# Data Privacy in Mobile and Pervasive Systems

*Claudio Bettini - Università degli Studi di Milano*

# Outline

- What is data privacy?
- Privacy threats in mobile apps and services
- Privacy threats in pervasive systems
- Privacy enhancing techniques for mobile and pervasive computing

# Privacy: what and why

## What

[privacy] «The right to be let alone»

Samuel Warren and Louis Brandeis, "The Right to Privacy", Harvard Law Review, 1890.

[personal data] Any information which are related to an identified or identifiable natural person (EU GDPR Art. 4)

[data privacy] The ability to control the release, use and distribution of own personal data

(Lack of the latter may put the former at risk...)



# Privacy: what and why

## Why

Lack of data privacy may bring to

- Deprivation of civil rights
- Discrimination
- Stalking
- Spam
- ...

# Privacy threats in mobile apps and services

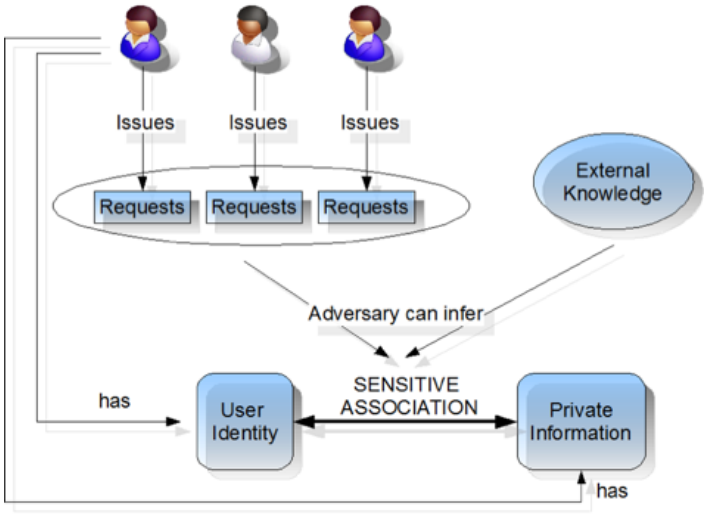


# Linking identity with sensitive data

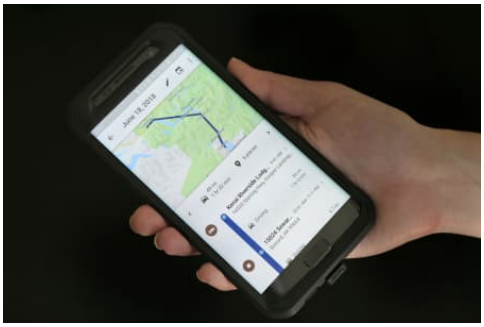
Location/context data inside requests can be exploited in two ways:

- Reveal identity

- Reveal Private Info



# Re-identifying through location data



Example: Alice, while being at her new office, queries a location based service (LBS) for a close-by vegetarian restaurant. She thinks to be anonymous since she is registered using a pseudonym and no identifying data; she is also using a dynamic/masked IP. However, the LBS provider, based on the location and on a public directory of personnel assigned to each office can re-identify Alice and understand she is vegetarian while she did not like to reveal it).

NOTE: Re-identification can also be based on frequent patterns of locations.



# Location as sensitive information

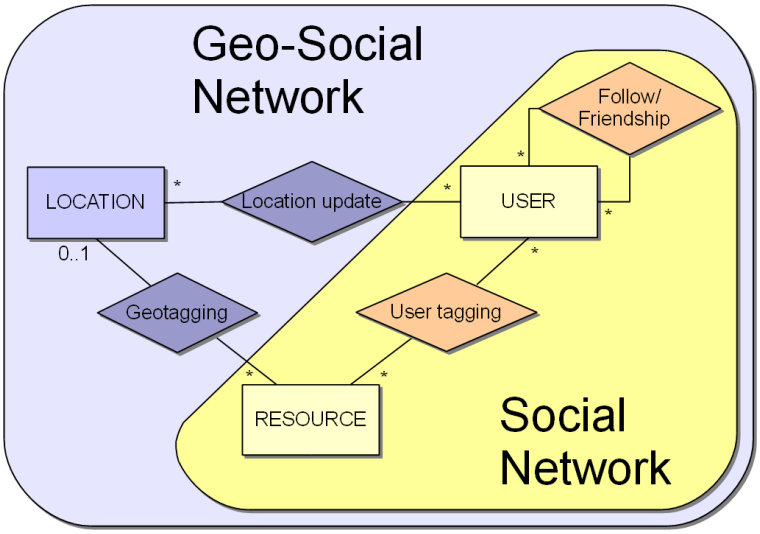

Example: Alice, while being in a specialised cancer treatment facility, issues a non-sensitive LBS request (e.g., localised weather forecasts, or closer ATM) using a service for which she is registered with her identity. By analysing the location information in the request and consulting public map services, the LBS provider can guess that Alice has a specific health condition.

NOTE: Re-identification can also be based on frequent patterns of locations (e.g., home-work).

