# Automatic Privacy and Utility Preservation of Mobility Data: A Nonlinear Model-Based Approach

Sophie Cerf, Sara Bouchenak, Bogdan Robu, Nicolas Marchand, Vincent Primault, Sonia Ben Mokhtar, Antoine Boutet, Lydia Y. Chen

# Automatic Privacy and Utility Preservation for Mobility Data: A Nonlinear Model-Based Approach

Sophie Cerf[•], Sara Bouchenak[⋆], Bogdan Robu[•], Nicolas Marchand[•],
Vincent Primault[II], Sonia Ben Mokhtar[⋆], Antoine Boutet[†], Lydia Y. Chen[◇]
[•] Univ. Grenoble Alpes – GIPSA-Lab – CNRS, Control Theory Research Group, Grenoble, France
[⋆] INSA Lyon – LIRIS – CNRS, Distributed Systems Research Group, Lyon, France
[II] University College London – Dept. Computer Science – Information Security Group, London, United Kingdom
[†] INSA Lyon – CITI – Inria, Privatics Research Group, Lyon, France
[◇] IBM Zurich Research Lab, Zurich, Switzerland
[•]{firstname.lastname}@gipsa-lab.fr,[⋆],[†]{firstname.lastname}@insa-lyon.fr,[II]v.primault@ucl.ac.uk,[◇]yic@zurich.ibm.com

***Abstract*—The widespread use of mobile devices and location-based services has generated massive amounts of mobility databases. While processing these data is highly valuable, privacy issues can occur if personal information is revealed. The prior art has investigated ways to protect mobility data by providing a large range of Location Privacy Protection Mechanisms (LPPMs). However, the privacy level of the protected data significantly varies depending on the protection mechanism used, its configuration and on the characteristics of the mobility data. Meanwhile, the protected data still needs to enable some useful processing. To tackle these issues, we present PULP, a framework that finds the suitable protection mechanism and automatically configures it for each user in order to achieve user-defined objectives in terms of both privacy and utility. PULP uses nonlinear models to capture the impact of each LPPM on data privacy and utility levels. Evaluation of our framework is carried out with two protection mechanisms of the literature and four real-world mobility datasets. Results show the efficiency of PULP, its robustness and adaptability. Comparisons between LPPMs' configurator and the state of the art further illustrate that PULP better realizes users' objectives and its computations time is in orders of magnitude faster.**

***Index Terms*—D.4.6 Security and Privacy Protection; D.4.8.b Modeling and prediction; H.2.0.a Security, integrity, and protection; J.9.a Location-dependent and sensitive; D.2.16.b Configuration control**

## I. Introduction

Geolocation data is increasingly used to improve the quality of services, leading to the surge of Location Based Services (LBS) such as navigators or nearest places recommendation applications. Hence, a large amount of mobility data is generated and currently used by companies and researchers. Indeed, the processing of mobility data can reveal many valuable information that may be used for a broad range of applications, e.g., traffic congestion management, urban development and etc. However, the processing of location data also comes with threats on the privacy of the recorded users. As motivation for privacy protection, one can cite the publication of mobility dataset by Strava in 2018 that revealed the maps of unknown US military bases [30]; or the new regulations that are enforced by the governments, such as the European GDPR [14]. The most common threats on privacy are (i) reidentification attacks where the identity of an anonymous user is guessed based on previously recorded data [11], [19], (ii) mobility prediction that anticipates users' next moves based on their habits [31], [13], (iii) extraction of user's places of interest ( e.g., home, workplace [12], and place of worship [10]) and (iv) inference of social relationships (e.g., partners, and coworkers) [4]. In this work, we will focus on the privacy threat that consists in identifying a user's Points Of Interest (POI), as it is often the very first step to infer users' other information [24].

