Privacy Preservation in Mobility Data

Information Systems Laboratory,

University of Piraeus http://infolab.cs.unipi.gr

Despina Kopanaki (dkopanak@unipi.gr)

January 2010

Contents

- Introduction
- Opportunities, Privacy Threats and Law Directions
- K –anonymity in Relational Databases
- Anonymity techniques in mobility data analysis
  - Sequence Hiding
  - Sequencial pattern hiding
- Privacy and anonymity in Location Based Services

Contents

- Introduction
- Opportunities, Privacy Threats and Law Directions
- K –anonymity in Relational Databases
- Anonymity techniques in mobility data analysis
  - Sequence Hiding
  - Sequential pattern hiding
- Privacy and anonymity in Location Based Services

The Wireless Network

- The pervasiveness of mobile and ubiquitous technologies is increasing day after day
  - GSM wireless phone networks
    - 1,5 billion in 2005, still increasing at a high speed
    - Italy: #mobile phones = # inhabitants
  - GPS and Galileo positioning systems
  - Wi-Fi and Wi-Max wireless networks
  - RFID's and sensor networks
- Positioning accuracy
  - Location technologies capable of providing increasingly better estimate of user location

Traces

- Our every day actions leave digital traces
  - Credit cards, e-transactions, e-banking
  - Electronic administrative transactions and health records
  - Shopping transactions with loyalty cards, etc.
- Wireless phone networks gather highly informative traces about the human mobile activities.
- Traces are stored because are worth being remembered.
- Precious knowledge may be revealed.

Contents

- Introduction
- Opportunities, Privacy Threats and Law Directions
- K –anonymity in Relational Databases
- Anonymity techniques in mobility data analysis
  - Sequence Hiding
  - Sequencial pattern hiding
- Privacy and anonymity in Location Based Services

Which new opportunities?

- Location based services
  - A certain service that is offered to the users based on their locations.
- Mobility data analysis:
  - Discovering knowledge from the digital traces of our mobile activity to support decision making in mobility related issues.
    - How people move around in the town?
    - Are there typical movement behaviours?
    - How are people movement habits changing in this area in last decade - year – month – day?

Individuals vs Enterprises

- Having so much information available about entities
  - provides many new and interesting ways to conduct research.
  - but makes it increasingly difficult to provide personal privacy.
- Privacy is an important issue today
  - Individuals feel
    - Uncomfortable: ownership of information
    - Unsafe: information can be misused
  - Enterprises need to
    - Keep their customers feel safe
    - Protect themselves from any legal dispute

Privacy in Mobility Data and Services

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

Law Directions

- In Greece
  - Law 2472/1997: protecting individuals from analyzing their private data by defining that the individual must be informed about who, when, where, how and why his data are being analyzed.
  - Law 2774/1999: protects human rights and private life from telecommunication data analyzing.
- In Europe
  - 95/46/EC: Goal is to ensure free flow information. Forbids sharing data with states that don't protect privacy
  - 2002/58/EC: protection of analyzing private data and private life in the domain of electronic communication.

Privacy Directive

- Laws are not directly enforceable.
- Practically they only remove user's identity.
- In certain cases, it is possible to reconstruct the exact identities from the released data, even when identities have been removed and replaced by pseudonyms.
- A famous example of re-identification by Sweeny.
  - She purchased the voter registration list for Cambridge Massachusetts – 54.805 people.
  - 69% of records: unique on zip code and date of birth.
  - 87% of records: unique on zip code, date of birth and sex.

Contents

- Introduction
- Opportunities, Privacy Threats and Law Directions
- K –anonymity in Relational Databases
- Anonymity techniques in mobility data analysis
  - Sequence Hiding
  - Sequencial pattern hiding
- Privacy and anonymity in Location Based Services

Link Private Information to Person

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

Quasi - identifiers

Sensitive Information

- **Quasi-identifiers: a set of attributes that may identify individuals.**
- **Sensitive attributes: information that individuals do not want to be published.**

The problem

- Transform a given dataset so that no one can
  - Associate a particular record with the corresponding data subject
  - Infer the sensitive information of any data subject

- Transformation must be minimal to preserve as much information as possible.
  - Minimize distortion of results.

The solution (Sweeney '01)

- K- anonymity:
  - any combination of values appears at least k times.
- The goal is to prevent linking a record from a set of released records to a specific individual.
- Under k-anonymity, there will be at least k individuals to whom a given record indistinctly refers.
- The k individuals appear in the released records.
- A lot of papers on k-anonymity in 2004-2006
  - (SIGMOD, VLDB, ICDE, ICDM)

Suppression - Generalization

Age	Location	Disease
α	β	Flu
α+2	β	Flu
δ	γ+3	Hypertension
δ	γ	Flu
δ	γ-3	Cold

Age	Location	Disease
*	β	Flu
*	β	Flu
δ	*	Hypertension
δ	*	Flu
δ	*	Cold

Original table

Zip	Gender	Age	Diagnosis
47918	Male	35	Cancer
47906	Male	33	HIV+
47918	Male	36	Flu
47916	Female	39	Obesity
47907	Male	33	Cancer
47906	Female	33	Flu

