Probabilistic Methods for Privacy and Secure Information Flow

Catuscia Palamidessi

Lecture 4

Utility

Let us start with an example. Suppose we have a medical database, and we want to use it to do research about a certain disease.

For instance, we want to ask queries like:

1. How many people in the DB have the disease?
2. What is the average age of the people with the disease?

Suppose we know that :

- there are 1000 people in the DB
- the maximum age is 120
- both queries are sanitised with DP

age	disease
41	no
45	yes
37	no
50	yes
...	...
20	no

Loss function

How to measure the quality of the reported answer?

Consider the first query: $f(x) =$ number of people with the disease.
Let $y = f(x)$ be the true answer, and z the reported answer.
Which of the following loss functions is better?

1. $\ell(y, z) = |z - y|$
2. $\ell(y, z) = (z - y)^2$
3. $\ell(y, z) = \begin{cases} 0 & \text{if } z = y \\ 1 & \text{if } z \neq y \end{cases}$
4. $\ell(y, z) = 0$
5. $\ell(y, z) = z + y$

Loss function

How to measure the quality of the reported answer?

Consider the first query: $f(x)$ = number of people with the disease.
Let $y = f(x)$ be the true answer, and z the reported answer.
Which of the following loss functions is better?

1. $\ell(y,z) = |z - y|$
2. $\ell(y,z) = (z - y)^2$
3. $\ell(y,z) = \begin{cases} 0 & \text{if } z = y \\ 1 & \text{if } z \neq y \end{cases}$
4. $\ell(y,z) = 0$
5. $\ell(y,z) = z + y$

(1), (2) and (3) are all reasonable loss functions, they all measure the "precision" of the answer. Which one is more suitable for our purposes depends on what we want to do.
On the other hand, (4) does not measure anything, and (5) does not make sense.

Monotonicity of the loss

In general, if $\mathcal{Y} \subseteq \mathcal{Z}$ and the domain $\mathcal{Z}$ is equipped with a notion of distance d, we want the loss to be *monotonic* w.r.t. d. Namely:

$$\ell(y, z) \leq \ell(y', z') \;\; \Leftrightarrow \;\; |z - y| \leq |z' - y'|$$

Utility as expected loss

Since there are many possible true answers, and even for the same true answer we have many possible reported answer, it is reasonable to define the utility as expectation.

Let π be the prior on $\mathcal{Y}$ (the true answers) and p the probability associated to the mechanism. We could define:

$$\begin{aligned} \mathcal{U}(\mathcal{K}, \pi) &= \mathbb{E}_{\pi,p}\ell(y,z) \\ &= \sum_{y,z} \pi(y)\, p(z|y)\, \ell(y,z) \end{aligned}$$

Utility as expected loss

Since there are many possible true answers, and even for the same true answer we have many possible reported answer, it is reasonable to define the utility as expectation.

Let π be the prior on $\mathcal{Y}$ (the true answers) and p the probability associated to the mechanism. We could define:

$$\begin{aligned} \mathcal{U}(\mathcal{K}, \pi) &= \mathbb{E}_{\pi,p}\ell(y,z) \\ &= \textstyle\sum_{y,z} \pi(y)\, p(z|y)\, \ell(y,z) \end{aligned}$$

Are we happy with this definition?

Utility as expected loss

Since there are many possible true answers, and even for the same true answer we have many possible reported answer, it is reasonable to define the utility as expectation.

Let π be the prior on $\mathcal{Y}$ (the true answers) and p the probability associated to the mechanism. We could define:

$$\begin{aligned} \mathcal{U}(\mathcal{K}, \pi) &= \mathbb{E}_{\pi,p}\ell(y,z) \\ &= \textstyle\sum_{y,z} \pi(y)\, p(z|y)\, \ell(y,z) \end{aligned}$$

Are we happy with this definition?

What if we get a negative answer? Or an answer greater than 1000, the number of people in the DB? (it could happen, for instance, with the geometric mechanism).

Utility as expected loss

Since there are many possible true answers, and even for the same true answer we have many possible reported answer, it is reasonable to define the utility as expectation.

Let π be the prior on $\mathcal{Y}$ (the true answers) and p the probability associated to the mechanism. We could define:

$$\begin{aligned} \mathcal{U}(\mathcal{K}, \pi) &= \mathbb{E}_{\pi,p}\ell(y,z) \\ &= \textstyle\sum_{y,z} \pi(y)\, p(z|y)\, \ell(y,z) \end{aligned}$$

Are we happy with this definition?

What if we get a negative answer? Or an answer greater than 1000, the number of people in the DB? (it could happen, for instance, with the geometric mechanism).

We are not going to believe these answers, so we could remap them in more likely values. For instance we could remap the negative values into 0, and those greater than 1000 into 1000

Remapping

We could use a remapping function defined as:

$$r(z) = \begin{cases} 0 & \text{if } z < 0 \\ z & \text{if } 0 \leq z \leq 1000 \\ 1000 & \text{if } z > 1000 \end{cases}$$

and define

$$\mathcal{U}(\mathcal{K}, \pi) \;=\; \sum_{y,z} \pi(y) p(z|y)\, \ell(y, r(z))$$

Remapping

We could use a remapping function defined as:

$$r(z) = \begin{cases} 0 & \text{if } z < 0 \\ z & \text{if } 0 \leq z \leq 1000 \\ 1000 & \text{if } z > 1000 \end{cases}$$

and define

$$\mathcal{U}(\mathcal{K}, \pi) \; = \; \sum_{y,z} \pi(y) p(z|y) \, \ell(y, r(z))$$

More in general, we assume that we exploit the prior knowledge, and the knowledge of the mechanism, to define and use the best possible remapping function:

$$\mathcal{U}(\mathcal{K}, \pi) \; = \; \min_{r} \sum_{y,z} \pi(y) p(z|y) \, \ell(y, r(z))$$

Notes about utility

- We saw a definition for discrete mechanisms. For continuous ones, like the Laplace, the definition is analogous except that the expectation has to be computed via integration
- The expected loss is not the only definition of utility that has been considered in the literature. There are others, for instance the worst-case loss, the expected ratio of ``good'' answers, etc. For the next results, however, we will assume that utility is defined as expected loss.

Optimal mechanisms

- Given a prior π, and a privacy level ε, an ε-differentially private mechanism K is called optimal if it provides the best utility among all those which provide ε-differential privacy
- Note that the privacy does not depend on the prior, but the utility (in general) does.
- In the finite case the optimal mechanism can be computed with linear optimization techniques, where the variables are the conditional probabilities p(z | y)
  where y is the exact answer and z is the reported answer
- A mechanism is universally optimal if it is optimal for all priors π

Counting Queries

- Counting queries are typical examples of discrete queries. They are of the form: How many individuals in the database satisfy the property $\mathcal{P}$?
  - Examples:
    - How many individuals in the DB are affected by diabetes?
    - How many diabetic people are obese?
- Question: what is the sensitivity of a counting query?

Privacy vs utility: two fundamental results

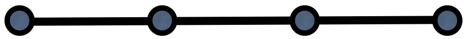
1. [Ghosh et al., STOC 2009]
   The geometric mechanism is universally optimal for counting queries and any monotonic loss function

Privacy vs utility: two fundamental results

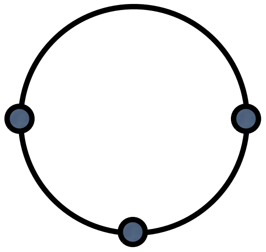
2. [Brenner and Nissim, STOC 2010] The counting queries are the only kind of queries for which a universally optimal mechanism exists

- This means that for other kind of queries one the optimal mechanism is relative to a specific user.

