Differential privacy and applications to location privacy

Catuscia Palamidessi
INRIA & Ecole Polytechnique

Plan of the talk

- General introduction to privacy issues
- A naive approach to privacy protection: anonymization
- Why it is so difficult to protect privacy: Focus on Statistical Databases
- Differential Privacy: adding controlled noise
- Utility and trade-off between utility and privacy
- Extensions of DP
- Application to Location Privacy: Geo-indistinguishability

Digital traces

In the "Information Society", each individual constantly leaves digital traces of his actions that may allow to infer a lot of information about himself

IP address ⇒ location.

History of requests ⇒ interests.

Activity in social networks ⇒ political opinions, religion, hobbies, . . .

Power consumption (smart meters) ⇒ activities at home.

Risk: collect and use of digital traces for fraudulent purposes.

Examples: targeted spam, identity theft, profiling, discrimination, . . .

Privacy via anonymity

Nowadays, organizations and companies that collect data are usually obliged to sanitize them by making them anonymous, i.e., by removing all personal identifiers: name, address, SSN, …

"We don't have any raw data on the identifiable individual. Everything is anonymous"
(CEO of NebuAd, a U.S. company that offers targeted advertising based on browsing histories)

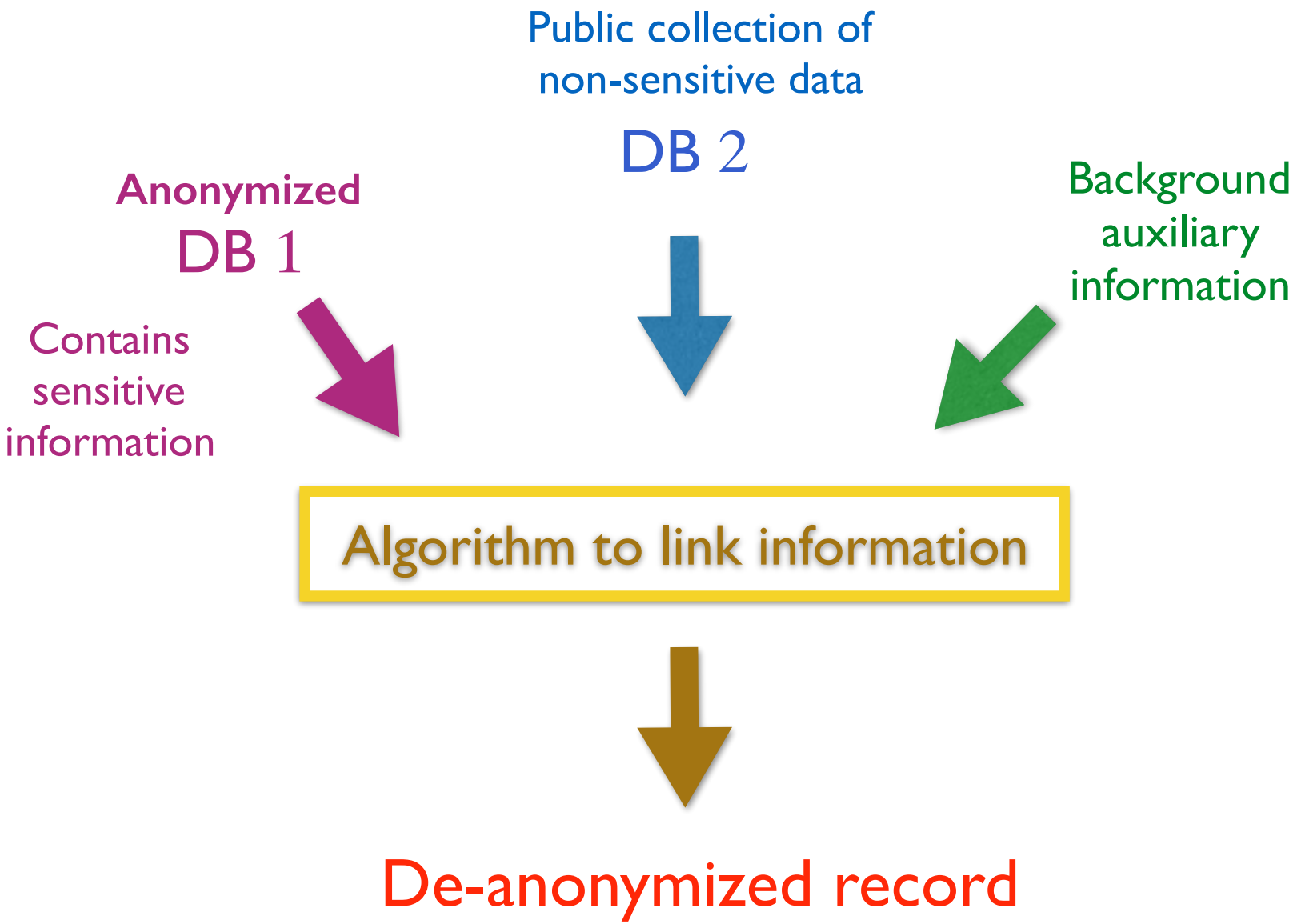
Similar practices are used by Facebook, MySpace, Google, …

Privacy via anonymity

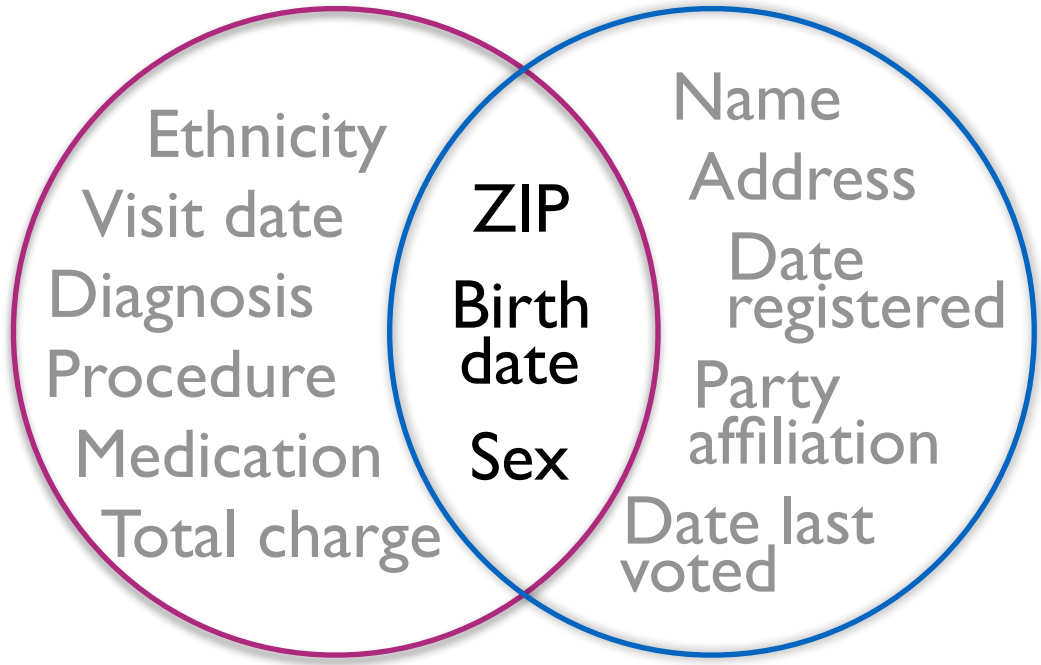
However, anonymity-based sanitization has been shown to be **highly ineffective:** Several de-anonymization attacks have been carried out in the last decade

- The quasi-identifiers allow to retrieve the identity in a large number of cases.
- More sophisticated methods (k-anonymity, ℓ-anonymity, …) take care of the quasi-identifiers, but they are still prone to composition attacks