2-anonymized version / 3-anonymized

Zip	Gender	Age	Diagnosis
4791*	Person	[35-39]	Cancer
4790*	Person	[30-34]	HIV+
4791*	Person	[35-39]	Flu
4791*	Person	[35-39]	Obesity
4790*	Person	[30-34]	Cancer
4790*	Person	[30-34]	Flu

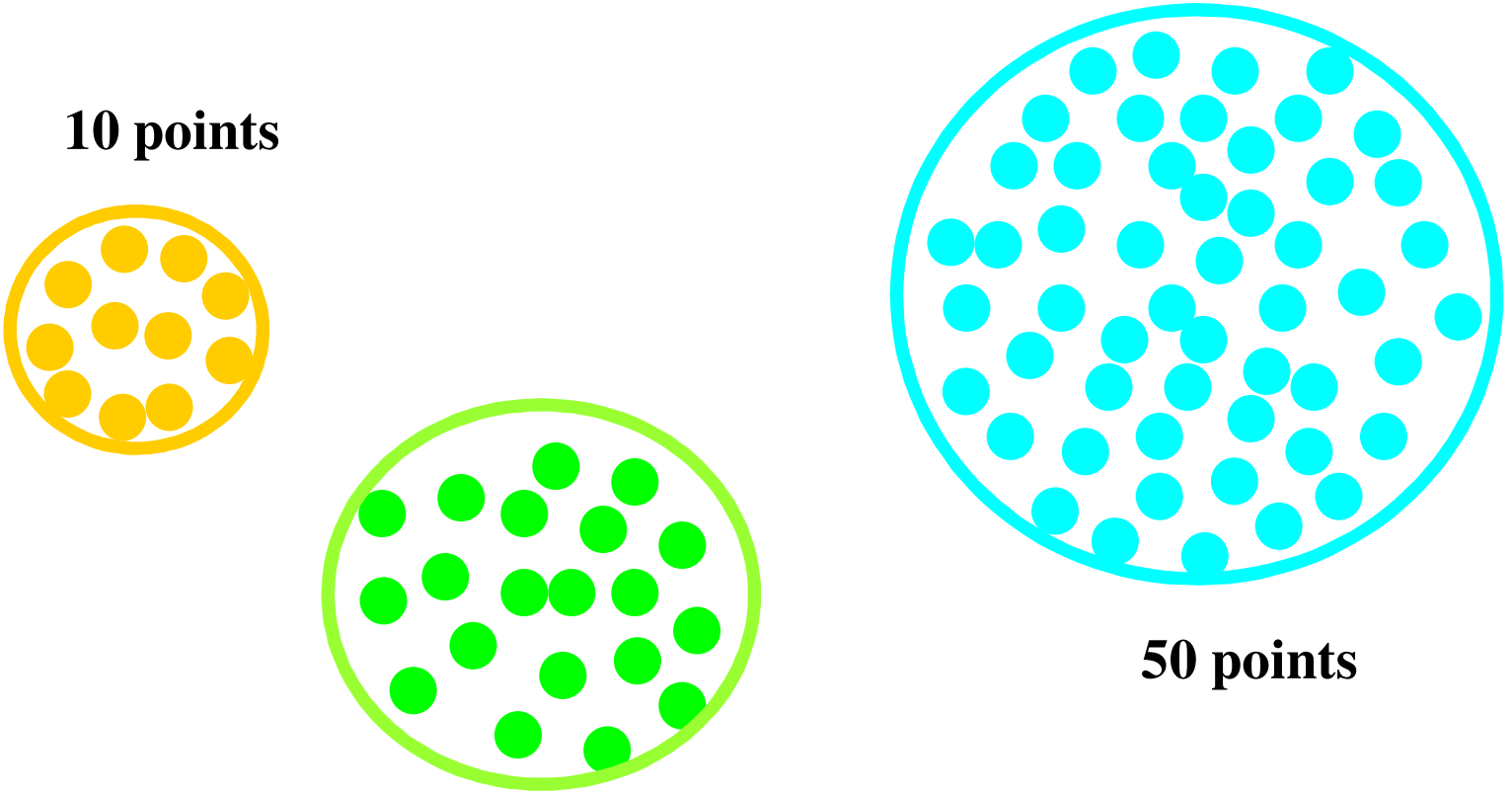
Advantages of Clustering

- Solve k-anonymity problems by using the advantages of clustering:
  - Clustering reduces the amount of distortion introduced as compared to suppressions /generalizations
- Suppression: changes the size of the data base, big information loss
- Generalization: Unnecessary generalization and which generalization is the best.