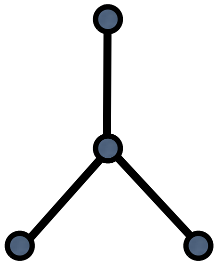
- The precise characterization is given in terms of the graph $(\mathcal{Y},\sim)$ induced by $(\mathcal{X},\sim)$



ok



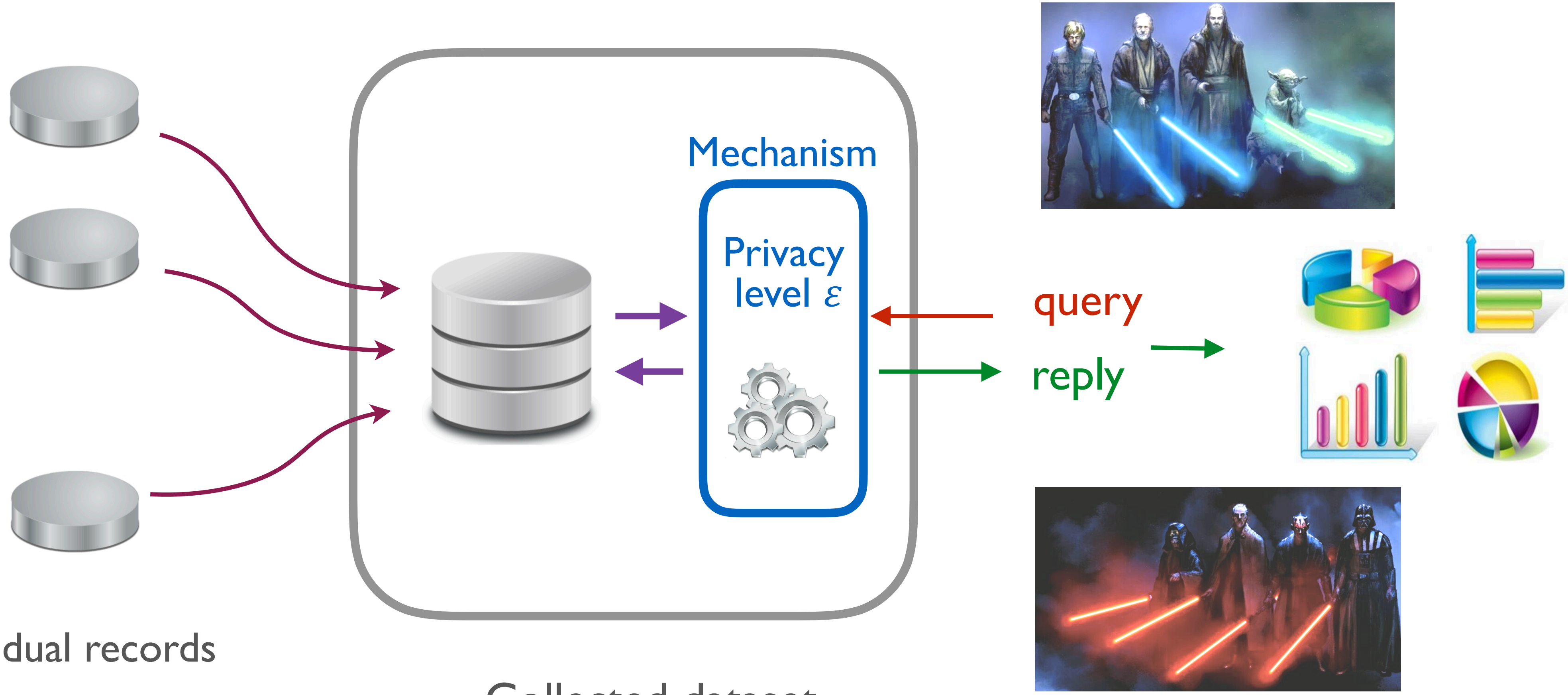
not ok



not ok

The Local Model

DP in the Global Model



Mechanism

Privacy level ε

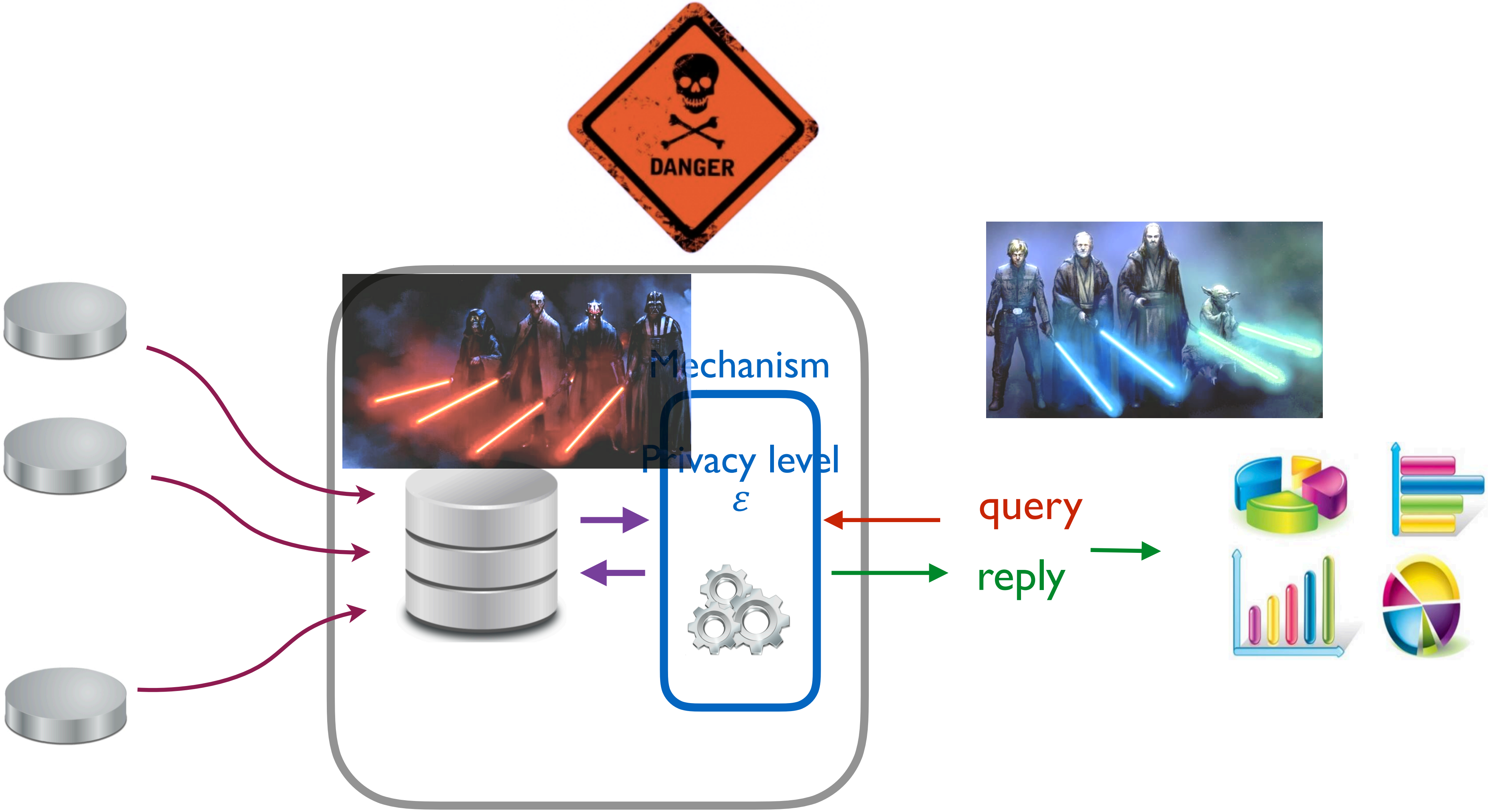
query

reply

Individual records

Collected dataset

Standard Differential Privacy



Mechanism

Privacy level
ε

query

reply

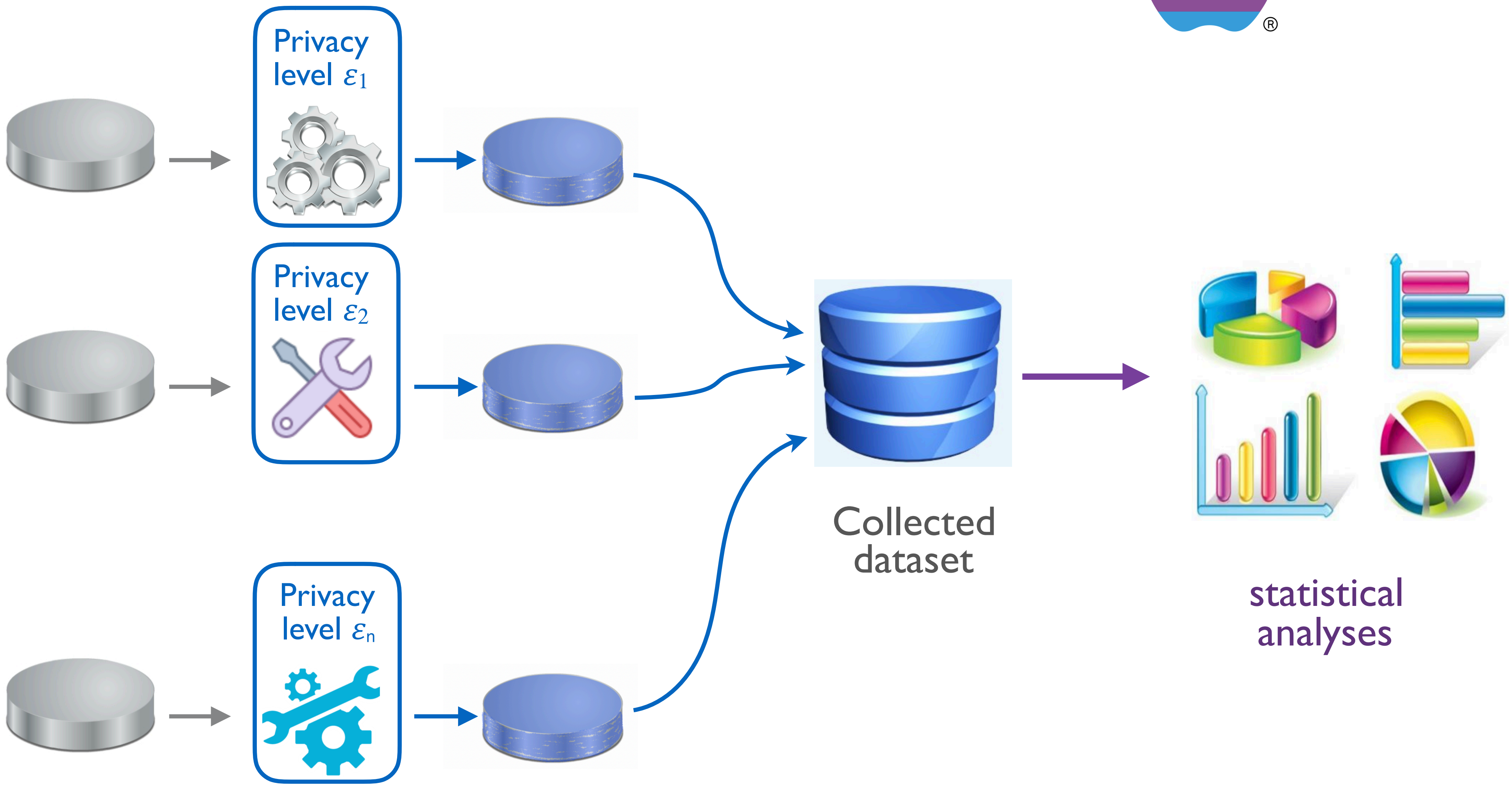
Individual records

Collected dataset

Local Differential Privacy

Google

Apple



Privacy level ε_1

Privacy level ε_2

Privacy level ε_n

Collected dataset

statistical analyses

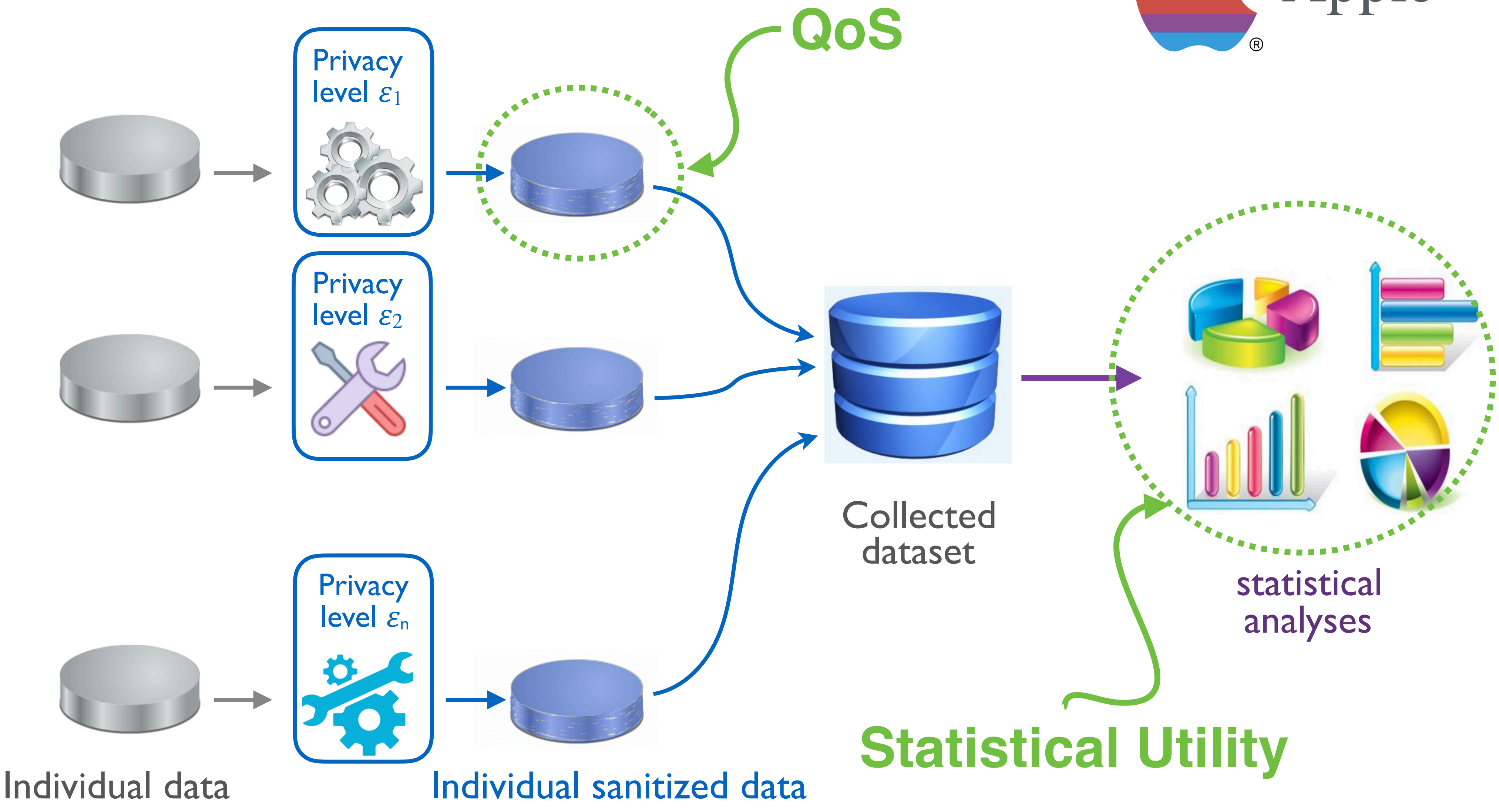
Individual data

Individual sanitized data

Local Differential Privacy

Google

Apple