Sweeney's de-anonymization attack by linking

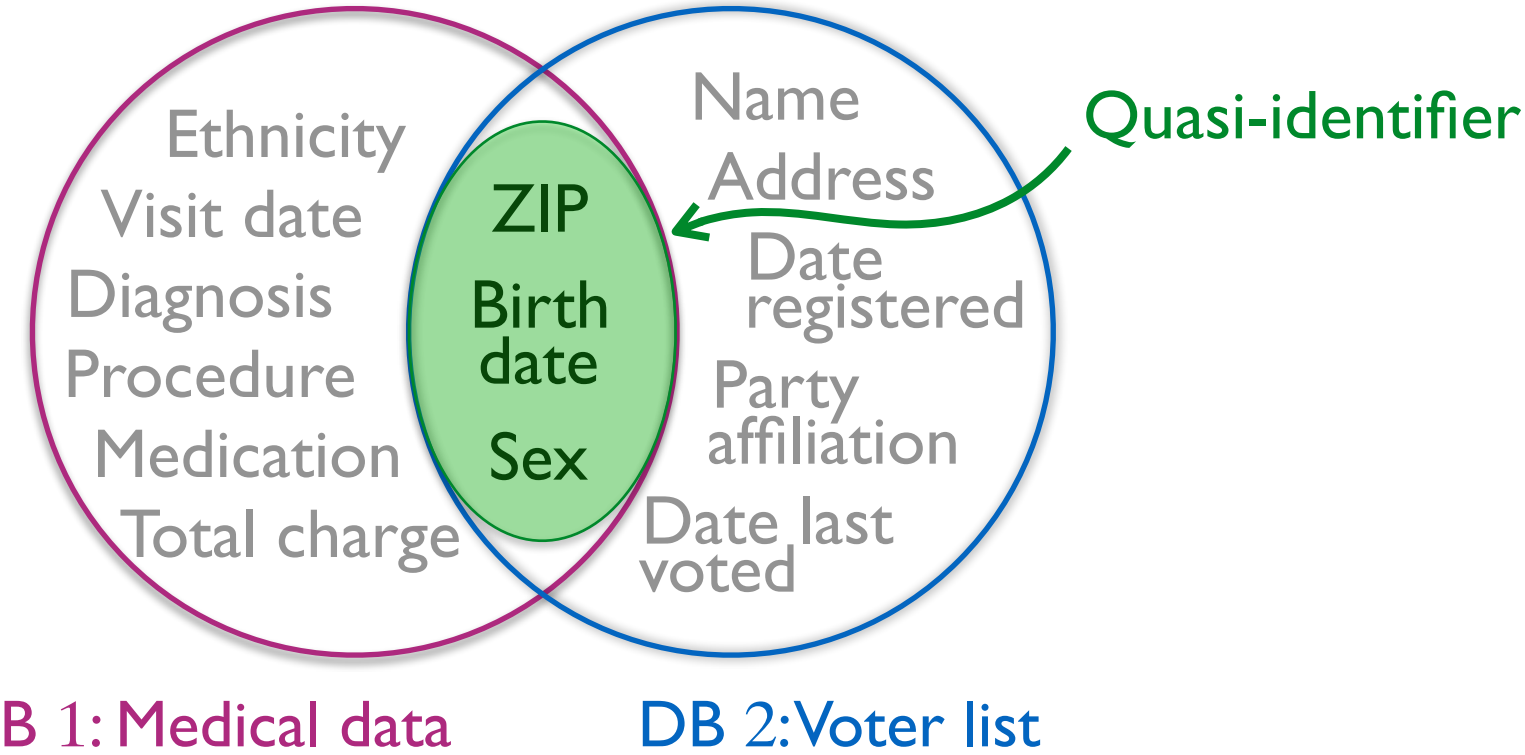


Sweeney's de-anonymization attack by linking



87 % of the US population is **uniquely identifiable** by ZIP, gender, DOB

Sweeney's de-anonymization attack by linking



87 % of the US population is **uniquely identifiable** by ZIP, gender, DOB

K-anonymity

- **Quasi-identifier:** Set of attributes that can be linked with external data to uniquely identify individuals
- **K-anonymity approach**: Make every record in the table indistinguishable from a least $k-1$ other records with respect to quasi-identifiers. This is done by:
  - suppression of attributes, and/or
  - generalization of attributes, and/or
  - addition of dummy records
- In this way, linking on quasi-identifiers yields at least k records for each possible value of the quasi-identifier

K-anonymity

Example: 4-anonymity w.r.t. the quasi-identifier {nationality, ZIP, age}

achieved by suppressing the nationality and generalizing ZIP and age

	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

Figure 1. Inpatient Microdata

	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

Figure 2. 4-anonymous Inpatient Microdata

Composition attacks I

Showed the limitations of K-anonymity

Robust De-anonymization of Large Sparse Datasets. Arvind Narayanan and Vitaly Shmatikov, 2008.

They applied de-anonymization to the **Netflix Prize dataset** (which contained anonymous movie ratings of 500,000 subscribers of Netflix), in combination with the **Internet Movie Database** as the source of background knowledge. They demonstrated that an adversary who knows only a little bit about an individual subscriber can identify his record in the dataset, uncovering his apparent political preferences and other potentially sensitive information.

Composition attacks 2

De-anonymizing Social Networks. Arvind Narayanan and Vitaly Shmatikov, 2009.

By using only the network topology, they were able to show that a third of the users who have accounts on both **Twitter** (a popular microblogging service) and **Flickr** (an online photo-sharing site), can be re-identified in the anonymous Twitter graph with only a 12% error rate.

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
- For instance, medical databases are often used for research purposes. Typically we are interested in studying the correlation between certain diseases, and certain other attributes: age, sex, weight, etc.
  - A typical query would be: "*Among the people affected by the disease, what percentage is over 60 ?* "
  - Personal queries are forbidden. An example of forbidden query would be: " *Does Don have the disease ?* "

The problem

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

- 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 should correspond to at least k individuals

Alice	Bob
Carl	Don
Ellie	Frank

The problem

Unfortunately, it 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

A better 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

Noisy mechanisms

- The mechanisms reports an approximate answer, typically generated randomly on the basis of the true answer and of some probability distribution
- The probability distribution must be chosen carefully, in order to not destroy the utility of the answer
- A good mechanism should provide a good trade-off between privacy and utility. Note that, for the same level of privacy, different mechanism may provide different levels of utility.

Differential Privacy

Definition [Dwork 2006]: a randomized mechanism $\mathcal{K}$ provides ε-differential privacy if for all databases x, x' **which are adjacent** (i.e., differ for only one record), and for all $z \in \mathcal{Z}$, we have

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

- The answer by K does not change significantly the knowledge about X
- Differential privacy is robust with respect to composition of queries
- The definition of differential privacy is independent from the prior

Typical implementation of differential privacy: add Laplacian noise

- Randomized mechanism for a query $f: \mathcal{X} \rightarrow \mathcal{Y}$.
- **Add Laplacian noise.** If the exact answer is y, the reported answer is z, with a probability density function defined as:

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

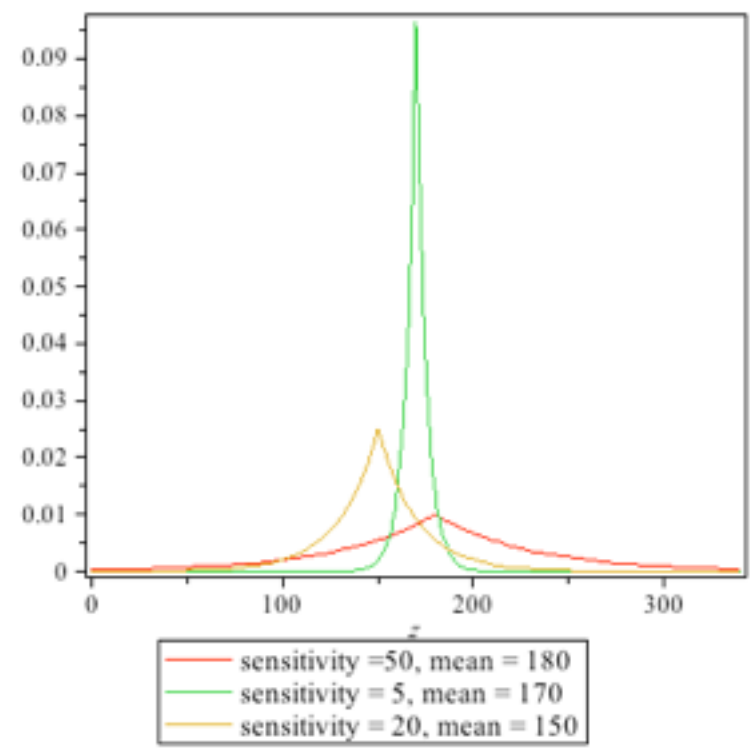
where Δf is the *sensitivity* of f:

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

($x \sim x'$ means x and x' are adjacent, i.e., they differ only for one record)

and c is a normalization factor:

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



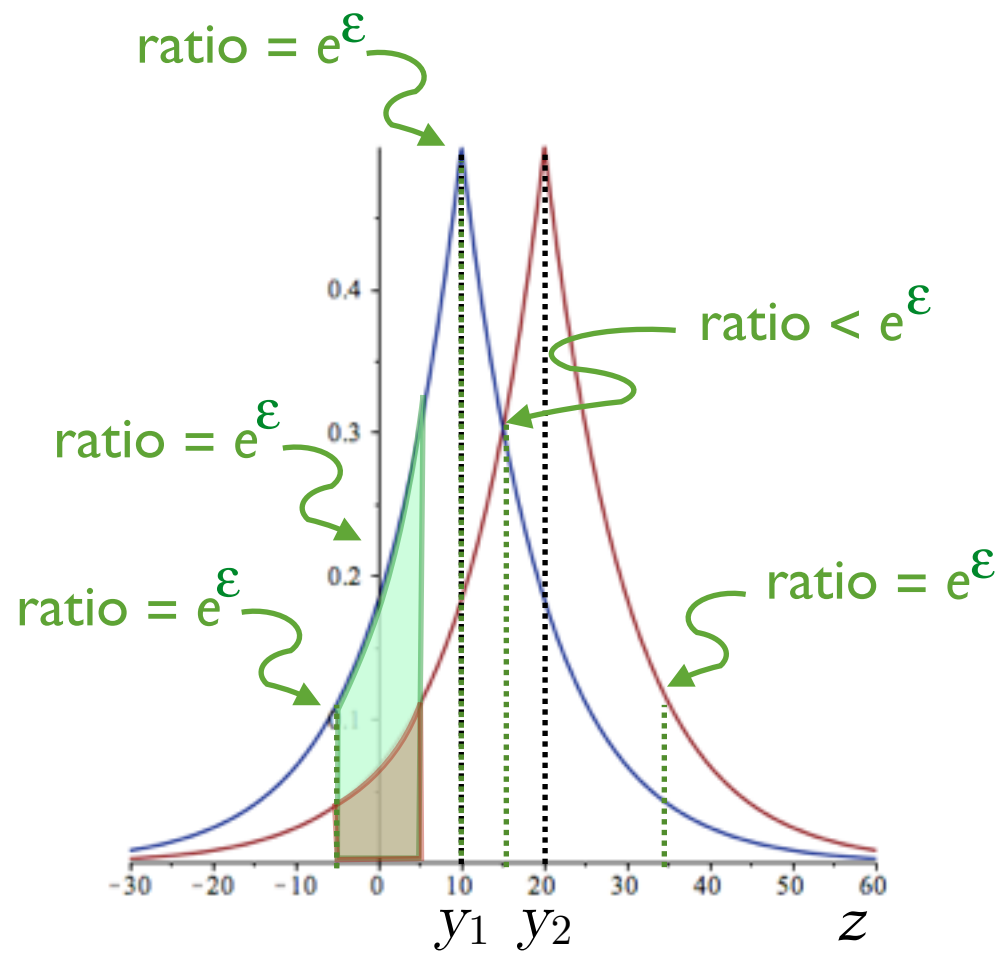
Intuition behind the Laplace distribution

Assume for example

- $\Delta_f = |f(x_1) - f(x_2)| = 10$
- $y_1 = f(x_1) = 10$, $y_2 = f(x_2) = 20$

Then:

- $dP_{y_1}(z) = \frac{\varepsilon}{2 \cdot 10} e^{\frac{|z-10|}{10}\varepsilon}$
- $dP_{y_2}(z) = \frac{\varepsilon}{2 \cdot 10} e^{\frac{|z-20|}{10}\varepsilon}$



The ratio between these distribution is

- $= e^{\varepsilon}$ outside the interval $[y_1, y_2]$
- $\leq e^{\varepsilon}$ inside the interval $[y_1, y_2]$

Note that the distance between y_1 and y_2 is greatest when y_1 and y_2 correspond to the sensitivity of f. In this case the ratio between the respective Laplaces is e^{ε}. In all other cases, the distance between y_1 y_2 is smaller, and therefore also the ratio is smaller. Similar considerations hold for the geometric mechanism.

Some prototypes implementing DP on DBs

- PINK
  http://research.microsoft.com/en-us/projects/pinq/
- FUZZ
  http://privacy.cis.upenn.edu/software.html
- AIRAVAT
  http://z.cs.utexas.edu/users/osa/airavat/
- GUPT
  https://github.com/prashmohan/GUPT

Some applications of DP

- The Census Bureau project *OnTheMap*, which allows to give researchers access to the data of the agency while protecting the privacy of the citizens
  http://www.scientificamerican.com/article/privacy-by-the-numbers-a-new-approach-to-safeguarding-data/

- Google' RAPPOR: Randomized Aggregatable Privacy Preserving Ordinal Response.
  Used for collecting statistics from end-user
  http://www.computerworld.com/article/2841954/googles-rappor-aims-to-preserve-privacy-while-snaring-software-stats.html

Extending differential privacy to arbitrary metrics

Differential Privacy:

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

$$\frac{p(z \mid x)}{p(z \mid x')} \leq e^{\varepsilon\, d_H(x,x')}$$

where d_H is the Hamming distance between x and x', i.e., the number of records in which x and x' differ