Clustering for Anonymity (Aggarwal ACM '06)

- Cluster Quasi-identifiers so that each cluster has at least r members for anonymity.
- Publish cluster centers for anonymity with number of point and radius.
- Tight clusters → Usefulness of data for mining.
- Large number of points per cluster→ Anonymity.

r-Gather Clustering



Minimize the maximum radius while ensuring that each cluster has at least r members

Example

Age	Location	Disease
α	β	Flu
α+2	β	Flu
δ	γ+3	Hypertension
δ	γ	Flu
δ	γ-3	Cold

Original table

Age	Location	Num Points	Disease
α+1	β	2	Flu Flu
δ	γ	3	Hypertension Flu Cold

2-gather clustering

L-Diversity (Machanavajjhala TKDD '07)

- Weaknesses in K-anonymous tables
- Homogeneity Attacks
  - k-Anonymity is focused on generalizing the quasi-identifiers but does not address the sensitive attributes which can reveal information to an attacker.
- Background Knowledge Attacks
  - Depending on other information available to an attacker, an attacker may have increased probability of being able to determine sensitive information.

Homogeneity Attacks

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

- Since Alice is Bob's neighbor, she knows that Bob is a 31-year-old American male who lives in the zip code 13053.
- Alice knows that Bob's record number is 9,10,11, or 12.
- She can also see from the data that Bob has cancer.

Background Knowledge Attacks

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

- Alice knows that Umeko is a 21 year-old Japanese female who currently lives in zip code 13068.
- Alice learns that Umeko's information is contained in record number 1,2,3, or 4.
- Umeko being Japanese and Alice knowing that Japanese have an extremely low incidence of heart disease.
- Alice can concluded with near certainty that Umeko has a viral infection.

L-divercity Principle

- A q*-block is l-diverse if contains at least l “well-represented” values for the sensitive attribute S. A table is l-diverse if every q*-block is l-diverse.

- The l-Diversity principle advocates ensuring well represented values for sensitive attributes but does not define what well represented values mean.

Contents

- Introduction
- Opportunities, Privacy Threats and Law Directions
- K –anonymity in Relational Databases
- Anonymity techniques in mobility data analysis
  - Sequence Hiding
  - Sequencial pattern hiding
- Privacy and anonymity in Location Based Services

Spatio-temporal linkage

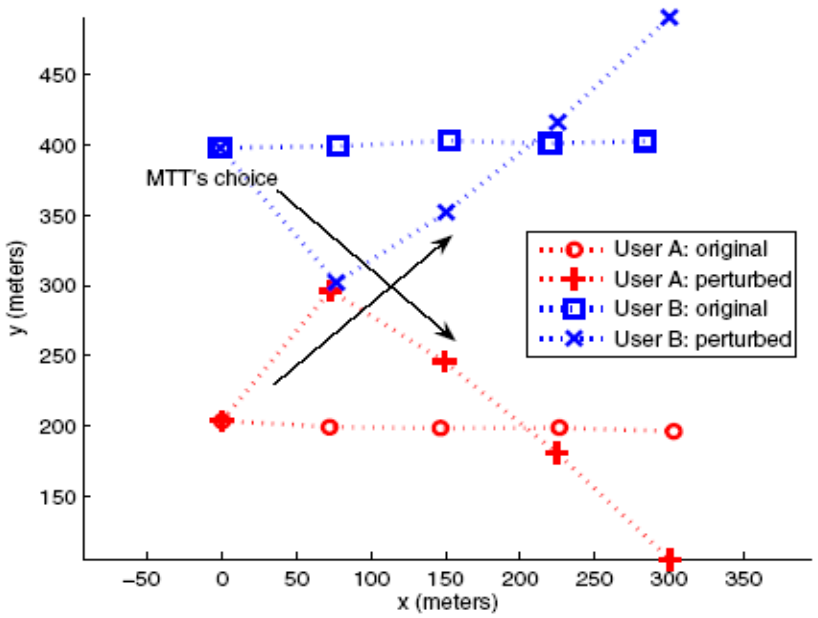
- An anonymous trajectory occurring every working day from location A to location B during the morning rush hours and in the reverse direction from B to A in the evening rush hours
  - The persons who live in A and work in B.
- If locations A and B are known, it is possible to identify specific persons.
- In mobility data, positioning in space and time is a powerful identifier.
- K-anonymity: anonymity set >= k
  - Strong k-anonymity allows multiple presence of the same user in the anonymity set.

Mobility Data Publishing

- Very little work on mobility data publishing.
- Main reasons
  - Data is not yet available due to privacy issues.
- Privacy – preserving techniques for data publishing exist for relational tables
  - They can easily extended to spatiotemporal data, but privacy concerns are not well-studied for these data.
  - Offline solutions would enable more accuracy while preserving anonymity of data donors.

Protecting Location Privacy through Path Confusion (Hoh-Gruteser SecureComm 2005)

- Idea of path crossing.
- Blue and red users move in parallel.
- Identify when two nonintersecting trajectories that belong to different users are reasonably close to each other and generates a fake crossing of these two.



Protecting Location Privacy through Path Confusion

- Achieved to prevent an adversary from tracking a complete user trajectory and thus identifying the corresponding user.
- Estimates the perturbed locations for each user such that their trajectories meet within a pre-specified time period.
- The radius that indicates the maximum allowable perturbation is the degree of privacy.

