni, Pedreschi, Pinelli.
  Trajectory pattern mining. In Proc. ACM SIGKDD 2007

T-Pattern discovery

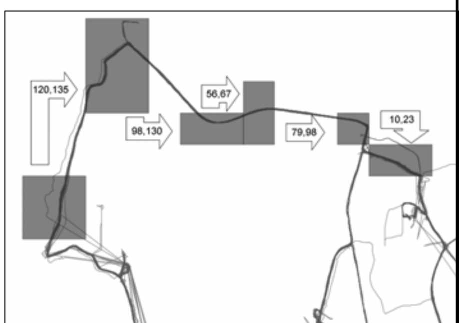


1- Find Regions of Interest

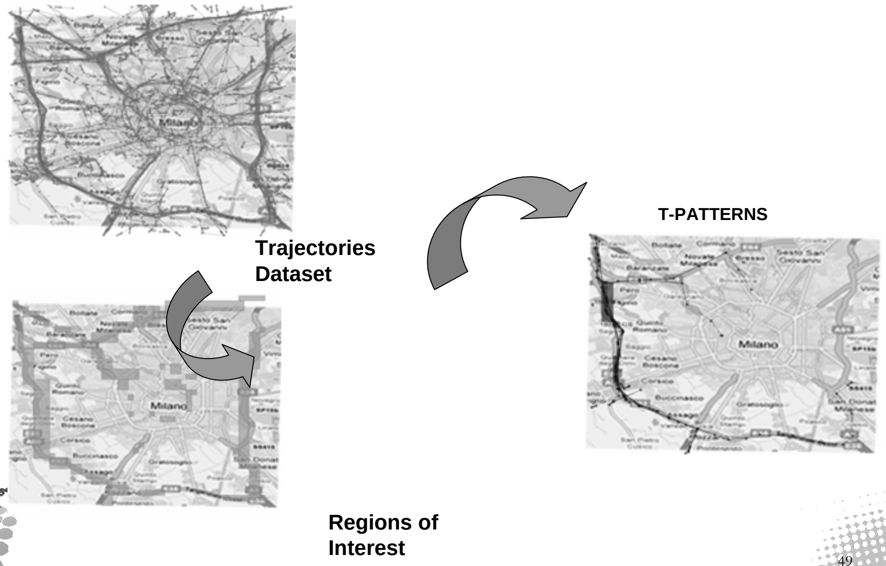
2- Find similar Trajectory in space and time



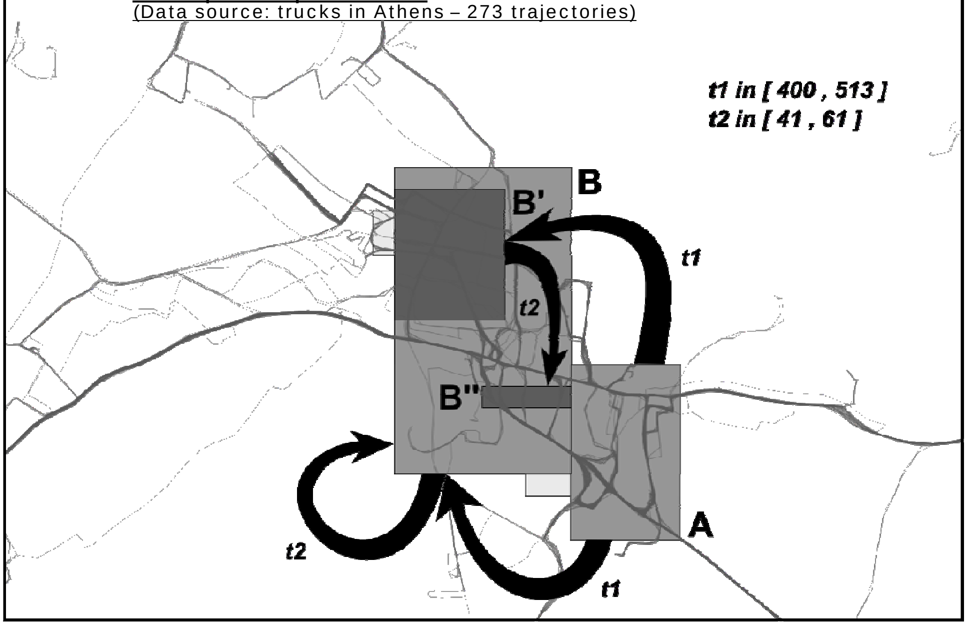
3- Extract patterns:



