# Three-way optimization of the privacy-utility trade-off

Catuscia Palamidessi
INRIA
Ecole Polytechnique

## Motivation and Objective

With the ever-increasing use of internet-connected devices, such as computers, smart grids, IoT appliances and GPS-enabled equipments, personal data are collected in larger and larger amounts, and then stored and manipulated for the most diverse purposes. Undeniably, the big-data technology provides enormous benefits to industry, individuals and society, ranging from improving business strategies and boosting quality of service to enhancing scientific progress. On the other hand, however, the collection and manipulation of personal data raises alarming privacy issues.

**Objective**

This project aims at developing the theoretical foundations, methods and tools to protect the privacy of the individuals while letting their data to be collected and used for statistical purposes. We aim in particular at developing a framework to study mechanisms that: (1) are *robust* with respect to combination of information from different sources, (2) can be *applied directly by the user*, thus avoiding the need of a trusted party, and (3) provide an *optimal trade-off between privacy and utility*, where utility is intended both as Quality of Service and as source of Statistical Information.

## State of the art

Until recently, the most used data sanitization technique was anonymization (removal of names) or slightly more sophisticated variants like $k$-anonymity and $\ell$-diversity. Unfortunately, these techniques have been proved ineffective, as several works have shown that individuals in anonymized datasets can be re-identified with high accuracy, and their personal information exposed. Most of these are inference attacks based on combining different sources of information.

In the meanwhile a new framework emerged and became very successful: *differential privacy* (DP) [4]. Together with its distributed version called *local differential privacy* (LDP) [3], they represents the cutting-edge of research on privacy protection, and a convincing alternative to anonymity. DP aims at protecting the individuals' data while allowing to answer queries on the aggregate information, and it achieves this goal by adding controlled noise to the query outcome. In contrast to anonymity, DP is robust to combination attacks (compositionality). LDP is a distributed variant in which the users obfuscate their personal data by themselves, before letting them be collected. In this way, the data collector can only see, stock and analyze the already sanitized data. Like DP, LPD is compositional, i.e., robust to combination attacks. Furthermore it has the following further advantages: (a) it allows each user to choose his own level of privacy, (b) it does not need to assume a trusted third party, and (c) since all stored records are individually sanitized, there is no risk of privacy breaches due to security attacks. LDP is having a considerable impact, and various companies have adopted it to collect their customers's data for statistical purposes, including Google, that has developed its own approach to LDP with the project RAPPOR [5].

## Research plan

As anticipated above, in this project we aim at developing methods and tools which are robust, flexible, controllable by the user, and preserve the utility of data. We plan to follow the principle of differential privacy to achieve robustness, and in particular its local variant that can be applied and managed by the user and that allows each user the flexibility to choose his own level of privacy.

One of the main problems in the development of privacy mechanisms is the preservation of the utility. In the case of local privacy, namely when the data are sanitized by the user before they are collected, the notion of utility is twofold:

**Utility as quality of service (QoS):** The user usually gives his data in exchange of some service, and in general the quality of the service depends on the precision of such data. For instance, consider a scenario in which Alice wants to use a LBS (Location-Based Service) to find some restaurant near her location $x$. The LBS needs of course to know Alice's location, at least approximately, in order to provide the service. If Alice is worried about her privacy, she may send to the LBS an approximate location $x'$ instead of $x$. Clearly, the LBS will send a list of restaurants near $x'$, so if $x'$ is too far from $x$ the service will degrade, while if it is too close Alice's privacy would be at stake.

**Utility as statistical quality of the data (Stat):** Bob, the service provider, is motivated to offer his service because in this way he can collect Alice's data, and quality data are very valuable for the big-data industry. We will consider in particular the use of the data collections for statistical purposes, namely for extracting general information about the population (and not about Alice as an individual). Of course, the more Alice's data are obfuscated, the less statistical value they have.