# Privacy in Geo-Social networks

Most modern SN associate posts with location and time

- Foursquare
- Facebook
- Twitter






# Privacy in Geo-Social networks

- Users share theirs as well as others' location to multiple users
- Geo-SNs expose many re-identifying shared data
- In a social context co-location may become private information
- Protection of location and absence privacy becomes trickier

# Privacy threats in smart environments



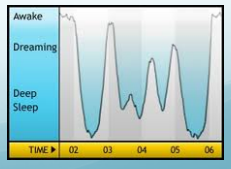
# Privacy Threats with emerging technologies


WHAT HEXOSKIN MONITORS

- Heart Rate, HRV (allowing to estimate stress and fatigue), Heart Rate Recovery, and ECG
- Breathing Rate (RPM), Minute Ventilation (L/min)
- Activity intensity, peak acceleration, steps, cadence and sleep positions

# Well being apps

- Various apps for smartphones and smartwatches
- Track activities, movements, stress, moods and sleep
- Data are usually sent to the service provider, but in most cases there would be no reason to do that…








# Adversaries and sensitive data (1)

| Category | Adversary | Sensitive Data |
|---|---|---|
| LBS | Service Provider | Location+time, absence, co-location, trajectories |
| Mobile Advertisement | Service provider, merchant | |
| GeoSN | Service provider, other users | |
| Participatory sensing | Data Collector, service users | Location, sensed data |

# Adversaries and sensitive data (2)

| Category | Adversary | Sensitive Data |
|---|---|---|
| eHealth, Quantified Self | Cloud and service Provider | bio-physical data, activity, habits, illness |
| Vehicular apps and smartcity services | Other drivers, city and road authorities | Trajectories, driving behavior, habits |
| SmartHome and Smart utilities | home automation providers, utility companies | Occupancy, habits, activities |





# Pervasive impact

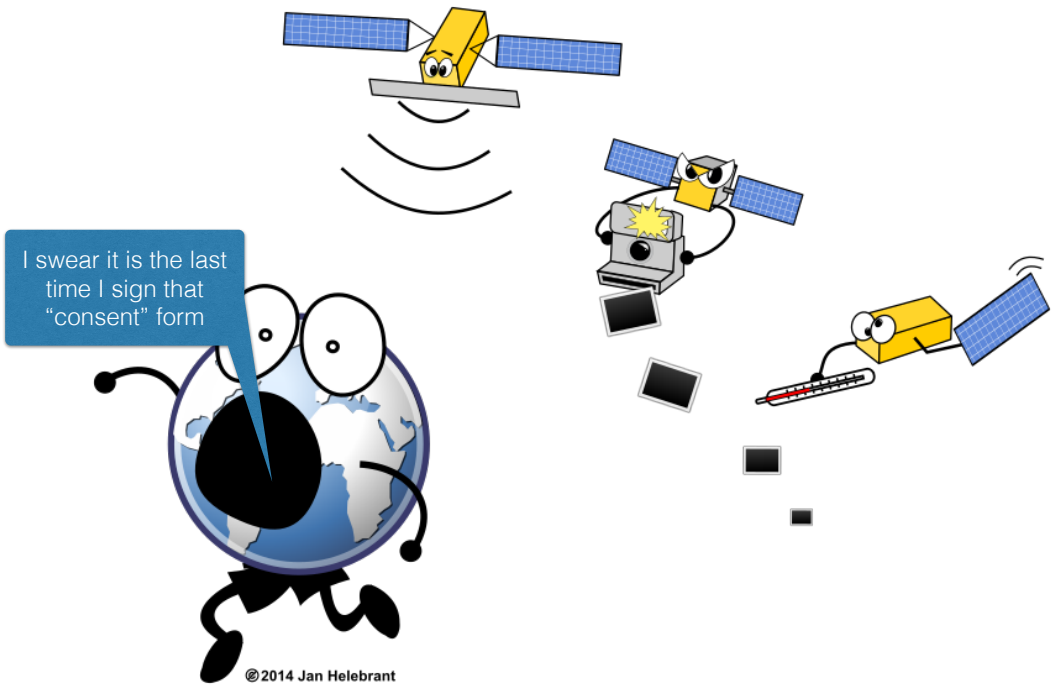
- More (personal) data —> more exposure and need for scalability
- New type of data —> new ways to re-identify and new ways to derive sensitive data
- Continuous stream —> protection from sequential/continuous release

# Pervasive impact

- Less awareness —> **more difficult to express consent and to monitor**
- Lack of interfaces —> **more challenging to control**





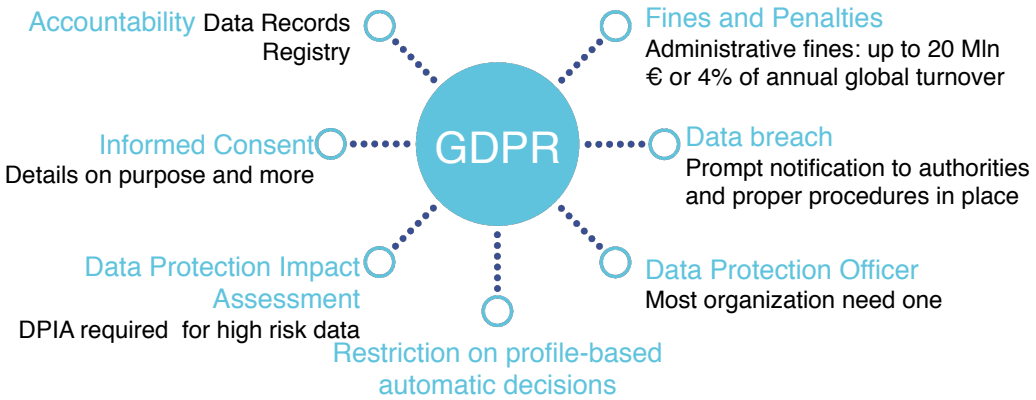



(How) can we protect?