T-Pattern: Extraction Process

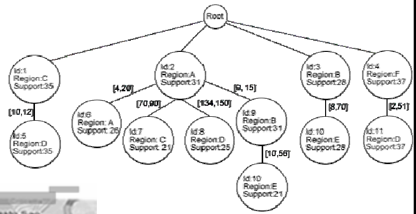


Sample T-patterns
(Data source: trucks in Athens – 273 trajectories)



Location Prediction based on T-patterns

- T-Pattern extracts a set of local patterns from a global set of data. Can we use these patterns to build a global model to predict the next location?



Local patterns
(T-pattern)

Global model
(Ptree)

Related works on T-patterns

- H. Cao, N. Mamoulis, and D. W. Cheung. Mining frequent spatio-temporal sequential patterns. ICDM'05.
- P. Kalnis, N. Mamoulis, and S. Bakiras. On discovering moving clusters in spatio-temporal data. SSTD'05.
- N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, and D. Cheung. Mining, indexing, and querying historical spatiotemporal data. KDD'04.
- J. Yang and M. Hu. TrajPattern: Mining sequential patterns from imprecise trajectories of mobile objects. EDBT'06.
- H. Cao, N.Mamoulis, and D.W. Cheung. Discovery of collocation episodes in spatiotemporal data.ICDM'06.

Related works on location prediction

- B. Xu and O. Wolfson. Time-series prediction with applications to traffic and moving objects databases. MobiDE, 2003.
- G. Yavas, D. Katsaros, O. Ulusoy, Y. Manolopoulos. A data mining approach for location prediction in mobile environments. Data Knowl. Eng., 54(2):121–146, 2005.
- M. Morzy. Prediction of moving object location based on frequent trajectories. ISCIS 2006, LNCS 4263 Springer.
- M. Morzy. Mining frequent trajectories of moving objects for location prediction. MLDM 2007, LNCS 4571 Springer.
- H. Jeung, Q. Liu, H. T. Shen, and X. Zhou. A hybrid prediction model for moving objects. ICDE, 2008.

Mobility data mining

Frequent Pattern Mining

Trajectory Clustering

Which distance between trajectories?

- **Average Euclidean distance**

$$D(\tau_1, \tau_2)|_T = \frac{\int_T \boxed{d(\tau_1(t), \tau_2(t))}\, dt}{|T|}$$

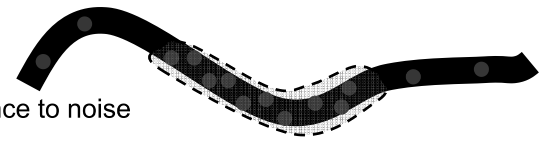
distance between moving objects τ1 and τ2 at time *t*

- "Synchronized" behaviour distance
  - Similar objects = almost always in the same place at the same time
- Computed on the whole trajectory
- Computational aspects:
  - Cost = O(|τ1| + |τ2|)   (|τ| = number of points in τ)
  - It is a metric => efficient indexing methos allowed

55

Which kind of clustering?

