Geo-Indistinguishability

A principled approach to location privacy

Catuscia Palamidessi
INRIA

joint work with
Nicolas Bordenabe, Konstantinos Chatzikokolakis and Marco Stronati

ICDCIT 2015, Bhubaneswar, India

Plan of the talk

My talk will be broader than the title:
I will first speak of privacy in general

- The problem of information leakage and privacy
- Randomized protection mechanisms
- Differential privacy
- Generalization of differential privacy to arbitrary metric spaces
- Application to location privacy: Geo-indistinguishability

Leakage of information / privacy threats

BBC

News Sport Weather Capital Future Shop

NEWS TECHNOLOGY

Home US & Canada Latin America UK Africa Asia Europe Mid-East Business Health Sci/Environ

13 March 2014 Last updated at 21:23 ET

Share

Mark Zuckerberg 'confused and frustrated' by US spying

Mr Zuckerberg said that the internet needed to be made more secure for users

Facebook founder Mark Zuckerberg has said he has called President Barack Obama to "express frustration" over US digital surveillance.

The 29-year-old said in a blog post the US government "should be the champion for the internet, not a threat".

Related Stories

'Spying setting fire to internet'

Trust in the internet

Bits

MARCH 13, 2014, 7:45 AM | Comment

Daily Report: Europe Moves to Reform Rules Protecting Privacy

By THE NEW YORK TIMES

E-MAIL
FACEBOOK
TWITTER
SAVE
MORE

The European Parliament passed a strong new set of data protection measures on Wednesday prompted in part by the disclosure by Edward J. Snowden, a former contractor at the United States National Security Agency, of America's vast electronic spying program, David Jolly reports.

theguardian

News | US | World | Sports | Comment | Culture | Business | Money

News › Society › NHS

NHS England patient data 'uploaded to Google servers', Tory MP says

Health select committee member Sarah Wollaston queries how data was secured by PA Consulting and uploaded to servers outside UK

Police will have 'backdoor' access to health records

Target says it declined to act on early alert of cyber breach

BY JIM FINKLE AND SUSAN HEAVEY

BOSTON/WASHINGTON Thu Mar 13, 2014 6:39pm EDT

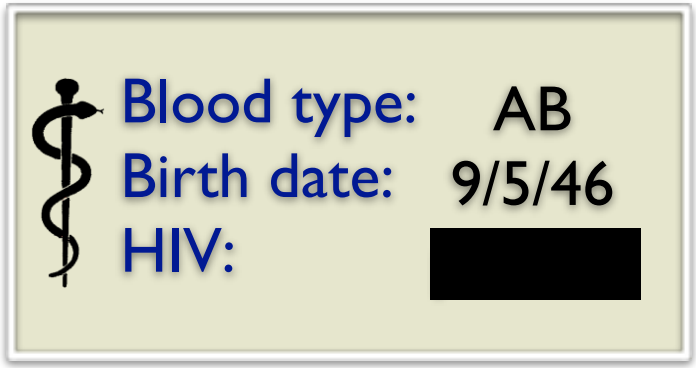
5 COMMENTS | Tweet 45 | Share 21 | Share this | 8+1 12 | Email | Print

Merchandise baskets are lined up outside a Target department store in Palm Coast, Florida, December 9, 2013.

CREDIT: REUTERS/LARRY DOWNING

Protection of sensitive information

- Protecting the confidentiality of sensitive information is a fundamental issue in computer security

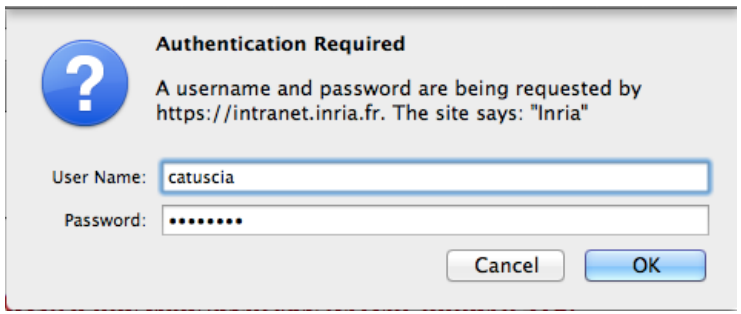


- Access control and encryption are not sufficient! Systems could leak secret information through the correlation with public information (observable).

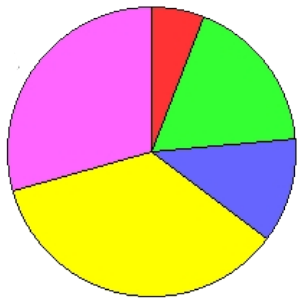
- The notion of "observable" is subtle and crucial.
  - It depends on the power of the adversary
  - It may be combined from different sources

Leakage through correlated observables

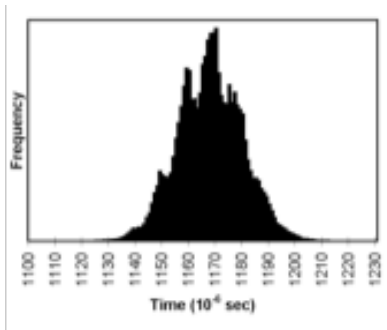
Password checking



Election tabulation



Timings of decryptions



Reasoning about information leakage: Quantitative approaches

- It is usually impossible to prevent leakage completely. Hence we have to reason about the **amount** of leakage. This is usually related to the probability that the adversary discovers the secret

- Many methods to protect information use randomization to obfuscate the link between secret and observable. Hence the correlation itself may have a probabilistic nature.

Various notions of leakage

- The choice of an appropriate measure of leakage depends on many factors
- In particular, we need to choose whether to consider the **worst case**, or the **average** leak: individuals are usually interested in the first, while companies may prefer the second.
- This talk will focus on the worst-case, and on the point of view of the user (**Privacy**)
- More specifically, this talk will focus on **differential privacy** and on the related approach of **geo-indistinguishability** for location privacy

Differential Privacy

- Differential privacy [Dwork et al.,2006] is a notion of privacy originated from the area of **Statistical Databases**

- **The problem:** we want to use databases to get statistical information (aka aggregated information), but without violating the privacy of the people in the database

The problem

- Statistical queries should not reveal private information, but it is not so easy to prevent such privacy breach.

- Example: in a medical database, we may want to ask queries that help to figure the correlation between a disease and the age, but we want to keep private the info whether a certain person has the disease.

name	age	disease
Alice	30	no
Bob	30	no
Don	40	yes
Ellie	50	no
Frank	50	yes

Query:
What is the youngest age of a person with the disease?

Answer:
40

Problem:
The adversary may know that Don is the only person in the database with age 40

The problem

- Statistical queries should not reveal private information, but it is not so easy to prevent such privacy breach.

- Example: in a medical database, we may want to ask queries that help to figure the correlation between a disease and the age, but we want to keep private the info whether a certain person has the disease.

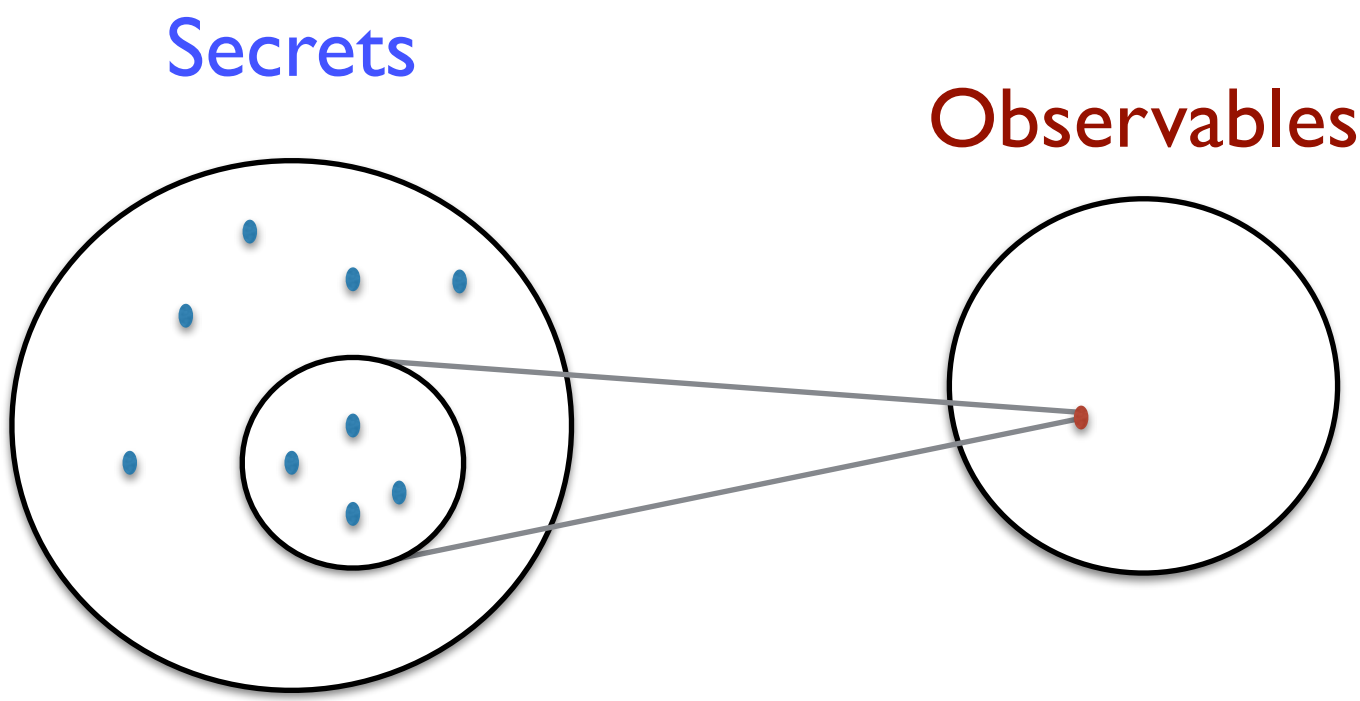
name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

k-anonymity: the answer always partition the space in groups of at least k elements

Alice	Bob
Carl	Don
Ellie	Frank

Correlation: Many-to-one

- Principle: Ensure that there are **many** secret values that correspond to **one** observable
- This is the general principle of most deterministic approaches to protection of confidential information (group anonymity, k-anonymity, ℓ-anonymity, cloacking, etc.)



