Privacy Protection through Anonymity in Location-based Services

Claudio Bettini

Data, Knowledge, and Web Engineering Lab. - Dip. di Informatica e Comunicazione
Università di Milano, Italy

Bolzano 2007

Location Based Service (LBS)

Location based service:

- internet service;
- provides information based on issuers location.

Example

"Give me the closest vegetarian restaurant".

Commercial impact of LBS

Currently: car navigation is the most popular LBS.
Future: more than 300 millions of users in 2011 (ABI research).

The intuitive reason

The technologies on which LBSs are based will become less expensive and more reliable:

- mobile device
- wireless communication
- positioning systems (e.g., dead reckoning, GPS)

Legal recognition of privacy

Privacy recognized as a human right

European Convention on Human Rights, Article 8

"Everyone has the right to respect for his private and family life"

National legislations provide directives to privacy protection.

- In Italy: legge 675/1996.

Privacy in LBS explicitly identified as a particular kind of privacy.

- In the USA: "Location Privacy Protection Act of 2001".

Directives on how to manage sensitive data:

- the HIPAA specifications.

Users' view of privacy

Social studies report that users:

- are becoming more aware about their privacy;
- perceive location information as particularly sensitive

Will privacy concerns limit the diffusion of LBSs?

Objective

Ultimate objective of this research field: allow each user to enjoy LBSs while protecting his/her privacy.

Current research efforts

One basic idea: obfuscate data in the request through a generalization algorithm ensuring a user-specified level of privacy and an acceptable quality of service.

- centralized anonymizer. Gruteser et Al. [Mobisys-03], Gedik and Liu [ICDCS-05], Mokbel et Al. [VLDB-06], Kalnis et Al. [TR-06]
- distributed anonymizer. Ghinita et Al. [WWW-07]

Current research efforts (2)

Other techniques/ideas:

- generate fake requests (Kido et Al. [ICDE-05])
- mix-zones (Beresford et Al. [PC-03])

Problems

- Informal description of attacks
- Unclear properties of proposed defense algorithms.

Our Project goals

We aim at providing:

- **Unifying formal framework for LBS context-aware privacy**
- **New methodology** to design generalization algorithms.
- **Classification of existing solutions** based on formal results.
- **New generalization algorithms**, proved to be correct through the framework.
- **Performance evaluation** through extensive experiments.

Partners and Sponsors

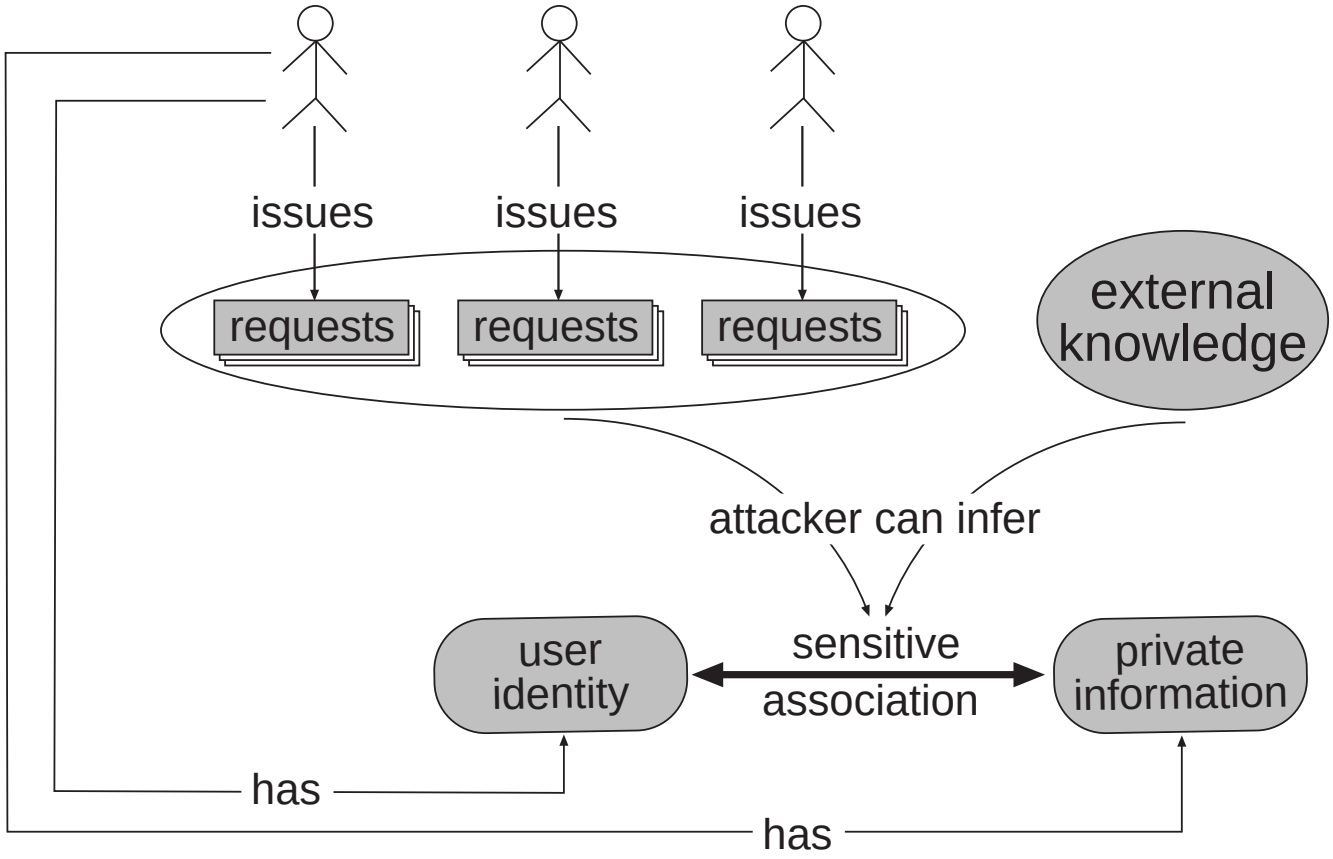
Joint Project with CSIS-GMU and CS-UVM, funded by NSF for the next three years. Mobility funded by MiUR Interlink project.

Impact on other areas

This research topic can also have impacts in the following areas:

- Release of database tables;
- Privacy preserving data mining.

General privacy threat in LBS



Private information

Examples of private information:

- political affiliations, health status, religious beliefs, sexual orientations, sensitive locations ...

Private information can be:

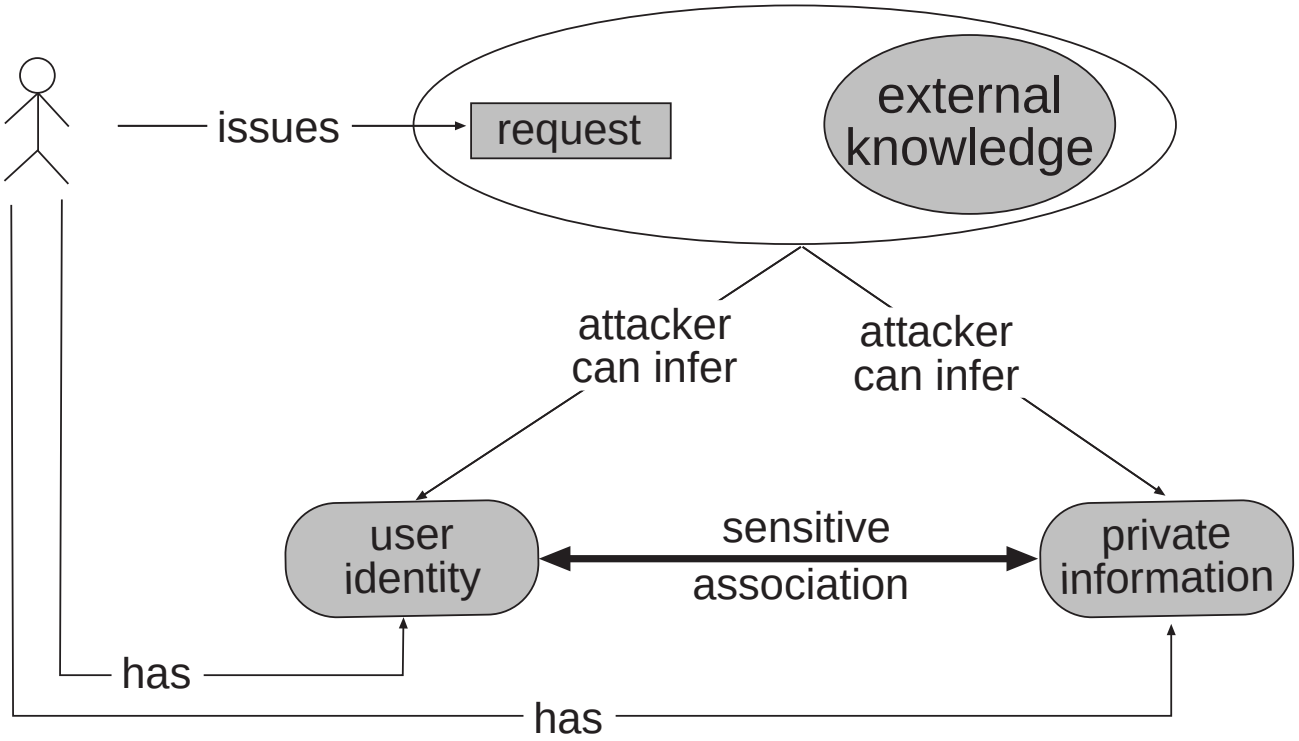
- part of the service parameters:
  - e.g.: "where is the closest religious building of religion X?"
- part of user's location
  - e.g.: user issuing a request while being in the red light district;
- inferred from parameters and/or location.