To overcome these privacy issues, many efforts in the literature aim to develop protection mechanisms. The protection efforts are not only motivated by cautious companies and researchers but is more and more forced by national and international governments and organizations. The so-called Location Privacy Protection Mechanisms (LPPM) modify the location information of users to improve their privacy level. The principle behind each LPPM varies; for instance Geo-Indistinguishability (GEO-I) adds noise to the spatial information of a user data [3], PROMESSE modifies timestamps in order to smooth the user speed [25], and CloakDroid assigns the value of a location point to its closest location on a grid [20]. LPPMs need fine tuning of their parameters that may require sophisticated knowledge and experience. The choice of these configurations (e.g., the amount of noise and data granularity) significantly impacts the level of protection of the obfuscated data. The obfuscation should be carried out carefully to make sure that the utility of the protected data is preserved. Indeed, the mobility data aims at being processed to retrieve some high level information ( e.g.,road frequencies and means of transportation). Dealing with both privacy and utility simultaneously is not straightforward given the natural trade-off that exists between the two. As privacy enhancing mechanisms alters the original datasets to hide information, the data usability by definition decreases. Using a LPPM may result into various levels of privacy and utility depending on the properties of the user mobility.

In order to enable the feasibility and practicability of these protection mechanisms for end-users, some configuration mechanisms have been proposed in the literature. In [1], the authors present a heuristic-based mechanism that iteratively modifies the configuration of a spatial cloaking LPPM to meet a privacy-oriented objective, while considering the utility of

data. In [8], the authors adapt the configuration of GEO-I to the density of the surrounding area, assuming that the less noise is needed to add to the original data for privacy protection when there are people around. In [26], the authors propose an iterative greedy mechanism that evaluates privacy and utility of obfuscated data to refine the parameters of a certain LPPM configuration. These solutions are often computing intensive as they are heuristics-based. They do not always explicitly take into account the usability of the protected data and lack of objective-driven formulation that enables a user to define her privacy and utility requirements. Moreover, these works focus only specific protection mechanism, hindering its applicability to comparing across mechanisms.

There is a strong need for a solution that enables choosing between LPPMs and configuring the chosen one in order to meet user-defined objectives in terms of privacy and utility. In this paper we present *PULP*, standing for Privacy and Utility through LPPM Parametrization. *PULP* a framework which automatically selects a LPPM among different ones, and determines the best configuration of the LPPM based on each user's objective. The core of *PULP* is user-specific modeling that captures the impact of every considered LPPM on the privacy and utility level of the obfuscated data. A model is built by measuring the privacy and utility levels of obfuscated data after a few profiling runs of applying the LPPM with a set of configuration parameters. Based on each user objectives, the behavioral model is used to chose and configure a LPPM for each one. Four objective formulations are considered, for various combinations of objectives in terms of privacy and utility: (i) ensure a given ratio between privacy and utility, (ii) guarantee minimal levels of privacy and utility, (iii) keep privacy above a given level while increasing utility as much as possible, and (iv) guarantee a minimal utility level while improving privacy as much as possible.

The evaluation of *PULP* is carried out using four mobility databases of 770 users in total. Two LPPMs, i.e., GEO-I and PROMESSE, are considered in the evaluation, and a POI-based privacy metric and a spacial-based utility are considered. Results show that *PULP* accurately models the impact of LPPMs on users' data and thus enables to recommend a LPPM and its configuration so as to satisfy the objectives, when possible. Results highlight the importance of tuning the LPPM and its configuration for individual users. Moreover, the use of models enables a significant reduction of the computing time compared to state of the art.

The contributions of the paper are then:

- Accurate, robust and adaptive modeling of LPPMs with different configuration parameters,
- Computing-efficient objective-based recommendation and configuration laws of protection mechanisms.

The rest of this paper is organized as follows. First some background information is given in Section II. Then, *PULP* framework is described in Section III. Sections IV and V describe *PULP*'s automatic LPPM configuration laws. Experimental validation and analysis are carried out in Section VI, followed by a review of the state of the art in Section VII. Conclusion and perspectives end the paper in Section VIII.

## II. Background and Motivation

In this section, we first provide detailed description of the mobility traces, LPPMs considered and the formal definitions of privacy and utility metrics, followed by a motivating example of why no single LPPM solution fits all users.