The problem

Unfortunately, the many-to-one approach is not robust under **composition**:

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

Alice	Bob
Carl	Don
Ellie	Frank

The problem of composition

Consider the query:

What is the minimal weight of a person with the disease?

Answer: 100

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

The problem of composition

Combine with the two queries:

minimal weight and the minimal age of a person with the disease

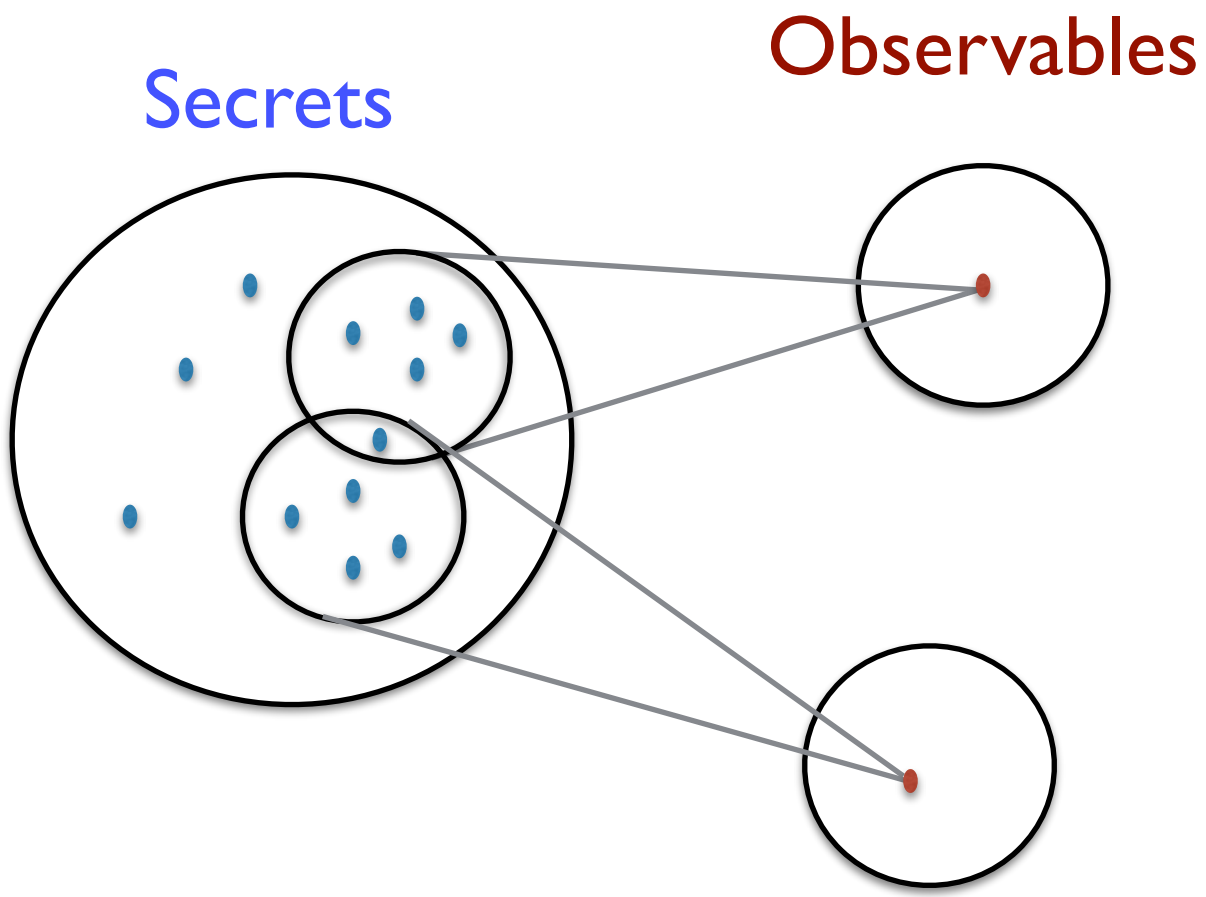
Answers: 40, 100

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

This is a general problem of the deterministic approaches (based on the principle of many-to-one): the combination of observations determines smaller and smaller intersections on the domain of the secrets, and eventually result in singletones



Composition attacks

Composition attacks are real!
For instance, in a recent paper, Narayanan et Smatikov showed that by combining the information of two popular social network (Twitten and Flickr) they were able to de-anonymize a large percentage of the users (about 80%) and retrieve their private information with only a small probability of error (12%).

De-anonymizing Social Networks, Arvind Narayanan and Vitaly Shmatikov. Security & Privacy '09.

Solution

Introduce some probabilistic noise on the answer, so that the answers of minimal age and minimal weight can be given also by other people with different age and weight

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

Noisy answers

minimal age:

40 with probability 1/2

30 with probability 1/4

50 with probability 1/4

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

Alice	Bob
Carl	Don
Ellie	Frank

Noisy answers

minimal weight:
100 with prob. 4/7
90 with prob. 2/7
60 with prob. 1/7

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

Noisy answers

Combination of the answers

The adversary cannot tell for sure whether a certain person has the disease

name	age	disease
Alice	30	no
Bob	30	no
Carl	40	no
Don	40	yes
Ellie	50	no
Frank	50	yes

name	weight	disease
Alice	60	no
Bob	90	no
Carl	90	no
Don	100	yes
Ellie	60	no
Frank	100	yes

Alice	Bob
Carl	Don
Ellie	Frank

Differential Privacy

- There have been various attempts to formalize the notion of privacy, but the most successful one is the notion of Differential Privacy, recently introduced by Dwork
- **Differential Privacy** [Dwork 2006]: a randomized function $\mathcal{K}$ provides ε-differential privacy if for all databases x, x' **which are adjacent (**i.e., differ for only one individual), and for all $z \in \mathcal{Z}$, we have

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

- The idea is that the likelihoods of x and x' are not too far apart, for every S
- Differential privacy is robust with respect to composition of queries
- The definition of differential privacy is independent from the prior (but this does not mean that the prior doesn't help in breaching privacy!)

Differential Privacy: alternative characterization

- Perhaps the notion of differential privacy is easier to understand under the following equivalent characterization.
- In the following, X_i is the random variable representing the value of the individual i, and $X_{\neq i}$ is the random variable representing the value of all the other individuals in the database

- **Differential Privacy, alternative characterization:** a randomized function $\mathcal{K}$ provides ε-differential privacy if and only if:

$$\text{for all } x \in \mathcal{X}, z \in \mathcal{Z}, p_i(\cdot)$$

$$\frac{1}{e^{\epsilon}} \leq \frac{p(X_i = x_i | X_{\neq i} = x_{\neq i})}{p(X_i = x_i | X_{\neq i} = x_{\neq i} \wedge K = z)} \leq e^{\epsilon}$$

A typical differentially-private mechanism

- Randomized mechanism for a query $f: \mathcal{X} \to \mathcal{Y}$.
  Instead of the exact answer to the query, the curator gives a randomized answer $\mathcal{K}: \mathcal{X} \to \mathcal{Z}$ ($\mathcal{Z}$ may be different from $\mathcal{Y}$)

- A typical randomized method: **the Laplacian noise.** If the exact answer is y, the reported answer is z, with a probability density function defined as:

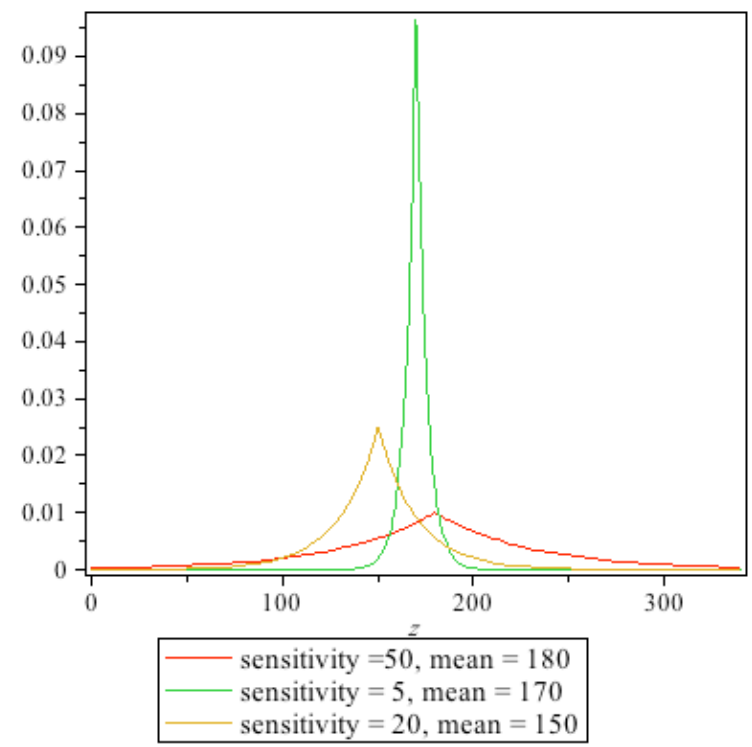
$$dP(z) = c\, e^{-\frac{|z-y|}{\Delta f}}$$

where Δf is the *sensitivity* of f:

$$\Delta f = \max_{x \sim x' \in \mathcal{X}} |f(x) - f(x')|$$

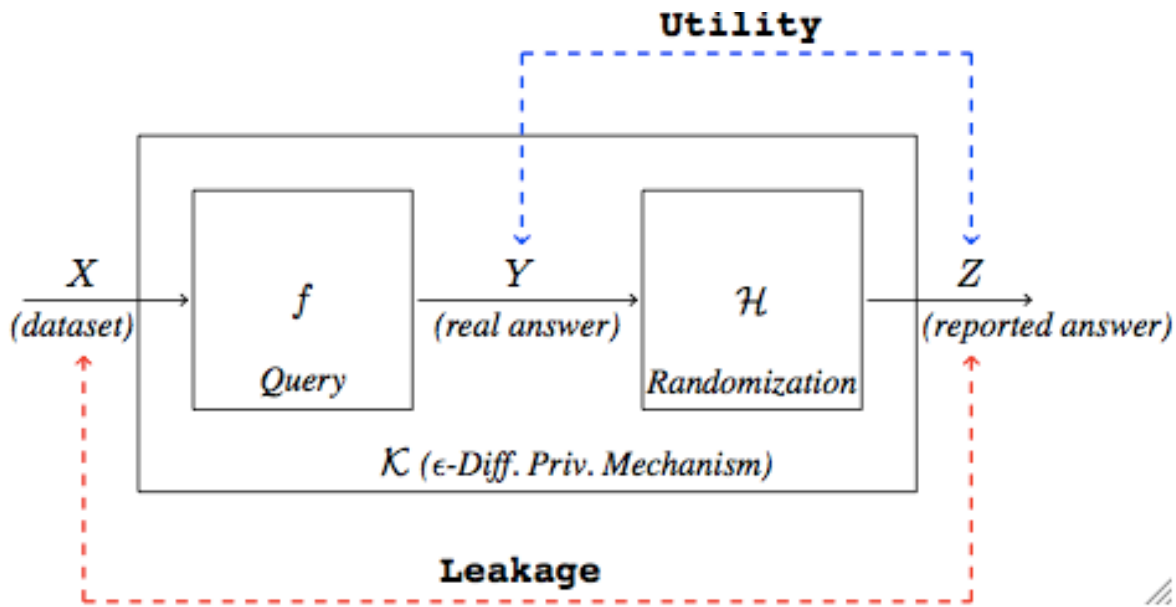
and c is a normalization factor:

$$c = \frac{1}{2\,\Delta f}$$



Privacy and Utility

- The two main criteria by which we judge a randomized mechanism:
  - **Privacy:** how good is the protection against leakage of private information
  - **Utility:** how useful is the reported answer
- Clearly there is a trade-off between privacy and utility, but they are not the exact opposites: privacy is about the individual data, while utility is about the aggregate data.



Privacy and utility

- There may be other differences between privacy and utility, depending on the application domain: one may be worst-case and the other average-case, one may take into account the prior information and the other not, etc.

- The construction of mechanisms that optimize the trade-off between privacy and utility is an active field of research

Privacy vs utility: two fundamental results

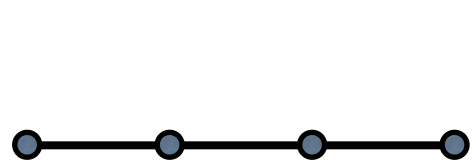
1. [Ghosh et al., STOC 2009]
   The discretized laplacian (aka geometric mechanism) is universally optimal for counting queries

   - Counting queries are of the form "how many individuals in the database satisfy the property P ?"
   - Optimal means that it provides the best utility for a fixed degree of differential privacy.
   - Utility is defined as the average gain of a Bayesian user, which may use some side (prior) knowledge. The gain function is any anti-monotonic function of the distance between the real answer and the guessed answer (obtained combining the answer produced by the mechanism and the side knowledge)
   - Universally optimal means that it is optimal for any user

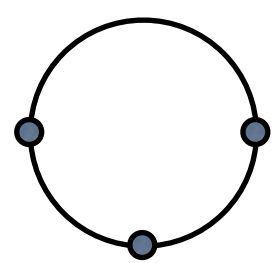
Privacy vs utility: two fundamental results

2. [Brenner and Nissim, STOC 2010] The counting queries are the only kind of queries for which a universally optimal mechanism exists

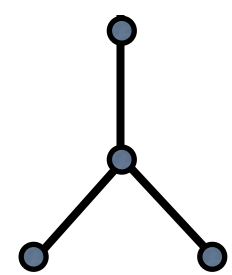
- This means that for other kind of queries one the optimal mechanism is relative to a specific user.
- The precise characterization is given in terms of the graph $(\mathcal{Y}, \sim)$ induced by $(\mathcal{X}, \sim)$



ok



not ok



not ok

Extending differential privacy to arbitrary metrics

Extending differential privacy to arbitrary metrics

Differential Privacy:

Extending differential privacy to arbitrary metrics

Differential Privacy:

A mechanism is ϵ-differentially private iff for every pair of databases x, x' we have:

$$p(Z = z|X = x) \leq e^{\epsilon\, d_H(x,x')} p(Z = z|X = x')$$

where d_H is the Hamming distance between databases:
$d_H(x, x')$ = number of individuals in which x and x' differ.

Extending differential privacy to arbitrary metrics

Differential Privacy:

A mechanism is ϵ-differentially private iff for every pair of databases x, x' we have:

$$p(Z = z|X = x) \leq e^{\epsilon\, d_H(x,x')} p(Z = z|X = x')$$

where d_H is the Hamming distance between databases:
$d_H(x, x') =$ number of individuals in which x and x' differ.

Generalization:

Extending differential privacy to arbitrary metrics

Differential Privacy:

A mechanism is ϵ-differentially private iff for every pair of databases x, x' we have:

$$p(Z = z|X = x) \leq e^{\epsilon\, d_H(x,x')} p(Z = z|X = x')$$

where d_H is the Hamming distance between databases:
$d_H(x, x')$ = number of individuals in which x and x' differ.

Generalization:

On a generic domain $\mathcal{X}$ provided with a metric d:

$$p(Z = z|X = x) \leq e^{\epsilon\, d(x,x')} p(Z = z|X = x')$$

d-privacy

Extending differential privacy to arbitrary metrics

Differential Privacy:

A mechanism is ϵ-differentially private iff for every pair of databases x, x' we have:

$$p(Z = z|X = x) \leq e^{\epsilon\, d_H(x,x')} p(Z = z|X = x')$$

where d_H is the Hamming distance between databases:
$d_H(x, x')$ = number of individuals in which x and x' differ.