QoS

Privacy level ε_1

Privacy level ε_2

Privacy level ε_n

Collected dataset

statistical analyses

Statistical Utility

Individual data

Individual sanitized data

Local Differential Privacy

[Jordan &Wainwright '13]

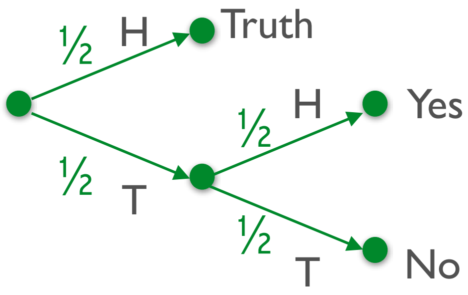
Definition Let $\mathcal{X}$ be a set of possible values and $\mathcal{Y}$ the set of noisy values. A mechanism $\mathcal{K}$ is ε-locally differentially private (ε-LDP) if for all $x_1, x_2 \in \mathcal{X}$ and for all $y \in \mathcal{Y}$

$$P[\mathcal{K}(x) = y] \leq e^{\varepsilon}\ P[\mathcal{K}(x') = y]$$

or equivalently, using the conditional probability notation:

$$p(y \mid x) \leq e^{\varepsilon}\ p(y \mid x')$$

For instance, the Randomized Response protocol is (log 3)-LPD



y

		yes	no
x	yes	¾	¼
	no	¼	¾

The flat mechanism (aka k-RR)

