Metrics for Differential Privacy
Application to Location-based Systems

Kostas Chatzikokolakis
CNRS, INRIA, LIX Ecole Polytechnique

joint work with
Miguel Andrés, Nicolás Bordenabe,
Catuscia Palamidessi, Marco Stronati

[PETS'13], [CCS'13] and ongoing work

CAPPRIS meeting, Sep 10, 2013

Differential Privacy

- A notion of privacy from the area of statistical databases
- Goal: disclose aggregate information while protecting the value of an individual

Differential Privacy

Name/Id	age	weight	sex	epilepsy	...
Mario Rossi	65	82	M	yes	...
Daniele Bianchi	35	120	M	yes	...
Lucia Verdi	40	45	F	no	...
...	...	...	...	...	...

We want to disclose

- How many people have epilepsy?
- What is the average age and weight of men who have epilepsy?

Without revealing

- Does Daniele Bianchi have the disease?
- What is the weight of Lucia?

Differential Privacy

- Mechanism: add noise to the query result

$$K(x) = \text{distribution on reported values, for database } x$$

Differential Privacy

- Mechanism: add noise to the query result

$$K(x) = \text{distribution on reported values, for database } x$$

- Adjacency: $x \sim x'$ iff they differ in exactly one individual

$$x = \langle 32, 41, 27 \rangle$$
$$x' = \langle 21, 41, 27 \rangle$$

Differential Privacy

- Mechanism: add noise to the query result

$$K(x) = \text{distribution on reported values, for database } x$$

- Adjacency: $x \sim x'$ iff they differ in exactly one individual

$$x = \langle 32, 41, 27 \rangle$$
$$x' = \langle 21, 41, 27 \rangle$$

- ϵ-differential privacy:
  adjacent databases should produce similar results

$$K(x) \text{ and } K(x') \text{ should be similar } \quad \forall x \sim x'$$

  Protect the value of an individual

Differential Privacy

- Mechanism: add noise to the query result

$$K(x) = \text{distribution on reported values, for database } x$$

- Adjacency: $x \sim x'$ iff they differ in exactly one individual

$$x = \langle 32, 41, 27 \rangle$$
$$x' = \langle 21, 41, 27 \rangle$$

- ϵ-differential privacy:
  adjacent databases should produce similar results

$$d_{\mathcal{P}}(K(x), K(x')) \leq \epsilon \qquad \forall x \sim x'$$

  Protect the value of an individual

Motivation

Broaden the scope of Differential Privacy

Can differential privacy be adapted to different privacy requirements?

Can we use differential privacy on secrets that are not databases?

Differential Privacy

- Hamming dist. $d_h(x, x')$: # of elements in which x, x' differ

$$x = \langle 32, 41, 27 \rangle$$
$$x' = \langle 21, 52, 27 \rangle \qquad d_h(x, x') = 2$$

- Differential privacy:

$$d_{\mathcal{P}}(K(x), K(x')) \leq \epsilon \qquad \forall x \sim x'$$

Differential Privacy

- Hamming dist. $d_h(x, x')$: # of elements in which x, x' differ

$$x = \langle 32, 41, 27 \rangle$$
$$x' = \langle 21, 52, 27 \rangle \qquad d_h(x, x') = 2$$

- Differential privacy:

$$d_{\mathcal{P}}(K(x), K(x')) \leq \epsilon d_h(x, x') \quad \forall x, x'$$

Differential Privacy

- Hamming dist. $d_h(x, x')$: # of elements in which x, x' differ

$$x = \langle 32, 41, 27 \rangle$$
$$x' = \langle 21, 52, 27 \rangle \qquad d_h(x, x') = 2$$

- Differential privacy:

$$d_{\mathcal{P}}(K(x), K(x')) \leq \epsilon d_h(x, x') \quad \forall x, x'$$

- $\epsilon d_h(x, x')$: distinguishability metric between x, x'
- the closer two databases are,
  the more similar the outcome should be

Differential Privacy, generalization

- Arbitrary domain of secrets $\mathcal{X}$
- $d_{\mathcal{X}}(x, x')$: distinguishability metric between x, x'
- Many natural choices
  - numerical distance
  - geographical distance
  - ...
- It could be scaled by some ϵ: $d_{\mathcal{X}} = \epsilon d_{\mathcal{X}}{}'$