Generalization:

On a generic domain $\mathcal{X}$ provided with a metric d:

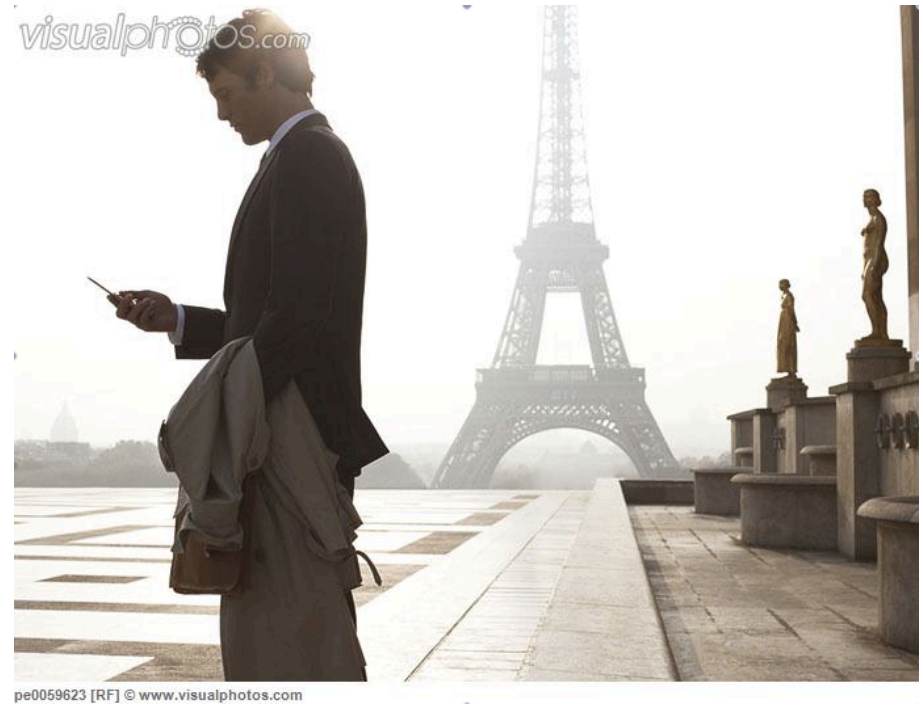
$$p(Z = z|X = x) \leq e^{\epsilon\, d(x,x')} p(Z = z|X = x')$$

d-privacy

Protection of the **accuracy** of the information

Application: Location Based Services

- Use an LBS to find a restaurant
- We do not want to reveal the exact location
- We assume that revealing an approximate location is ok



Example: Location Based Services

d : the Euclidean distance

x : the exact location

z : the reported location

d-privacy:

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

The characterization of d-privacy:

$$p(x|B_r(x), z) \leq e^{\epsilon r} p(x|B_r(x))$$

geo-indistinguishability

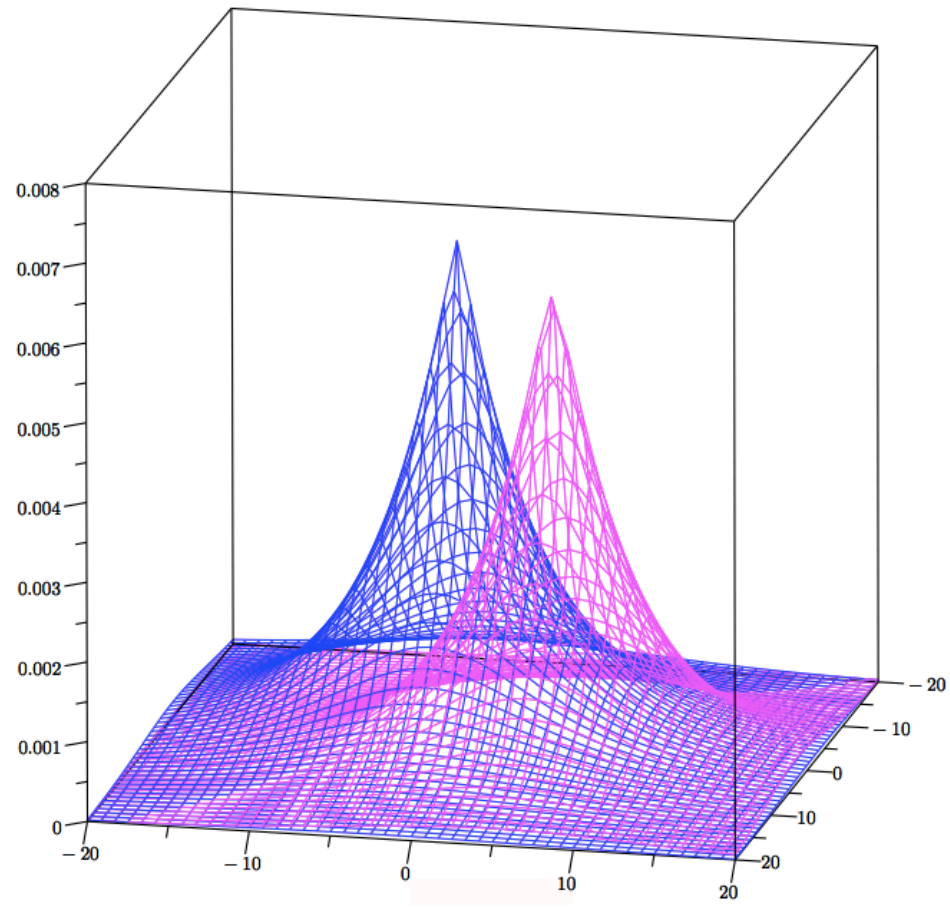
A *d*-private mechanism for LBS:
Planar laplacian

Bivariate Laplacian

$$dp_x(z) = \frac{\epsilon^2}{2\pi} e^{\epsilon\, d(x,z)}$$

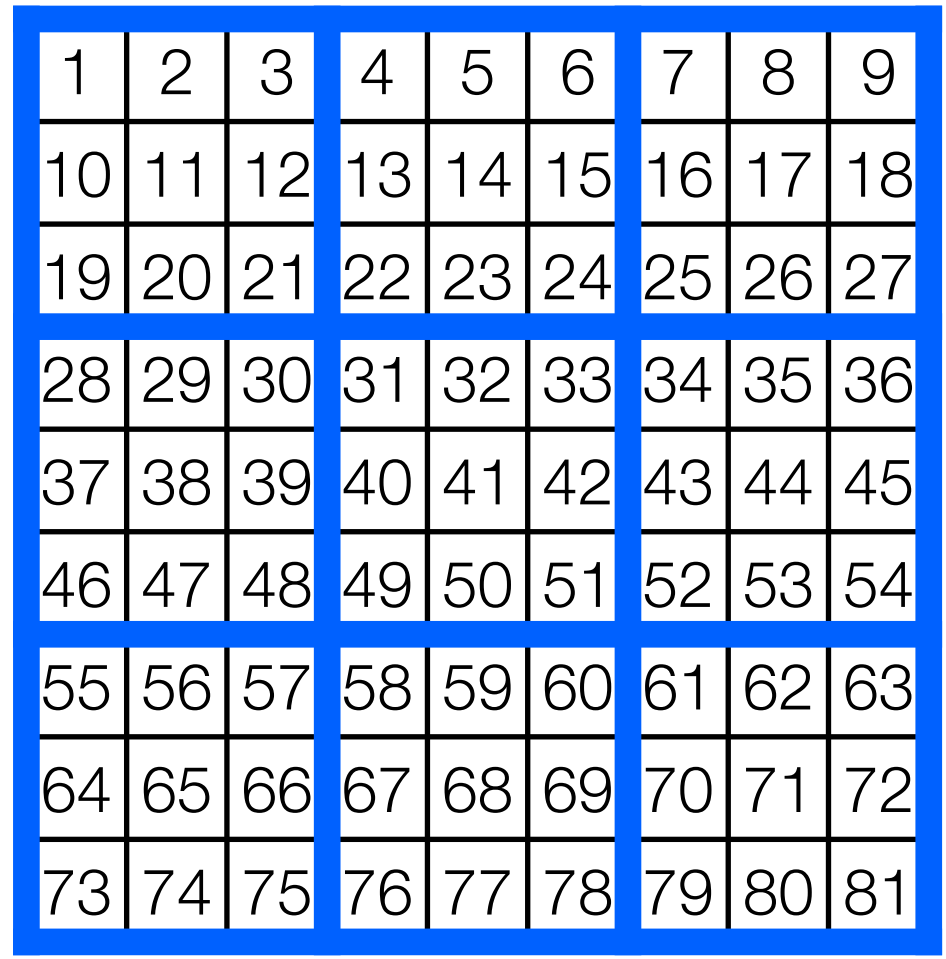
Efficient method to draw points based on polar coordinates

Some care needs to be taken when translating from polar to standard (latitude, longitude) coordinates. Degradation of the privacy level in single precision, but negligeable in double precision.



Privacy versus utility: evaluation

- 9x9 = 81 “points”.
- We compare 4 mechanisms.
- Configured to the **same utility**.