Privacy Preservation in the publication of trajectories (Terrovitis Mamoulis MDM 2008)

- Sequences of places that each user has visited in the course of her movement.
- No other information of spatial or temporal nature is provided.
- This technique removes some of the places that were visited by specific users to protect identity from adversaries.
- Operates an iterative fashion to min the probability of a given adversary to associate a place that in the publicized data (side effects).

id	trajectory
t_1	$a_1 \rightarrow b_1 \rightarrow a_2$
t_2	$a_1 \rightarrow b_1 \rightarrow a_2 \rightarrow b_3$
t_3	$a_1 \rightarrow b_2 \rightarrow a_2$
t_4	$a_1 \rightarrow a_2 \rightarrow b_2$
t_5	$a_1 \rightarrow a_3 \rightarrow b_1$
t_6	$a_3 \rightarrow b_1$
t_7	$a_3 \rightarrow b_2$
t_8	$a_3 \rightarrow b_2 \rightarrow b_3$

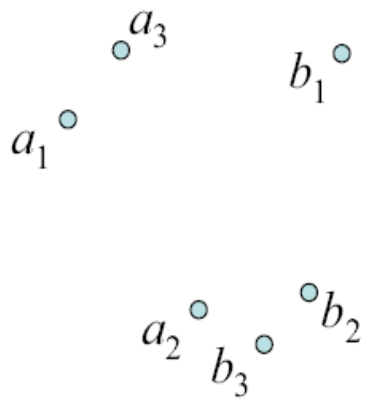
(a) exact data (T)

id	trajectory
t_1^A	$a_1 \rightarrow a_2$
t_2^A	$a_1 \rightarrow a_2$
t_3^A	$a_1 \rightarrow a_2$
t_4^A	$a_1 \rightarrow a_2$
t_5^A	$a_1 \rightarrow a_3$
t_6^A	a_3
t_7^A	a_3
t_8^A	a_3

(b) A's knowledge (T_A)

id	trajectory
t'_1	$a_1 \rightarrow b_1 \rightarrow a_2$
t'_2	$a_1 \rightarrow b_1 \rightarrow a_2$
t'_3	$a_1 \rightarrow b_2 \rightarrow a_2$
t'_4	$a_1 \rightarrow a_2 \rightarrow b_2$
t'_5	$a_3 \rightarrow b_1$
t'_6	$a_3 \rightarrow b_1$
t'_7	$a_3 \rightarrow b_2$
t'_8	$a_3 \rightarrow b_2$

(c) transformed database (T')



(d) the map of locations

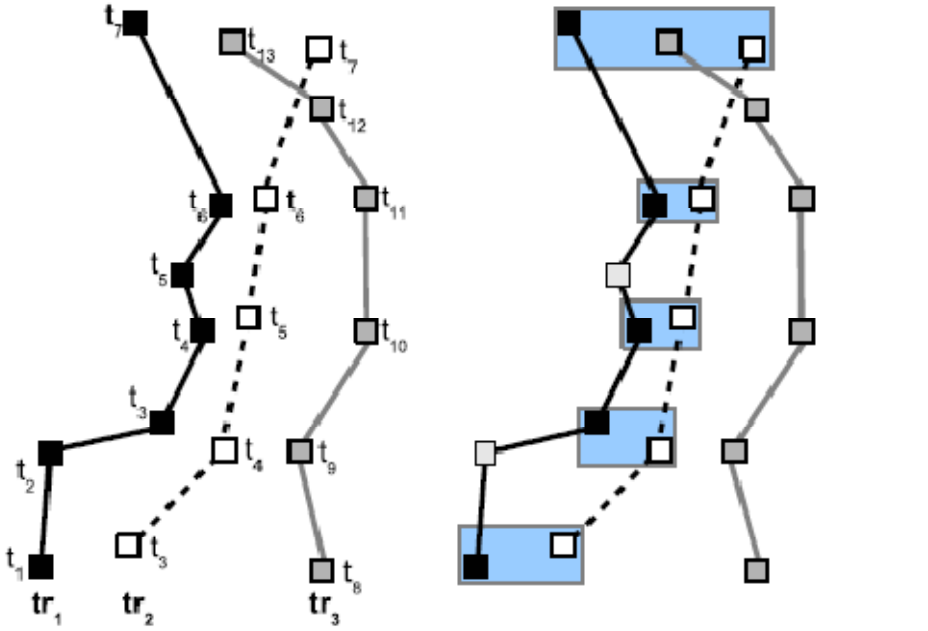
Towards Trajectory Anonymization (Nergiz ACM GIS '2008)

- Generates a sanitized dataset that consists only of K-anonymous sequences in two phases.
- 1st: trajectories incorporated into k groups based on a similarity measure that quantifies the cost optimal anonymization.
- 2nd: It computes a matching point between the points of the pair of the trajectories that have been clustered. The matched points in a pair of trajectories are placed by their Minimum Bounding Rectangle (MBR) while the unmatched points are suppressed.