### A. Mobility Traces

The base of this work is mobility datasets collected in the wild: the Cabspotting (CABS) [23], the PRIVAMOV [6], the GEOLIFE [32], and the Mobile Data Challenge (MDC) datasets [17], amounting to a total of 770 users. These datasets contain mobile information about users during their daily life. To have homogeneous datasets, we align the length period of the four datasets to the one of the shortest (i.e., CABS which has 30 days of mobility data) by selecting their most active period.

We index each user by the subscript of $i$, and each LPPM by $j$. We denote the mobility trace of user $U_i$ by $T_i$ when the mobility data has not been obfuscated, and by $T'_{ij}$ after applying $LPPM\ j$ on the trace $T_i$. Both $T_i$ and $T'_{ij}$ are sets of records chronologically ordered. A record is a tuple $\langle lat, lng, t\rangle$ that indicates for user $U_i$ her location on the surface of the Earth defined by latitude-longitude coordinates (i.e., $lat, lng$), at a given time $t$.

### B. Location Privacy Protection Mechanisms (LPPMs)

Roughly speaking, state-of-art LPPMs alter the spatial information of user mobility data or its temporal information. In the following, we present two examples of LPPMs: GEO-I that focuses on spatial distortion, and PROMESSE that adds temporal disturbance to the data.

**GEO-I.** Geo-Indistinguishability protects a user location data by adding spatial noise drawn from a Laplace distribution to each record of the actual mobility trace [3]. GEO-I has a configuration parameter $\epsilon$, expressed in meters$^{-1}$ varying in $\mathbb{R}^+$, which quantifies the amount of noise to add to raw data. The lower the $\epsilon$ is, the more noise is added. GEO-I is a state of the art LPPM that follows the differential privacy model [9]; that is, it allows to calibrate noise in order to increase privacy while reducing the impact on data utility. Therefore, in the following we consider GEO-I as one underlying LPPM to validate our *PULP*'s approach.

**PROMESSE.** PROMESSE has been developed in order to prevent the extraction of Points-Of-Interest (users' stop places) while maintaining a good spatial accuracy [25]. Its principle is to distort timestamps of location traces as well as remove and insert records in a user's trace in order to keep a constant distance between two events of the trace (parameterized by $\epsilon$ in meters). PROMESSE adds temporal noise to a trace while GEO-I introduces spatial noise.

Although we specifically consider GEO-I and PROMESSE, the proposed methodology in the following sections is general for any LPPM that works for every user independently and that has a single configuration parameter. For some LPPMs, the computation of obfuscated trace is done accordingly the obfuscation of other users, k-anonymity for instance. *PULP* works only for LPPM for which the obfuscation for one user only depends on this user.

### C. Data Privacy and Utility Metrics

Protecting raw mobility data with LPPMs improves the user privacy but also risks the quality or the usability of the resulting data. There is no standard way of assessing, at a user level, these two complementary dimensions associated to LPPMs. We choose to define privacy by looking at a user's POI (i.e. significant stops) protection [12], and utility by evaluating the spatial accuracy of revealed locations [8]. Both metrics evaluate the gain of privacy and the loss of utility of the obfuscated data compared to the raw data. The proposed metrics have parameters that enable to adjust the notion of privacy and utility to the considered LBS and to the user requirements. The next paragraphs define privacy and utility metrics and give an illustrated example of their computation.

*1) Privacy Metric:* To evaluate privacy of mobility traces, we first consider the retrieval of POIs. A Point Of Interest is a meaningful geographical area where a user made a significant stop. A POI is defined by the position of its centroid and by a diameter $d$, describing an area where the user stayed for at least $t$ minutes. We define $poi(T)$ as the set of POIs retrieved from the mobility trace $T$.