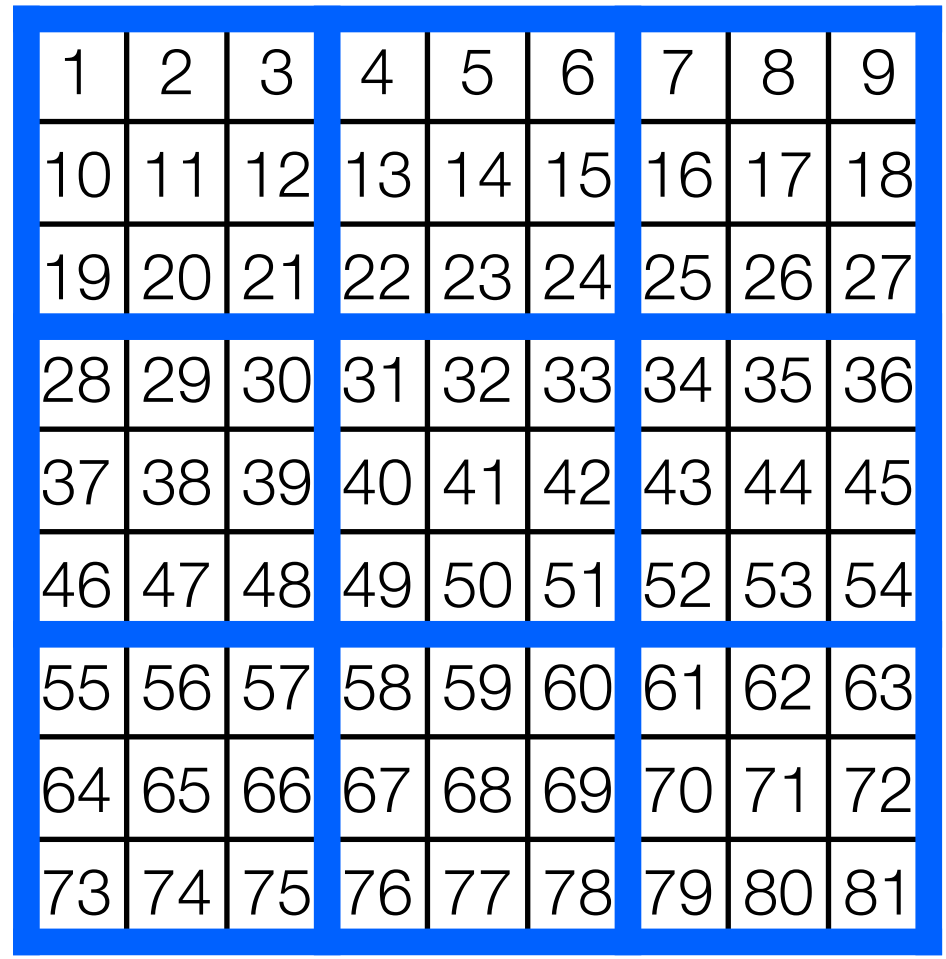
Privacy versus utility: evaluation

- 9x9 = 81 "points".
- We compare 4 mechanisms.
- Configured to the **same utility**.
- Optimal mechanism by [Shroki et al., S&P 2012] for the **corresponding prior**.
- Three **prior independent**:

  Planar Laplacian (discretized).

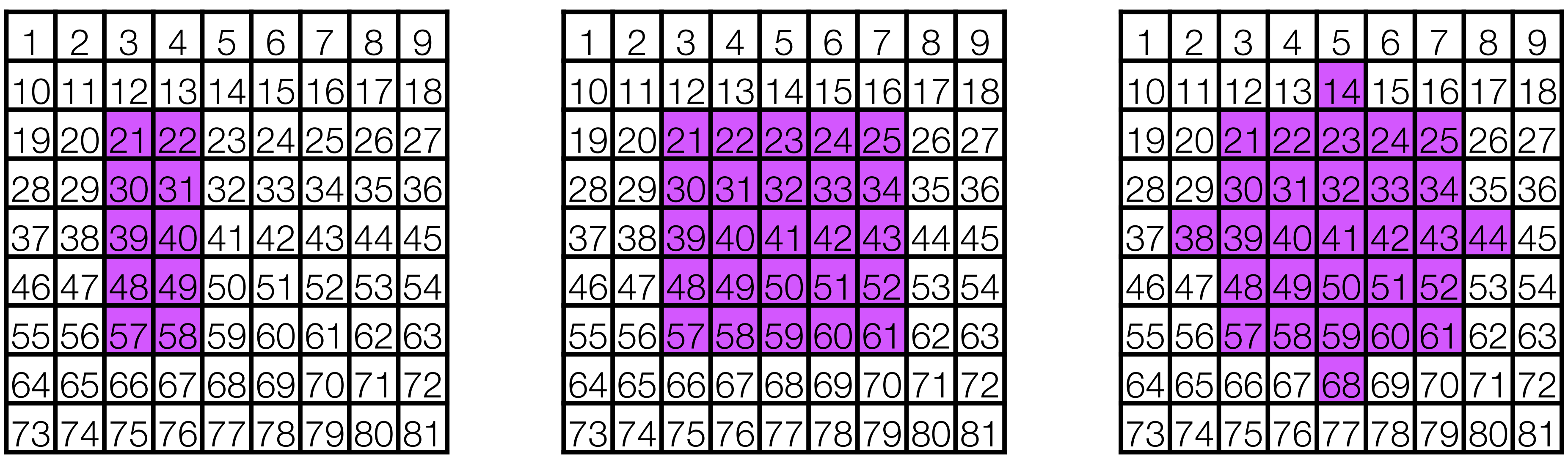
  Optimal under uniform prior.

  Simple cloaking.



Privacy versus utility: evaluation

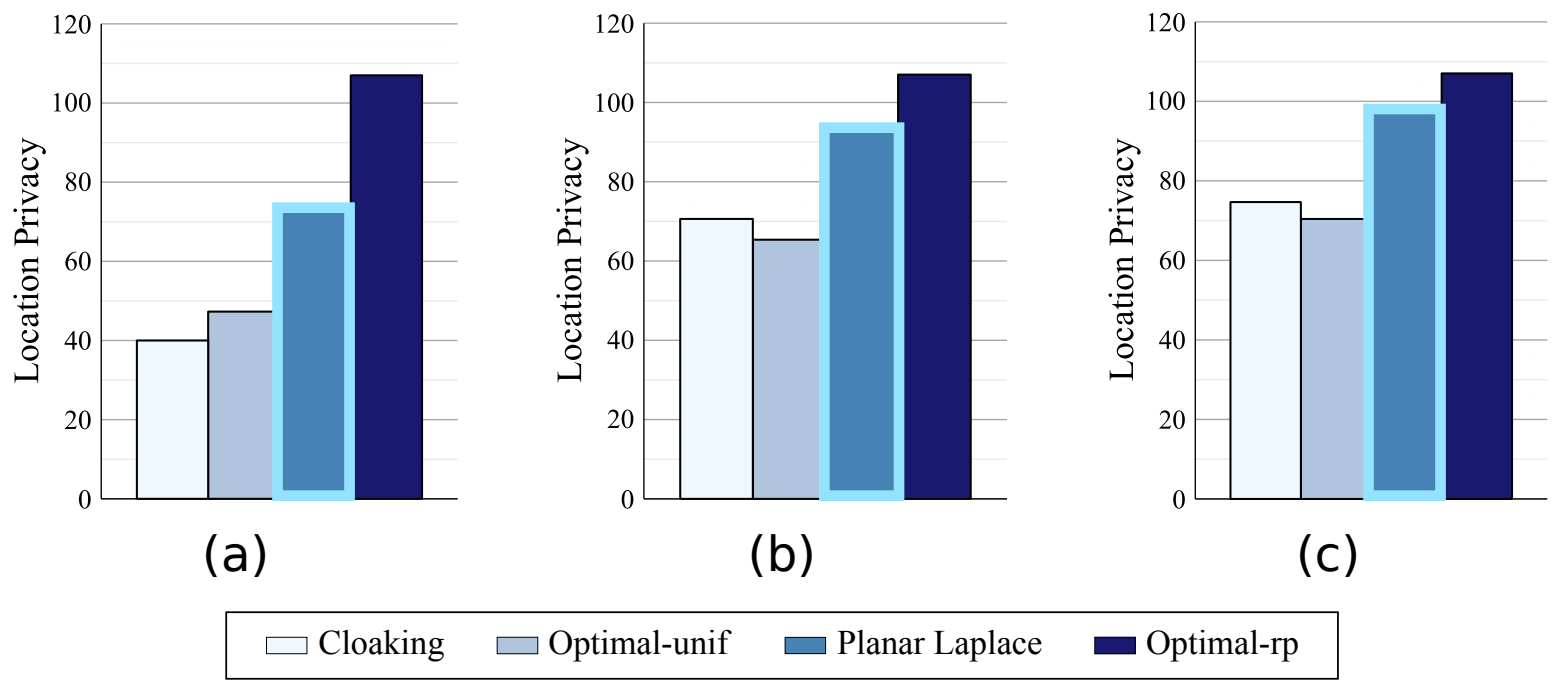
- We fix the utility and measured the privacy.
- Utility measured as the **expected error of the LBS** [Shroki et al., S&P 2012]
- Privacy measured as the **expected error of the attacker** (using prior information) [Shroki et al., S&P 2012]
- Priors: uniform over colored regions