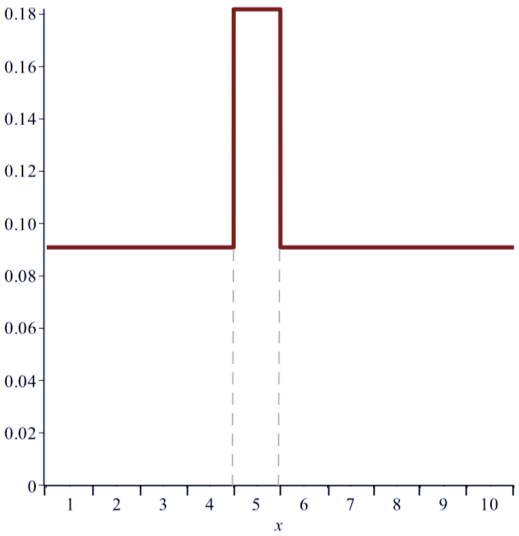
[Kairouz et al, '16]

The flat mechanism is the simplest way to implement LPD. It is defined as follows:

$$p(y|x) = \begin{cases} c\, e^{\varepsilon} & \text{if } x = y \\ c & \textit{otherwise} \end{cases}$$

where c is a normalization constant.

namely $c = \dfrac{1}{k - 1 + e^{\varepsilon}}$ where k is the size of the domain



Privacy Properties:

- Compositionality
- Independence from the side knowledge of the adversary

What about Utility ?

- Statistical Utility
- QoS

d-privacy: a generalization of DP and LDP [Chatzikokolakis et al., '13]

d-privacy

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

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

generalizes

Differential Privacy

- x, x' are databases
- *d* is the Hamming distance

Local Differential Privacy

- *d* is the discrete distance

Properties

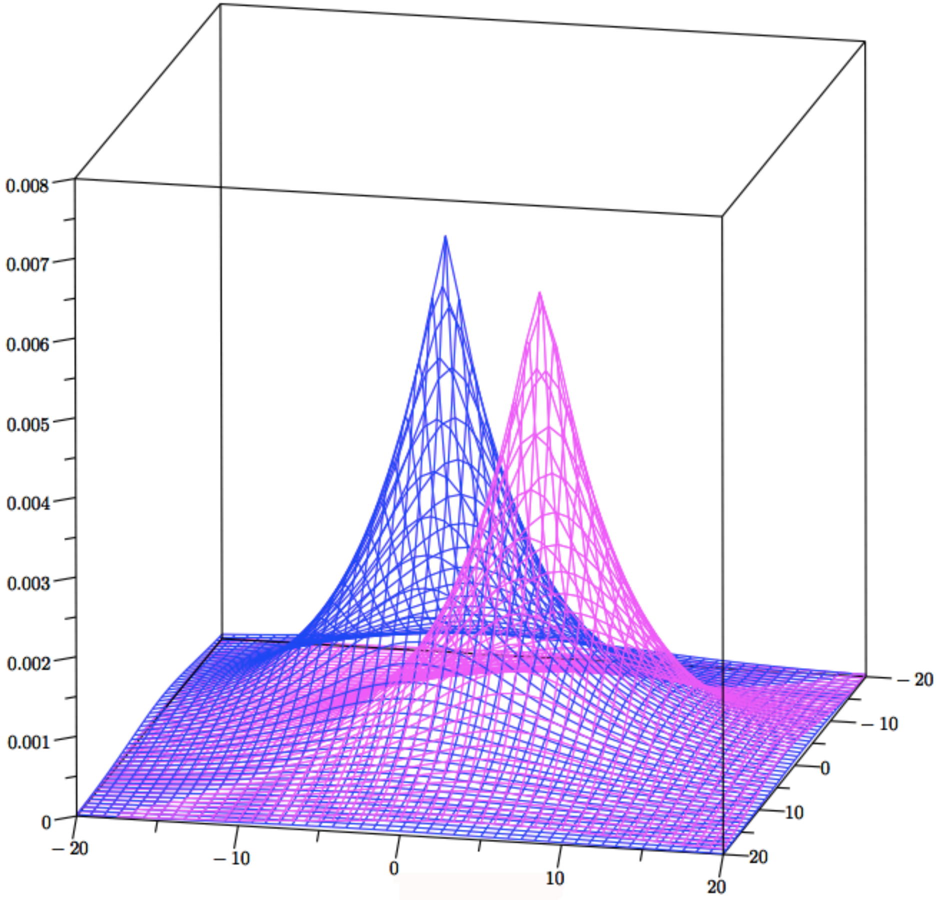
- Like LDP, it can be applied at the user side
- Like DP and LDP, it is compositional

Typical *d*-private mechanisms

Laplace, Geometric, and their Planar versions

Planar Laplace

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



Used especially for location privacy, where *d*-privacy is called *geo-indistinguishability*

Statistical Utility

Statistical utility: The matrix inversion method
[Kairouz et al, '16]

- Let C be the stochastic matrix associated to the mechanism
- Let q be the empirical distribution (derived from the noisy data).
- Compute the approximation of the true distribution as $r = q\, C^{-1}$

Example Assume $q(Yes) = \frac{6}{10}$ and $q(No) = \frac{4}{10}$. Then:

$$\frac{3}{4}\, p(Yes) + \frac{1}{4}\, p(No) = \frac{6}{10}$$

$$\frac{1}{4}\, p(Yes) + \frac{3}{4}\, p(No) = \frac{4}{10}$$

From which we derive $p(Yes) = \frac{7}{10}$ and $p(No) = \frac{3}{10}$

y

x

	yes	no
yes	¾	¼
no	¼	¾

Statistical utility: The matrix inversion method

Problem 1: C must be invertible

Problem 2: Assume $q(Yes) = \frac{4}{5}$ and $q(No) = \frac{1}{5}$. Then:

$$\frac{3}{4}\,p(Yes) + \frac{1}{4}\,p(No) = \frac{4}{5}$$

$$\frac{1}{4}\,p(Yes) + \frac{3}{4}\,p(No) = \frac{1}{5}$$

		y	
		yes	no
x	yes	¾	¼
	no	¼	¾

From which we derive $p(Yes) = \frac{11}{10}$ and $p(No) = -\frac{1}{10}$

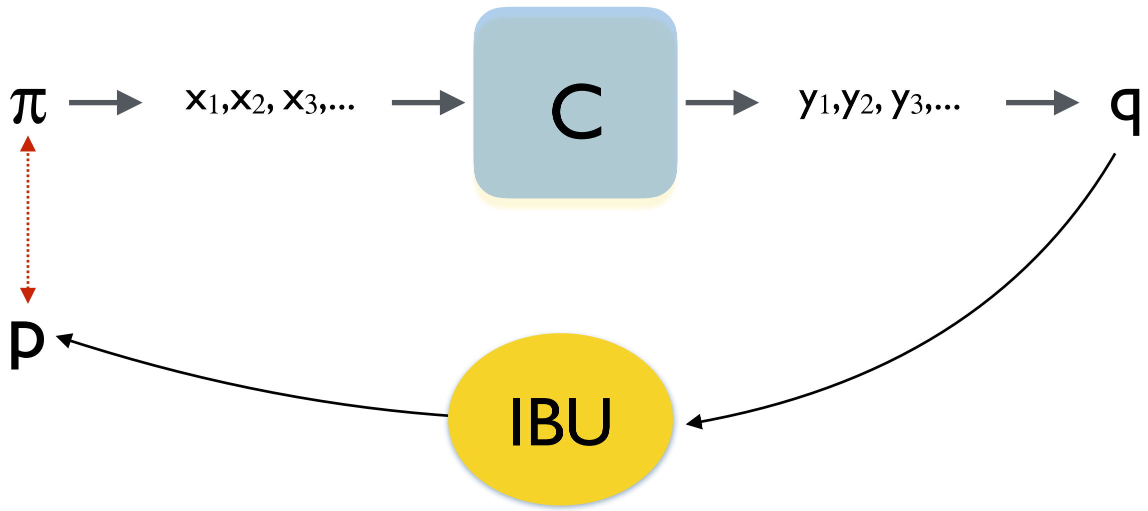
Statistical utility: The matrix inversion method

$r = q\, C^{-1}$ may not be a distribution because it may contain negative elements. In order to try to obtain the true distribution π we can either:

- set to 0 all the negative elements, and renormalize, or
- project r on the simplex.