We intend to consider both kinds of utility, and study the "three way" optimization problem in the context of $d$-privacy, our approach to local differential privacy. Namely we want to develop methods for producing mechanisms that offer the best trade-off between $d$-privacy, QoS and Stat, at the same time. In order to achieve this goal, we will need to investigate various issues. In particular:

- how to best reconstruct (and approximation of) the original distribution from a collection of noisy data, in order to perform the intended statistical analysis
- what metrics to use for assessing the statistical value of a distributions (for a given application), in order to reason about Stat, and
- how to compute in an efficient way the best noise from the point of view of the trade-off between $d$-privacy, QoS and Stat.

More in detail, the research will be organized along the following lines:

## Local differential privacy on metric domains

There are many cases in which the input domain comes equipped with a notion of distance. For instance, location data, energy consumption in smart meters, age and weight in medical records, etc. Usually, when these data are collected for statistical purposes, the accuracy of the distribution is measured also with respect to the same notion of distance. We believe that the trade-off between privacy and utility can be greatly improved by exploiting the concept of approximation intrinsic in metrics. Following this intuition, we plan to develop a variant of local differential privacy based on the notion of $d$-privacy (where $d$ represent the distance function on the data domain $X$) [2] that we have proposed in our team COMETE: An obfuscation mechanism $\mathcal{K}$ is $\varepsilon d$-private if for every $x, x', y \in X$ we have $P[\mathcal{K}(x) = y] \leq e^{\varepsilon d(x,x')} P[\mathcal{K}(x') = y]$, where $P[a]$ represents the probability of the event $a$. In constrast, standard LPD requires $P[\mathcal{K}(x) = y] \leq e^{\varepsilon} P[\mathcal{K}(x') = y]$. Hence, $d$-privacy relaxes the privacy requirement of LDP by allowing two data to become more and more distinguishable as their distance increases. Nevertheless, $d$-privacy shares the pleasant properties of DP and LDP: it is compositional, it does not need a trusted third party, it is not subject to security risks, and it allows each user to choose his own level of privacy.

We have already developed $d$-privacy for the case of location privacy, and the resulting notion, called *geo-indistinguishability*, has been quite successful. Indeed, geo-indistinguishability and its typical implementation mechanism, the Planar Laplace noise, have been adopted as the basis of several tools and frameworks for location privacy, including: Location Guard [9], LP-Guardian [7], LP-Doctor [6], a system for secure nearby-friends discovery [10], and the SpatialVision QGIS plugin [11].

In this project, we intend to study the general properties of $d$-privacy in abstract metric spaces, and investigate in depth other application domains provided with a natural notion of distance, like for instance spatio-temporal trajectories, activity traces, and relational graphs.

## Statistics estimation from noisy data

In order to reconstruct as precisely as possible the original distribution from the noisy data, we plan to adapt the *Iterative Bayesian Update* (IBU) [1], which is a statistical inference method that iteratively estimates the distribution until convergence to a fixed point. It is known that the fixed points of this method are exactly the distributions which have *maximal likelihood* (given the noisy data) to be the distribution on the original data, but depending on the noise that has been applied, there could be more than one fixed points, i.e., more than one distributions with maximal likelihood and if there is more than one, then there are infinitely many because they form a vector space.

We intend to study the properties of the IBU in the case in which the noisy data are obtained using a laplacian density function or other implementations of $d$-privacy, and stopping conditions that allow to obtain sufficiently accurate estimations of the fixed point distribution. In particular, we want to identify the conditions (on the noise function) under which the fixed point is unique, which is a necessary and sufficient condition to guarantee that the result of the IBU converges to the distribution on the original data as the size of the dataset grows.

Looking at each iteration as a transformation on distributions, we will study the topological properties of this transformation, and provide precise guarantees on the stopping criteria. In particular, we plan to identify the conditions under which the transformation is a *contraction* (i.e. a distance-shinkring transformation), which will ensure the uniqueness of the fixed-point. We plan to use, for this purpose, the so-called Kantorovich distance on distributions induced by the underlying distance $d$ on $X$.

### Three-way optimization of the trade-off privacy-utility

We will study how to tune the parameters so to obtain mechanisms that offer the best trade-off between $d$-privacy, QoS and Stat, at the same time. We plan to use convex optimization techniques for this purpose, such as Lagrange multipliers.