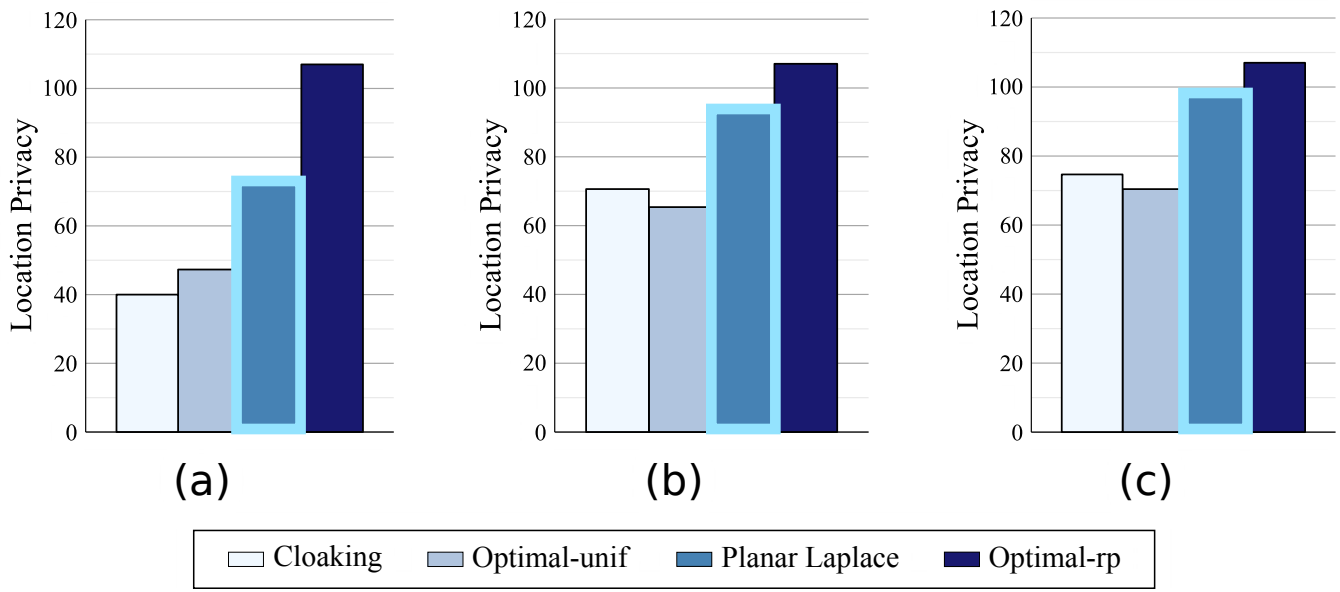
Privacy versus utility: evaluation

The four mechanisms:

- Cloaking,
- Optimal by [Shroki et al. S&P 2012] for the uniform prior
- Ours (Planar Laplacian)
- Optimal by [Shroki et al. S&P 2012] for the given prior



Privacy versus utility: evaluation

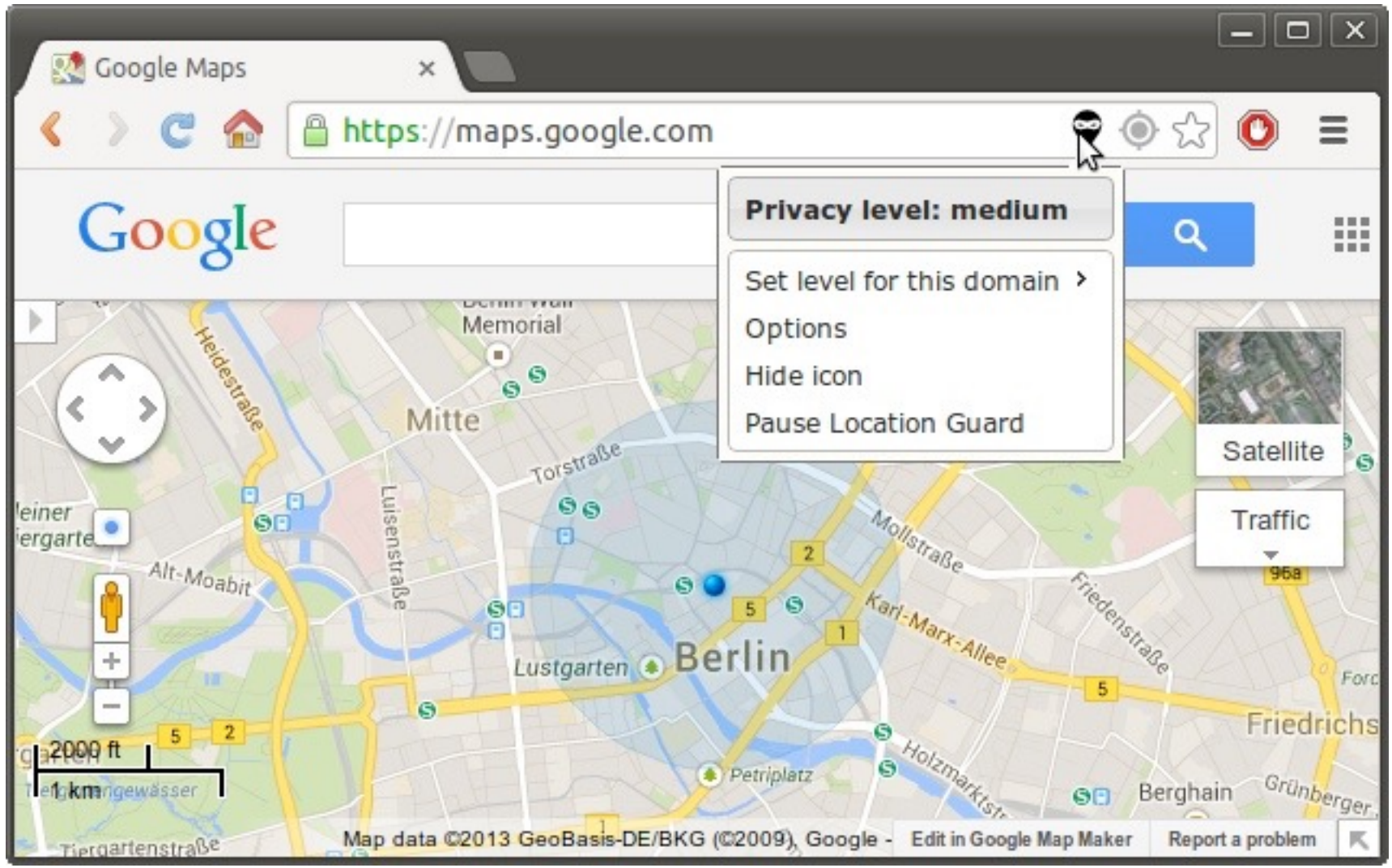


With respect to the privacy measures proposed by [Shokri et al, S&P 2012], our mechanism performs better than the other mechanisms proposed in the literature which are independent from the prior (and therefore from the adversary)

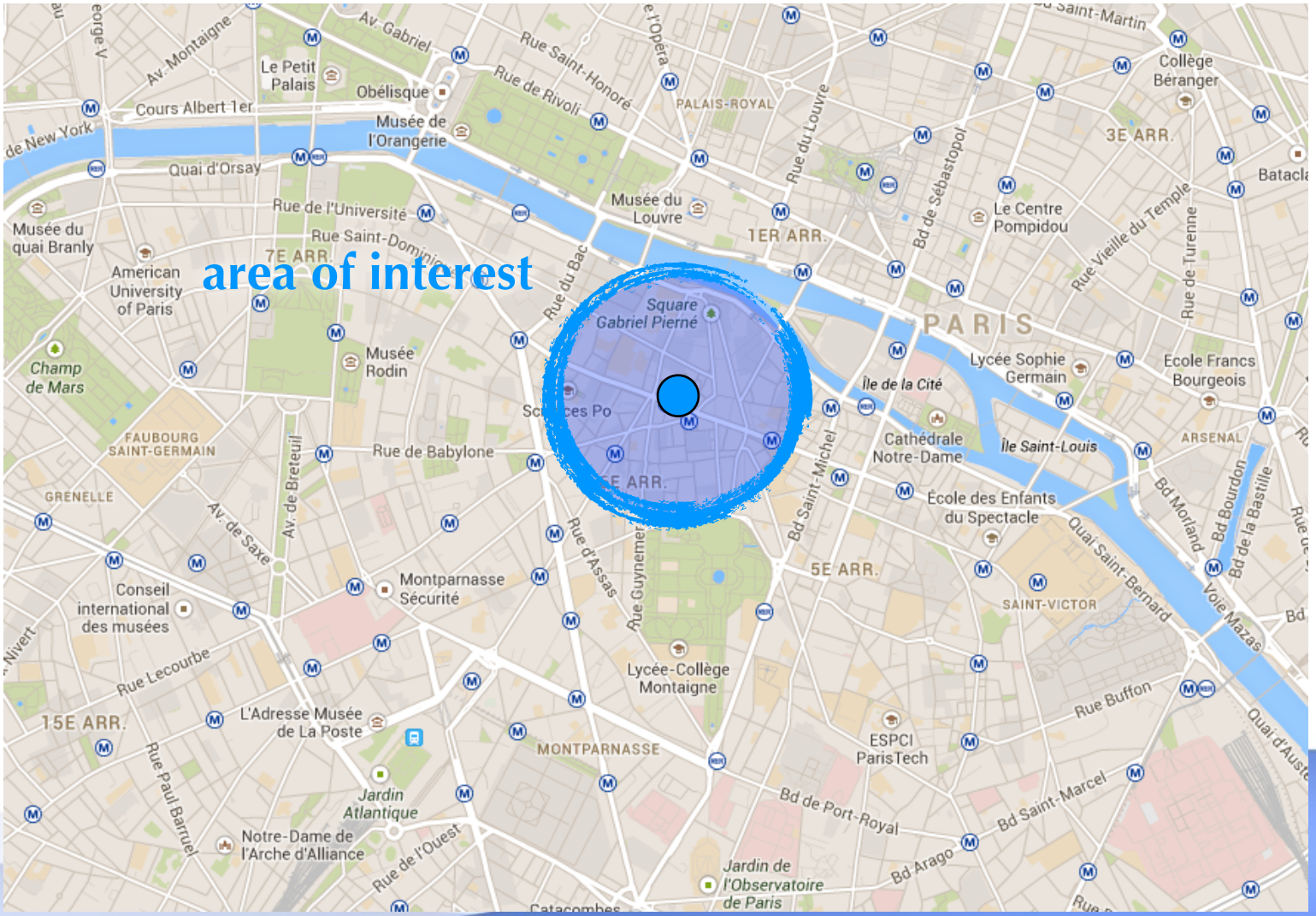
The only mechanism that outperforms ours is the optimal by [Shokri et al, S&P 2012] for the given prior, but that mechanism is adversary-dependent