User's identity

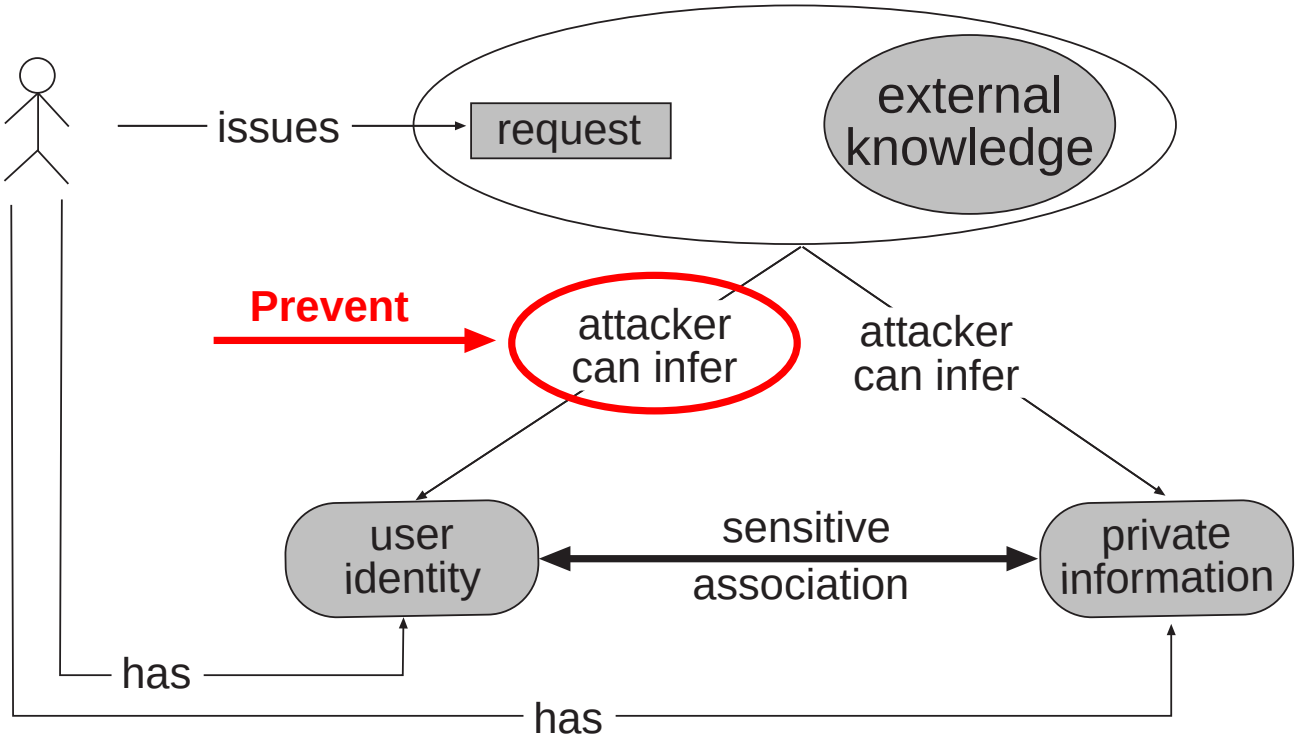
User's identity can be:

- explicitly specified in the request;
- inferred from:
  - the service parameters;
  - user's location;
  - a pattern involving one or both of the above.