Trajectory Anonymization (AWO)

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

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



Trajectory Anonymization (AWO)

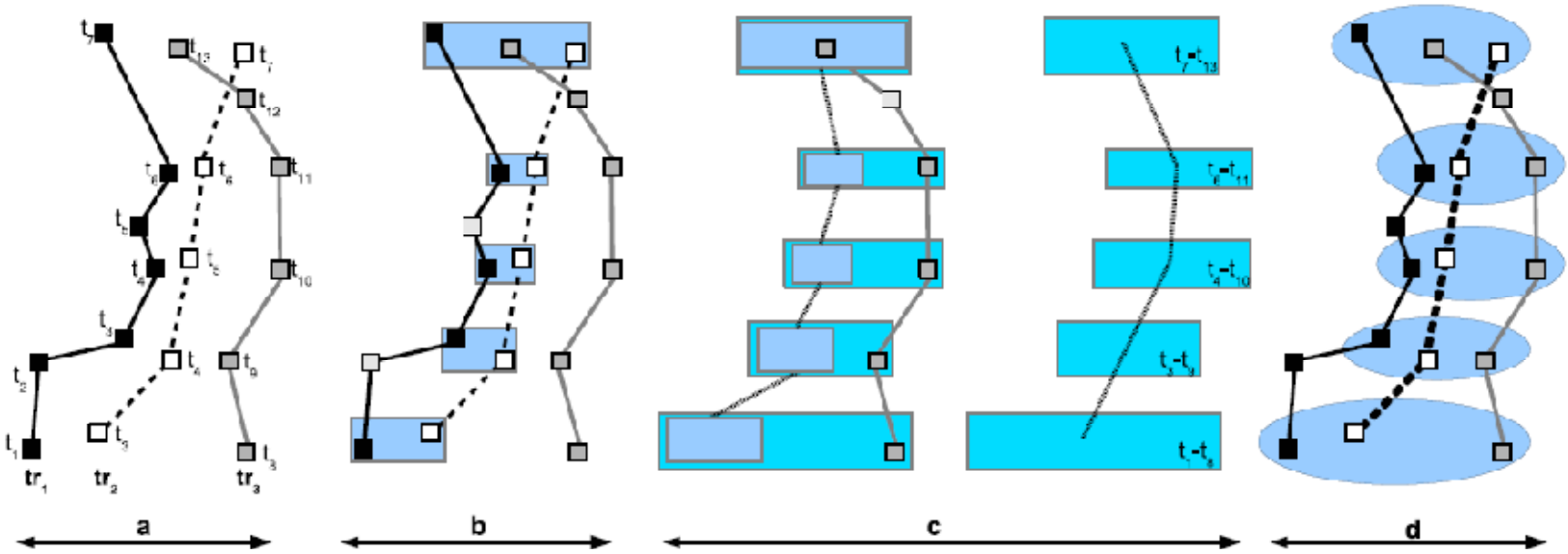


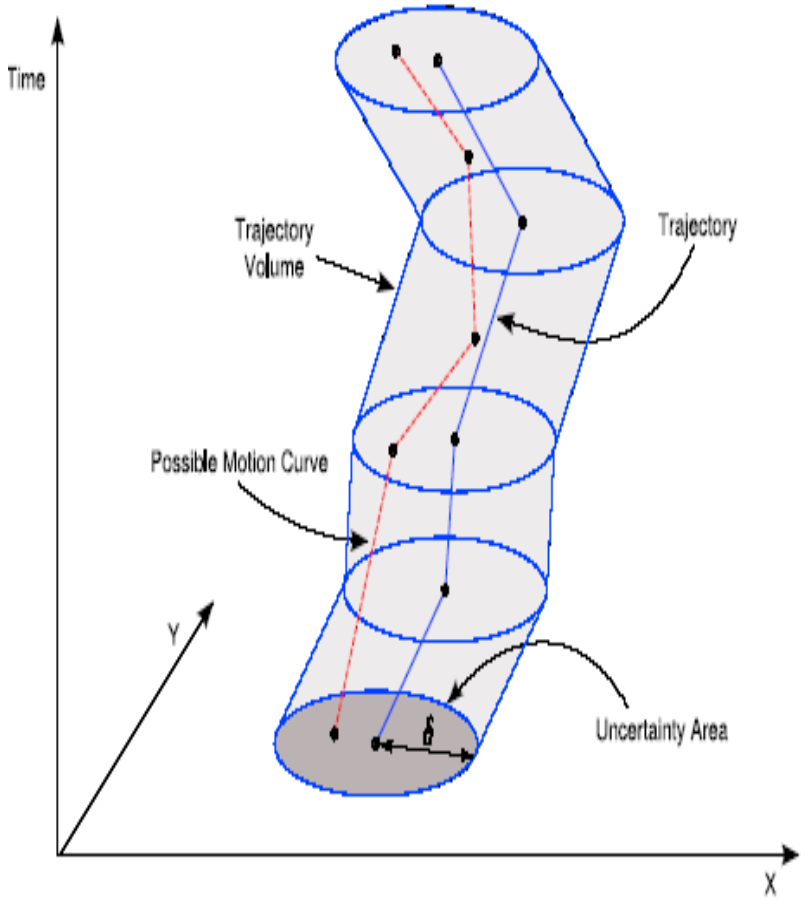
Figure 1. Anonymization Process
a. trajectories tr_1,tr_2, and tr_3; b. anonymization tr^* of tr_1 and tr_2; c. anonymization of tr^* and tr_3; d. point matching used in the anonymization of tr_1,tr_2, and tr_3. Matching contains five point links