- General requirements:
  - Non-spherical clusters should be allowed
    - E.g.: A traffic jam along a road = "snake-shaped" cluster

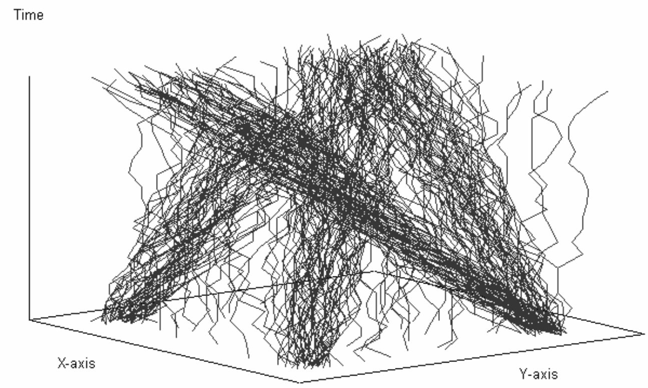


  - Tolerance to noise
  - Low computational cost
  - Applicability to complex, possibly non-vectorial data
- A suitable candidate: Density-based clustering
  - T(rajectory)-OPTICS

56

A sample dataset

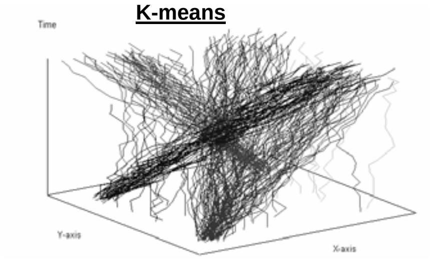
- Set of trajectories forming 4 clusters + noise (synthetic)



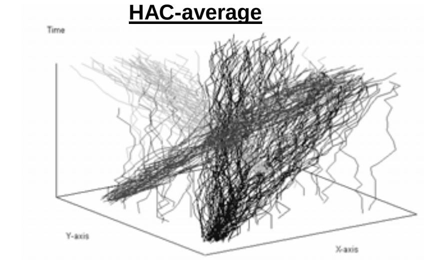
57

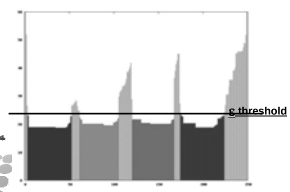
T-OPTICS vs. HAC & K-means

K-means



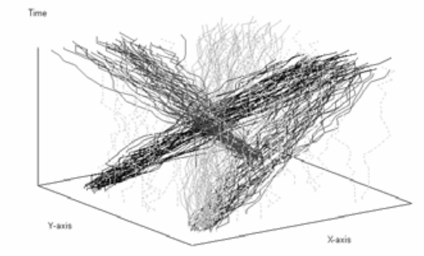
HAC-average





Reachability plot
(= objects reordering for distance distribution)

T-OPTICS



Related works on Trajectory Clustering

- Gaffney, S. and Smyth, P., Trajectory Clustering with Mixtures of Regression Models, ACM SIGKDD 1999.
- Gaffney, S., Robertson, A., Smyth, P., Camargo, S., and Ghil, M., Probabilistic Clustering of Extratropical Cyclones Using Regression Mixture Models, Tech. Rep. UCI-ICS 06-02, 2006.
- Nanni, M., Pedreschi, D. Time-focused clustering of trajectories of moving objects. J. of Intelligent Information Systems, 2006.
- Lee, J.-G., Han, J., and Whang, K.-Y., Trajectory Clustering: A Partition-and-Group Framework, SIGMOD 2007.
- Rinzivillo, Pedreschi, Nanni, Giannotti, Andrienko, Andrienko. Visually-driven analysis of movement data by progressive clustering. J. of Information Visualization, 2008

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
  - …

Privacy-preserving mobility data querying & mining

The naive scientist's view

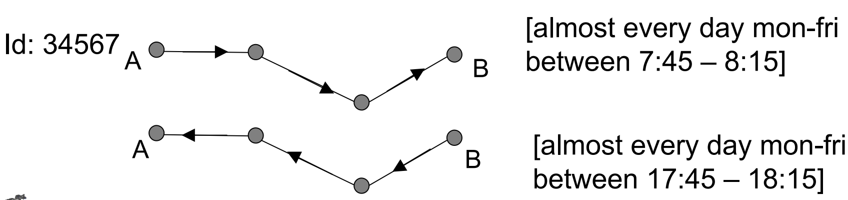
- Knowing the exact identity of individuals is not needed for analytical purposes
  - De-identified mobility data are enough to reconstruct aggregate movement behaviour, pertaining to groups of people.
- Reasoning coherent with European data protection laws: personal data, once made anonymous, are not 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 examples 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