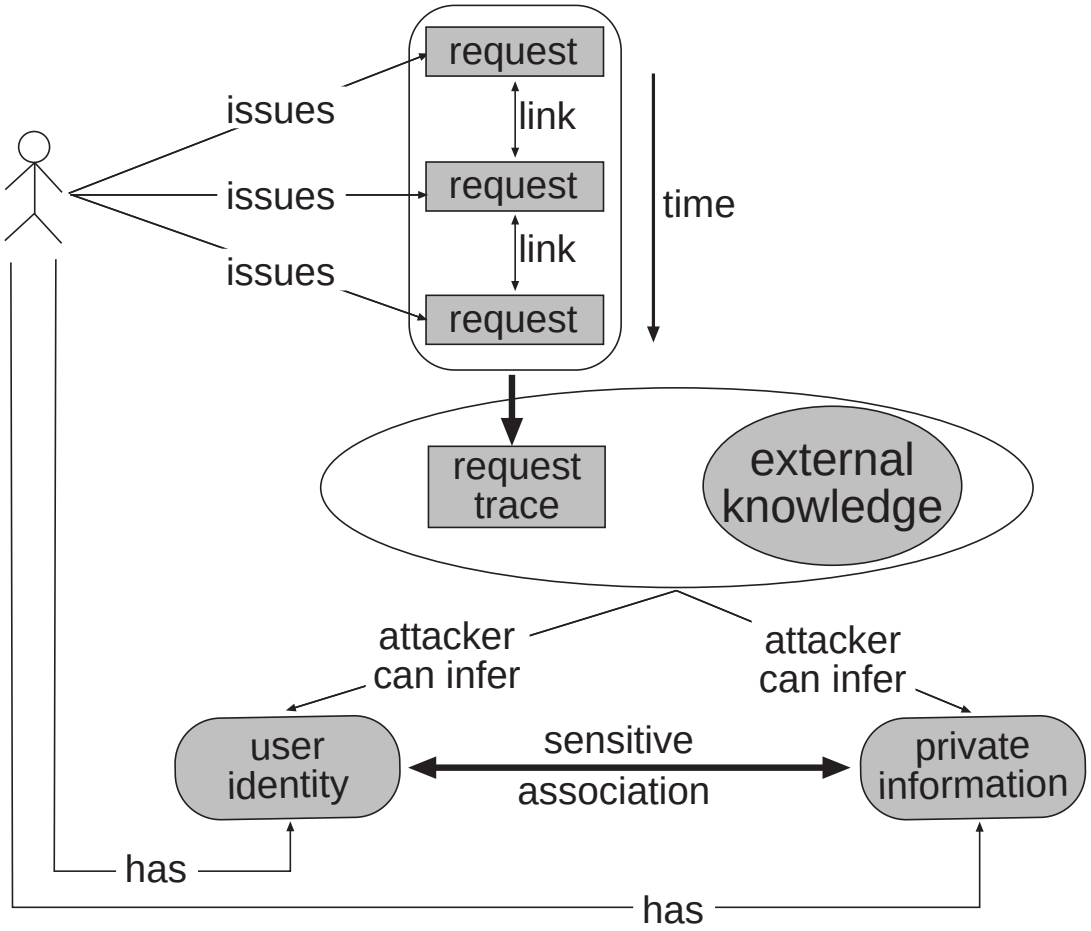
Static, single-issuer case



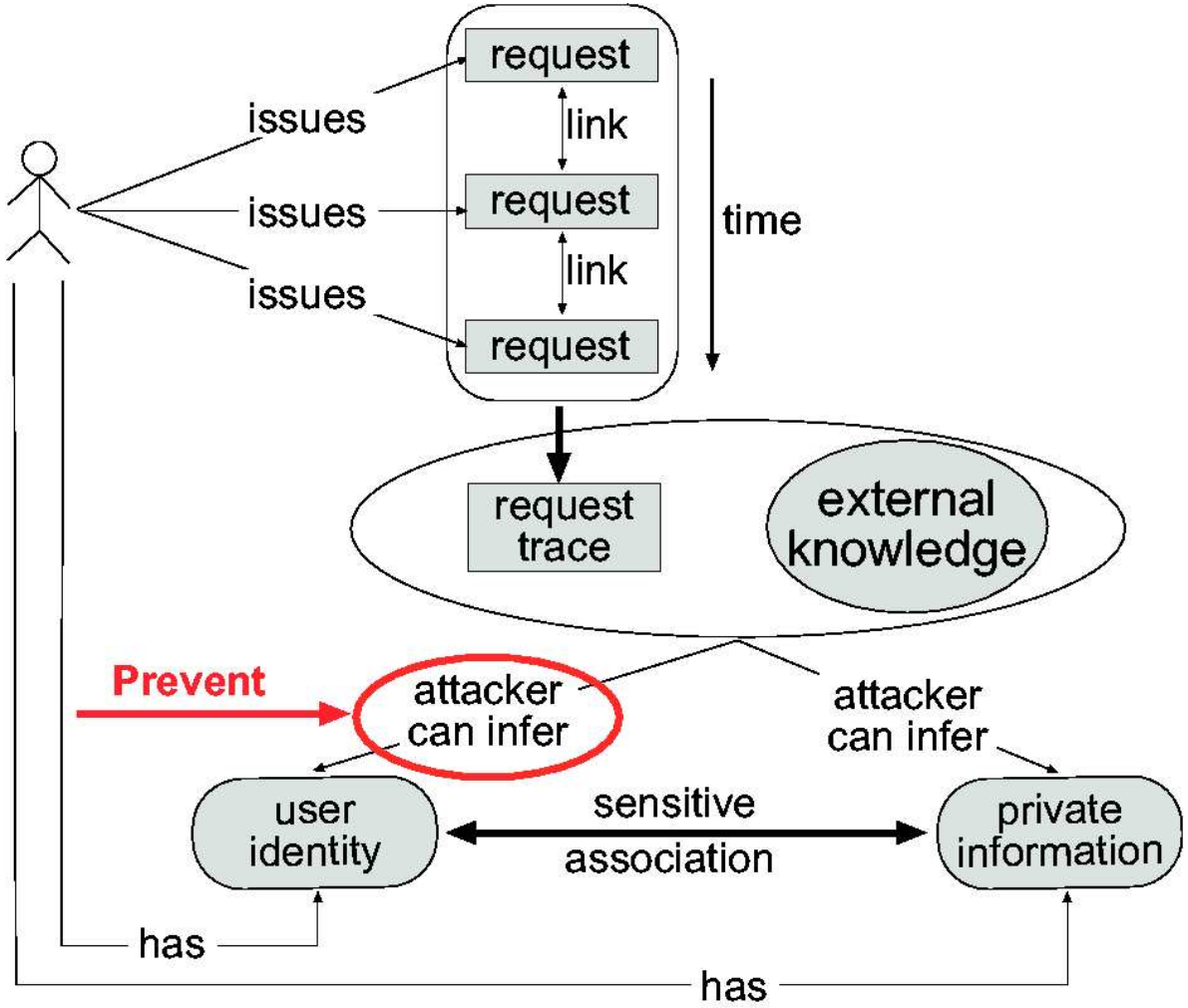
Static, single-issuer case



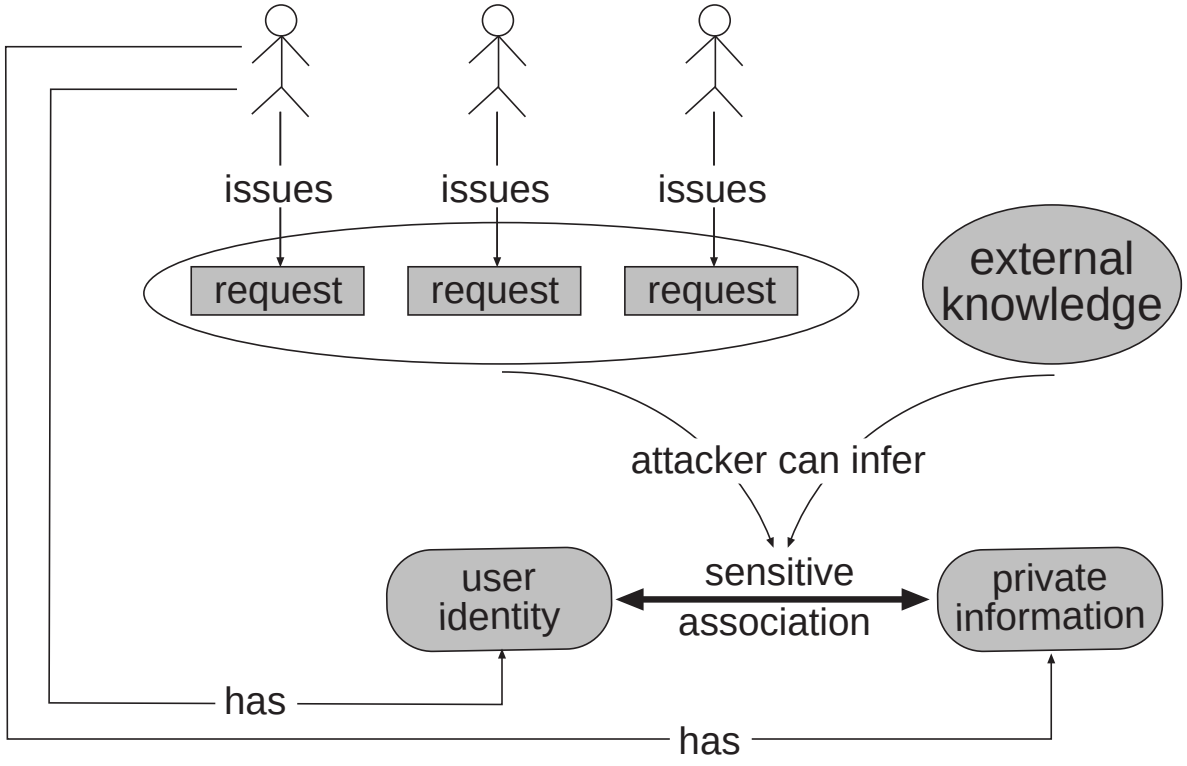
Dynamic, single-issuer case



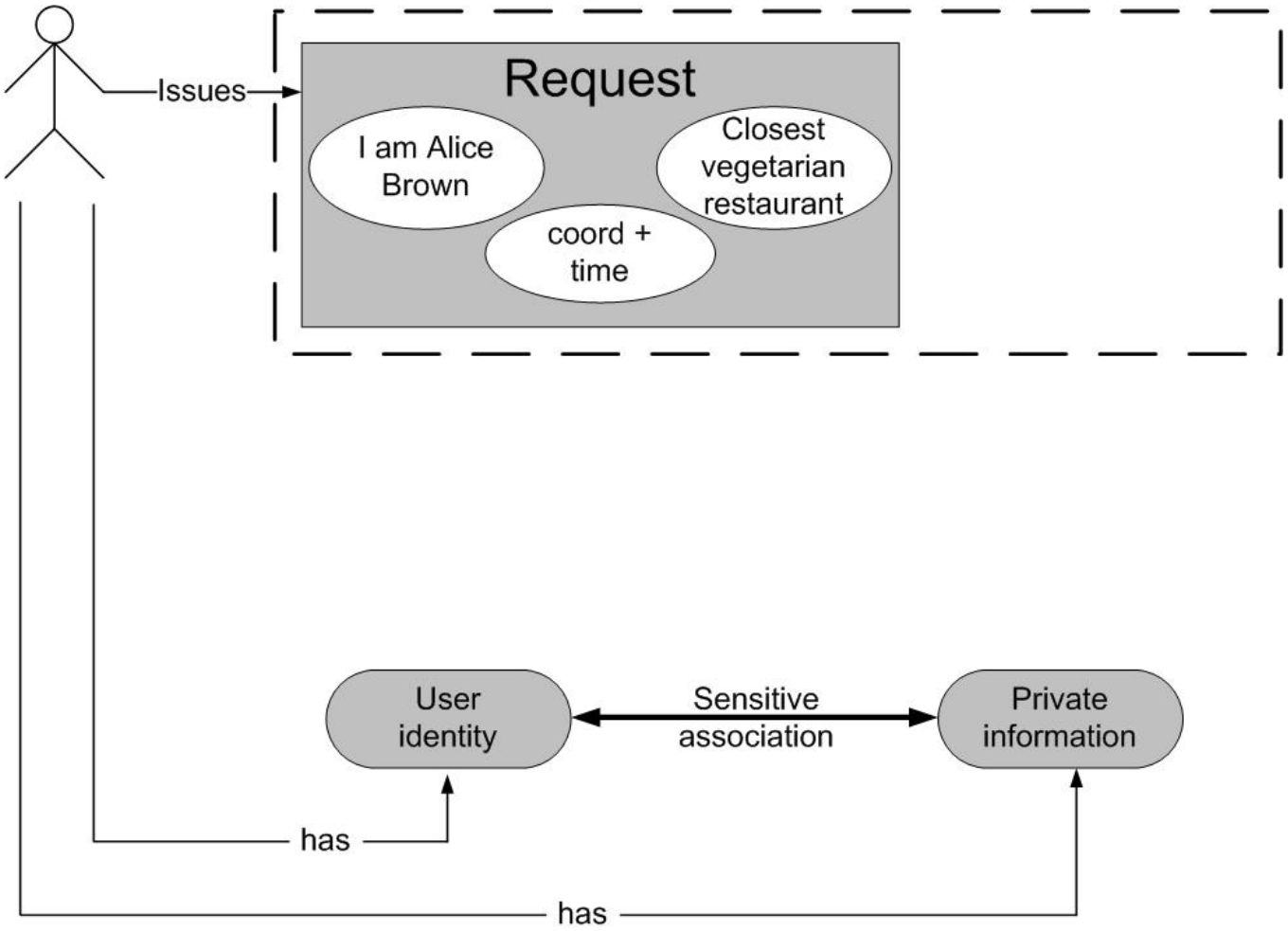
Dynamic, single-issuer case



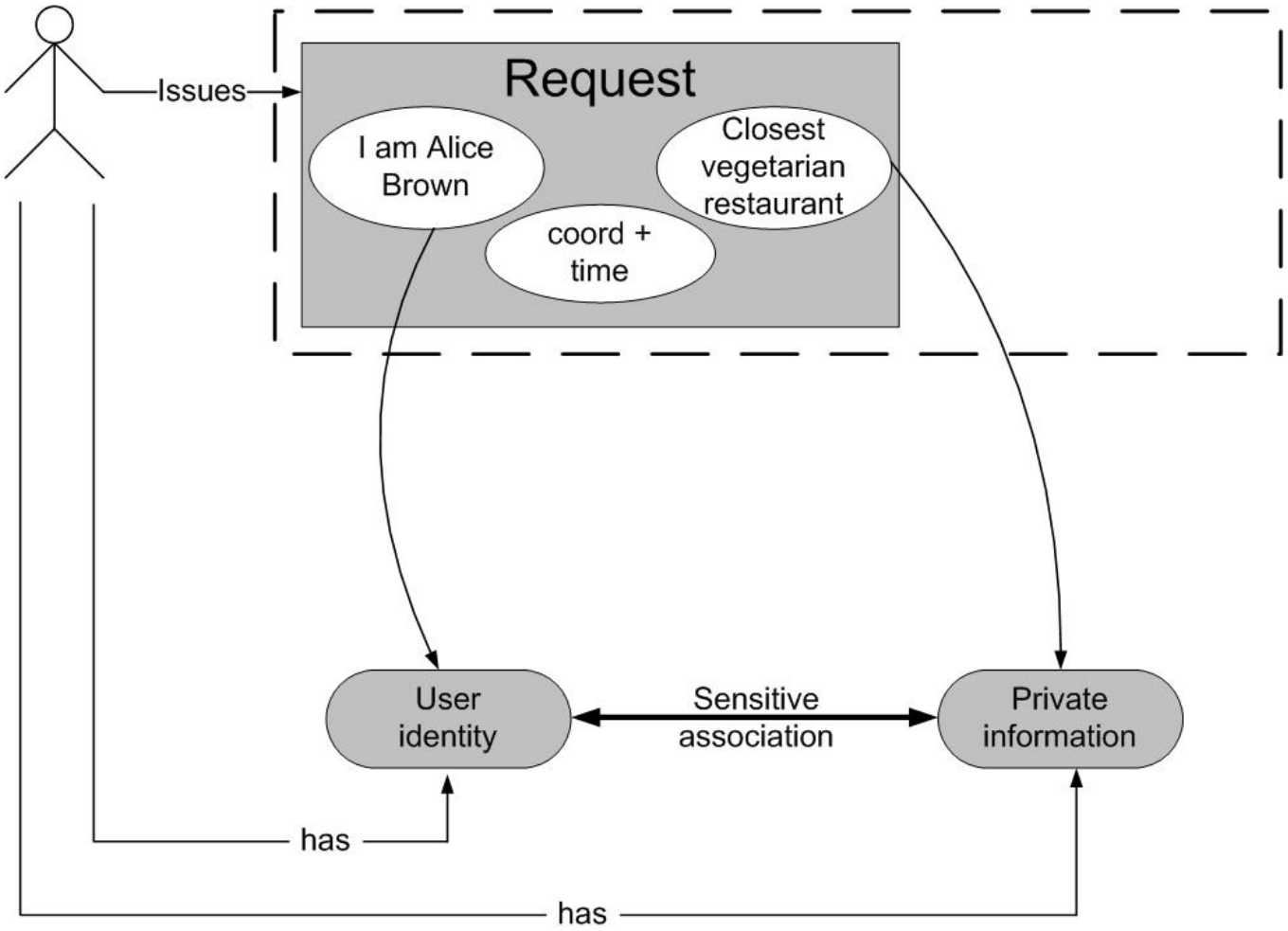
The static, multiple-issuer case