Never Walk Alone (Abul et al ICDM '08)

- Basic Idea
  - To exploit the inherent uncertainty of moving objects position for enforcing anonymity with less information loss.
- Main contribution
  - Concept of (k,δ)-anonymity

Uncertainty and Trajectories

- The trajectory of the moving object is within a cylinder.
- But we do not know exactly where.
- If another object moves within the same cylinder they are indistinguishable from each other



Anonymity and Trajectories

- NWA is developed along three main phases:
  - Pre-processing: aimed at enforcing larger equivalence classes of trajectories w.r.t. same time span.
  - Clustering: based on GC method and enhanced with techniques to keep low the radius of produced clusters, at the price of suppressing some outlier trajectories.
  - Space Translation: transforming each cluster found into a (k, δ)-anonymity set.
- Limitations
  - Only trajectories starting and ending at the same time can be clustered together.

Contents

- Introduction
- Opportunities, Privacy Threats and Law Directions
- K –anonymity in Relational Databases
- Anonymity techniques in mobility data analysis
  - Sequence Hiding
  - Sequencial pattern hiding
- Privacy and anonymity in Location Based Services

Pattern Hiding

- So far we have seen how to anonymize the original data.
- But what happened if we try to anonymize the patterns that have been extracted from the original data?
- A few approaches have been proposed based on anonymizing the extracted patterns.

Sequential Frequent Patterns (Pensa '08)

- Pattern hiding methodology that removes all the infrequent subsequences from the original dataset
- First generates a prefix tree based on the sequences of the original dataset.
- The infrequent subsequences are pruned away from the tree in order to anonymize the prefix tree.
- The subsequences that have been removed are re-appended to the prefix tree. As a result the support of each frequent sequence decreases.
- Finally, the algorithm generates the sanitized dataset from the sequences of the prefix tree.

Sequential Frequent Patterns

Dataset: D

B C
A B C D
A B C D
B C E
B C D

Minimum support = 3

SFP (D): S

B
C
D
B C
B D
C D
B C D

A : occurs only 2 times in D

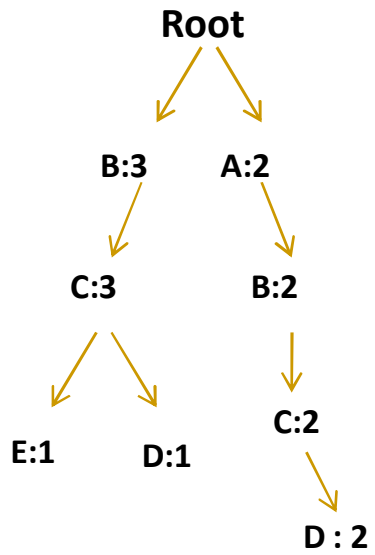
C B: does not occur (order is important!)