### Comparison of the trade-off utility-privacy

We will compare our proposal based on $d_X$-privacy with the other LPD mechanisms proposed in literature, like the K-ary Randomized Response (K-RR) [8] and Google's Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR) [5]. The comparison will be based on the evaluation of the trade-off between privacy and utility provided by each method. Each proposal is tailored to its own notion of privacy (the $\varepsilon$ used in $d$-privacy is to be multiplied by a distance, hence it has not the same meaning as the $\varepsilon$ used in LPD), hence, in order to be fair, the privacy comparison will have to be performed on neutral ground, i.e. we will have to find a unifying notion of privacy which can measure and normalize all the notions to be compared. The utility will be measured in terms of approximation of the true distribution, using a notion of distance between distributions that takes into account the underlying distance on the data domain, like the Kantorovich metric.

### Implementation

We intend to implement the mechanisms for privacy protection based on $d$-privacy, using an extended notion of Laplacian noise, and the statistical reconstruction methods based on IBU for the stochastic matrices generated by this mechanism. We will study the decomposition of the IBU iteration step in vector products, so to obtain a very fast implementation on the GPU cluster of the team COMETE. Furthermore we will study how to implement efficiently the Kantorovich metric. To this end, we plan to investigate the techniques developed in the field of computer vision, where the Kantorovich metric is known as Earth Mover's Distance.

*Cases study.* We plan to apply our methods to real-life cases, using data collections that are available publicly, like Geolife and Gowalla (for spatio-temporal trajectories). We will seek for data collections available in other domains, possibly from the industrial partners of INRIA and of the Ecole Polytechnique.

### Example: A preliminary study for the case of location data

We show here an example of what we want to achieve. We have experimented with data from the Gowalla dataset in an area of $3 \times 3$ Km$^2$ in San Francisco downtown. The dataset consists of about 10K records representing checkins from a community of users, where each record contains, among other information,

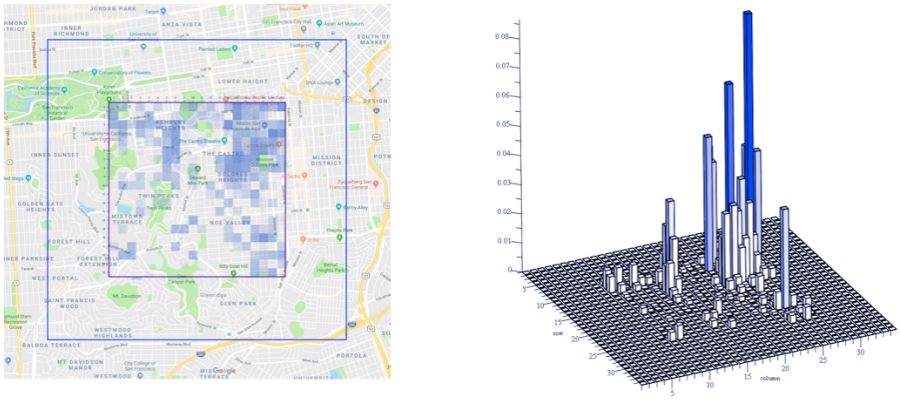


Figure 1: The upper left corner of the blue square in (a) corresponds to the upper left of the basis of (b).

the geographical coordinates of the location of the user at the moment he did the checkin. We have discretized the area using a grid of $20 \times 20$ cells of $150 \times 150$ m$^2$ each, lumping together the checkins from the various points of the same cell. The area and the density of the checkins are illustrated in Figure 1(a) (the intensity of the color blue represents, in logarithmic scale, the density), while Figure 1(b) depicts the same density normalized to a probability distribution (i.e., scaled uniformly in such a way that the total mass is 1), and in form of histogram.

We have experimented with two mechanisms: KRR, which is a typical mechanism for standard local differential privacy, and the planar laplacian, which is a typical implementation of our proposal, $d$-privacy. We have calibrated the respective privacy parameters so to provide a similar protection.