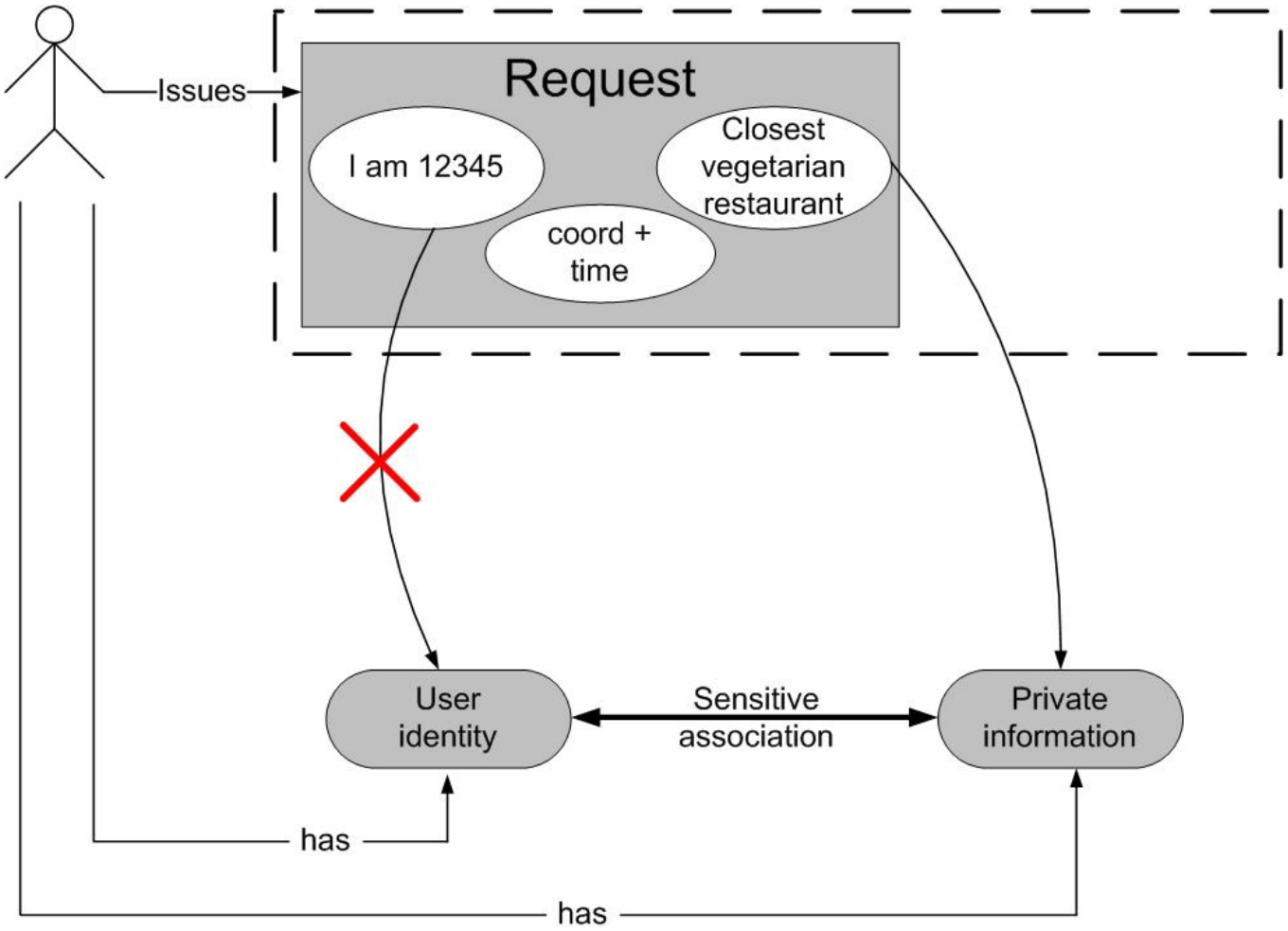
Example, the static case



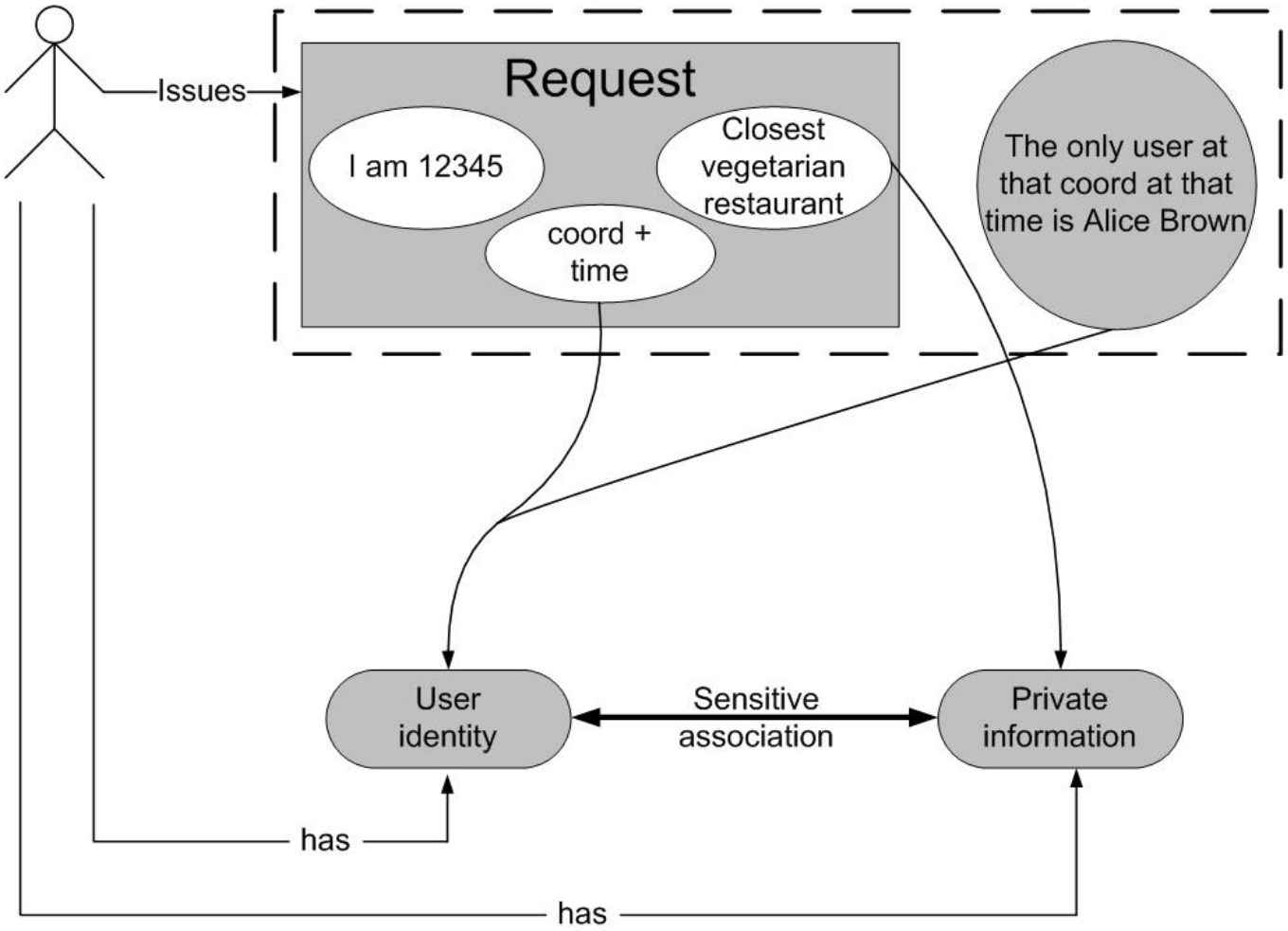
Example, the static case



Example, the static case



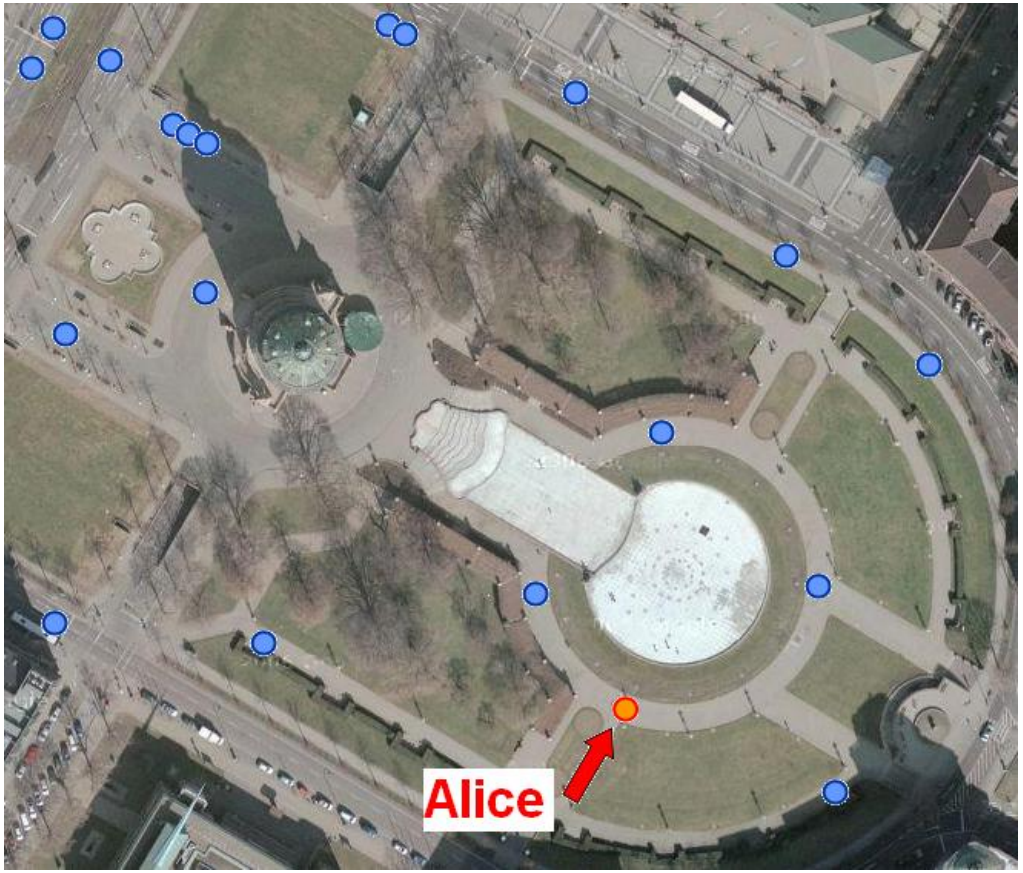
Example, the static case



Problem statement

How is it possible to guarantee that an attacker is not able to re-identify the issuer?