- 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…

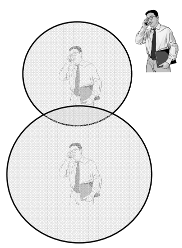


How do people (try to) stay anonymous?

- either by camouflage
  - pretending to be someone else or somewhere else
- or by hiding in the crowd
  - becoming indistinguishable among many others

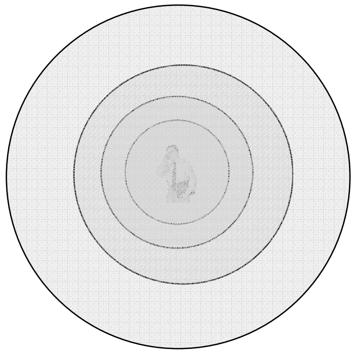
Location Perturbation – Randomization

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



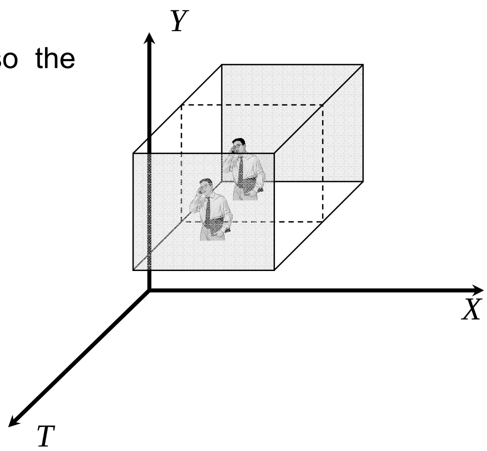
Spatial Cloaking – Generalization

- 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 region, but has no clue where the user is exactly located
- The area of the region achieves a trade-off between user privacy and accuracy



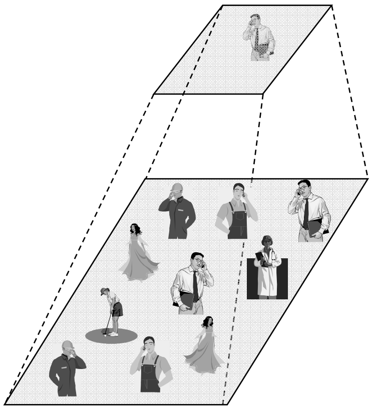
Spatio-temporal generalization

- In addition to the spatial dimension, generalize also the temporal dimension



k-anonymity

- User's position is generalized to a region containing at least k users
- The user is indistinguishable among other k users
- The area largely depends on the surrounding environment.
- A value of k =100 may result in a very small area downtown Hong Kong, or a very large area in the desert.