Generalization: d-privacy

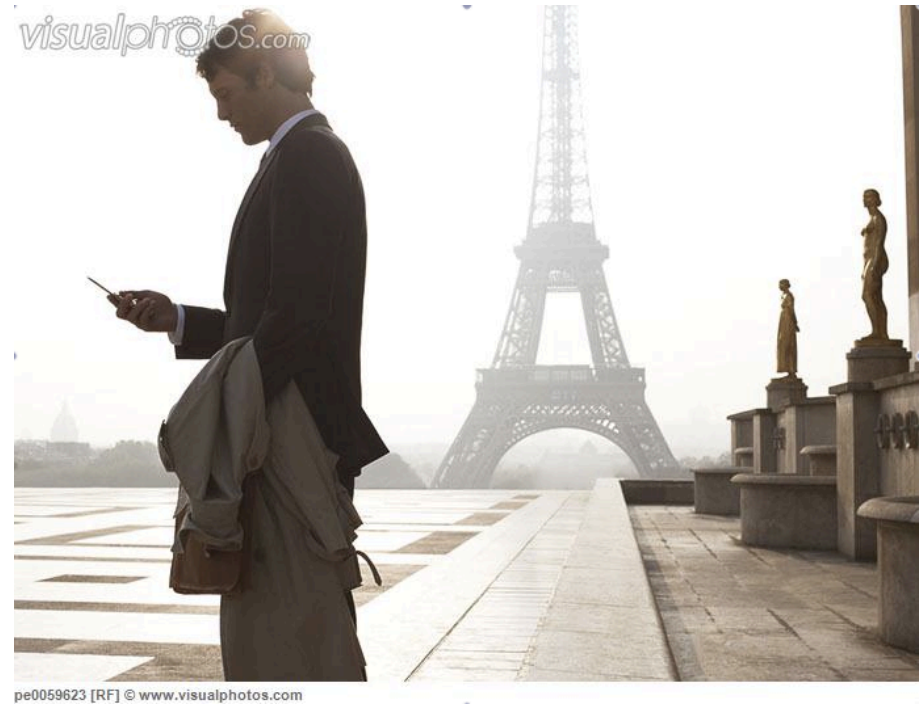
On a generic domain $\mathcal{X}$ provided with a distance d:

$$\forall x, x' \in \mathcal{X}, \forall z \quad \frac{p(z \mid x)}{p(z \mid x')} \leq e^{\varepsilon\, d(x,x')}$$

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

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

where r is the distance between x and x'

geo-indistinguishability

Alternative characterization

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

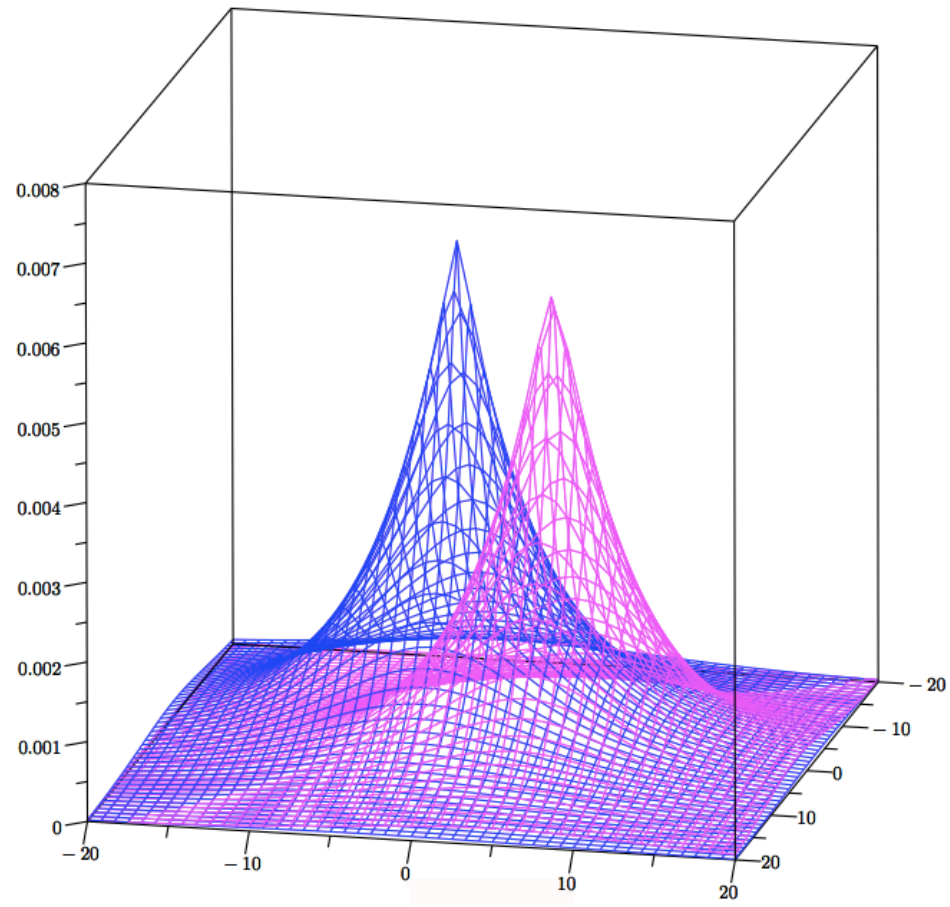
A *d*-private mechanism for LBS:
Planar laplacian

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

We have compared the trade off utility-privacy of our mechanism (Planar laplacian) with three other mechanisms in the literature:

- The Optimal Mechanism by Shroki et al., [S&P 2012]. Note that this mechanism is prior-dependent: it is specifically generated assuming a certain adversary (with a certain prior knowledge). Our mechanism, in contrast, is prior-independent. The Optimal Mechanism is obtained by linear programming techniques.

- Two prior-independent mechanisms:

  - Simple cloacking: We partition the area of interest in zones, and instead of reporting the point, we report the zone.

  - The mechanism of Shokri et al., generated assuming uniform prior.

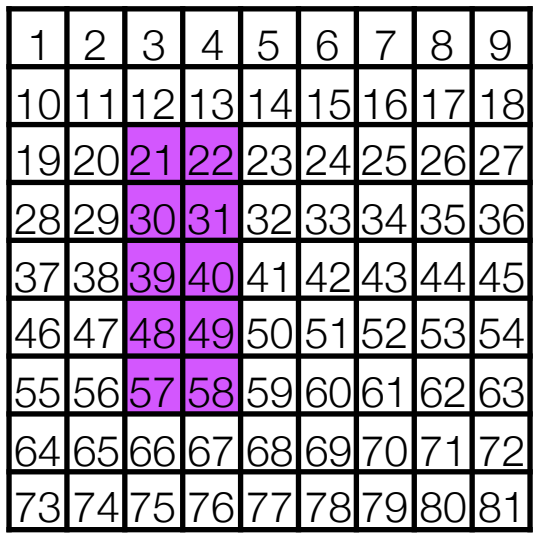
Privacy versus utility: evaluation

- We have designed an ``area of interest'' containing 9x9 = 81 "locations".
- For the cloaking mechanism, we have partitioned the area in 9 zones, indicated by the blue lines

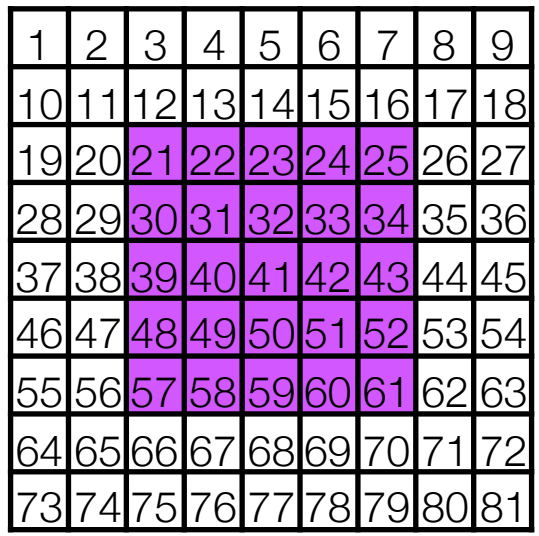
1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18
19	20	21	22	23	24	25	26	27
28	29	30	31	32	33	34	35	36
37	38	39	40	41	42	43	44	45
46	47	48	49	50	51	52	53	54
55	56	57	58	59	60	61	62	63
64	65	66	67	68	69	70	71	72
73	74	75	76	77	78	79	80	81