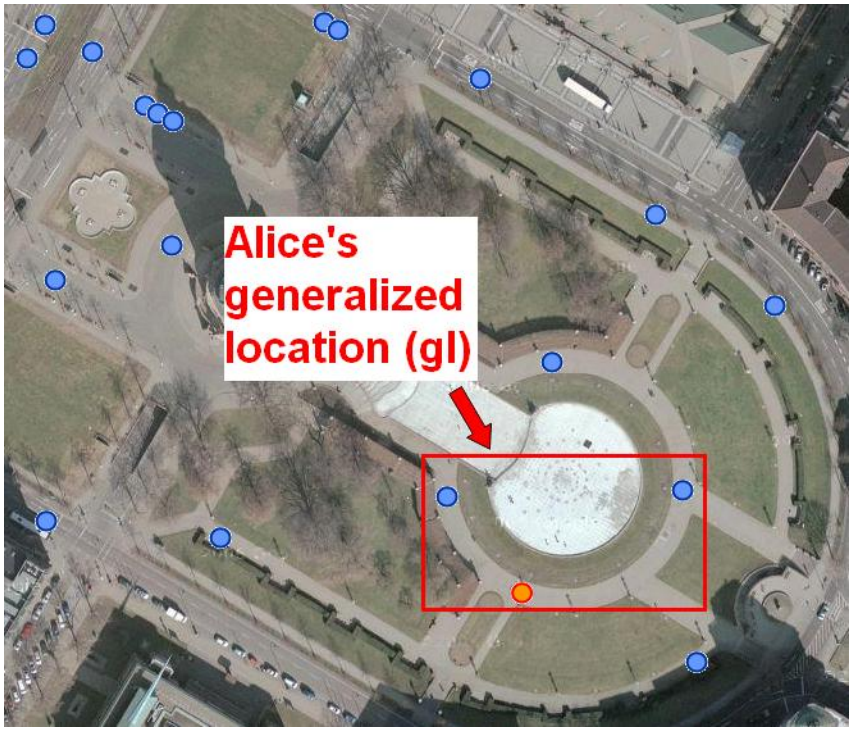
k-anonymity extended to LBS: example



Example (cont.)

Solution: **Generalize** user's location to a region with k users.

Alice's request: "the vegetarian restaurant closest to gl".



The attacker cannot identify the issuer in a group of 3 users.

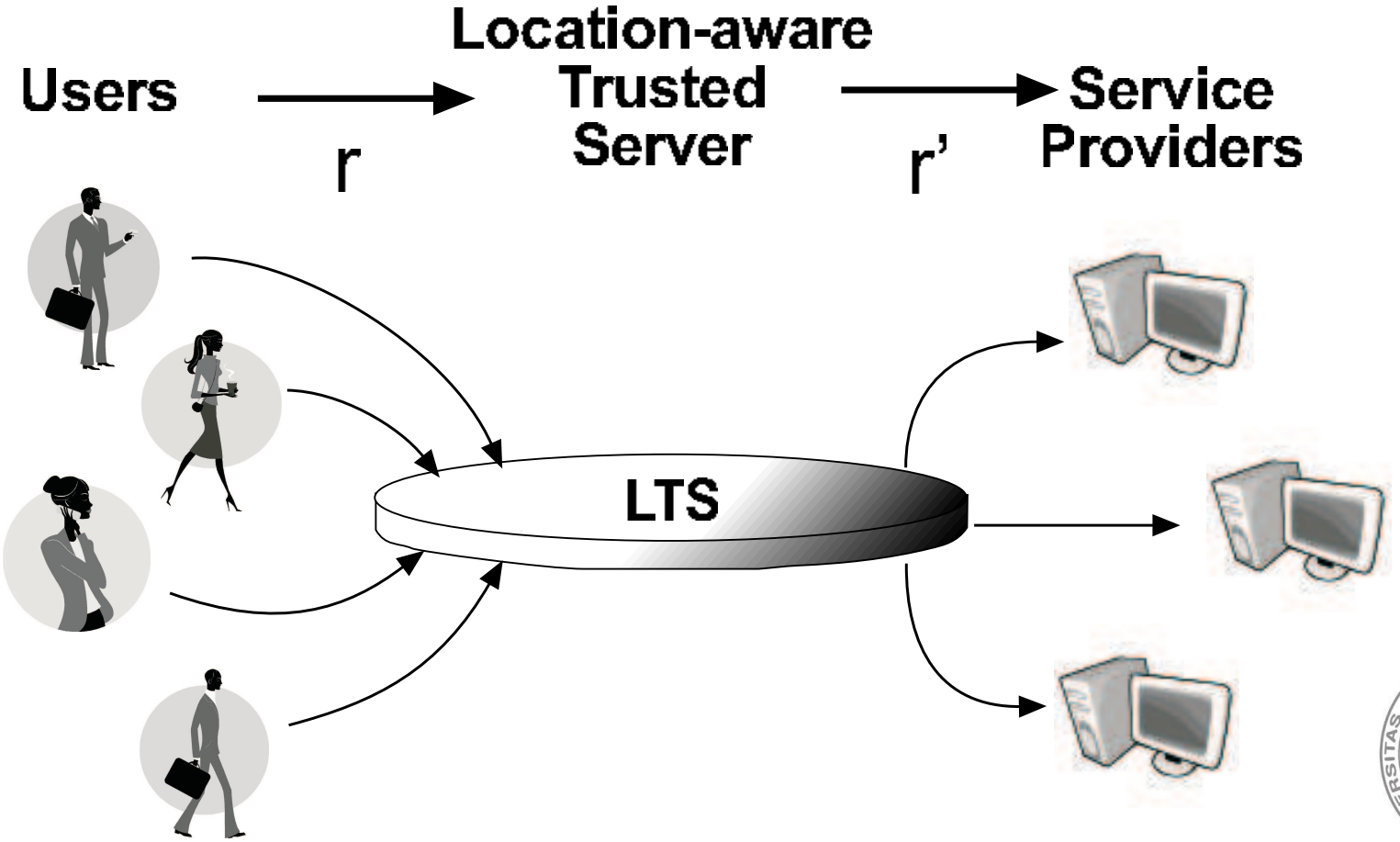
Example (cont. 2)

The response contains the set of vegetarian restaurants that are the closest to each point of gl.

- If gl is small, the result is similar to what obtained providing the exact location.
- If gl is too coarse, the set of items in the result may be large:
  - network overhead;
  - if results are filtered on the client, this implies computational overheads;
  - if results are not filtered, the user may be provided with many useless results.

Objective: produce small gl.

Scenario


Users
r
Location-aware
Trusted
Server
r'
Service
Providers
LTS

Explicit assumptions about attacker's knowledge

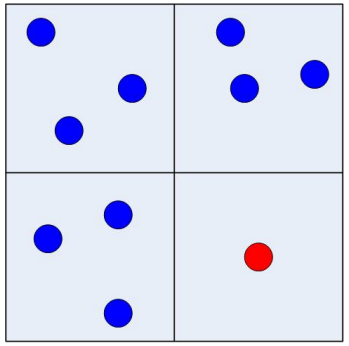
A context specifies the assumptions about attacker's knowledge and reasoning abilities.

Idea

Context must be explicit if the correctness of the generalization algorithm is to be proved.

Problems arising when context is not explicit

Beresford [PhDThesis-2005] showed a counterexample to Gruteser's defense called "the outlier problem".



The problem:

- previous papers implicitly assumes that the attacker knows users' location;
- Beresford's counterexample assumed the attacker also knows the generalization function.

The attack

An attack $Att_C(r', i)$ is the likelihood the attacker has about the fact that a user i issued a request r'.

- the context C specifies which knowledge and reasoning abilities the attacker has.

Since the attack is a probabilistic distribution among the set of users in I:

$$\sum_{i \in I} Att_C(r', i) = 1$$

The defense

A request r' is **safe** against an attack Att_C with threshold h if the attack cannot recognize the correct issuer of r' with likelihood greater than h.

$$Att_C(r', issuer(r')) \leq h$$

A function that transforms all input requests into safe requests is a **defense function**.

An algorithm that computes a defense function is a **defense algorithm**.

The static case

In the **static case** the attacker is not able to understand that a set of requests is issued by the same (anonymous) user.

Example

A service in which no pseudo-identification is required and in which requests are sporadic.

Specification of context C_{st}

Many papers (implicitly) considered context C_{st}:

- the attacker knows the exact location of each user at each time instant.

Why C_{st}:

- attacker may know the identity of the users that are in **some** locations AND
- the LTS does not know **where** the attacker can identify the users AND
- we want a **conservative** approach THEREFORE:
- assume the attacker knows the identity of each user in **each** location.

Our contribution in context C_{st}

We prove, in terms of our framework, the correctness of the existing generalization algorithms:

- a k-anonymous request is safe against $Att_{C_{st}}$ with threshold $1/k$

We designed the "optimal" algorithm that computes the generalization having the smallest generalized area;

- not practically applicable, but useful as a benchmark in the experiments.

We implemented the algorithms and obtained experimental results.

Specification of context C_{st+g}

In the "outlier problem" it is implicitly assumed that the attacker knows the generalization function.

In context C_{st+g} the attacker knows:

- the exact location of each user at each time instant;
- the procedure used by the LTS to compute the generalization.