Differential Privacy, generalization

- Arbitrary domain of secrets $\mathcal{X}$
- $d_{\mathcal{X}}(x, x')$: distinguishability metric between x, x'
- Many natural choices
  - numerical distance
  - geographical distance
  - ...
- It could be scaled by some ϵ: $d_{\mathcal{X}} = \epsilon d_{\mathcal{X}}{}'$

$d_{\mathcal{X}}$-privacy

$d_{\mathcal{P}}(K(x), K(x')) \leq d_{\mathcal{X}}(x, x') \quad \forall x, x'$

- the closer two secrets are,
  the more similar the outcome should be

Differential Privacy, generalization

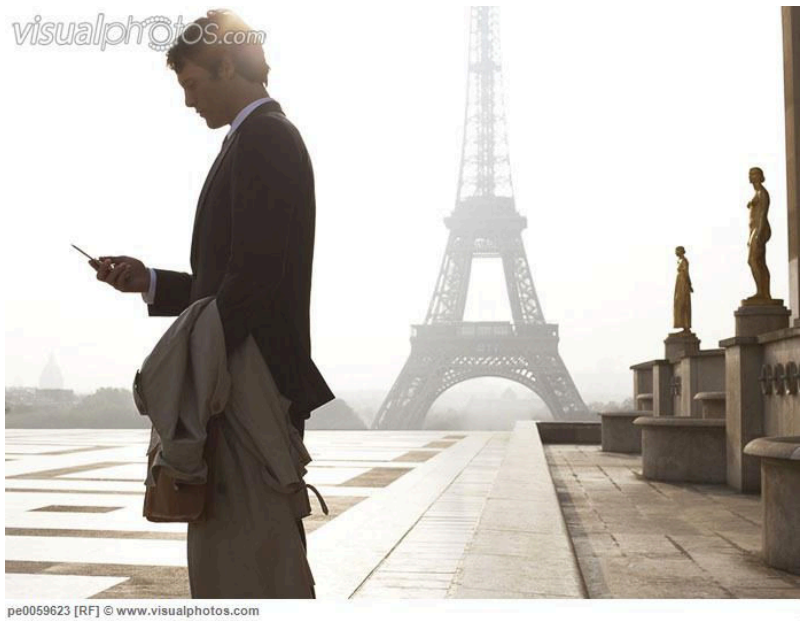
[PETS'13]

- Develop the general theory of $d_{\mathcal{X}}$-privacy
  - Characterization results
  - Optimality results
- What can we achieve with metrics?
  - Strengthen differential privacy
    (and get some surprising results)
  - Protect the accuracy, not the complete value
  - Location privacy