# Privacy protection regulation and techniques



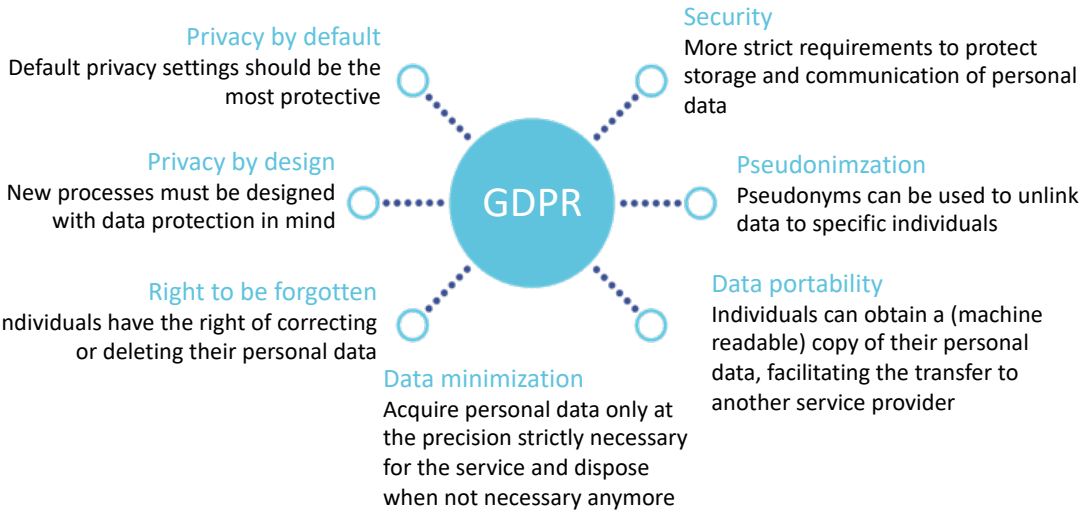

## GDPR: The EU General Data Protection Regulation



Every Ware Lab

# GDPR: the new EU regulation on data protection

## Technical novel aspects:



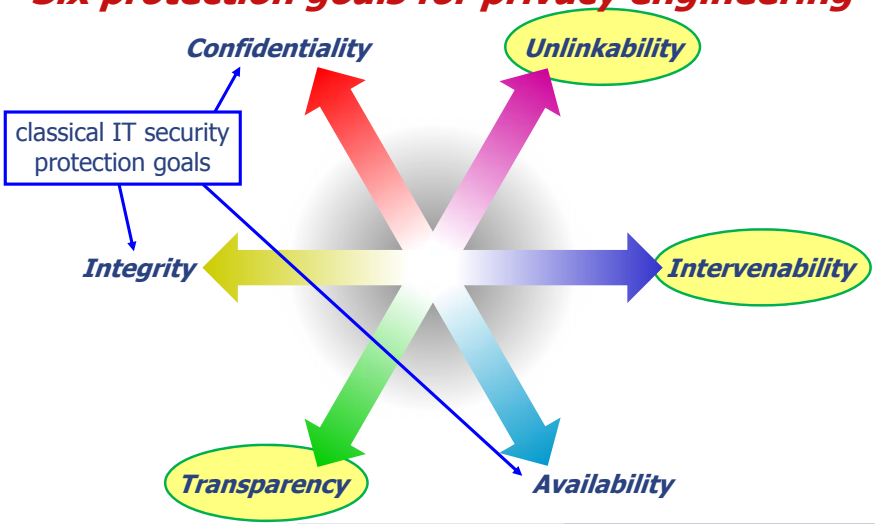

**Enforce on companies through accountability**

Technical expertise is required for proper assessment



Every Ware Lab

# Six protection goals for privacy engineering




**From Marit Hansen (ULD) talk at DPPT'15**

# Security principles

(and techniques to ensure them)

- **Confidentiality:** Only authorized parties can access data (authentication, encryption of data in transfer, at rest, and in use)
- **Integrity:** Data should not be altered without authorization (cryptographic hashing, checksums)
- **Availability:** Whenever needed, the system/data should be available (automatic recovery, replication of HW/data/processes)



# Privacy principles

- **Transparency:** all privacy-relevant data processing (legal, technical, organisational) can be understood and reconstructed at any time (Blockchain may be an enabling technology)
- **Unlinkability:** privacy-relevant data cannot be linked across domains that are constituted by a common purpose and context. (de-identification, pseudonimization, generalization, data minimization, separation, access control)
- **Intervenability:** intervention is possible concerning all ongoing or planned privacy-relevant data processing (see next slide)