Privacy versus utility: evaluation

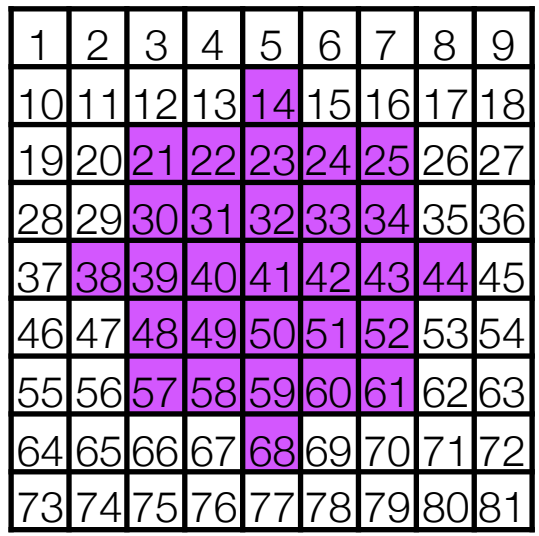
- We configured the four mechanisms so to give the same utility, and we measured their privacy.
- **Utility**: **expected distance** between the true location and the reported one (utility loss) [Shroki et al., S&P 2012]
- **Privacy**: **expected error of the attacker** (using prior information) [Shroki et al., S&P 2012]. Note that we could not use differential privacy, because our mechanism is the only one that provide differential privacy
- Priors: concentrated over colored regions



(a)



(b)

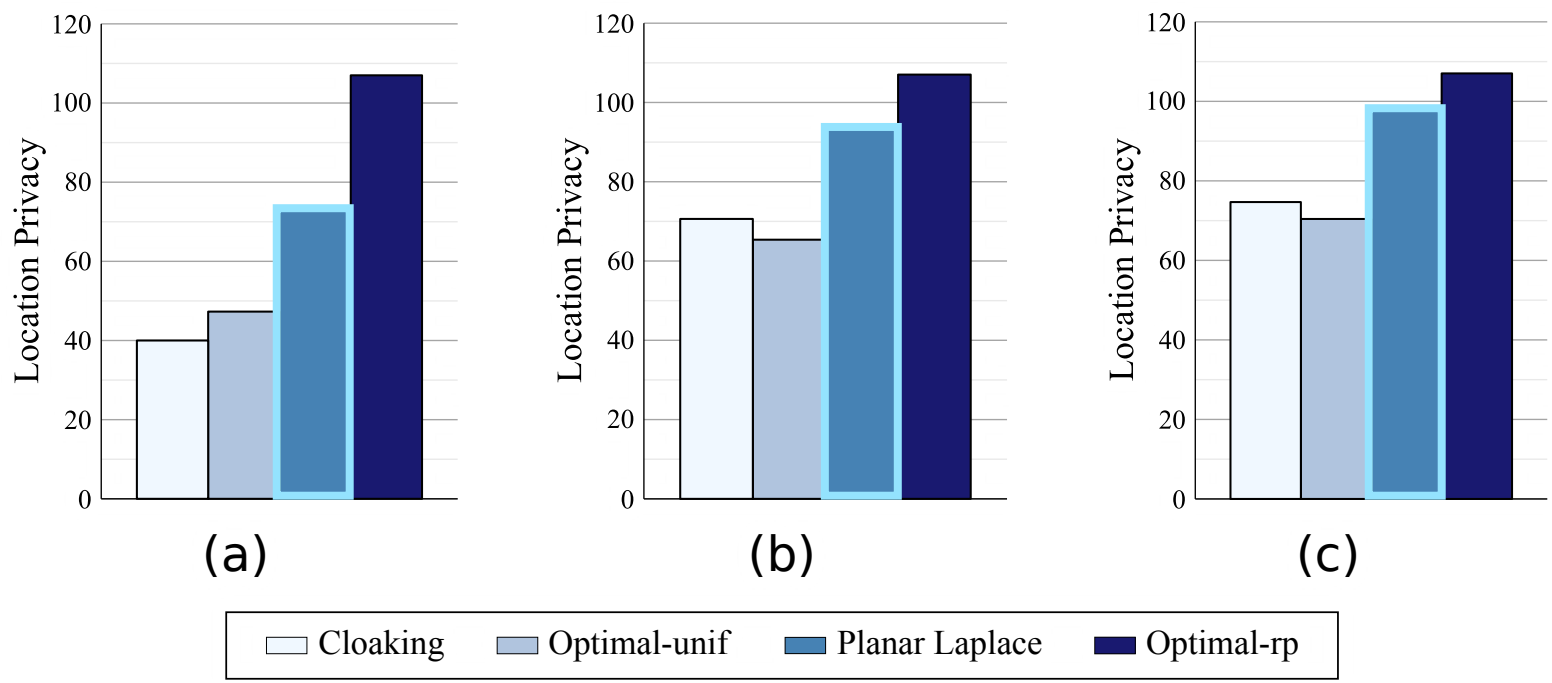


(c)

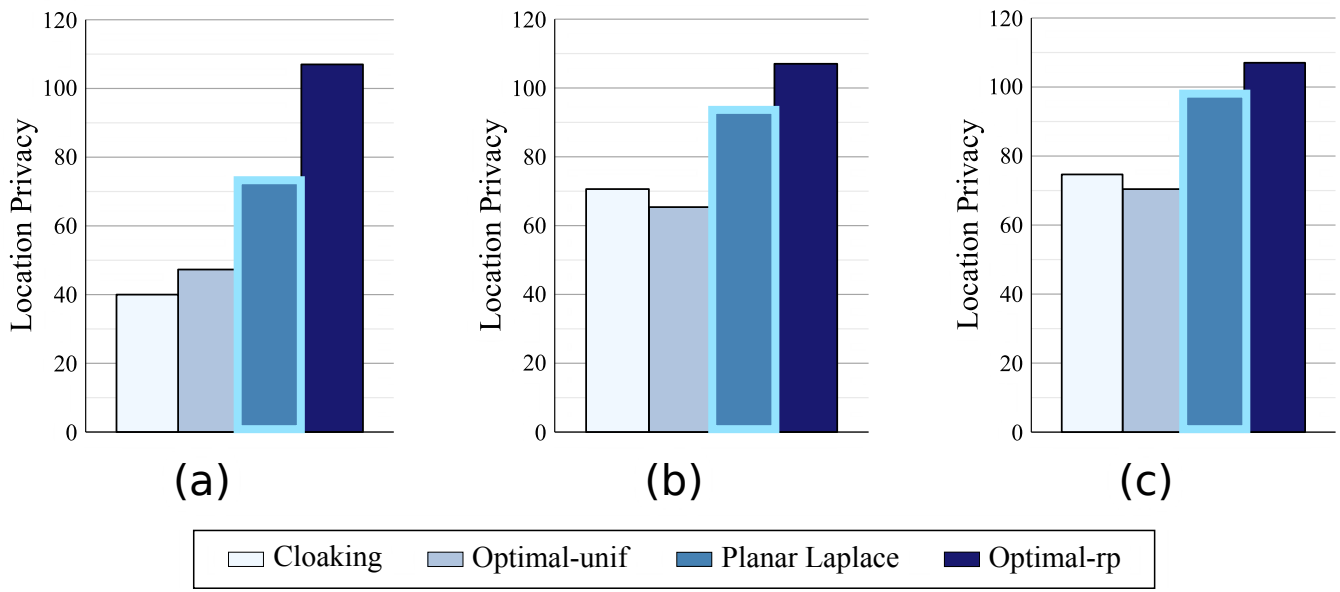
Privacy versus utility: evaluation

The four mechanisms:

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



Privacy versus utility: evaluation



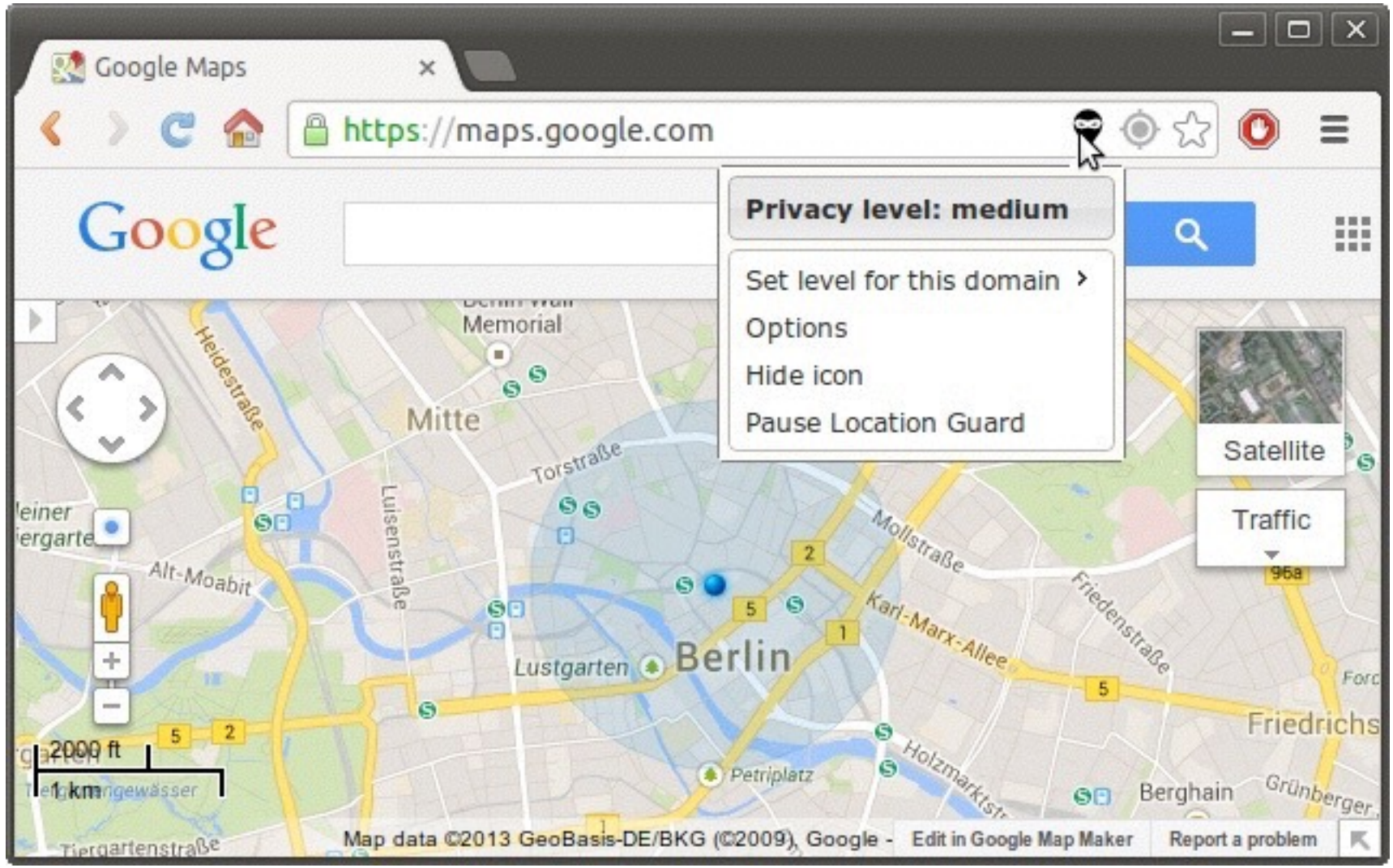
With respect to the privacy measures proposed by [Shokri et al, S&P 2012], our mechanism performs better than the other mechanisms proposed in the literature which are independent from the prior (and therefore from the adversary)

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

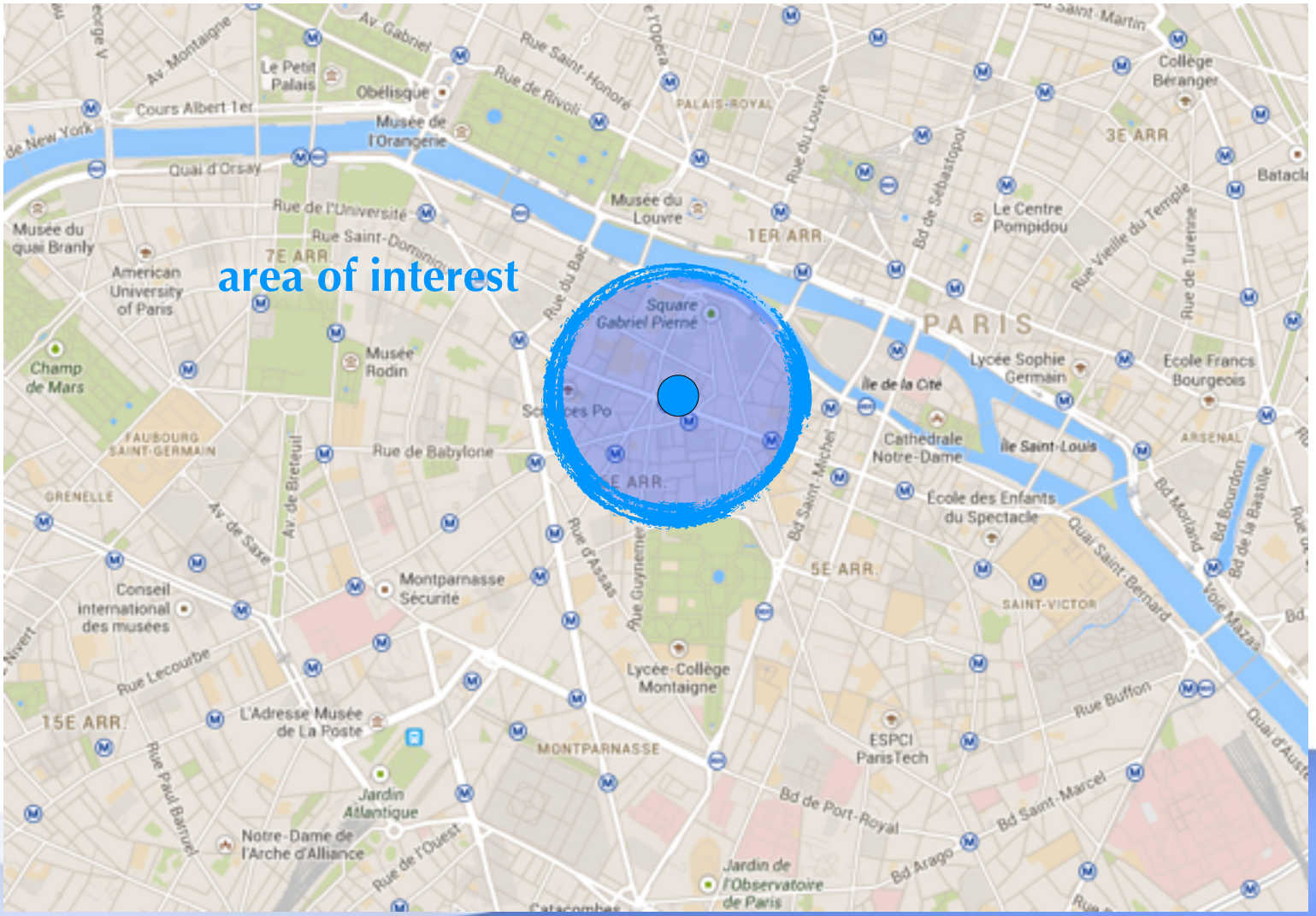
Tool: "Location Guard"