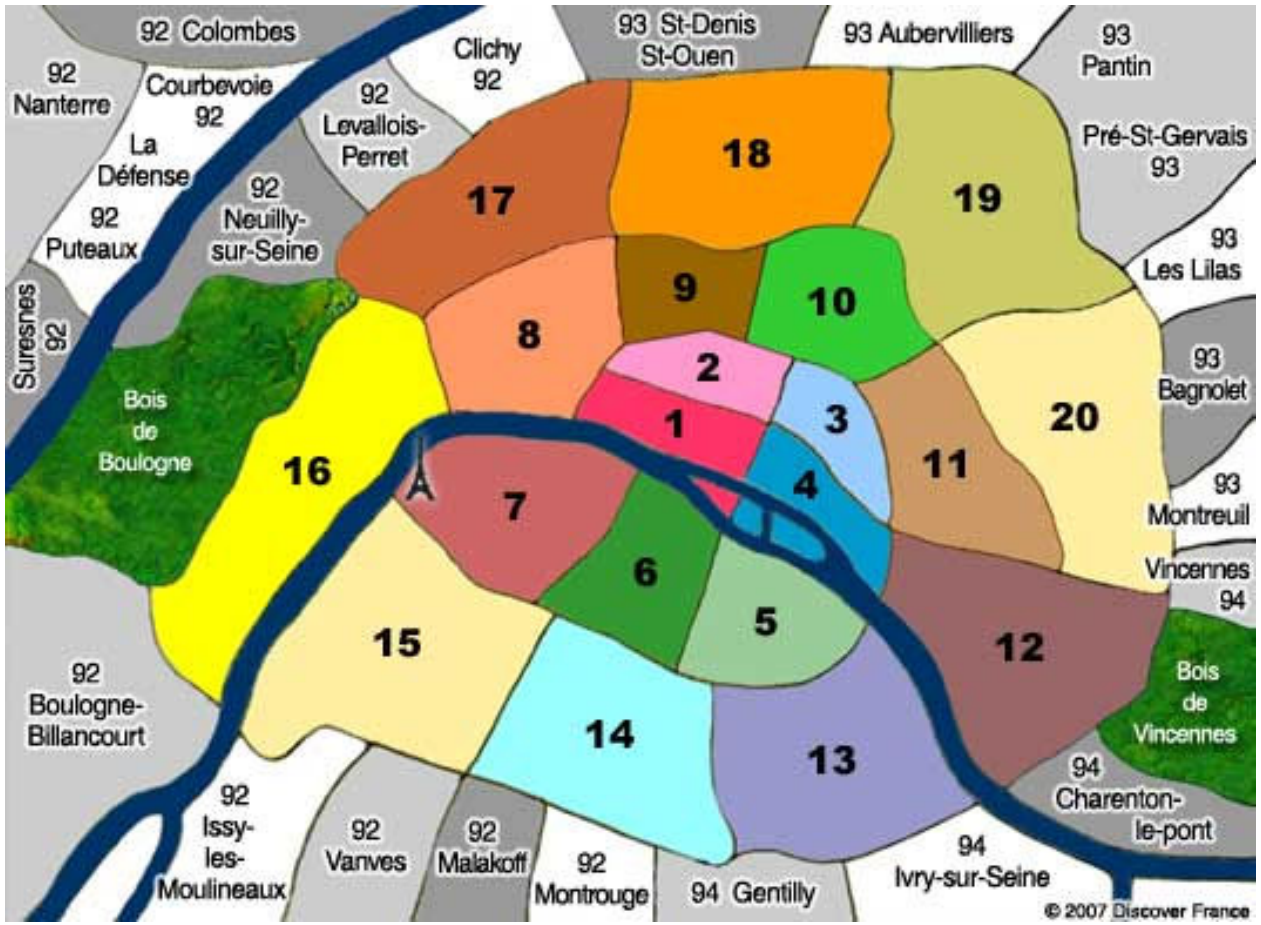
Location Based Services

Goal:

- Find a nearby restaurant, ATM, ...
- Hide the user's location (not identity) from the service provider
- the coarse location should be revealed, but not the exact one



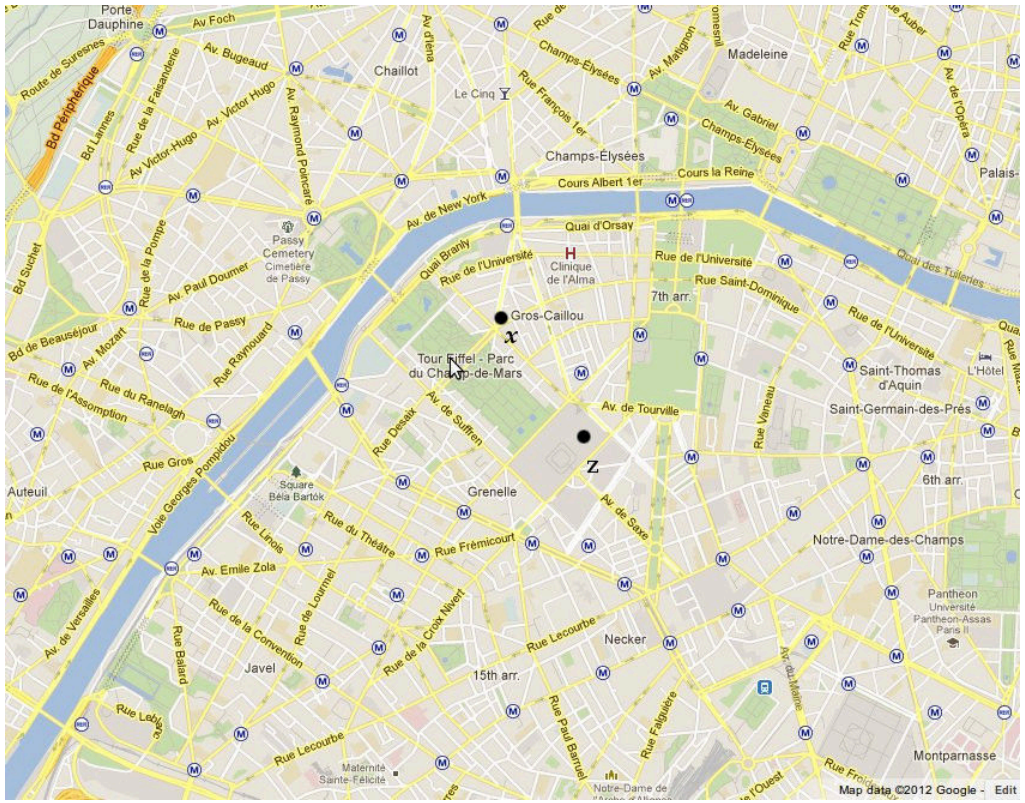
A simple cloaking "solution"