Prefix Tree Construction

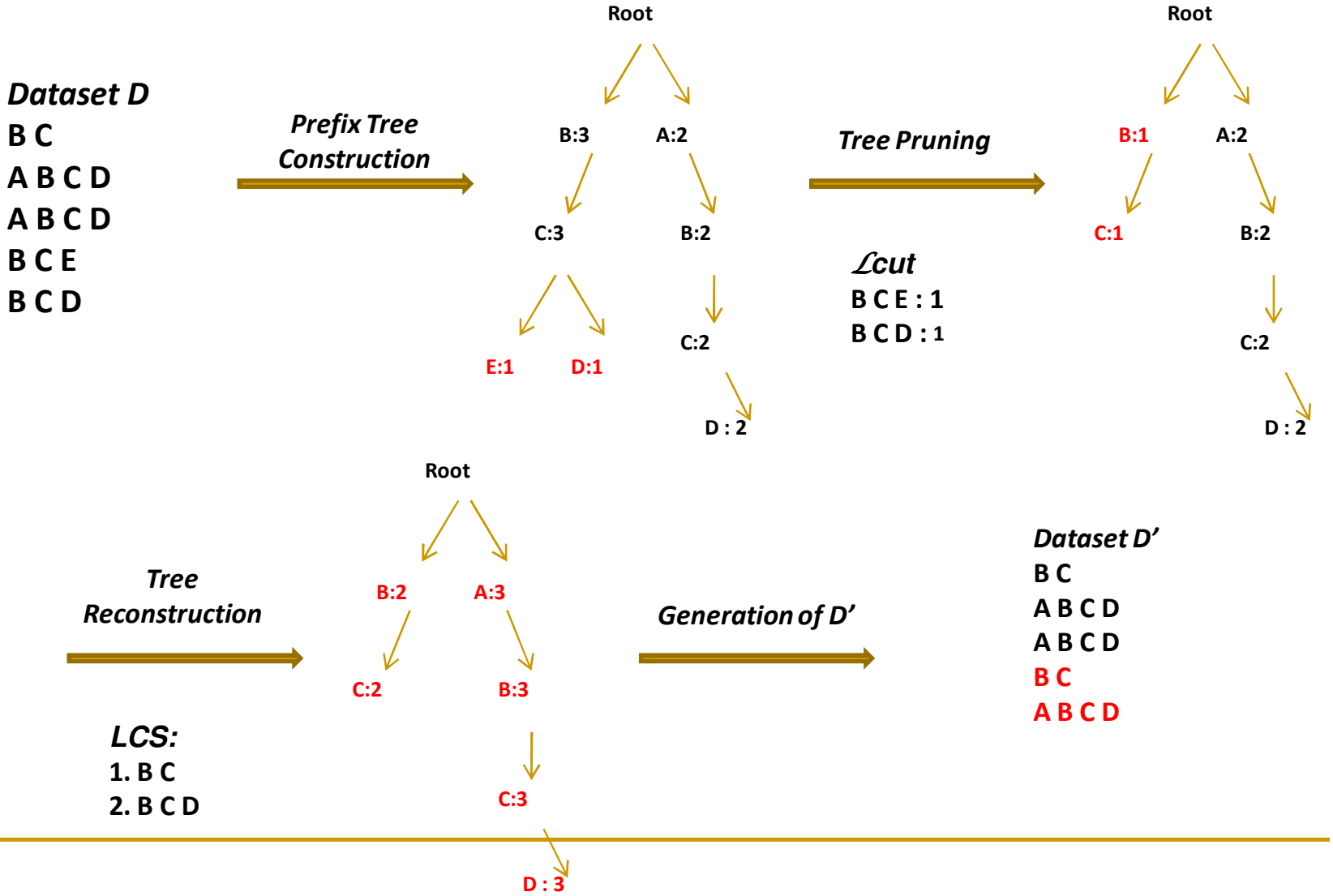
Dataset D

B C
A B C D
A B C D
B C E
B C D

Prefix Tree Construction



Running example: k = 2

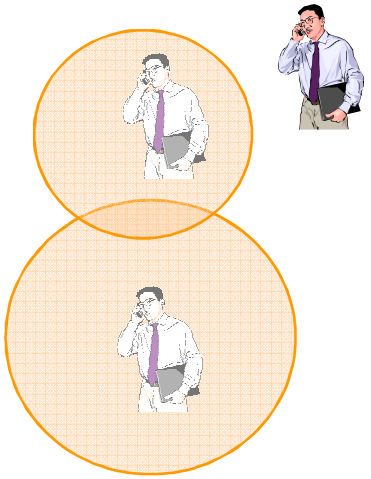


Contents

- Introduction
- Opportunities, Privacy Threats and Law Directions
- K –anonymity in Relational Databases
- Anonymity techniques in mobility data analysis
  - Sequence Hiding
  - Sequencial pattern hiding
- Privacy and anonymity in Location Based Services

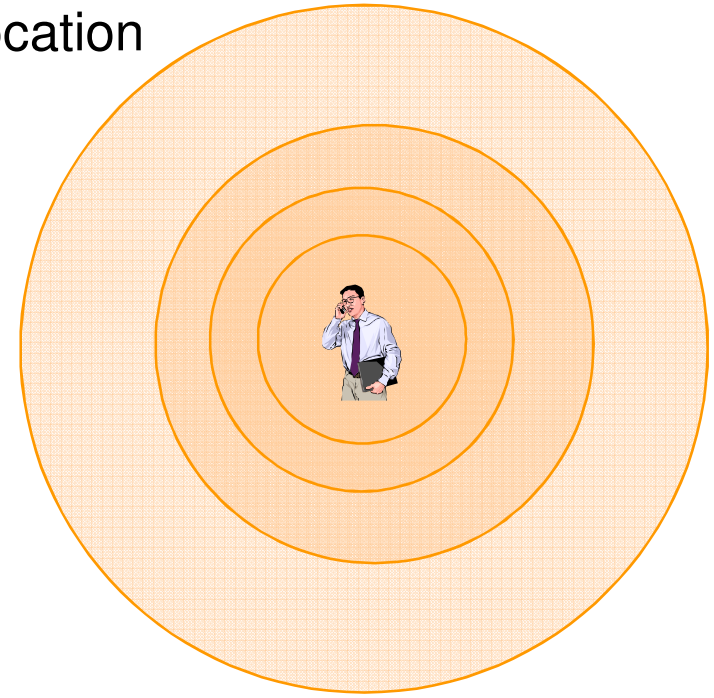
Concepts for Location Privacy

- The aim is to provide the service without learning user's exact position, and the data can also be forgotten once that the service has been provided.
- The user is represented with a wrong value.
- The privacy is achieved from the fact that reported location is false.
- The accuracy and the amount of privacy depends on how far the reported location from the exact location.