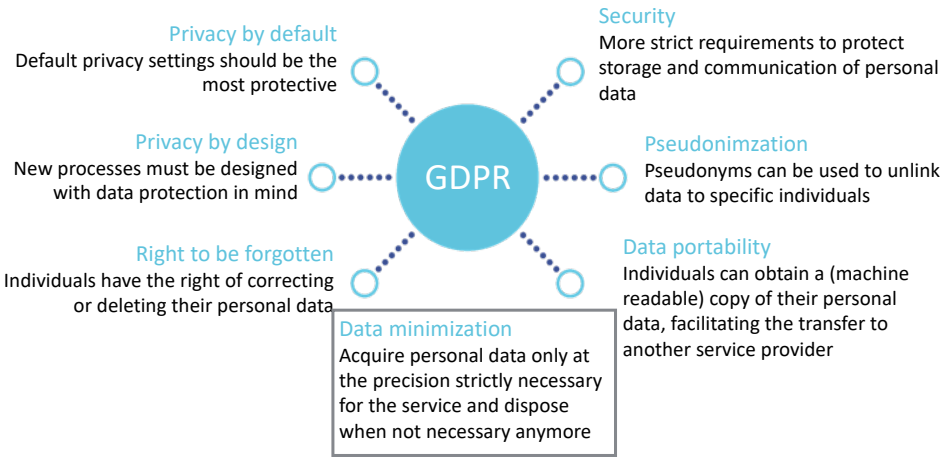
# Interveneability

- Data rectification, access, erasure
- Break Glass procedures (i.e., override access control in emergency)
- Choice of deactivation (i.e., stop data acquisition if violating someone's privacy policy)
- Manual override of automated decisions



## GDPR: the new EU regulation on data protection

Technical novel aspects:




Technical expertise is required for proper assessment

# Not every app requires precise location (and time)

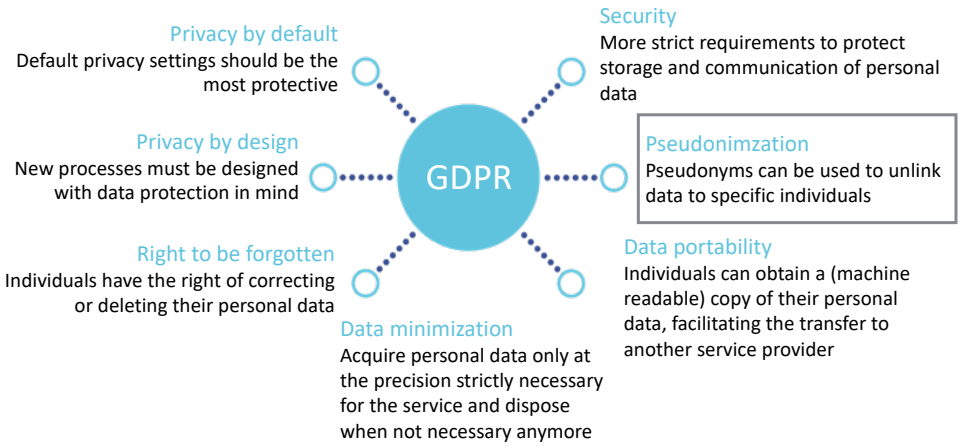
| LBS type | Req. Precision | Continuous | Explicit Identity | Adversaries |
|---|---|---|---|---|
| POI services | high | No | No | SP |
| Weather/news | low | No | No | SP |
| Navigation | high | Yes | No | SP |
| GeoSN posts | high | No | Yes | SP, users |





## GDPR: the new EU regulation on data protection

Technical novel aspects:




Technical expertise is required for proper assessment

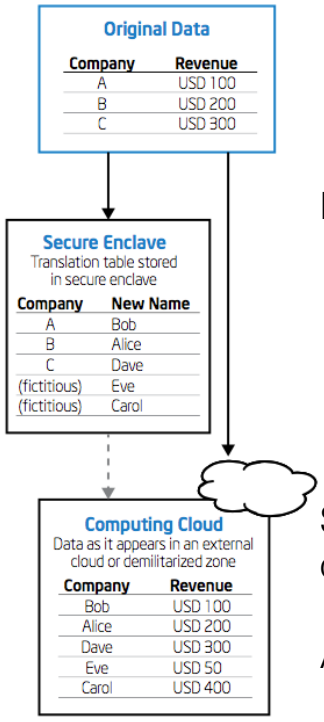
# Pseudonymization

- Explicitly mentioned in the GDPR as personal data protection technique
- If correctly implemented, it separates sensitive data from the data respondents keeping the mapping between them accessible only to selected authorised entities
- It is NOT anonymisation. Anonymised data, in principle, cannot be re-identified (by anybody no matter what information they can access). Anonymised data is not subject to the GDPR.





**Figure: Intel White Paper: Enhancing Cloud Security Using Data Anonymization, 2012.**




# Achieving Unlinkability

De-identify sensitive information

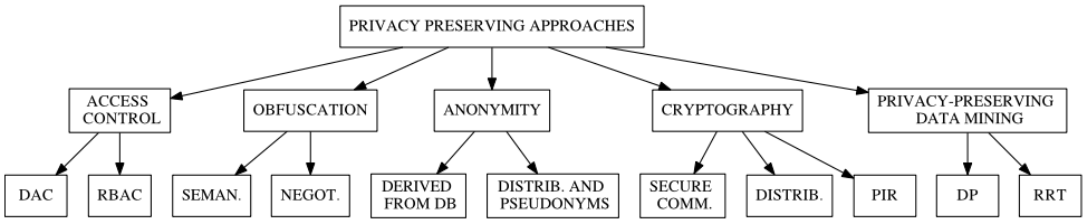
Anonymization/Pseudonymization:
Suppress, generalize, truncate QID

Statistical perturbation

Sensitive data minimisation and/or obfuscation

Access control

# A wide range of techniques




[Bettini-PMCJ-14]



# Unlinking by protecting anonymity

Can we apply a concept similar to k-anonymity to avoid re-identification through location data by the adversary?