Fails when the user moves between regions

Obfuscation

- Add noise to x, report z
- Can be seen as probabilistic cloaking

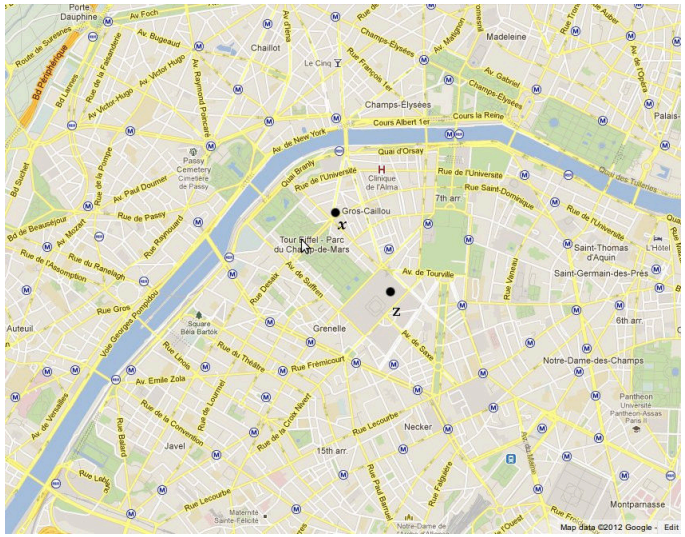


Measuring location privacy

Shokri et al, S&P'11, CCS'12

Privacy metric

Expected error of a Bayesian adversary



Construction of an optimal mechanism

- given a prior and a utility constraint
- find the mechanism with the best privacy

Drawback: depends on the prior information on locations

Differential privacy

adjacent databases should be indistinguishable

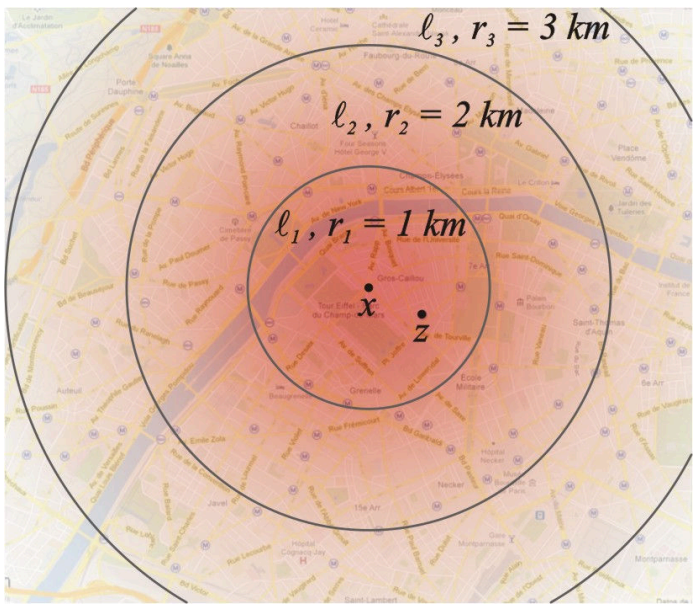
But we only have a single individual
Such a mechanism would have zero utility

We actually want some locations to be distinguishable

Geo-indistinguishability [CCS'13]

the closer (geographically) two secrets are
the more indistinguishable they should be

$$d_{\mathcal{P}}(K(x), K(x')) \leq \epsilon d(x, x') \qquad \forall x, x'$$



Provides privacy for any radius r with a proportional level $\epsilon \cdot r$

Characterization

Characterization

Hiding function $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ applied before K

Characterization

Characterization

Hiding function $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ applied before K

the adversary's conclusions should be "similar"

regardless of whether ϕ has been applied or not

$$K(x) \quad \text{gives "as much" information as} \quad K(\phi(x))$$

Characterization

Characterization

Hiding function $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ applied before K

the adversary's conclusions should be "similar"