Using the concept of POI and $poi(\cdot)$ set, we aim to quantify user' privacy level by looking at how POIs retrieved from the obfuscated data (under LPPM $j$) match successfully to the POIs retrieved from the non-obfuscated data, i.e., comparison between $poi(T_i)$ and $poi(T'_{ij})$ sets. We define the function $\mathrm{Matched}(poi(T'_{ij}), poi(T_i))$ that, given two sets of POIs, derive the subset of $poi(T'_{ij})$ containing the POIs that match with POIs in the second set $poi(T_i)$. Two POIs are considered as *matched* if they are sufficiently close one to the other ($d_{max}$ being the maximal distance threshold). To formally define privacy, one can use either measurement of precision ($P_{pr}(i,j)$) which defines the ratio between the number of obfuscated trace's POIs successfully matched with real POIs and the number of obfuscated POIs,

$$P_{pr}(i,j) = \frac{|\mathrm{Matched}(poi(T'_{ij}), poi(T_i))|}{|poi(T'_{ij})|},$$

or recall ($R_{pr}(i,j)$) which defines the ratio between the number of obfuscated trace's POIs successfully matched with real POIs and the number of real POIs,

$$R_{pr}(i,j) = \frac{|\mathrm{Matched}(poi(T'_{ij}), poi(T_i))|}{|poi(T_i)|}.$$

The precision function assesses the accuracy of the matching while the recall function evaluates the completeness. We advocate to use Fscore to reconcile the both measurements of precision and recall.

We formally write the privacy metric, showing the normalized percentage of successfully hidden (non-matched) POIs, after applying LPPM $j$ on user $i$ as:

$$Pr(i,j) = 1 - \frac{2 \cdot P_{pr}(i,j) \cdot R_{pr}(i,j)}{P_{pr}(i,j) + R_{pr}(i,j)}. \quad (1)$$

This privacy metric is defined in range of $[0,1]$ where a higher value reflects a better protection.

Leveraging the POI diameter $d$, its minimal duration $t$ and the matching threshold distance $d_{max}$ enable to clearly define a user's conception of her privacy. For instance, a user willing to hide her home and work place with a high accuracy should choose a large $t$, and small $d$ and $d_{max}$. However, if a user wants to hide most of the places she goes to in order to dissimulate her hobbies, she should set a small $t$ and a rather large $d_{max}$.

*2) Utility Metric:* To evaluate data utility of users' traces, we resort to the comparison between the area coverage of the original mobility trace and the one of the obfuscated trace. We define the area coverage using the concept of map cells. A cell is said visited (or covered) by a user if the mobility trace of the user contains at least one record with coordinates in this cell. We first define $cell(T_i)$ and $cell(T'_{ij})$ as the sets of cells visited by the mobility trace of user $i$, before and after applying the LPPM $j$. To enable the comparison of cell coverage across a user's trace, we use precision and recall; formally defined as

$$P_{ut}(i,j) = \frac{|cell(T_i) \cap cell(T'_{ij})|}{|cell(T'_{ij})|},$$

$$R_{ut}(i,j) = \frac{|cell(T_i) \cap cell(T'_{ij})|}{|cell(T_i)|}.$$

Similarly to privacy metric, we finally define the utility metric of user $i$ obfuscated with LPPM $j$, $Ut(i,j)$, by the Fscore which gathered precision and recall of cell coverage

$$Ut(i,j) = \frac{2 \cdot P_{ut}(i,j) \cdot R_{ut}(i,j)}{P_{ut}(i,j) + R_{ut}(i,j)}. \quad (2)$$