For the KRR mechanism we have choosen $\varepsilon = \ln(16)$, which means that the true cell is 16 times more likely to be reported than any other cell in the grid [1]. We have applied the mechanism independently on each checkin of the Gowalla dataset, and we have collected the resulting noisy locations. The distribution computed from the noisy data is illustrated in Figure 2(a). Then we have applied the IBU to try to retrieve the original distribution. The result obtained after 400 iterations is shown in Figure 2(b). As we can see, it is quite different from the original distribution. This is not because 400 iterations are too few, nor is it the fault of the IBU: the resulting distribution *is* the most likely original distribution given the noisy data. The problem is that the noisy data have created too much confusion, and in particular the noisy checkins coming from the high peak in the original distribution have been scattered everywhere. The evolution of the distribution from (a) to (b) via the IBU can be visualized at http://www.lix.polytechnique.fr/%7Ecatuscia/temp/IBU_kRR_SF.gif.

The second mechanism we have considered is the planar laplacian noise, with $\varepsilon$ calibrated so to give the same level of $\ln(16)$-geo-indistinguishability in an area of $1.2 \times 1.2$ Km$^2$. To obtain this, we set the $\varepsilon = \ln(2)$. This means that the true location is 2 times more likely to be reported than any other cell in the grid *among those that are within an area of* $450 \times 450$ m$^2$ *centered on the true location.*

Again, we have applied the laplacian noise to each checkin independently, and collected the noisy data. The distribution computed on the noisy data is shown in Figure 3(a). As we can see, it is less confused than the one obtained with KRR, and in fact it has retained more information from the original distribution. In fact by applying the IBU the results converges after about 300 iterations and the result, shown is in

[1] One may think that we are using more noise than necessary, because the probability to report the true cell is only $16/415$, but with an easy calculation we can see that with as little as 4 checkins from the same cell the probability of reporting at least twice the true cell is almost 30%, while it is negligible for every other cell. Thus with less noise we would not have adequate protection in case of repeated checkins from the same location or locations nearby.

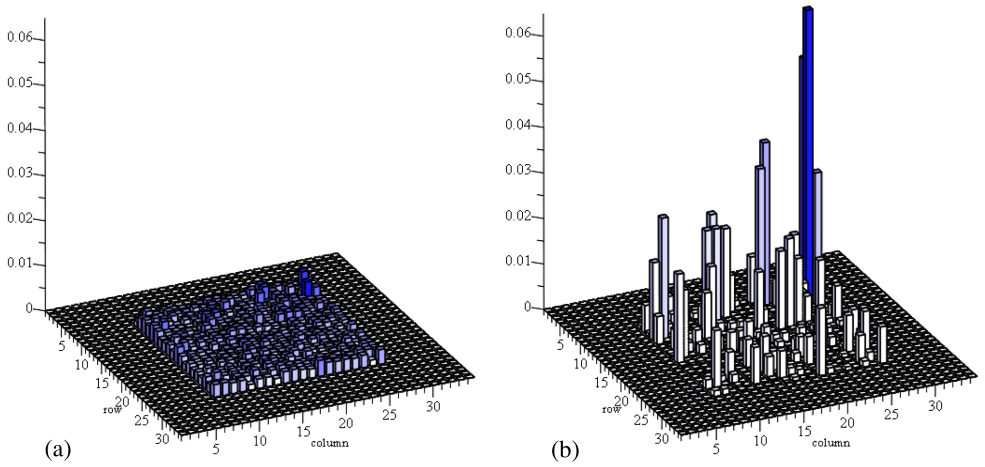


Figure 2: (a): Distribution after the application of the KRR mechanism with $\varepsilon = \ln(16)$. (b): Result after 400 iterations of the IBU.