# k-anonymity in tabular data

**Main Idea: each released record cannot be associated to less than k possible respondents.**

| Zip | Age | Disease |
|---|---|---|
| 130▪ | 2▪ | Heart disease |
| 130▪ | 2▪ | Heart disease |
| 130▪ | 2▪ | Heart disease |
| 130▪ | 2▪ | Viral infection |
| 130▪ | 3▪ | Cancer |
| 130▪ | 3▪ | Cancer |

**Figure: Intel White Paper: Enhancing Cloud Security Using Data Anonymization, 2012.**



**Key problem**: Find Quasi-Identifiers (QID)



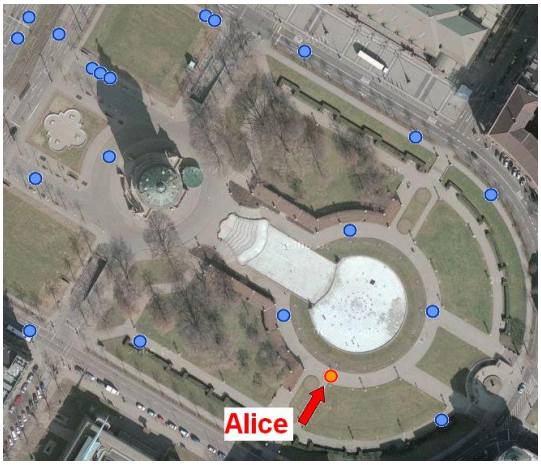
# Location data may act as QID

**Goal**: making the user indistinguishable among many users that are *in the same geographical region* while preserving service utility

**Location K-anonymity solution**: Use pseudonyms and spatio-temporal cloaking (extended to traces) through a trusted server that knows users' locations

**NOTE**: Sensed data in pervasive applications may also act as QID!

# Anonymization in Location Based Services (LBS)



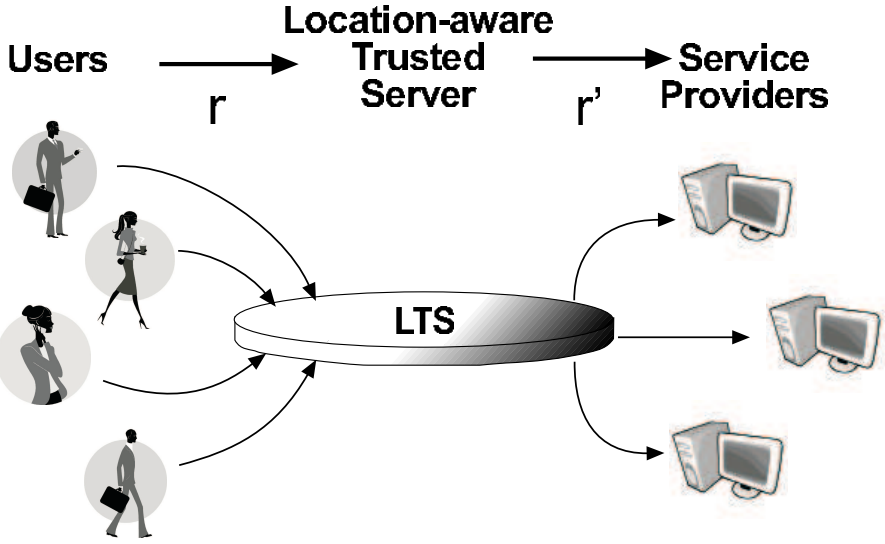

- Alice issues a LBS request for a veg restaurant
- Priv Info: she is vegetarian
- her exact location may reveal her identity



# Location k-anonymity

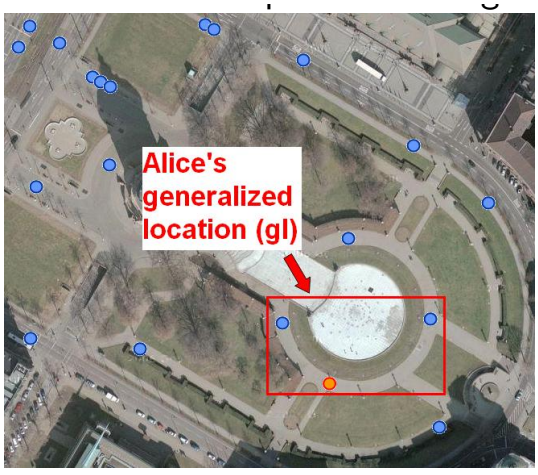
- The principle of k-anonymity requires that the individual (here Alice) must be indistinguishable from k – 1 other potential issuers of the LBS request.
- Applying k-anonymity in the above example requires the geographic user position to be enlarged to **a cloaking region including k – 1 other users** before sending the request to the LBS.
- This region can be computed by a trusted anonymization service knowing many user positions (e.g. a mobile operator) . Then, even in the worst case in which the untrusted LBS provider can identify all the k users in the reported area, he can only tell that one of them searched for a veg restaurant, and there is only a chance of 1/k that this user was Alice.