10-anonymity

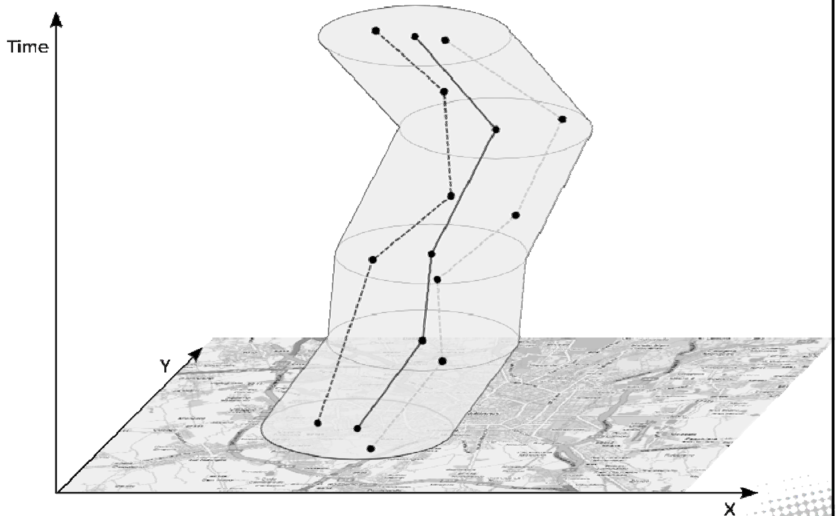
Trajectory anonymization

- Several variants developed in GeoPKDD:
  - "Never Walk Alone" by Abul, Bonchi, Nanni (Pisa KDD LAB)
  - "Always Walk with Others" by Nergiz, Atzori, Saygin (Sabanci Univ. + Pisa KDD LAB)
- Common goal: construct an anonymized version of a trajectory dataset, preserving some target analytical properties
- Different techniques adopted

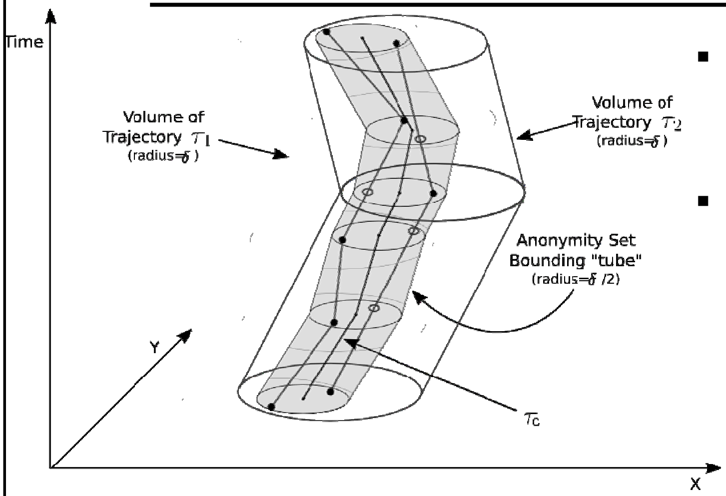
Never Walk Alone

- Bonchi, Abul, Nanni. Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases. ICDE 2008
- Basic ideas:
  - Trade uncertainty for anonymity: trajectories that are close up the uncertainty threshold are indistinguishable
  - Combine k-anonymity and perturbation
- Two steps:
  - Cluster trajectories into groups of k similar ones (removing outliers)
  - Perturb trajectories in a cluster so that each one is close to each other up to the uncertainty threshold

Trajectory cluster



(K,δ) –anonymity set



- K = minimum number of trajectories in the set
- δ = uncertainty threshold (e.g., measurement error of GPS device)

Quality of anonymized datasets

- For reasonable values of K and δ, some interesting analytical properties of the original dataset are preserved by the anonymized trajectories :
  - density (aggregate count of mobile users in the spatio-temporal dimension)
  - Clustering (to some extent …)
  - T-patterns: NOT!
- Prototype trajectory anonymity toolkit available

Key open challenges

- Define an acceptable formal measure of anonymity protection:
  - Probability of re-identification (in a given context)
  - A (technically supported) juridical issue!
- Sampling: a necessity and an opportunity!
  - Necessary for performance/feasibiliy of data mining from massive mobility datasets
  - Good for anonymity (re-identification probability decreases)

Conclusions

Conclusions

- Challenge: UbiComp will flood us with new complex data (in a decentralized setting)
  - data miners have only begun to scratch the surface of this problem
- "Privacy-preserving Mobility Data Mining" =
  - Obtaining the advantages of collective mobility knowledge without disclosing inadvertently any individual mobility knowledge.

Acknowledgements

- We are grateful to all the GeoPKDD (2005-09) and MODAP (2009-12) researchers!

- Links:
  - www.geopkdd.eu
  - www.modap.org
  - infolab.cs.unipi.gr