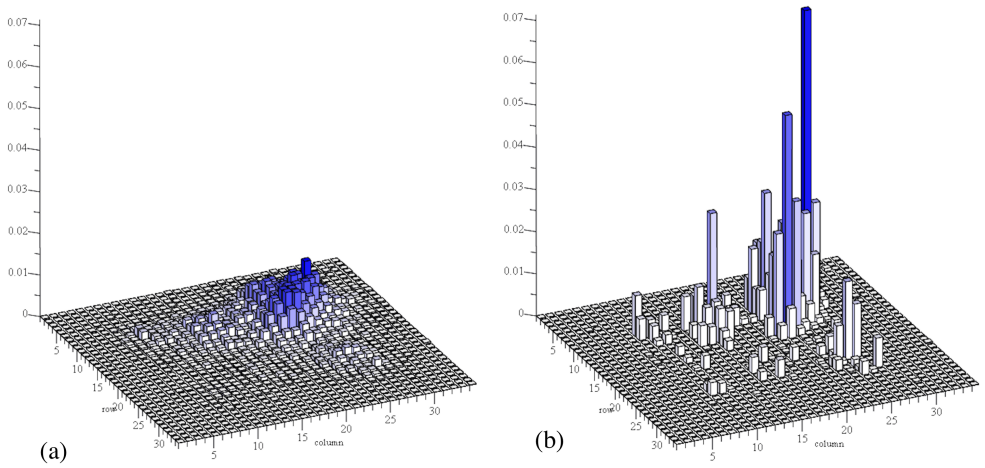


Figure 3: (a): Distribution after the application of the Planar Laplacian mechanism with $\varepsilon = \ln(2)$. (b): Result after 300 iterations of the IBU.

Figure 3(b), is quite similar to the original distribution. The evolution from (a) to (b) can be visualized at http://www.lix.polytechnique.fr/%7Ecatuscia/temp/IBU_PlanarLaplacian_SF.gif.

## Novelties of the proposal

The major novelty of our proposal consists in considering the optimization between privacy and two notions of utility: the Quality of Service and the Statistical Quality of the collected data. The mechanisms for LDP considered so far in the literature only consider the trade-off between privacy and the statistical utility.

A second contribution consists in enhancing local differential privacy with the notion of distance. The mechanisms for LDP considered so far ignore any structure on the input domain. For this reason, we are convinced that our proposal will allow us to increase significantly the trade-off between utility and privacy.

A third significant novelty is the study of the topological properties of the Iterative Bayesian Update method, which will provide useful criteria for its application in practical problems of statical reconstruction. Despite of its proved performance, the theory underlying this method is largely unexplored.

Finally, a fourth important contribution will consist in the methods for comparing the reconstructed statistics on the basis of the Kantorovich distance, and the study of techniques for its efficient implementation in the case of statistics extracted from noisy data generated by privacy-protection mechanisms. The traditional methods in the literature of LPD typically use the total variation distance, the Hellinger distance, or the Kullback–Leibler divergence, none of which takes the structure of the domain into account.

We believe that the Kantorovich distance, being a lifting from the metric of the domain, is the right one to measure the quality of a distribution whose precision depends not only on the probability mass, but also on the approximation of the correct value.

# References


[1] Rakesh Agrawal, Ramakrishnan Srikant, and Dilys Thomas. Privacy Preserving OLAP. In *Proceedings of the 24th ACM SIGMOD Int. Conf. on Management of Data*, SIGMOD '05, pages 251–262. ACM, 2005.

[2] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás E. Bordenabe, and Catuscia Palamidessi. Broadening the scope of Differential Privacy using metrics. In *Proc. of PETS*, volume 7981 of *LNCS*, pages 82–102. Springer, 2013.

[3] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *Proc. of FOCS*, pages 429–438. IEEE Computer Society, 2013.

[4] Cynthia Dwork, Frank Mcsherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proc. of TCC*, volume 3876 of *LNCS*, pages 265–284. Springer, 2006.

[5] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In *Proc. of CCS*, pages 1054–1067. ACM, 2014.

[6] Kassem Fawaz, Huan Feng, and Kang G. Shin. Anatomization and protection of mobile apps' location privacy threats. In *Proc. of USENIX Security 2015*, pages 753–768. USENIX Association, 2015.

[7] Kassem Fawaz and Kang G. Shin. Location privacy protection for smartphone users. In *Proc. of CCS*, pages 239–250. ACM Press, 2014.

[8] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *The JMLR*, 17(1):492–542, 2016.

[9] Location guard. https://github.com/chatziko/location-guard.

[10] Changsha Ma and Chang Wen Chen. Nearby friend discovery with geo-indistinguishability to stalkers. *Procedia Computer Science*, 34:352 – 359, 2014.

[11] QGIS Processing provider plugin (Methods for anonymizing data for public distribution). https://github.com/SpatialVision/differential_privacy.