The resulting distribution however usually is not the best approximation of the original distribution.

Our approach: Iterative Bayesian Update



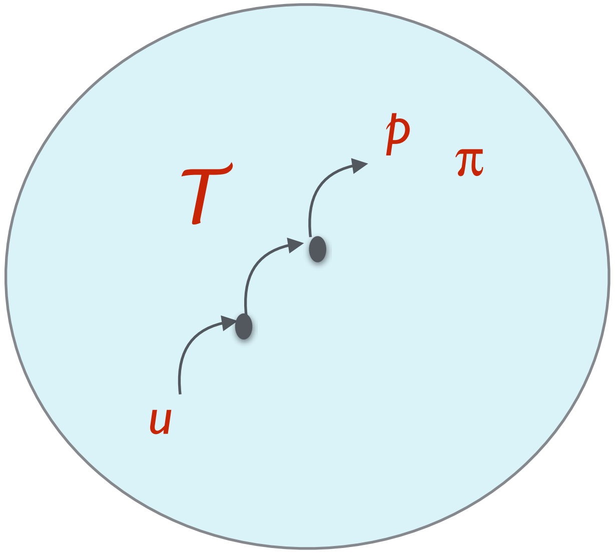
The IBU:

- is based on the **Maximization-Expectation** method
- produces a **Maximum Likelihood Estimator** p of the true distribution π
- If C is invertible, the MLE is unique and as the number of samples grows it converges to π

The Iterative Bayesian Update

- Define $p^{(0)}$ = any distribution (for ex. the uniform distribution)
- Repeat: Define $p^{(n+1)}$ as the Bayesian update of $p^{(n)}$ weighted on the corresponding element of q, namely:

$$p_x^{(n+1)} = \sum_y q_y \frac{p_x^{(n)} C_{xy}}{\sum_z p_z^{(n)} C_{zy}}$$



- Note that $p^{(n+1)} = T(p^{(n)})$
- When C is invertible, T has unique fix point (the MLE)
- Open problem: in some cases (with few samples) the MLE may not be the best estimation of the true distribution. We are trying to devise corrective methods.