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

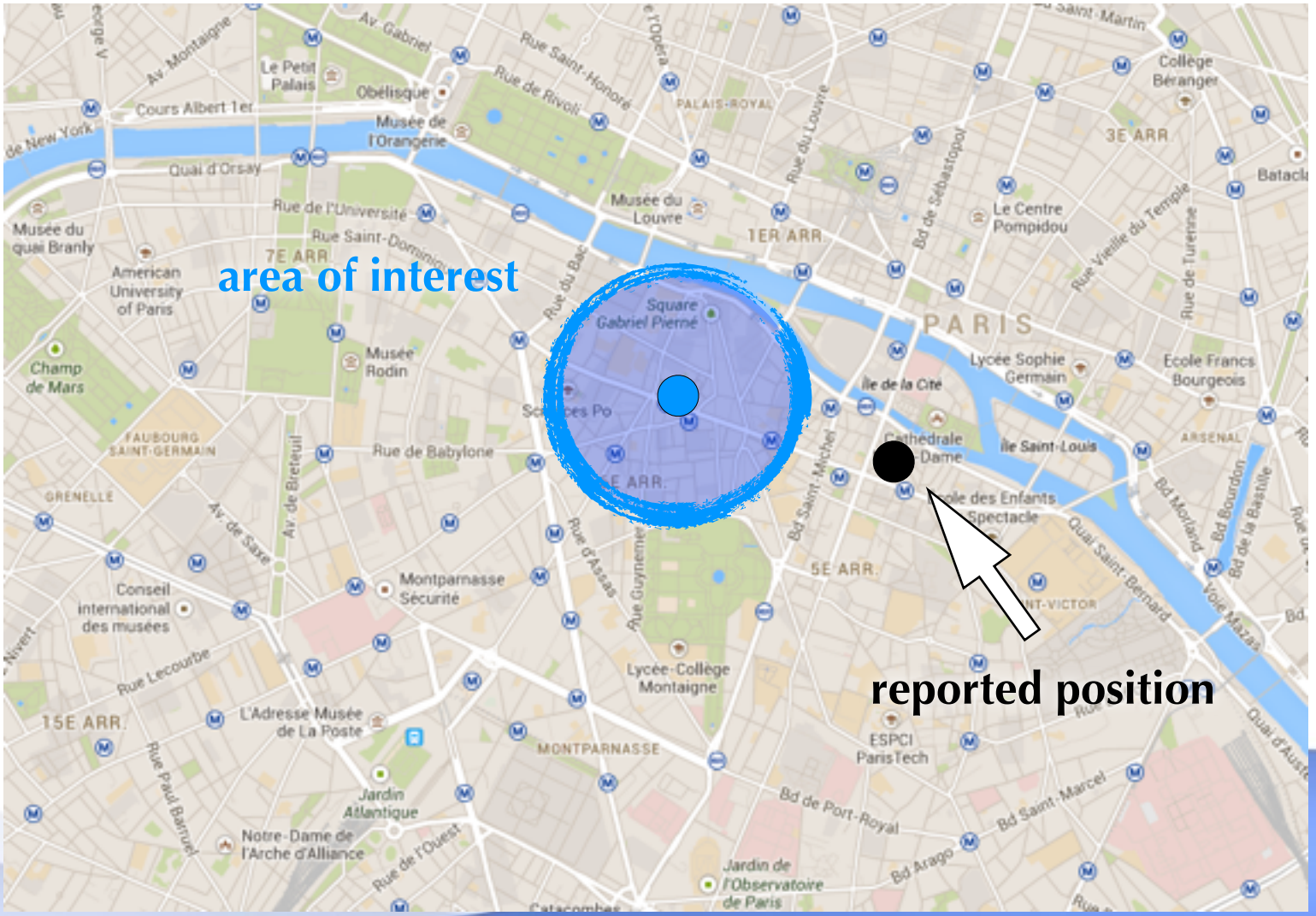
Extension for Firefox, Chrome, and Opera. It has been released about one year ago, and nowadays it has about 60,000 active users.