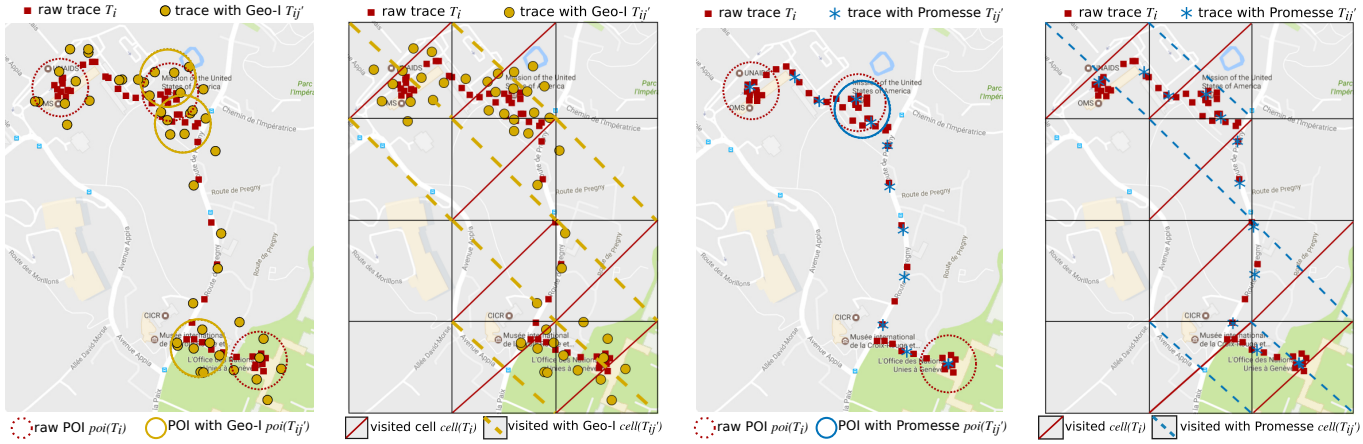
This utility metric is defined in the range $[0,1]$, where a higher value reflects a better utility, meaning a better spacial accuracy of the LBS results.

Playing with the cells' size enables to adapt to the LBS used. Some services require a really good spatial accuracy such as running training apps; and some are less demanding, such as news apps. For the first category of LBS, cells' size should be small (tenth of meters) while for the other the size can be way larger (more than a kilometer).

Note that the level of privacy and utility of a user depends not only on the LPPM used to protect her data but also of its configuration $\epsilon$. However, for sake of readability, we did not introduce $\epsilon$ here in our notations.

*3) Illustration of Privacy and Utility Metrics with LPPMs:* To better illustrate the definition of privacy and utility, we use a schematic example by applying GEO-I and PROMESSE on a synthetic user's trace, see Figure 1.

*Computing privacy metric:* In Figure 1 (a), the raw mobility trace of the user $T_i$ is represented with the small squares, each square being a location record. We overdraw the mobility trace after adding some noise with GEO-I ($T'_{ij}$). Their Points-of-Interest (POIs) are illustrated with large circles. The set of POIs of the original trace $poi(T_i)$ are the dashed circles, while POIs of the obfuscated trace $poi(T'_{ij})$ are the continuous ones. Based on those sets, we can compute the number of obfuscated POIs that match the real ones (the 2 top ones in this example). Thus, the precision and recall of the matching of POIs are $2/3$ and the level of privacy $1/3$.




(a) POI - GEO-I (b) Cell - GEO-I (c) POI-PROMESSE (d) Cell-PROMESSE

Fig. 1. Schematic examples of how POIs and cell coverage change for a single user after applying GEO-I and PROMESSE.

Figure 1 (c) is similar to Figure 1 (a) but here the considered LPPM is PROMESSE. In this case, the obfuscated data $T'_{ij}$ (the small stars) are spatially regularly distributed (timestamps are modified). In this illustration, all obfuscated POIs correspond to the real ones, the privacy precision is 1 but the recall is only $1/3$. The resulting privacy value is then $1/2$.

*Computing utility level:* Utility metric is illustrated in Figure 1 (b) for GEO-I and in 1 (d) for PROMESSE. In each case, the set $cell(T_i)$ is illustrated by the cells with the right diagonal (7 in total) while the sets $cell(T'_{ij})$ are the ones with left dashed diagonals. For GEO-I, the obfuscated trace covers 9 cells, the utility precision is $7/9$ and the recall 1, thus the utility level is 0.86. For PROMESSE, the obfuscated trace covers 6 cells, the precision and recall are respectively 1 and $6/7$, hence a utility of 0.92.