# General architecture for location k-anonymity






# Spatial cloaking for anonymization




- Alice's request origin is generalized to a region with k candidate issuers

The LBS result can be filtered by the trusted server to return precise result

# Limits of (k-)anonymity

- Re-identification by
  - linking with external knowledge
  - correlation between multiple releases
  - predictive models
  - shadow attacks
- Insufficient diversity of sensitive values (l-diversity and t-closeness as refinements)





# Limits of k-anonymity

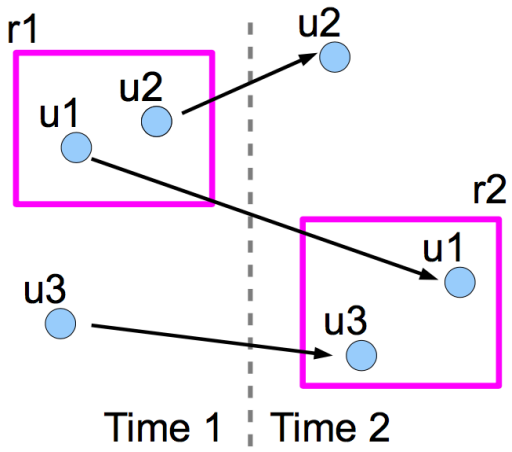
**Main Idea: each released record cannot be associated to less than k possible respondents.**

| Zip | Age | Disease |
|---|---|---|
| 130▪ | 2▪ | Heart disease |
| 130▪ | 2▪ | Heart disease |
| 130▪ | 2▪ | Heart disease |
| 130▪ | 2▪ | Viral infection |
| 130▪ | 3▪ | Cancer |
| 130▪ | 3▪ | Cancer |

**Figure: Intel White Paper: Enhancing Cloud Security Using Data Anonymization, 2012.**

What if I know that Bob is in its 30s and is in the data?

# Multiple requests: historical k-anonymity



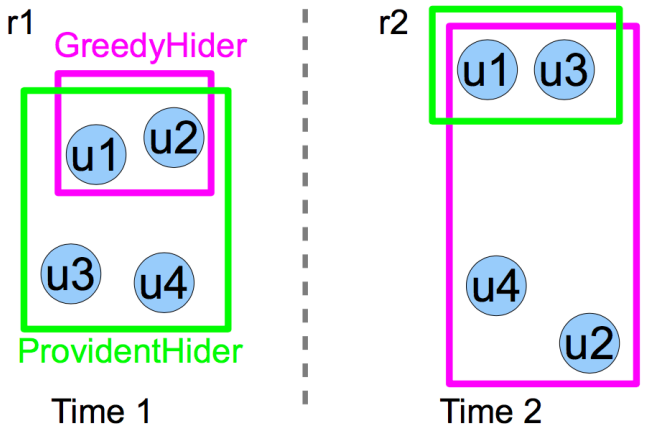

If the adversary knows that $r1$ and $r2$ are *linked* (i.e., issued by the same user), $u1$ is identified as the only possible issuer.

[Bettini-SDM05]



# Algorithms ensuring Historical k-anonymity

Goal: 2-anonymity.




[Mascetti-MDM09]

# Problems with de-identification

- Need to frequently change pseudonyms
- Need for a trusted server (or accurate predictive model of user distribution)
- No formal guarantees on avoiding linking based on adversary background knowledge





# Unlinking by reducing sensitiveness of location/time

When location/context information is important for the quality of service but it is sensitive, how can we still include it without compromising privacy?

# Location data may act as sensitive information

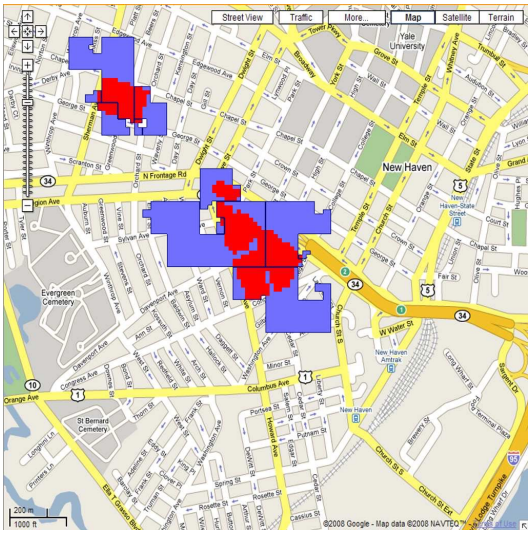

Example: My LBS requests (for car sharing for example) originate from a specialised hospital every Tuesday