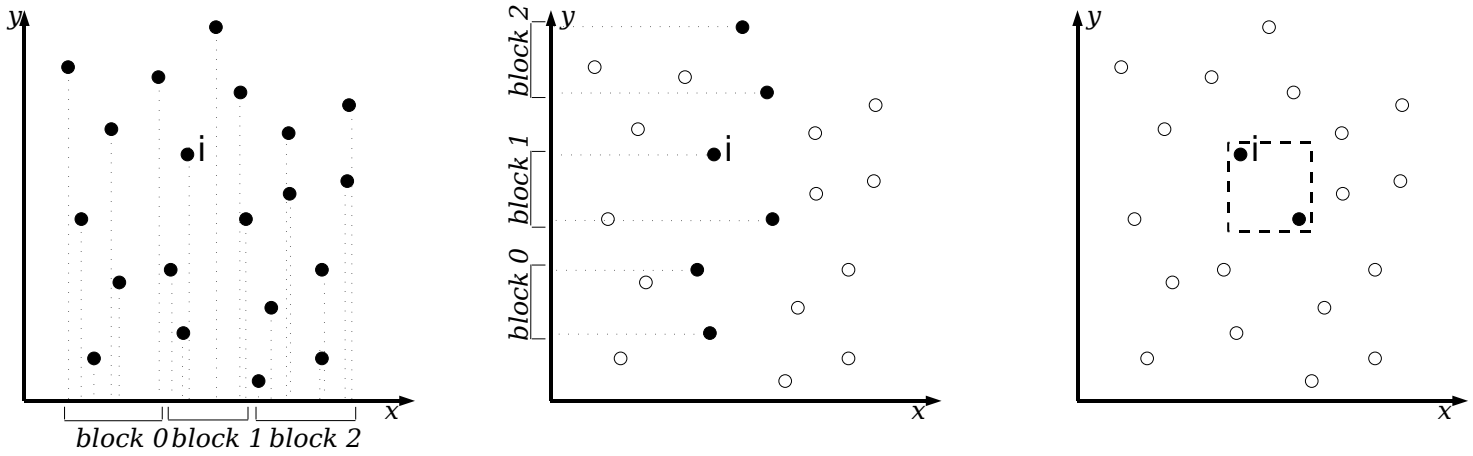
Our contribution in context C_{st+g}

We prove that the only generalization algorithm proposed in the literature that does not suffer the "outlier problem" is a defense algorithm against $Att_{C_{st+g}}$

We propose two defense algorithms and we prove their correctness.

We implemented the algorithms and obtained experimental results.

The Grid algorithm



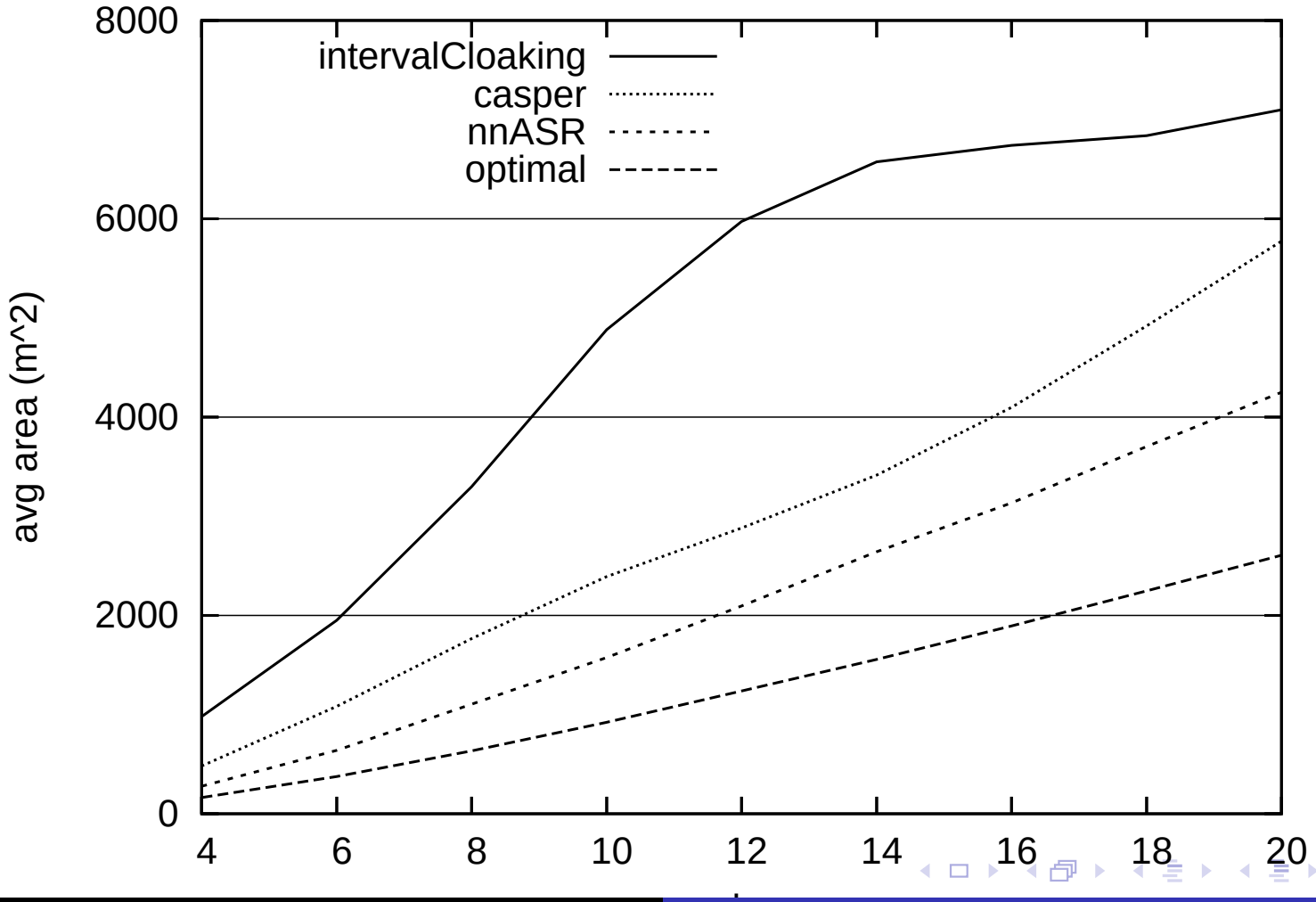
(a) First iteration (b) Second iteration (c) Third iteration

Experimental settings

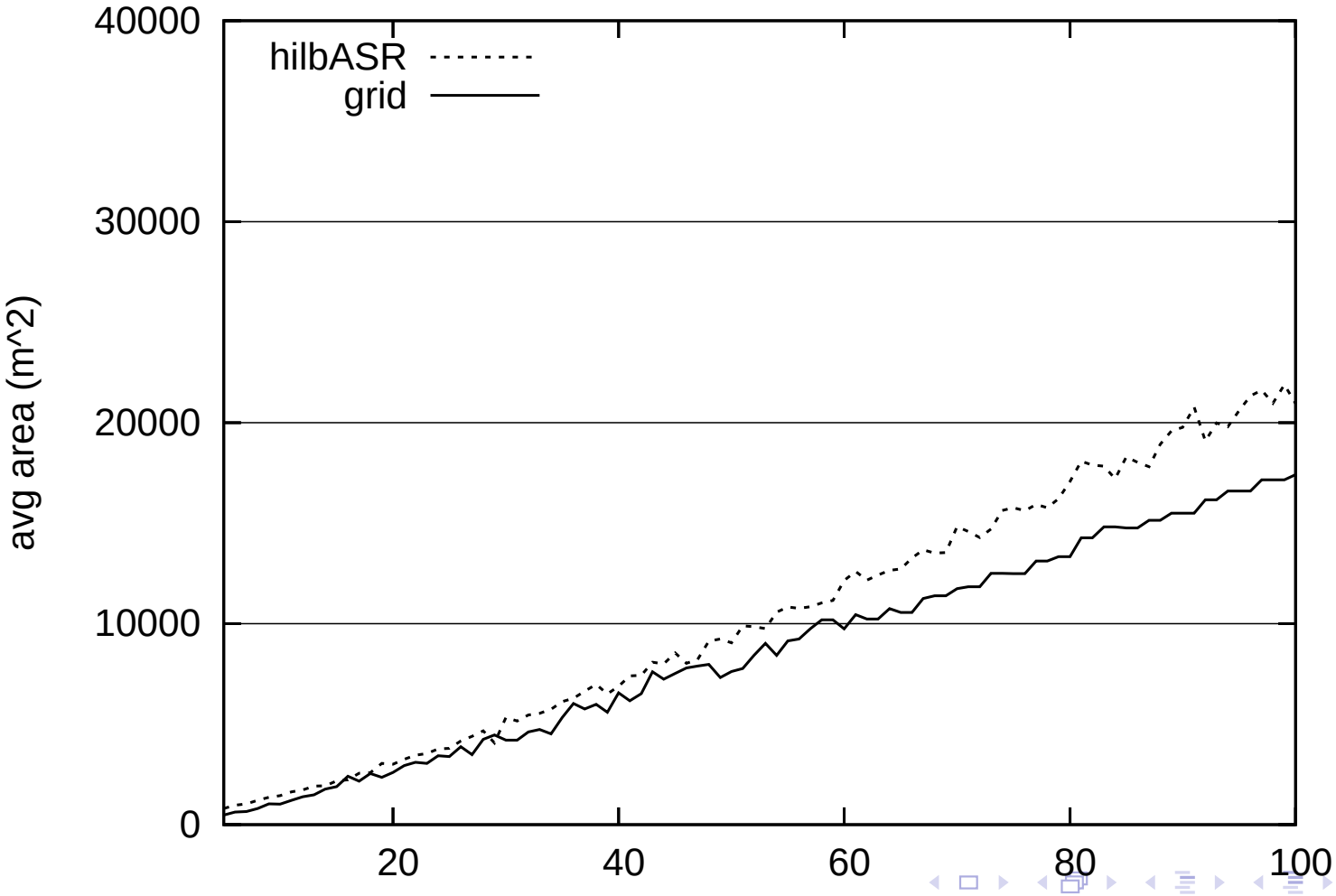
- **Total area:** 100 Km^2;
- **Number of users:** $500,000$;
- **Average density:** $5,000$ users / Km^2;
- Average values obtained through $1,000$ tests.

User's locations generated using a moving object generator [Brinkhoff-GeoInformatica2002].
Users move on the road network of the city of San Francisco.