Comparison:
C = Laplace ϵ = 0.1
Data domain: {0,1,...,99}

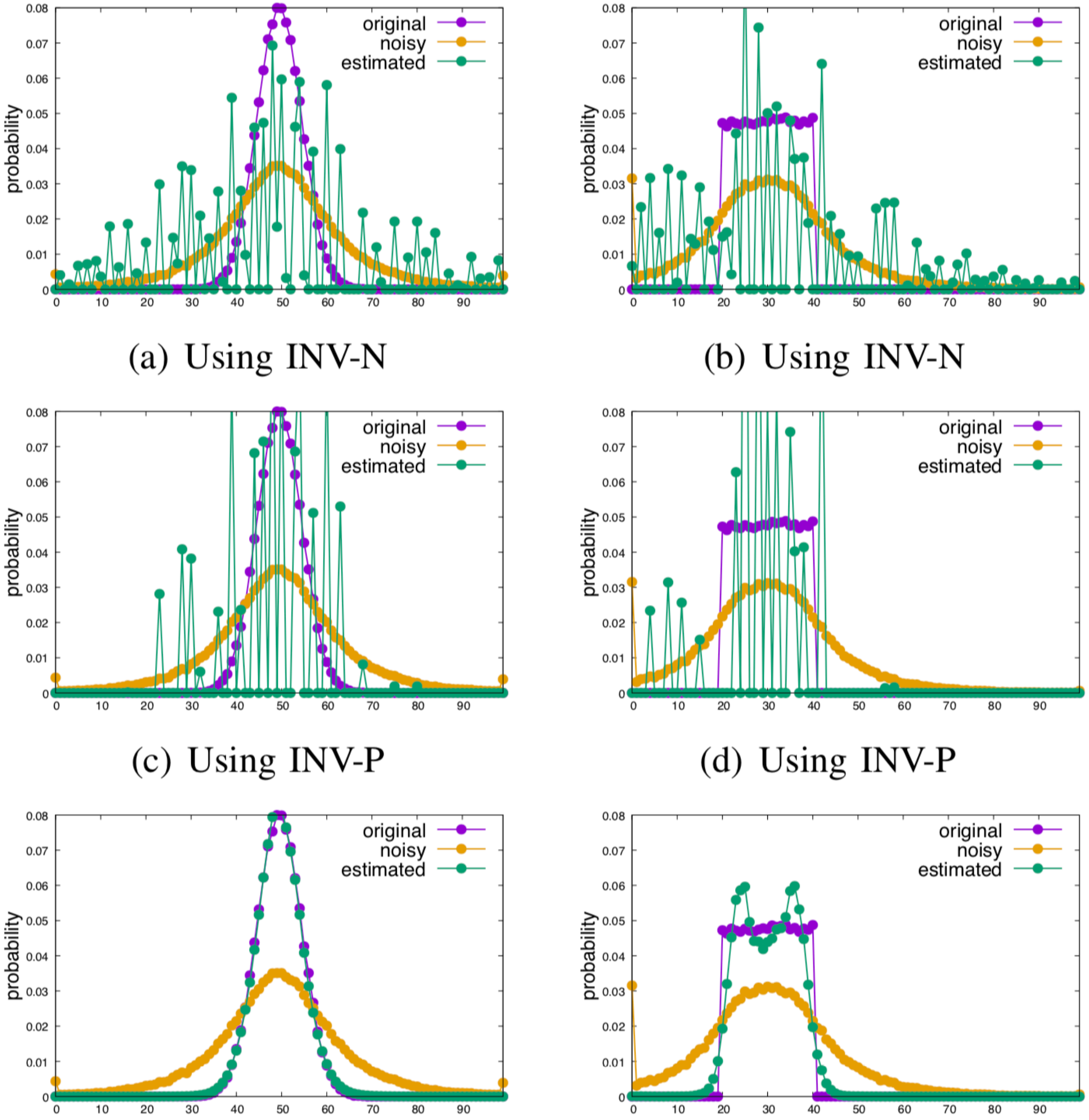
Gaussian

Uniform on an interval

INV-N

INV-P

IBU



(a) Using INV-N

(b) Using INV-N

(c) Using INV-P

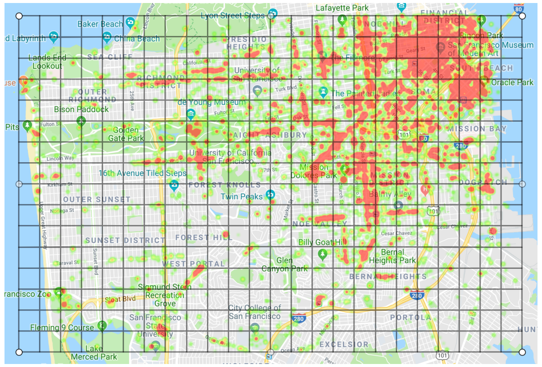
(d) Using INV-P

(e) Using EM

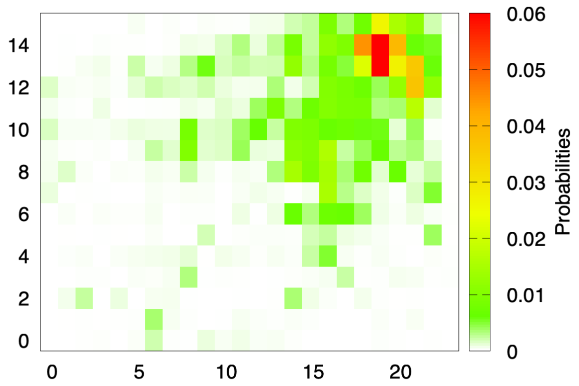
(f) Using EM

Comparison: C = Planar Laplace ϵ = 1
Gowalla Location Data in S. Francisco

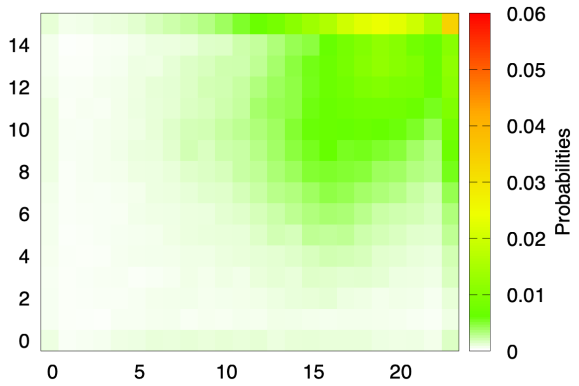
planar Laplace noise



San Francisco area

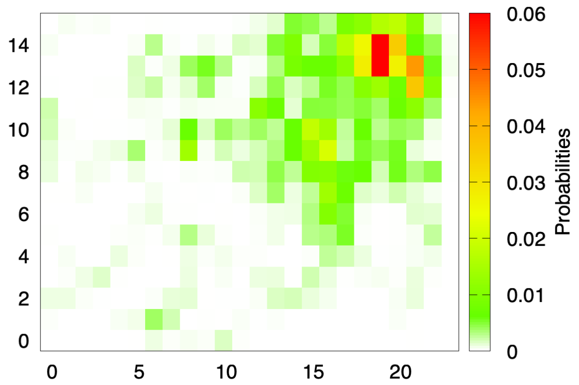


Original

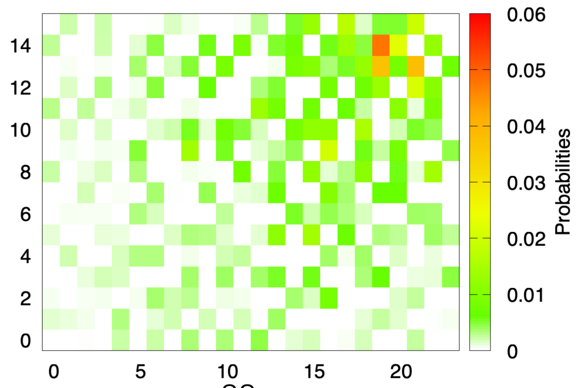


Empirical from noisy data

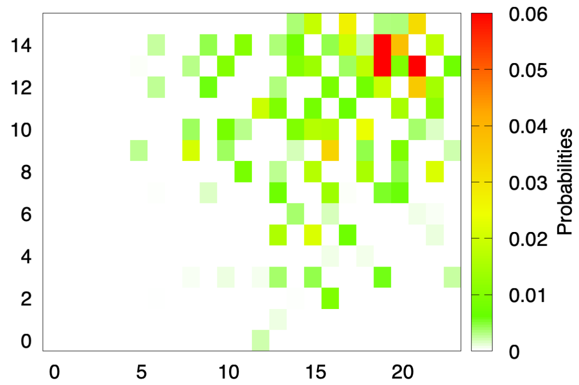
IBU



INV-N



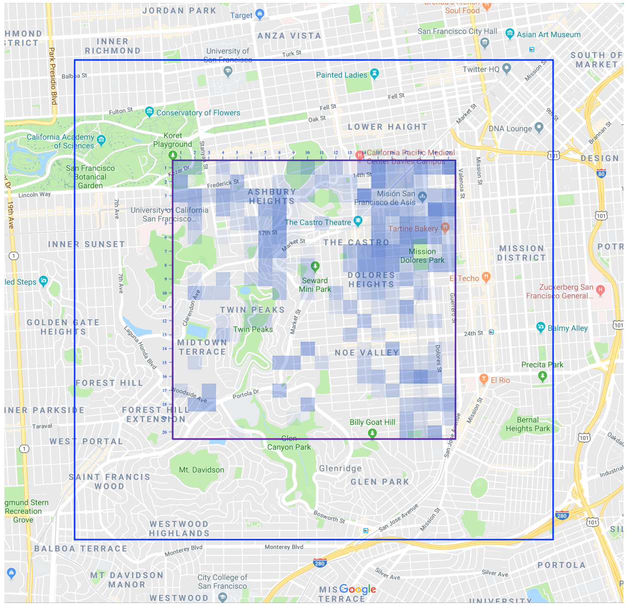
INV-P

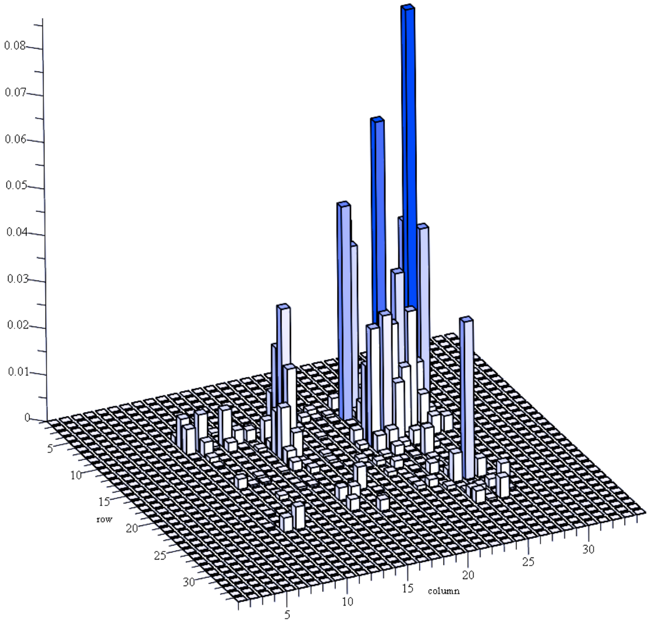


Comparison between LPD and *d*-privacy Experiments on the Gowalla dataset

- Gowalla is a dataset of geographical checkins in several cities in the world

- We have used it to compare the statistical utility of kRR and Planar Laplace with the respective ε calibrated so to satisfy the same privacy constraint:
  same level of privacy within about 1 Km^2