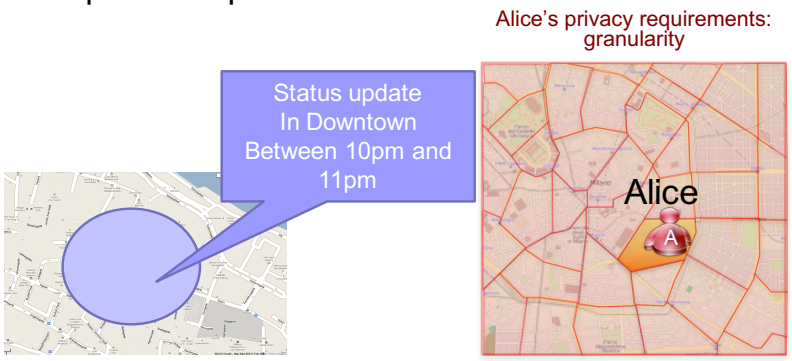
# Protecting location as sensitive information (in LBS)

- Location privacy, absence privacy, and co-location privacy
- No need for trusted server
- Possible technique: **query enlargement:** Instead of sending a position, **send an area including other possible non sensitive locations** of the user. Note that this is different from spatial cloaking for location k-anonymity.
- Examples: PROBE [Damiani-TDP10] and SafeBox [Mascetti-TDP14]



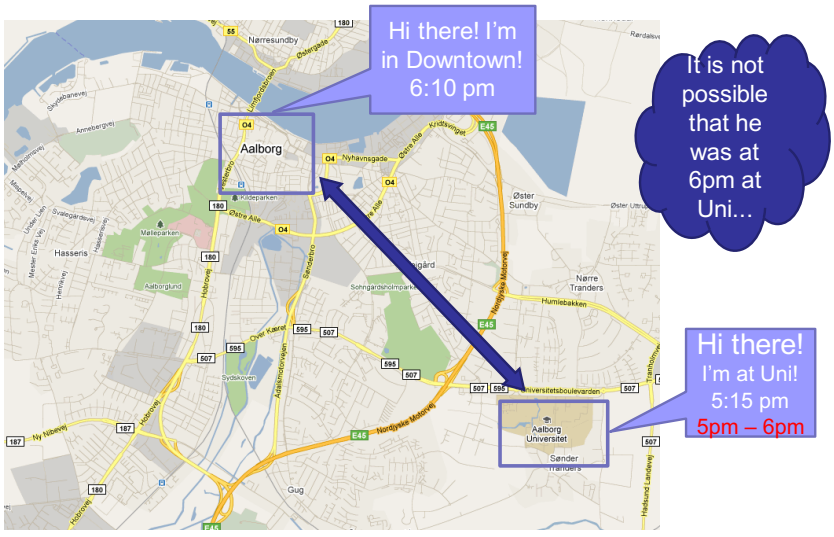
# User preferences (in geoSN)

- Minimum Uncertainty Region (MUR)
  - enforced if the adversary *cannot exclude* any point as the origin of the resource
  - Spatio-temporal






# Attacks on multiple requests



# Problems with cloaking/ enlargement

- Algorithms must prevent inversion (avoid reverse engineering when the adversary knows the algorithm)
- Cloaking/enlargement is not supported (or only partially supported) by current LBS and geo-SN



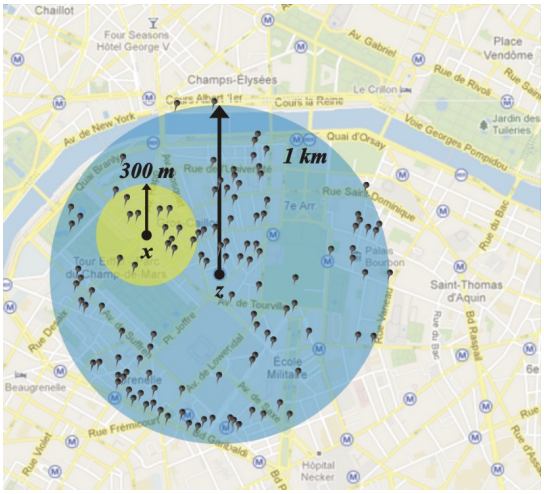


# An alternative: fake locations

- Simultaneous requests from n+1 locations (one of them being the real one) and discard n results
- Single request from one fake location and filter results
- Sequential requests with a strategy from a single fake location (e.g., SpaceTwist [Yiu-ICDE08])

# POI request from a fake location

**Problem with fake locations**: how to pick fake locations with privacy and utility guarantee? And a fake trajectory?




from [Andres-CCS13]





# Apple's solution for location privacy

**Introduced in 2020 with iOS14**

- reveal only an area including the user
- resistant to inversion (fixed regions)
- context-aware size
- unfrequent updates

# Beyond location: De-id/obfuscation in pervasive systems

| Data type | De-id/obfuscation Technique |
|---|---|
| video | selective blurring, shading, skeleton tracking |
| audio | voice transformation |
| location/time | spatial/temporal cloaking, noise |
| sensors | granular release, encryption |

**[S. Jana, A. Narayanan, V. Shmatikov: A Scanner Darkly: Protecting User Privacy from Perceptual Applications. IEEE Symposium on Security and Privacy 2013: 349-363]**



# Techniques with theoretical privacy guarantees

# Differential Privacy (DP)

***Goal***: Retrieve from a **statistical database** aggregated personal data with provable guarantees of privacy for individuals

Introduced for static tabular data (C. Dwork, ICALP06)

***Main idea***: Given DB1 and DB2 that differ only for DB2 including also a tuple about a person P, the information separately extracted from the two DBs with a DP method is not significantly different. Hence it cannot be used to violate P's privacy.