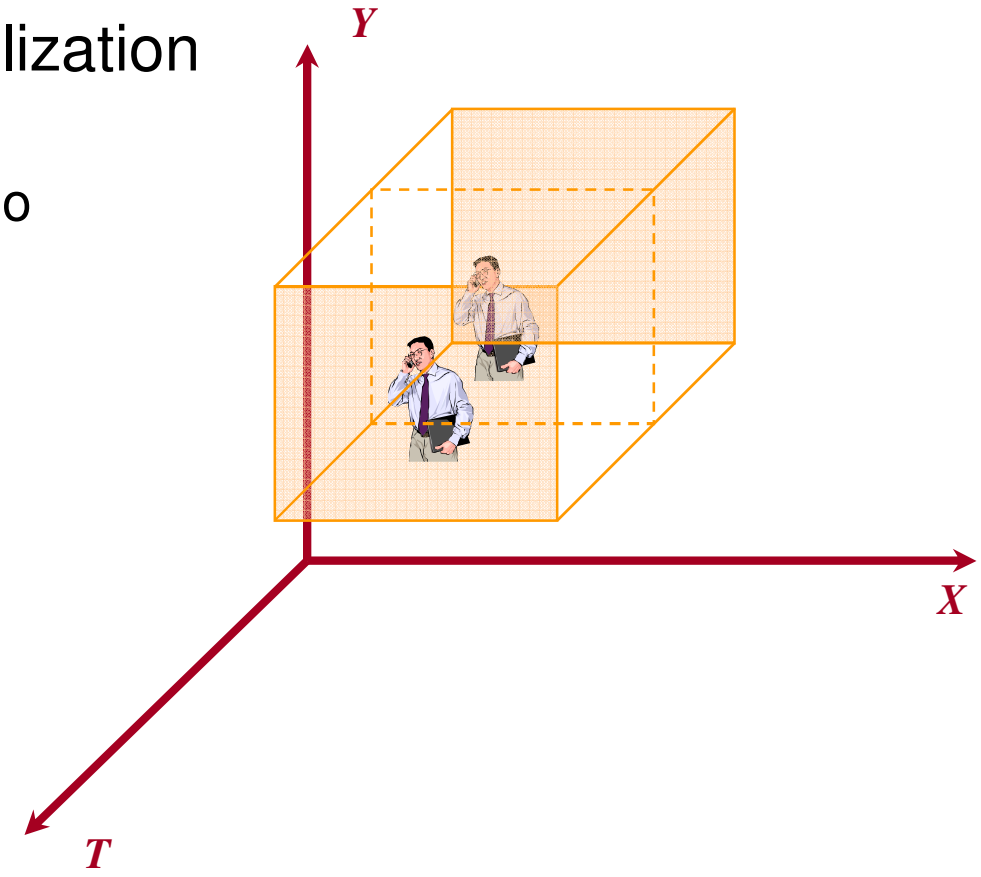
Concepts for Location Privacy

- 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



Concepts for Location Privacy

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



K-anonymity in LBS.

- The idea is to require that a user belongs in a group of at least K-1 others users prior to sending a continuous query for the provision of an LBS.
- Users may leave their groups upon completion of the requested service but no one is allowed to leave while a request in LBS is in progress.

Location Based Services

- Various methods have proposed, such as:
- A spatial subdivision in areas, and on *delaying the request* as long as the number of users in the specified area does not reach k.
- Allows each message to specify an independent anonymity value k.
- The area in which location anonymity is evaluated is divided into several regions and position data is delimited by the region.
  - Ubiquity: a user visits at least k regions (location anonymity).
  - Congestion: the number of users in a reagon to be at least k (local anonymity).
- A mix zone is an area where the location based service providers can not trace users' movements.
  - When a user enters a mix zone, the service provider does not receive the real identity of the user but a pseudonym that changes whenever the user enters a new mix zone.

An anonymous communication technique using dummies (Kido ICPS '05)

- Introduces several false position data (dummies) along with the true locations of the users to protect the privacy of the requesters of LBSs.
- The challenge is to achieve realistic dummy movements that will confuse an adversary regarding the true locations of the user.
- The location of the first dummies are decided randomly
  - Moving in a Neighborhood (MN): the communication device of the user memorizes the previous position of each dummy. Then the device generates dummies around the memory.
  - Moving in a Limited Neighborhood (MLN): the device generates dummies around the memory that are the same as the MN algorithm. If there are many users in the generated region, the device generates the dummy again.

HERMES++

- HERMES++ is a privacy-aware trajectory tracking query engine. (*Gkoulalas-Divanis et al. 2008)*.
- Offers strict guarantees about what can be observed by untrusted third parties.
- In order to achieve K-anonymity it produces K-R fake trajectories.
- The dummies are kept in the database for future convenience.
- Supports a variety of queries (range, landmark, route query).
- Deals with identification and sequential tracking attacks.

Architecture of HERMES++

- *Big picture of HERMES++*

	ID	Dummy	Age	Income	...	Movement
real	a	N	35	$36,000		
	b	N	38	$29,000		
	c	N	26	$25,000		
	...					
fake	f	Y	31	$33,000		
	g	Y	35	$30,000		
	h	Y	33	$27,000