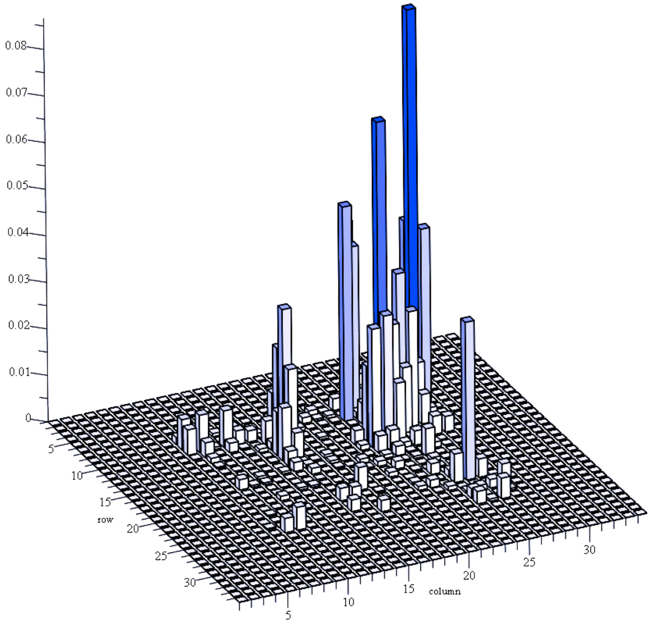




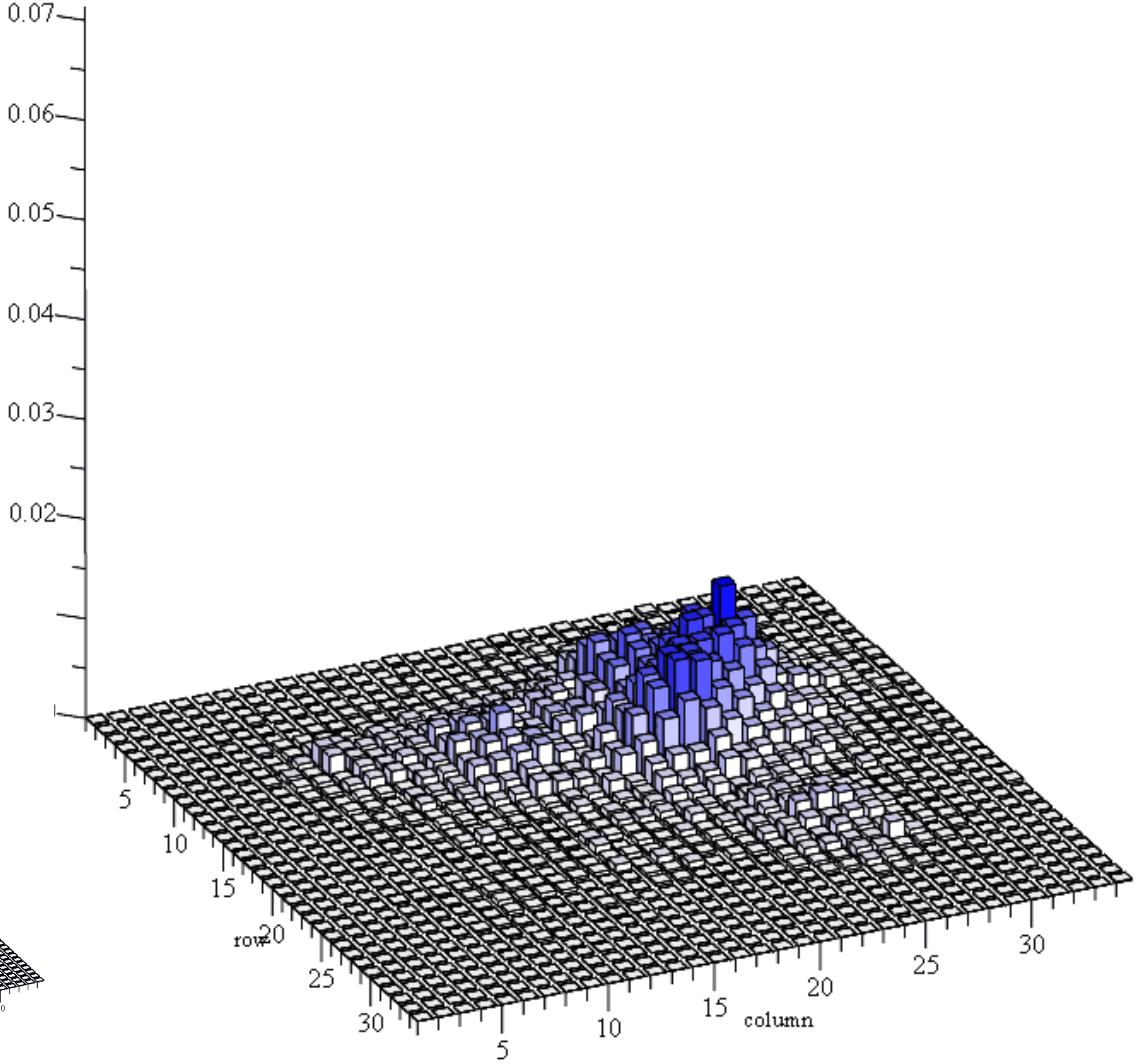
Gowalla checkins in an area of 3x3 km^2 in San Francisco downtown (about 10K checkins)

The Planar Laplace mechanism

ε = ln(2)



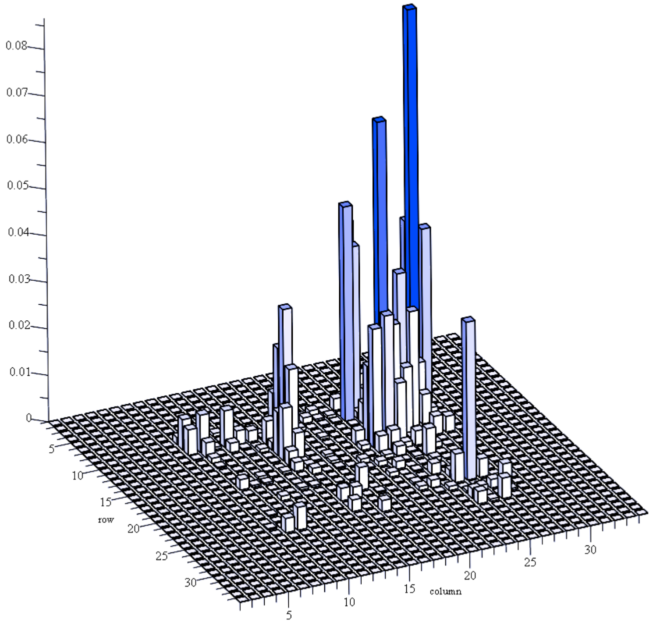
The real distribution



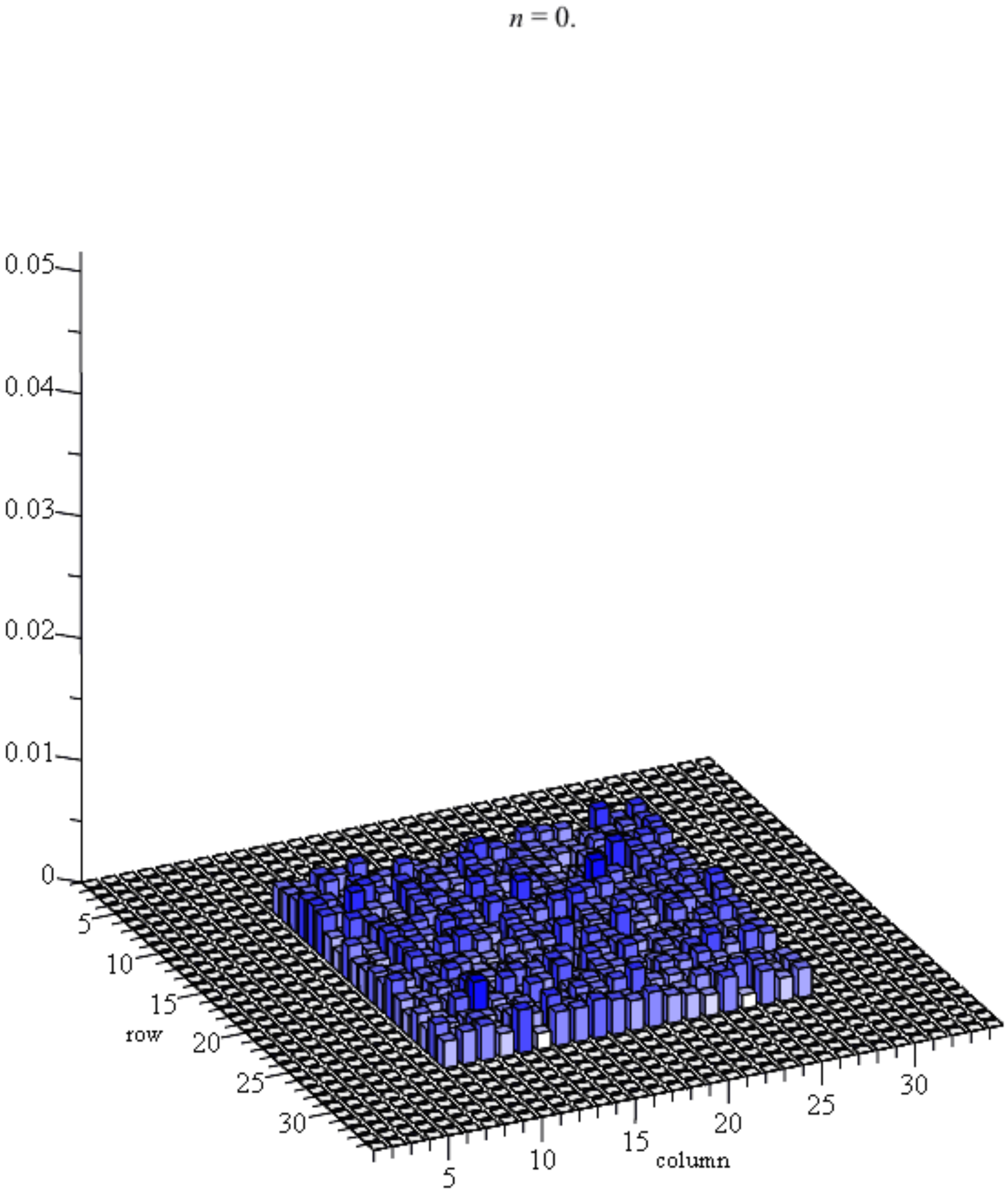
The noisy distribution and the result of the IBU (300 iterations)