Our tool: "Location Guard" for Chrome

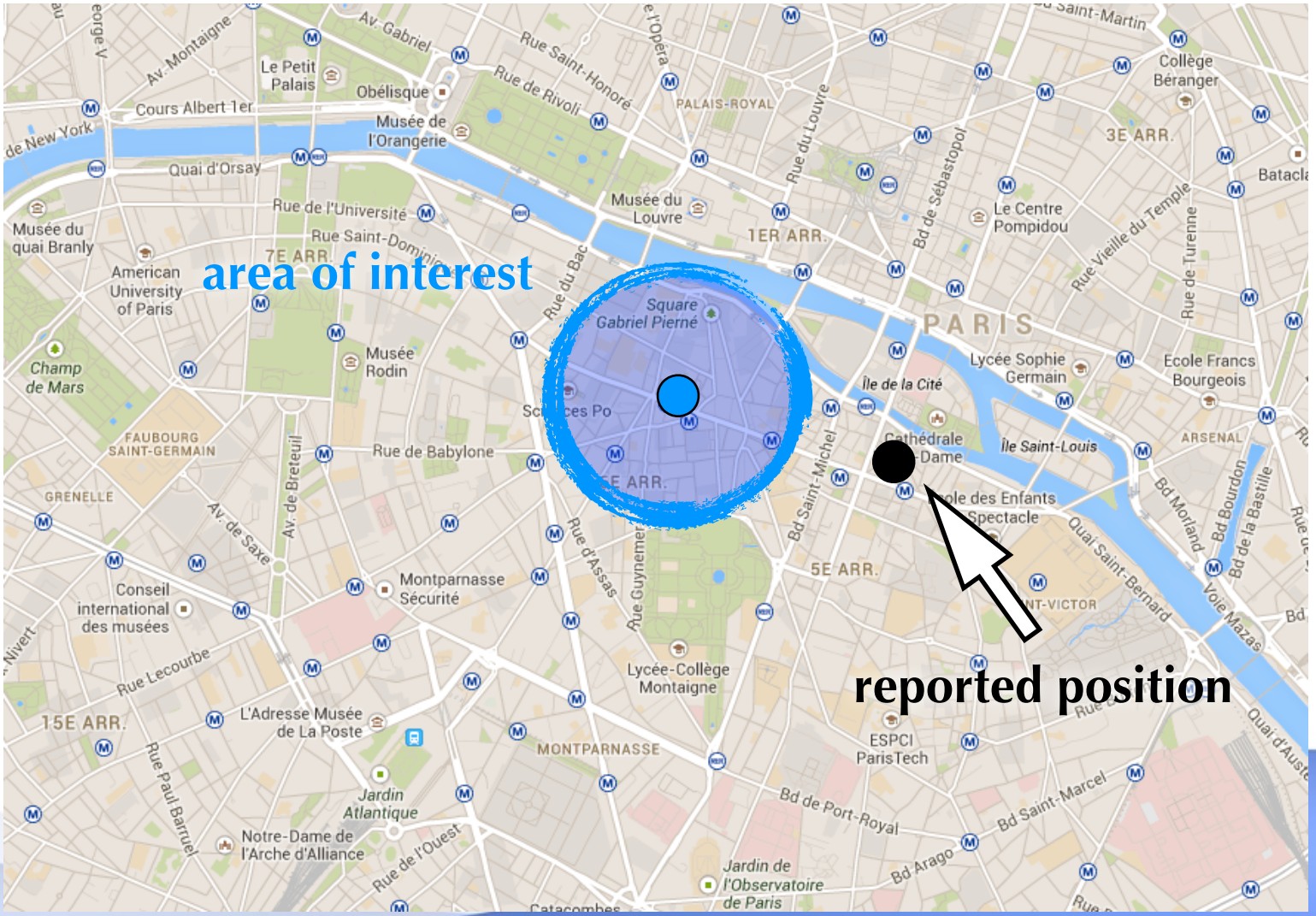
http://www.lix.polytechnique.fr/~kostas/software.html