Comparison of defense algorithms against $Att_{C_{st}}$



Comparison of defense algorithms against $Att_{C_{st+g}}$



The dynamic case

- Users make multiple requests to the same SP
- The attacker can obtain multiple requests and understand they are issued by the same anonymous user (e.g., by comparing pseudo-ids used for accounting/personalization)

Linked requests

Two or more requests are said to be linked if they can be associated by the attacker to the same issuer

Specification of contexts C_{st+pid} and $C_{st+g+pid}$

Contexts C_{st+pid} and $C_{st+g+pid}$ are the extension to the dynamic case of contexts C_{st} and C_{st+g}, respectively:

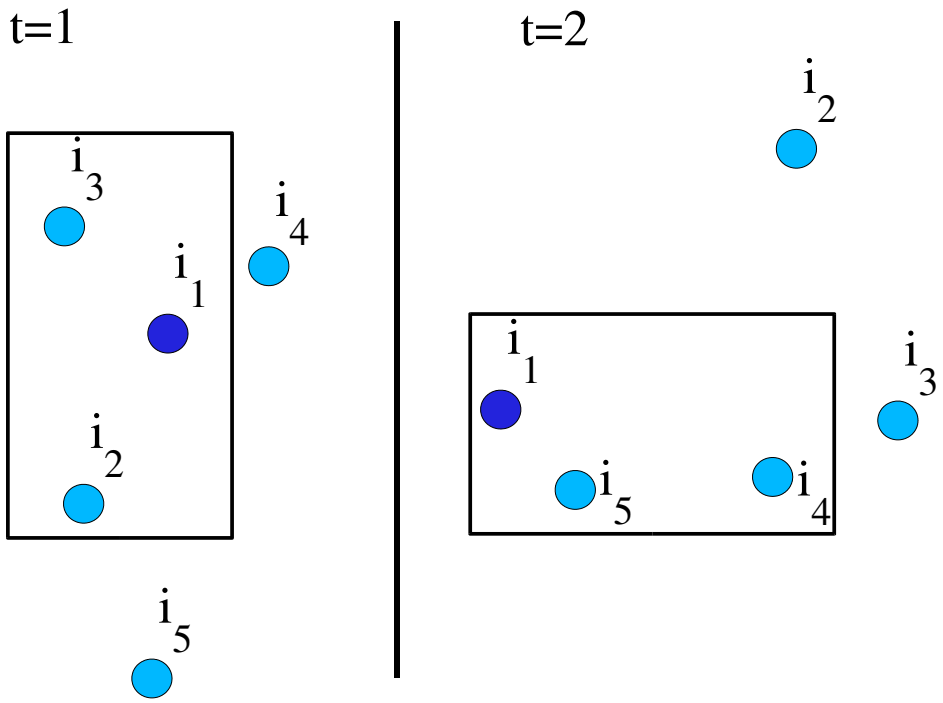
- the attacker can obtain and link the requests issued with the same pseudo-identifier.

Intuition

More extensive generalization is needed, since it is necessary to find a set of user that are "moving together" with the issuer.

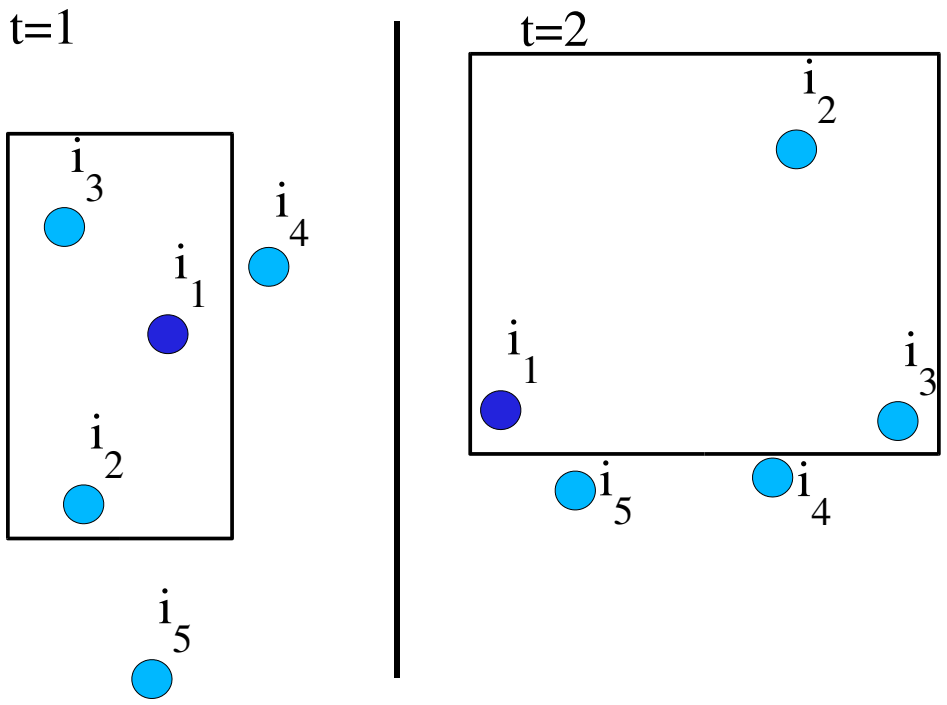
Example of attack under C_{st+pid}

The generalization of each request with a defense algorithm against $Att_{C_{st}}$ is not a defense against $Att_{C_{st+pid}}$ [BettiniEtAl-SDM05].



Example of attack under C_{st+pid}

The same users that are in the generalized region of the first request have to be in the generalized region of the second request.



Experimental settings

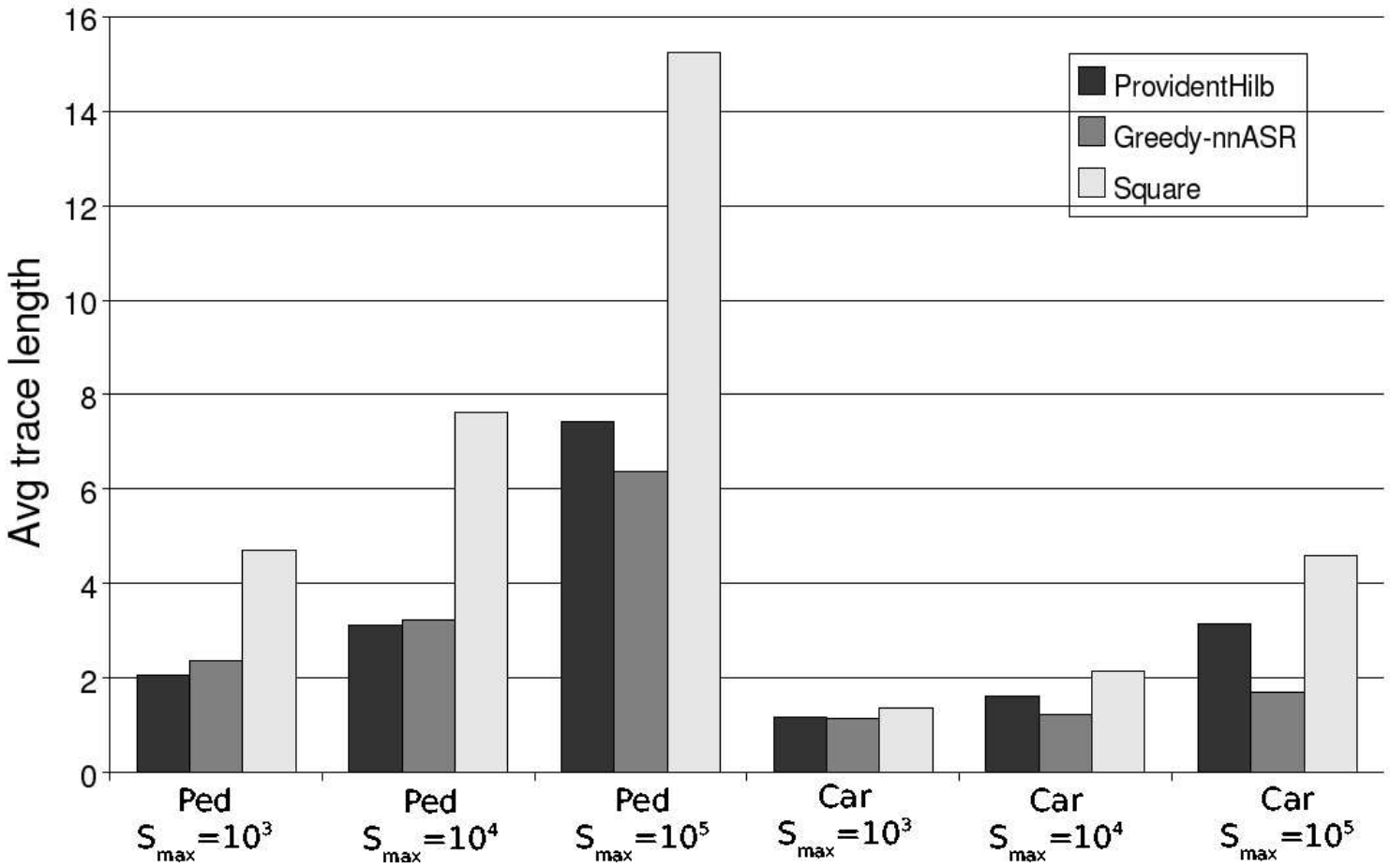
The generation of users' locations is similar to the static case.
In addition, movements of users are divided in 2 classes:

- car: moves up to 100 km/h;
- pedestrian: moves up to 4 km/h;