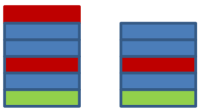




# Privacy Parameter ε

For every pair of inputs that differ in one row

For every output …



$D_1$ $D_2$ $O$

$$\Pr[A(D_1) = O] \leq e^{\varepsilon} \Pr[A(D_2) = O]$$

Controls the degree to which $D_1$ and $D_2$ can be distinguished. Smaller ε gives more privacy (and worse utility)

From A. Machanavajjhala et al.
Sigmod2017 Tutorial "DP in the Wild"

# DP and location

- DP can be directly applied when *publishing aggregate location data (macro data)*. E.g., in check-in collection, counting users in given areas, … (See systems used by Apple and Google)
- However the idea behind DP has been extended to location micro-data: locations instead of records. Adjacent “DBs” differ for one location. Compute a fake location to send to the LBS so that the risk of discovering the real location is quantified.





# Geo-indistinguishability (1)

The fake location ensuring privacy guarantee is picked by adding *planar Laplace* noise


[Andres-CCS13]

# Geo-indistinguishability (2)

A mechanism to generate a fake location z from a real location x such that for each x' at a distance at most r from x

$$\frac{p(z|x)}{p(z|x')} \leq e^{\epsilon r}$$

Issues:

- Unclear which privacy budget is appropriate (and which r)
- Practical? Still unclear





# What we did not cover

- Privacy preserving publication of (aggregate) location data
- Extension of techniques from location to other context data
- Privacy preserving crowdsourcing
- Cryptography based techniques (including multi-party secure computation, private information retrieval, and homomorphic encryption)
- …

# Conclusions

- Emerging technologies make privacy challenges harder, but the basic principles are the same
- We need to design **Privacy Enhancing Technologies specific to pervasive** data and new software architectures (e.g., edge)
- It is an **interdisciplinary issue** (IT-Law). A composition of techniques need to be complemented by appropriate regulation

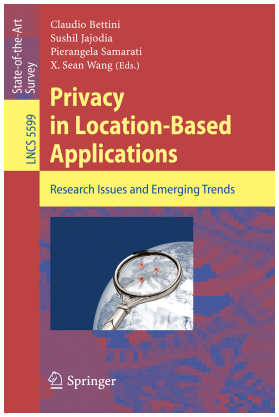


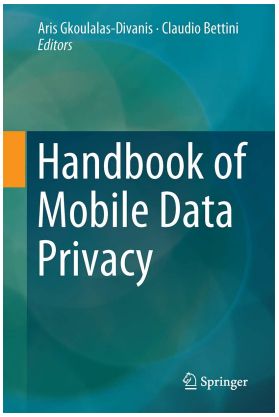


# Material

[Bettini-PMCJ-14] C. Bettini, D. Riboni, **Privacy Protection in Pervasive Systems: State of the Art and Technical Challenges**. Journal of Pervasive and Mobile Computing, Elsevier, 2014.

[Bettini-18] C. Bettini. **Privacy Protection in Location-Based Services: A Survey**. In Handbook of Mobile Data Privacy, Aris Gkoulalas-Divanis, Claudio Bettini (Eds.), Springer, 2018



# References

A comprehensive up-to-date online course on data privacy: https://www.kau.se/cs/pbd

**Literature specific to location privacy:**

[Andres-CCS13] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, C. Palamidessi. Geo-Indistinguishability: Differential Privacy for Location Based Systems. Proc. of CCS, 2013.

[Beresford-PervComp03] Beresford, Alastair R., and Frank Stajano. Location privacy in pervasive computing. *IEEE Pervasive computing* 2.1 (2003)

[Bettini-SDM05] C. Bettini, X. S. Wang, S. Jajodia. Protecting Privacy Against Location-based Personal Identification. Proc. of Secure Data Management (SDM), 2005.

[Damiani-TDP10] Maria Luisa Damiani, Elisa Bertino, Claudio Silvestri. The PROBE Framework for the Personalized Cloaking of Private Locations. Trans. Data Privacy 3(2): 123-148 (2010)





# References

[Mascetti-TDP14] Mascetti, S., Bertolaja, L:, Bettini, C.. SafeBox: adaptable spatio-temporal generalization for location privacy protection. Transactions on Data Privacy 7.2 (2014).

[Vicente-IC11] Carmen Ruiz Vicente, Dario Freni, Claudio Bettini, Christian S. Jensen. Location-Related Privacy in Geo-Social Networks. IEEE Internet Computing 15(3): 20-27 (2011)

[Yiu-ICDE08] Yiu, Man Lung, et al. Spacetwist: Managing the trade-offs among location privacy, query performance, and query accuracy in mobile services. In Proc. of IEEE ICDE 2008.

# Anonymisation and pseudonymisation tools and recommendations

- Pseodonymization techniques and best practices, ENISA Report 11/2019
  https://www.enisa.europa.eu/publications/pseudonymisation-techniques-and-best-practices/at_download/fullReport
- Data Anonymization tool. https://arx.deidentifier.org/
  (with pointers to other tools)
- IHSN Statistical Disclosure Control software
  http://www.ihsn.org/software/disclosure-control-toolbox