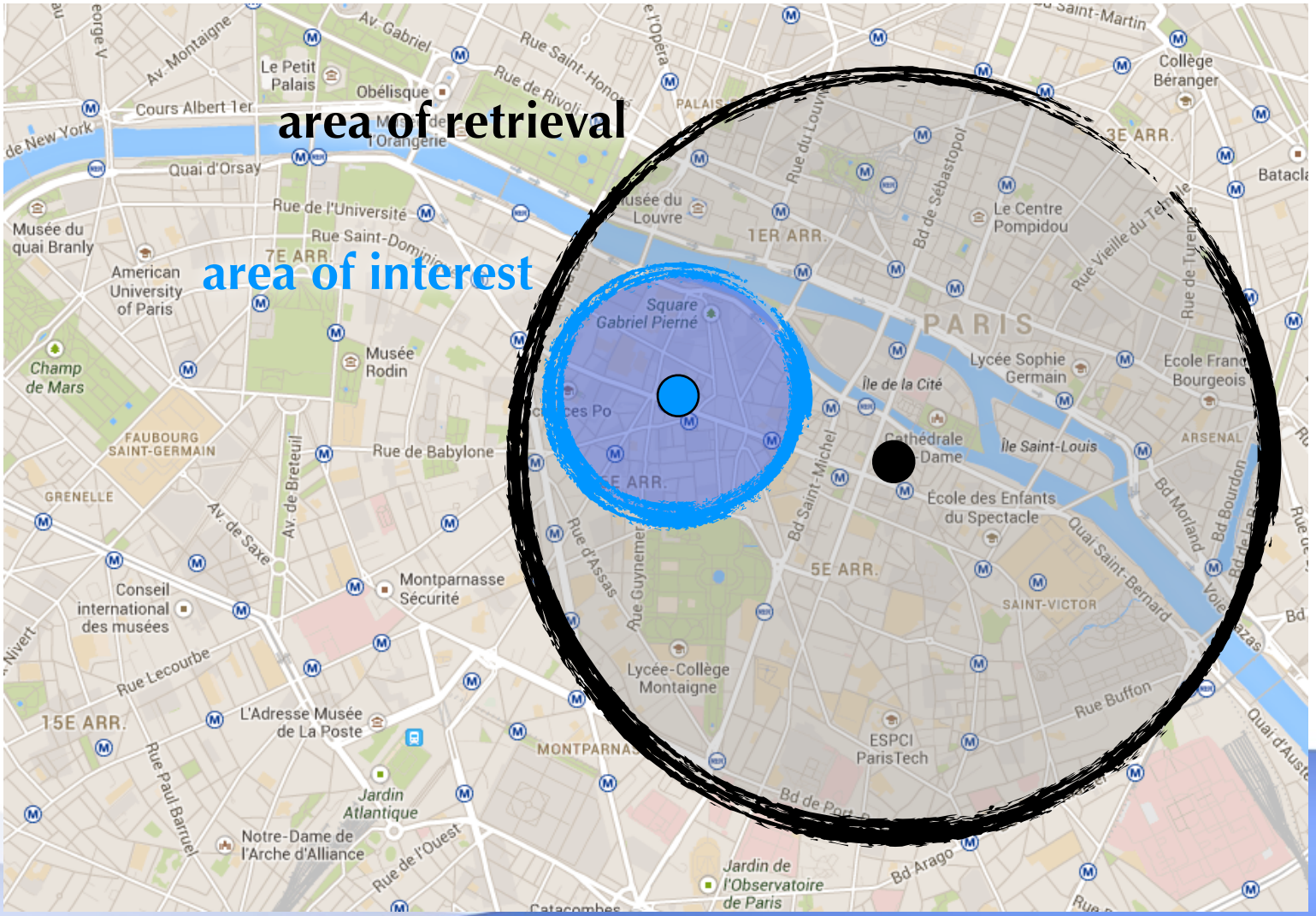
Location guard for Chrome



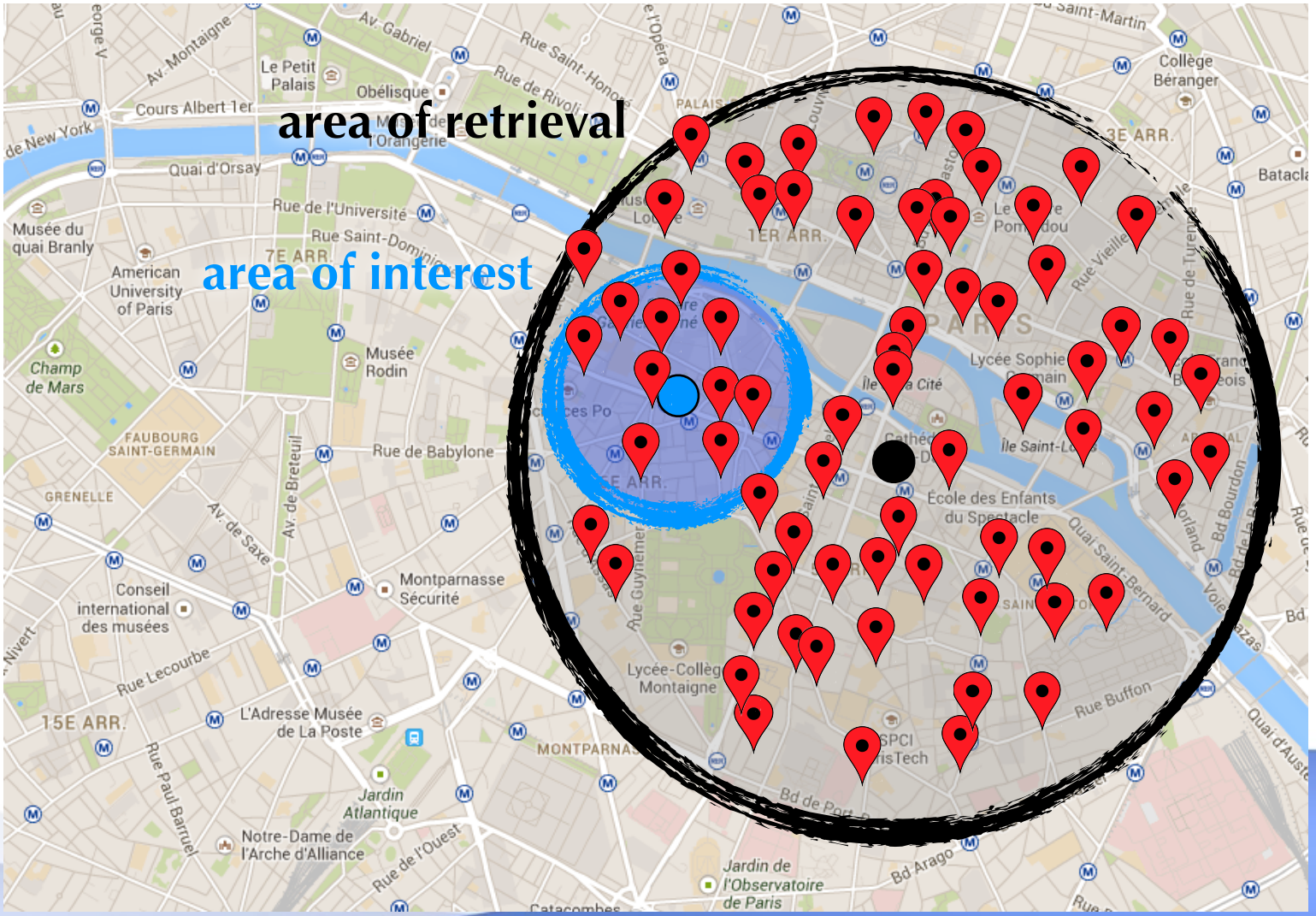
Location guard for Chrome



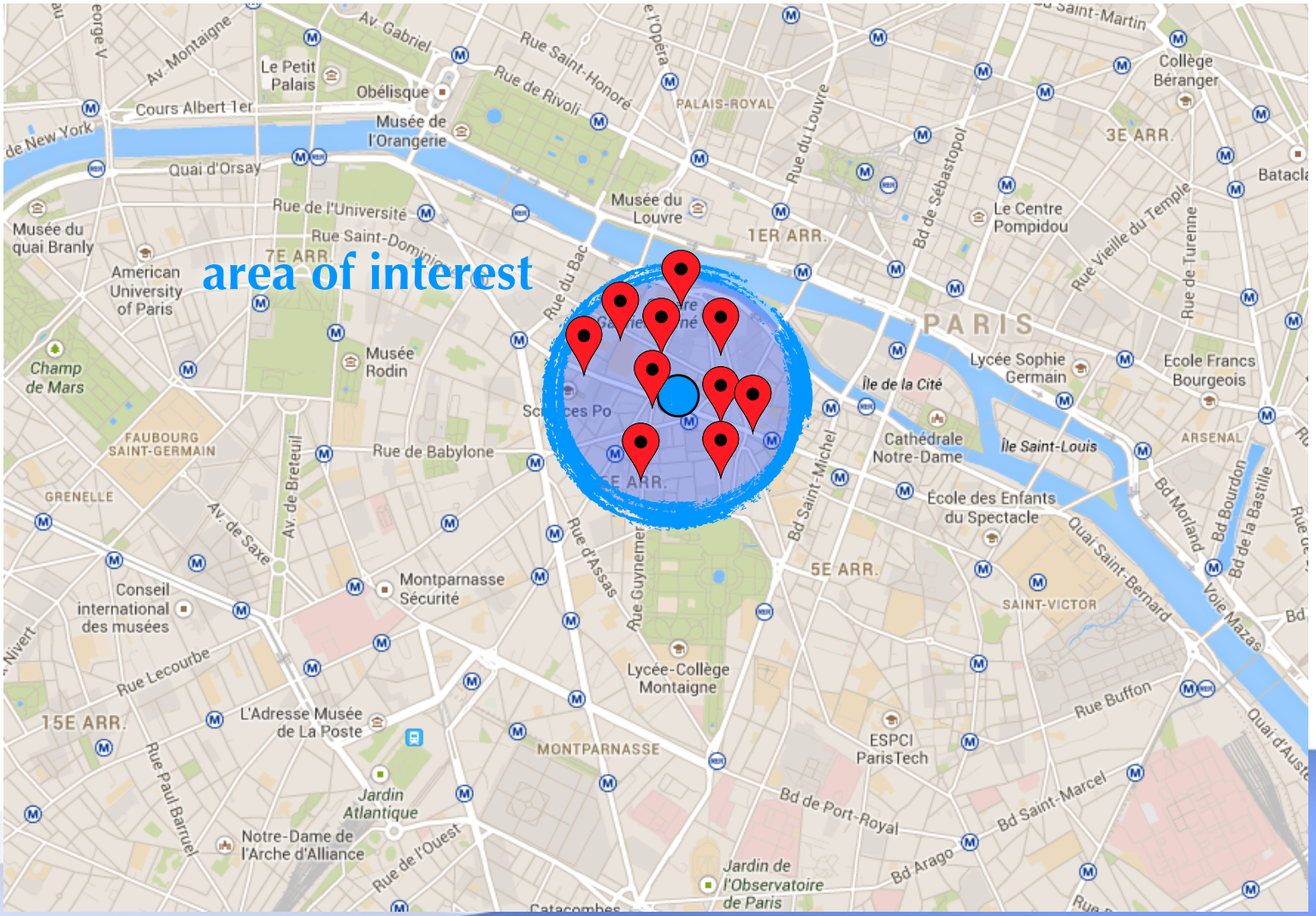
Location guard for Chrome



Location guard for Chrome



Location guard for Chrome



Conclusion

- Differential privacy
- Generalization of differential privacy to arbitrary metric domains: **d-privacy**
- Application to location privacy

Future directions

- It many be interesting to explore other applications of d-privacy. In general d-privacy can be useful when we aim at protecting the **precision** of the information.
  - Signatures of electrical appliances - smart meters
  - Privacy-friendly biological data

Thank you !

Some bibliography

C. Dwork. A firm foundation for private data analysis.
Communications of the ACM, 54(1):86–96, 2011.

M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, and C. Palamidessi.
Quantitative Information Flow and Applications to Differential Privacy.
In A. Aldini and R. Gorrieri, editors, Foundations of Security Analysis and Design VI – FOSAD Tutorial Lectures, volume 6858 of Lecture Notes in Computer Science, pages 211–230. Springer, 2011.

K. Chatzikokolakis, M. E. Andrés, N. E. Bordenabe, C. Palamidessi.
Broadening the scope of Differential Privacy using metrics.
Proc. of PETS 2013.

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

N. E. Bordenabe, K. Chatzikokolakis, C. Palamidessi.
Optimal Geo-Indistinguishable Mechanisms for Location Privacy.
Technical Report (Feb 2014).