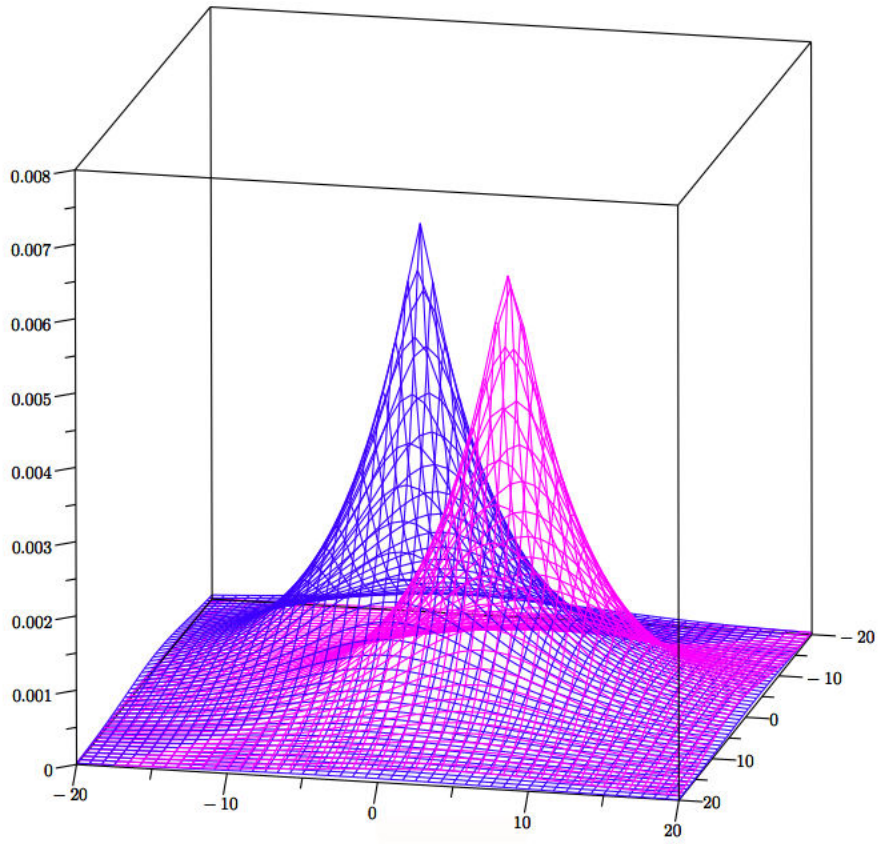
regardless of whether ϕ has been applied or not

$$K(x) \quad \text{gives "as much" information as} \quad K(\phi(x))$$

A second characterization is obtained by comparing prior and posterior knowledge.