The kRR mechanism

ε = ln(8)

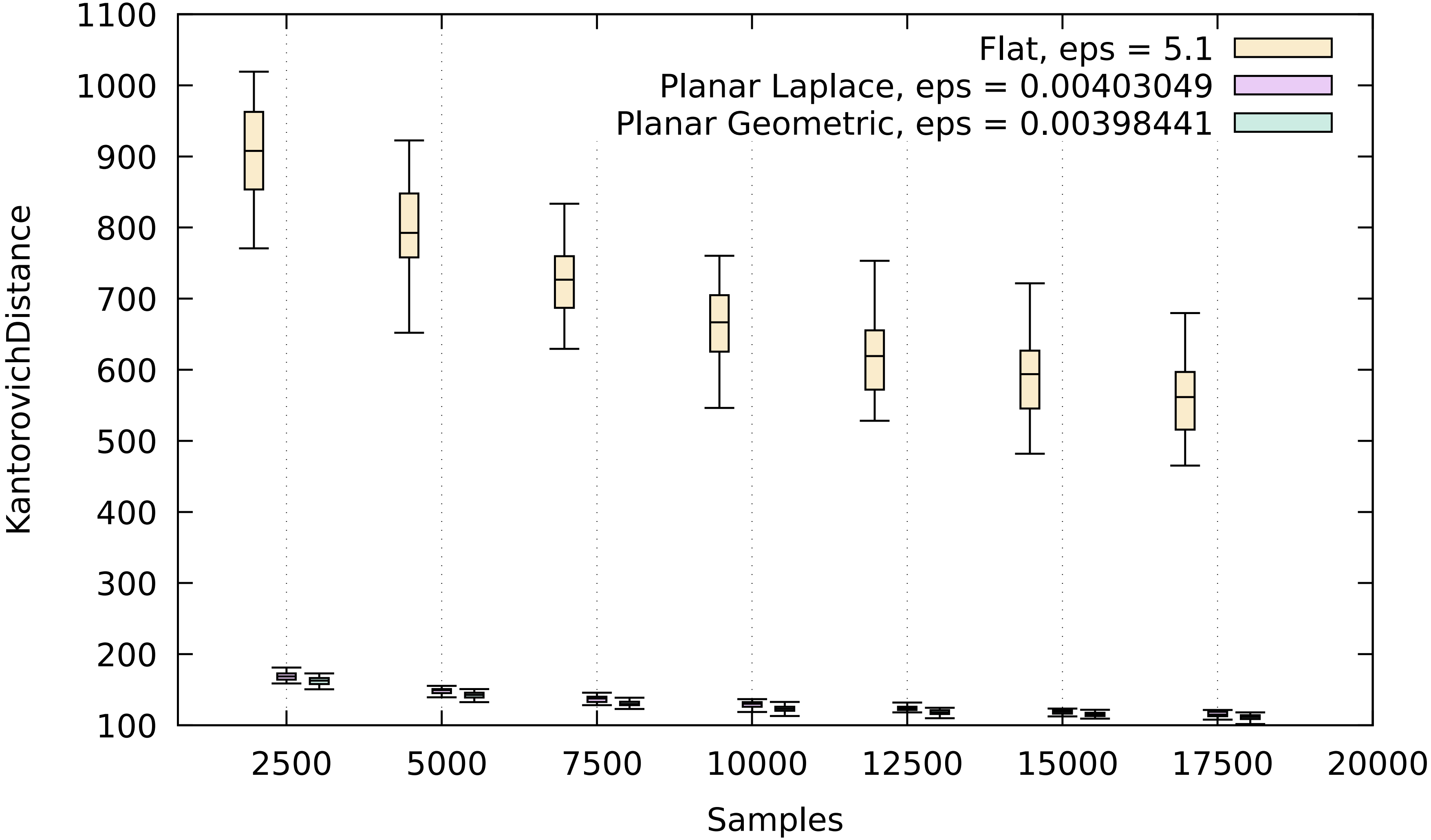


The real distribution

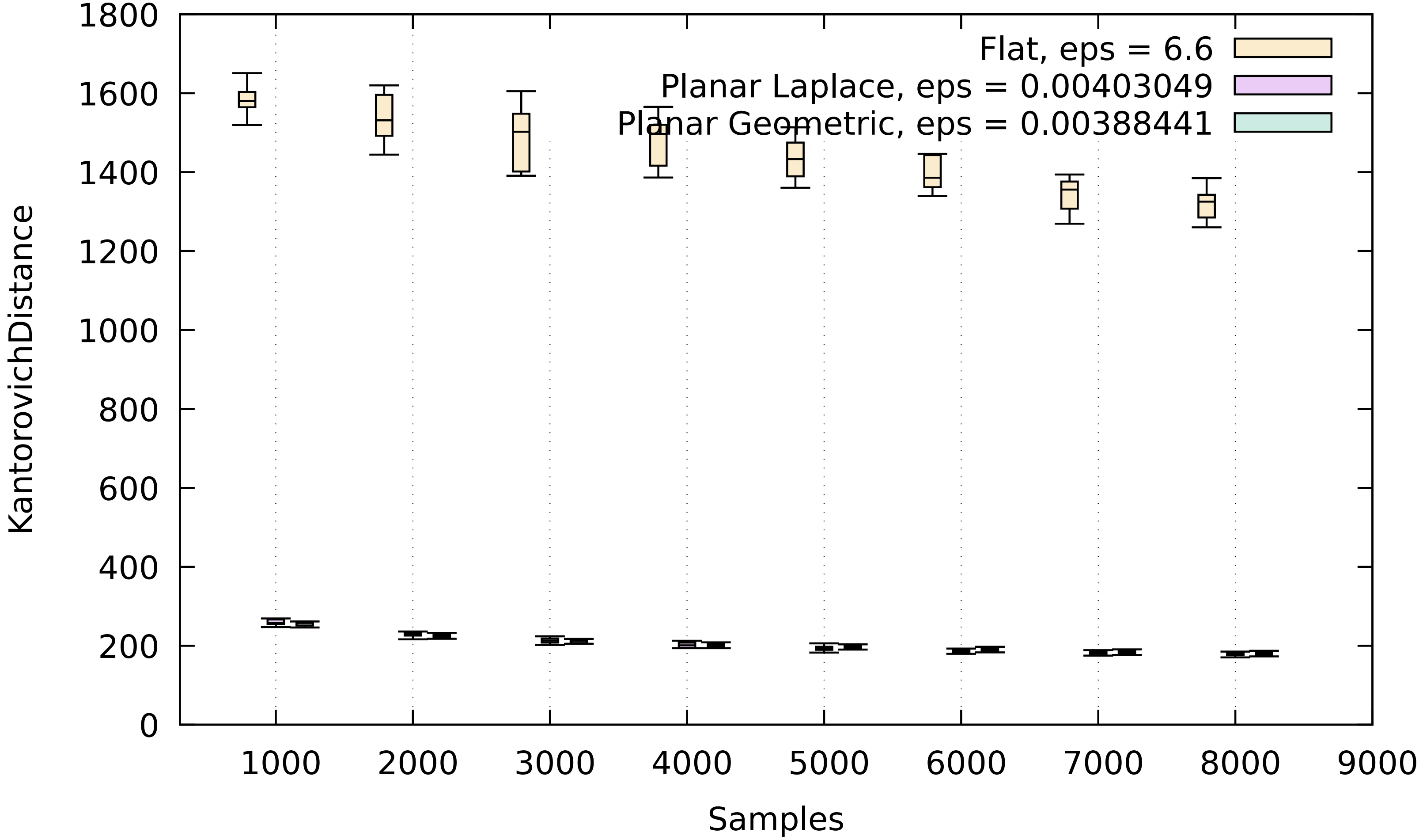


The noisy distribution and the result of the IBU (500 iterations)

Evaluation: San Francisco



Evaluation: Paris



Positions available:
Phd, Postdoc and Research Assistant

European Research Council
Established by the European Commission

HYPATIA:

- Statistical utility from noisy data
- Optimal privacy-utility trade-off
- Generation of optimal mechanism via ML

INSTITUTE DATAIA
Data Science, Intelligence & Society

- Analysis of privacy threats in ML

Thanks!

Questions?