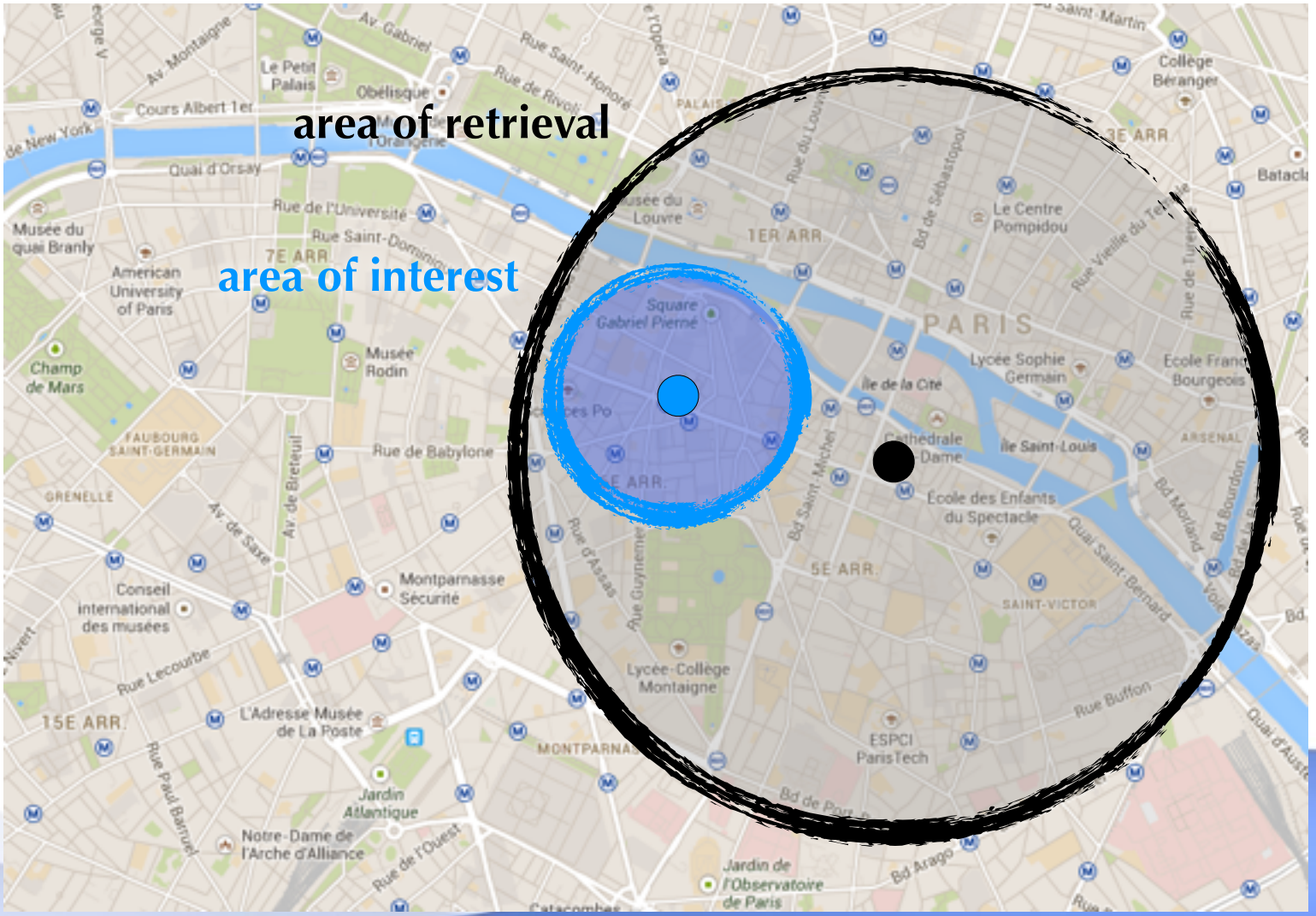
Location guard for Chrome



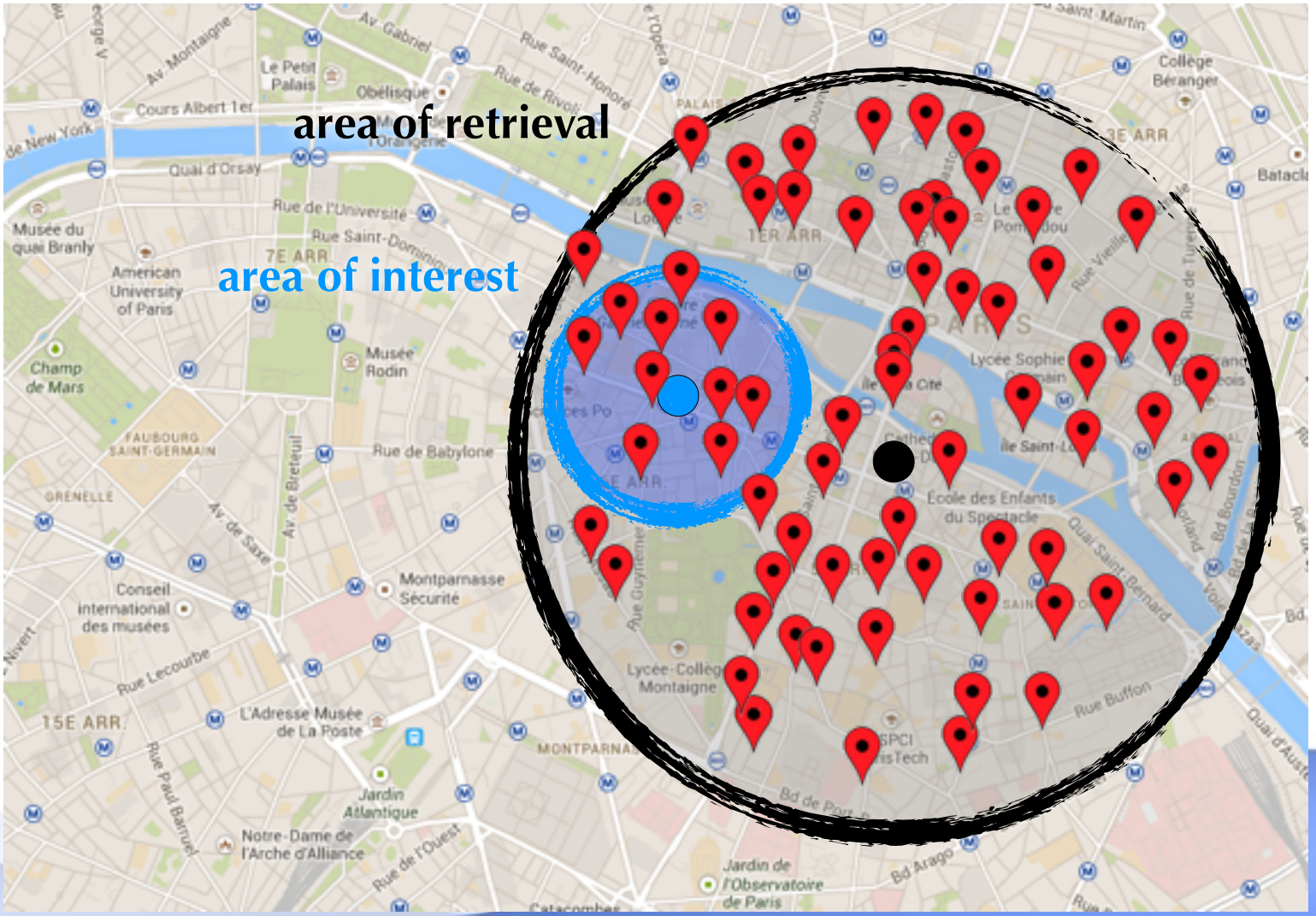
Location guard for Chrome



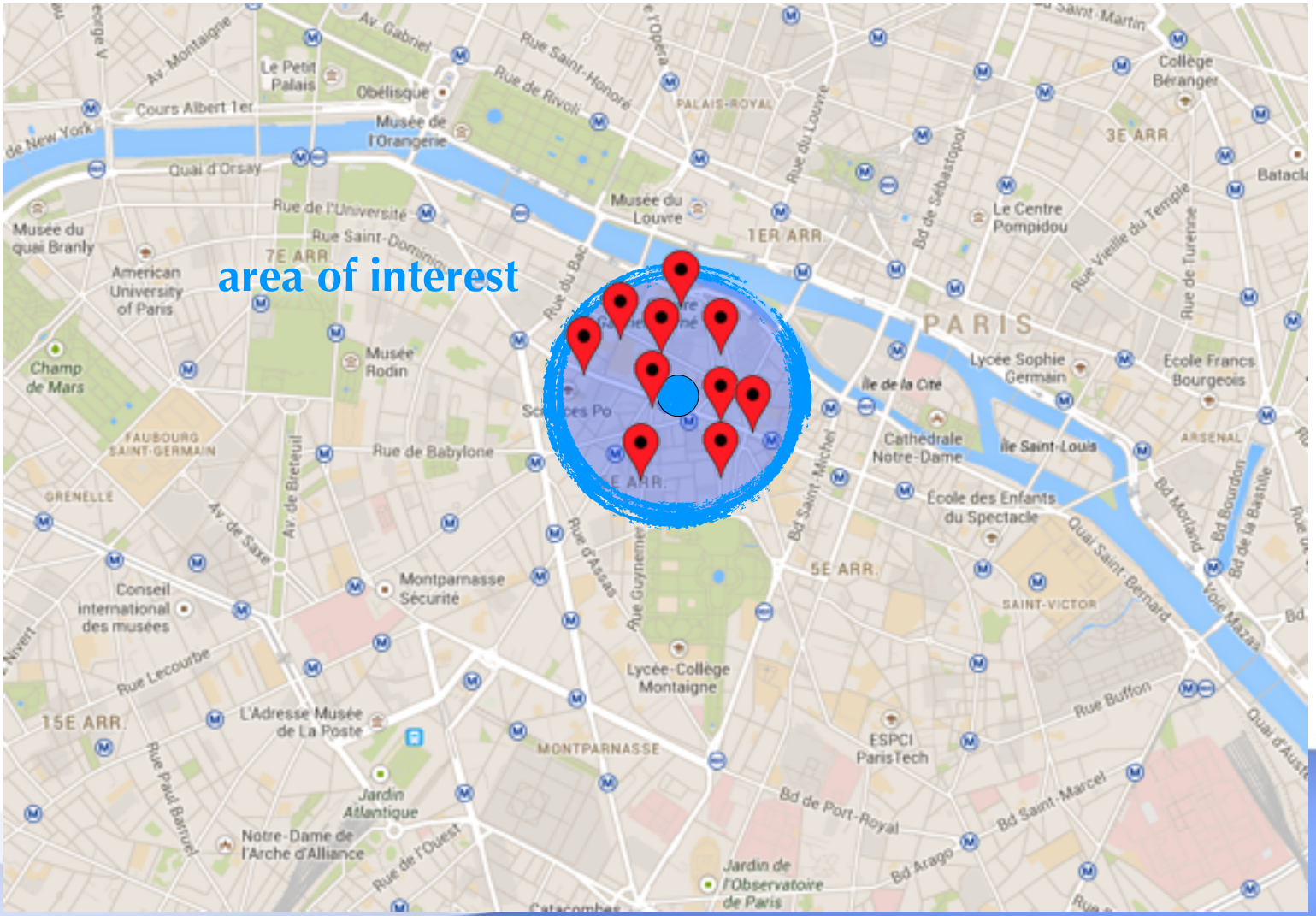
Location guard for Chrome



Location guard for Chrome



Location guard for Chrome



Thank you !