A mechanism for geo-indistinguishability

Add noise from a 2-dimensional Laplace distribution

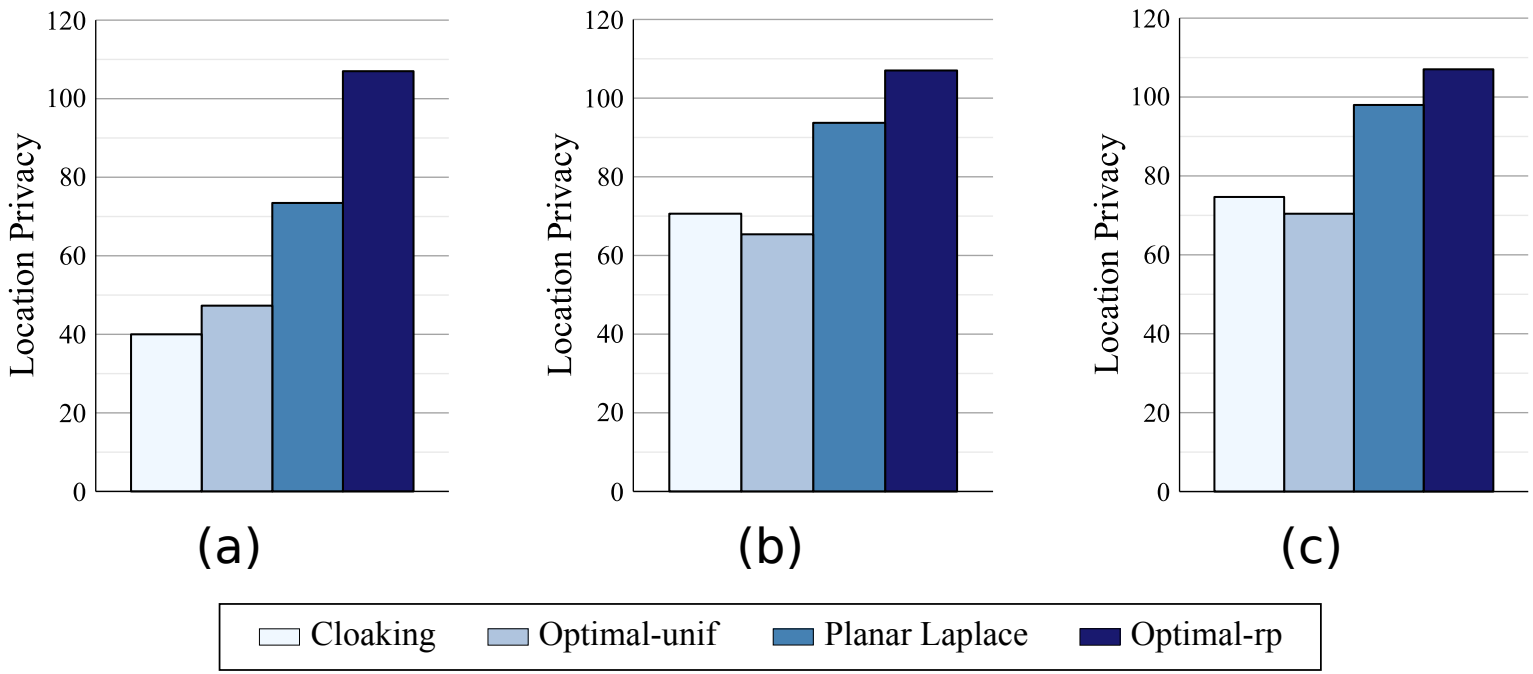


Efficient

Solve discretization and truncation issues

Comparison

- Using the privacy metric of Shokri et al
- Under various priors
- Against:
  - simple cloaking
  - optimal mechanism constructed using the correct prior
  - optimal mechanism constructed using a uniform prior

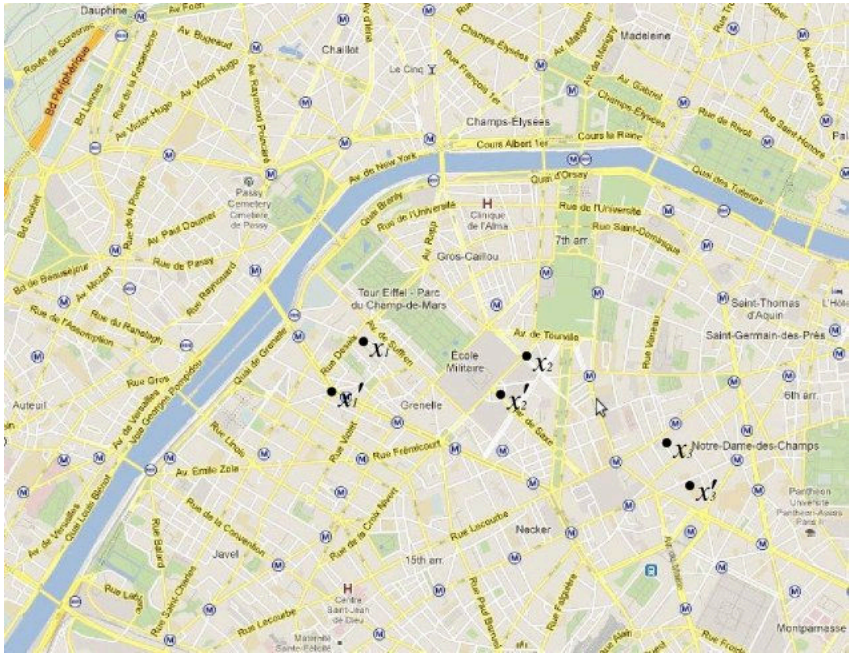


Protecting location traces

- Secrets are now tuples $\mathbf{x} = (x_1, \ldots, x_n)$
- Distance between tuples:

$$d_\infty(\mathbf{x}, \mathbf{x}') = \max_i d(x_i, x_i')$$

- Use ϵd_∞-privacy

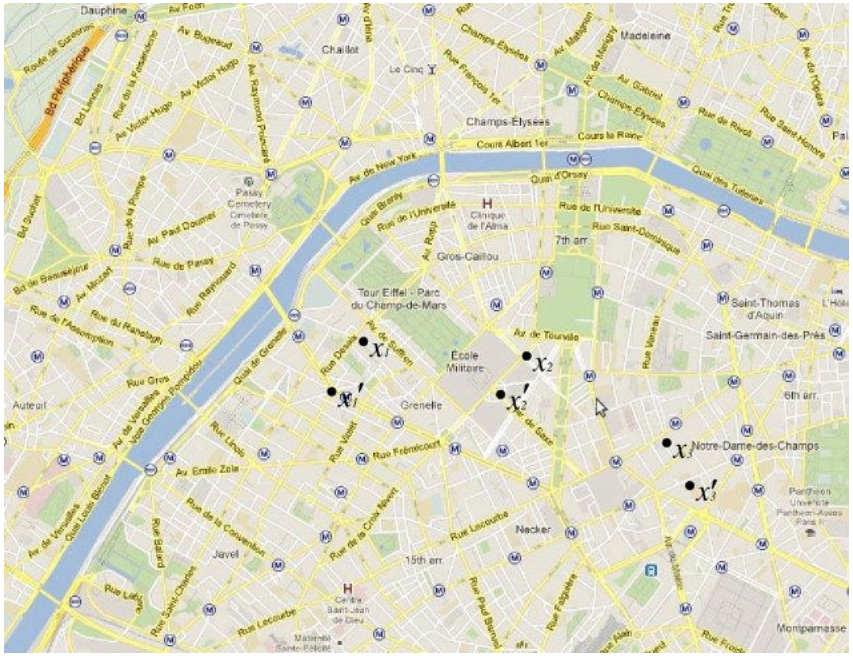


Protecting location traces

- Secrets are now tuples $\mathbf{x} = (x_1, \ldots, x_n)$
- Distance between tuples:

$$d_\infty(\mathbf{x}, \mathbf{x}') = \max_i d(x_i, x_i')$$

- Use ϵd_∞-privacy



Mechanism: add independent noise to each point

Problem: this only satisfies $n\epsilon$-geo-indistinguishability

Budget: we pay ϵ each time we generate noise

Predictive mechanism

- Basic intuition: if the user is not moving, he should always report the same obfuscated location z
- An obfuscated point z_n close to the current location x_n can be predicted from the previously reported locations $\mathbf{z}$.

Predictive mechanism

- Basic intuition: if the user is not moving, he should always report the same obfuscated location z
- An obfuscated point z_n close to the current location x_n can be predicted from the previously reported locations $\mathbf{z}$.

If you can predict it why pay for it?

Predictive mechanism

- Basic intuition: if the user is not moving, he should always report the same obfuscated location z
- An obfuscated point z_n close to the current location x_n can be predicted from the previously reported locations $\mathbf{z}$.

If you can predict it why pay for it?

Predictive mechanism

At step n compute a prediction $z_n^* = \Omega(\mathbf{z})$

- if z_n^* is close to x_n, report z_n^*
- otherwise, generate new z_n by adding noise to x_n

Predictive mechanism

At step n compute a prediction $z_n^* = \Omega(\mathbf{z})$

- if z_n^* is close to x_n, report z_n^*
- otherwise, generate new z_n by adding noise to x_n

Privacy:

- if the test is differentially private, then the mechanism satisfies geo-indistinguishability
- We pay
  - a (small) ϵ_t for each test
  - plus an ϵ each time we add noise
- If the prediction is good, the savings are big

Constructing an optimal mechanism

Shokri et al, CCS'12

- given a prior and a utility constraint
- find the mechanism with the best privacy (expected error)

Both privacy and utility depend on the prior.

Constructing an optimal mechanism

Shokri et al, CCS'12

- given a prior and a utility constraint
- find the mechanism with the best privacy (expected error)

Both privacy and utility depend on the prior.

Ongoing work:

- given a privacy constraint (geo-indistinguishability)
- find the mechanism with the best utility (for a given prior)
- Advantages:
  - privacy is a hard constraint
  - only utility depends on the prior
- graph-based technique for reducing the number of constraints

Conclusion

- Metrics allow to broaden the scope of Differential Privacy
  - Capture new privacy notions
  - Extend to new applications, where secrets are not databases
- Natural application: location privacy
- Ongoing work:
  - predictive mechanism for location traces
  - construction of optimal mechanism

Questions?