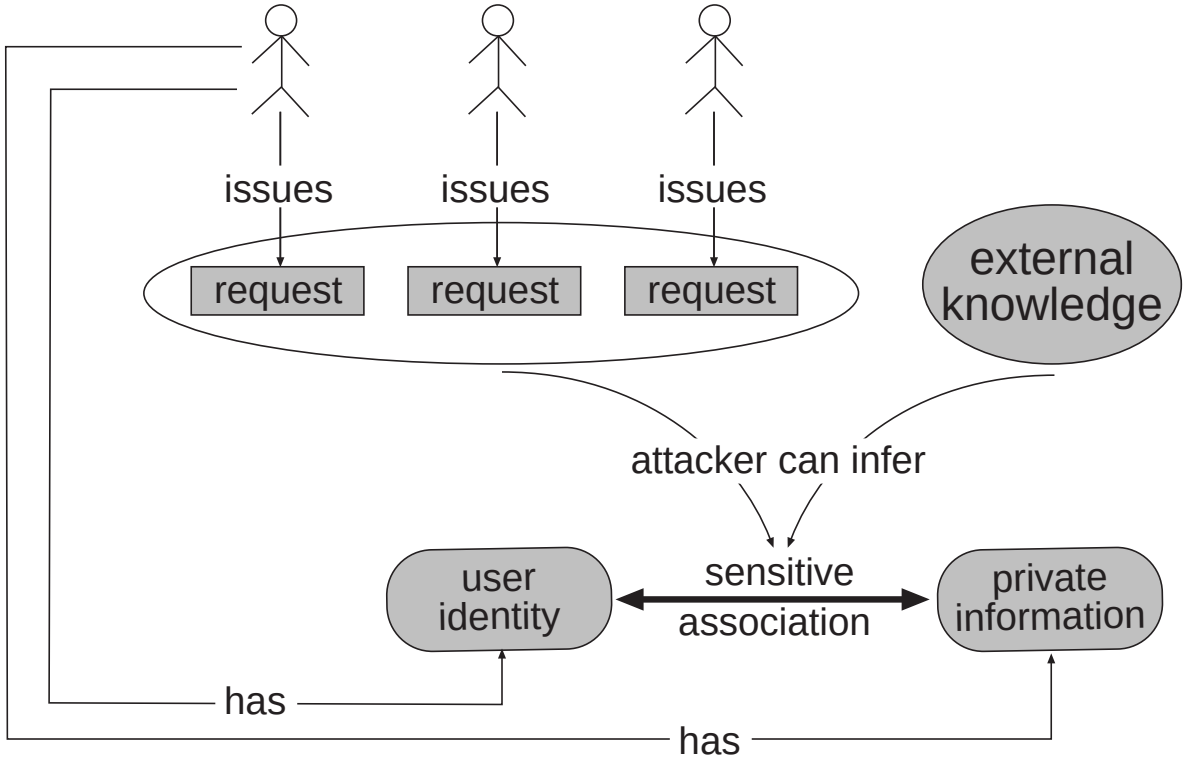
Each user issues a requests every minute.

Experimental results: Average length of traces

The value of h is fixed to $1/10$

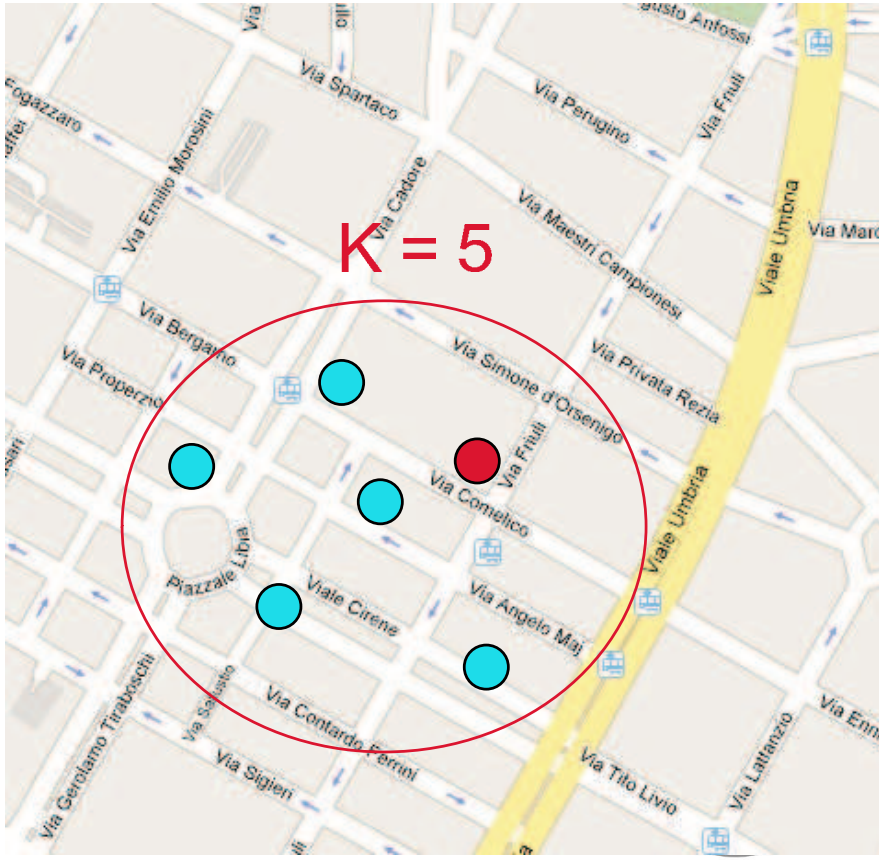


The static, multiple-issuer case



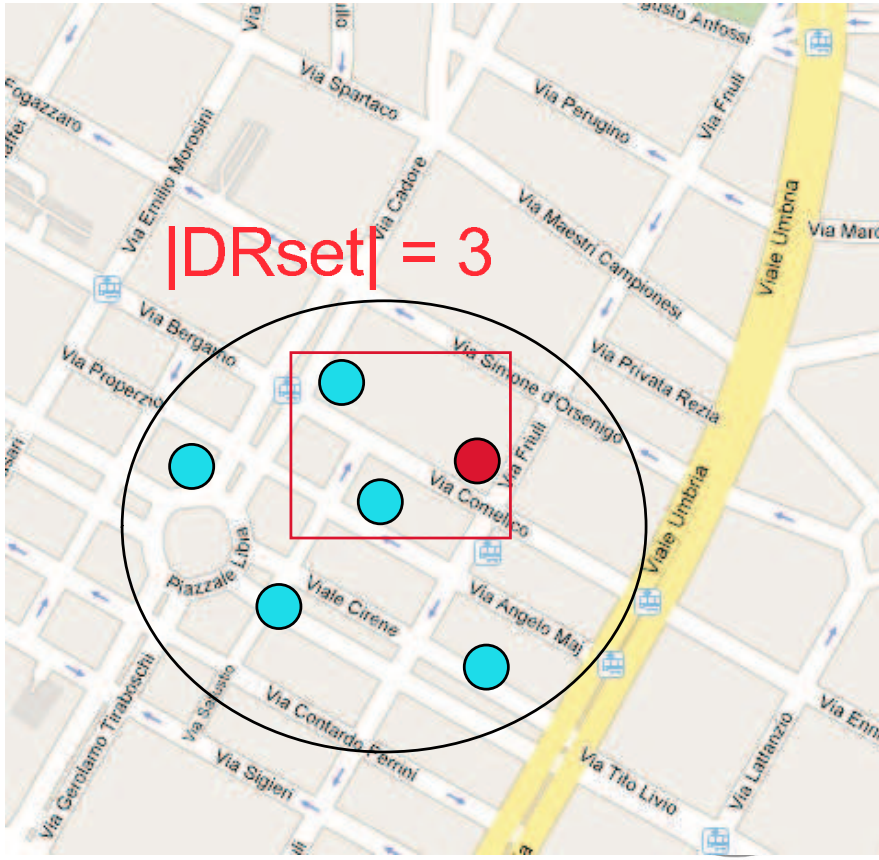
Anonymity and Diversity

- *Aset* = collection of users indistinguishable from the actual issuer
- *DRset* = collection of requests issued by users in the same *Aset*
  - requests in the *DRset* can be grouped according to the value of *PI* contained in the requests



Anonymity and Diversity

- *Aset* = collection of users indistinguishable from the actual issuer
- *DRset* = collection of requests issued by users in the same *Aset*
  - requests in the *DRset* can be grouped according to the value of *PI* contained in the requests

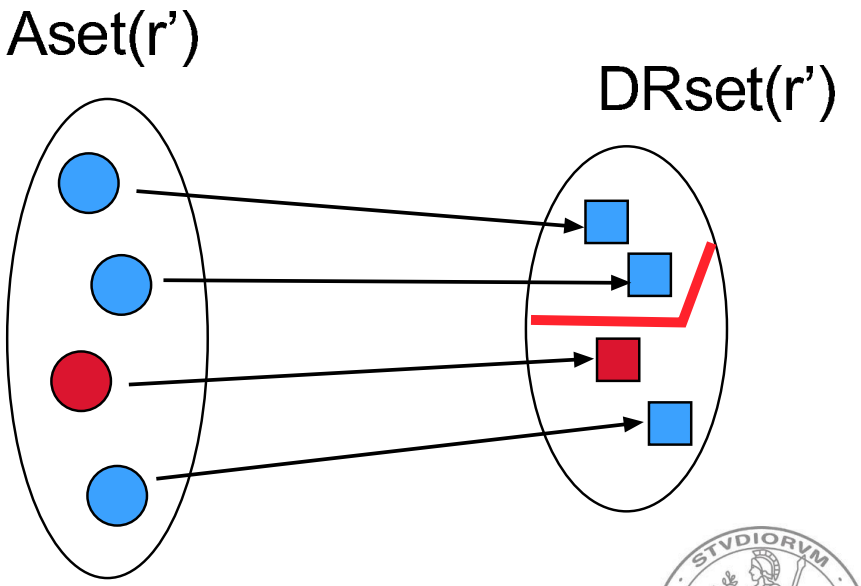


Anonymity and Diversity (2)

- **Homogeneity attack**: each user in the *Aset* issues a request AND there is no diversity among private information in the *DRset*
- **l-diversity** with $l \geq 2$ is necessary for providing privacy

Anonymity and Diversity (2)

- **Homogeneity attack**: each user in the *Aset* issues a request AND there is no diversity among private information in the *DRset*
- **l-diversity** with $l \geq 2$ is necessary for providing privacy



Conclusions

- There are several aspects to privacy in LBS and a lot of confusion in the current preliminary approaches
- The multiple-issuer and the dynamic cases have not yet received the necessary attention
- A formal framework is needed to evaluate safety of solutions with respect to specific attacks
- Solutions should also be empirically validated in terms of performance and quality of service. A big effort is required to obtain realistic simulations or useful real data.

Future Work

- Dealing with the general case: Dynamic, multiple issuers
- Extending techinques for LBS to general context-aware services
- Exporting results to (recurrent) publication of data from DB

References

- Bettini, Mascetti, Wang. Privacy Protection through Anonymity in Location-based Services. In **Digital Privacy: Theory, Technologies, and Practices**, Taylor and Francis, 2007.
- Bettini, Mascetti, Wang. Privacy Issues in Location-based Services. In **Encyclopedia of Geographical Information Science**, Springer, 2007.
- Sergio Mascetti. Privacy Protection through Anonymity in Location-based Services. PhD Dissertation, DICo, Università di Milano, 2007.
- Technical papers in conf. proceedings (SDM-2005, MDM-2007, PALMS-2007, PERCOM-2007)

End

Thank you for your attention.