### D. Problem Statement: No Single Solution Fits All

We now present a motivating example showing that applying LPPMs in an ad-hoc fashion can result into very different privacy and utility values for individual users. Particularly, we choose four users (selected to show diversity) and apply both GEO-I with $\epsilon 1 = 0.01m^{-1}$ and $\epsilon 2 = 0.005m^{-1}$, and PROMESSE with $\epsilon = 100m$ for all of them. Following definitions of eq. (1) and (2), we obtain the privacy and utility

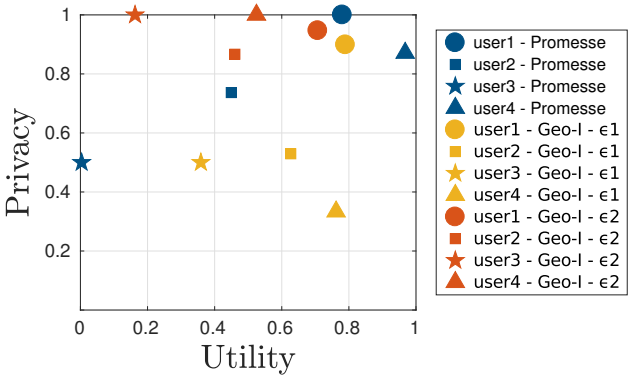



Fig. 2. Same LPPM can result into different privacy and utility metrics: examples from 4 users using PROMESSE with $\epsilon = 100m$ and GEO-I with two different configurations: $\epsilon 1 = 0.01m^{-1}$ and $\epsilon 2 = 0.005m^{-1}$.

metrics for all combinations of LPPMs, configurations, and users in Figure 2. Let us first analyze those metrics from the perspective of individual users. Both utility and privacy values of user 4 (triangles of all colors) differ when applying GEO-I or PROMESSE, showing the importance of LPPM choice. Such an observation can also be made for user 1, 2 and 3, with varying degrees of differences. Taking the perspective fixed LPPM, either GEO-I or PROMESSE, one can see that they offer different levels of privacy protection and utility preservation to different users (symbols of the same color). Figure 2 also illustrate that using one LPPM but with various configurations can lead to totally different privacy protection and service utility. In other words, it is impossible to find a single (configuration) solution that fits all users' privacy and utility objectives. All these observations highlight the complex interplay among privacy/utility metrics, the LPPM and its configuration. Moreover, to ensure the fulfillment of privacy and utility objectives for every user, it is deemed important and necessary to consider the impact of LPPMs and their configurations at the level of individual users.

## III. Design Principles of *PULP* Framework

### A. Overview

This section describes the methodology and design principles of *PULP*, a framework that efficiently chooses and configures LPPMs according to each users' privacy and utility objectives. *PULP* leverages a nonlinear modeling approach and provides several variants of automatic LPPM configuration laws. The key components of *PULP* framework are illustrated in Figure 3, and the configuration laws are summarized in Figure 4.

The profiler conducts off-line experiments to build users' privacy and utility profiles, with respect to the LPPMs considered and a set of values of their configuration parameter. For each user, the modeler bases on its off-line profile and extrapolates the privacy models and utility models which are non-linear functions in LPPM configuration parameter (one privacy model and one utility model for each LPPM). According to users' objectives and privacy & utility models, the configurator suggests the suitable LPPM and its configuration.

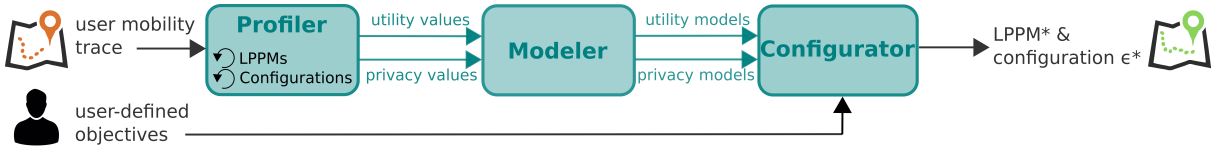



Fig. 3. *PULP* framework

*PULP* is effective for processing already collected databases, with protection mechanisms that work at the user level, with only one main configuration parameter. Indeed, *PULP* is not suitable for online processes, as it does not include temporal aspects in the decision making.

The rest of this section presents each component of *PULP*. Then, Sections IV and V describe *PULP*'s automatic LPPM configuration laws.

### B. Profiler

The aim of the profiler is to obtain the values of privacy and utility of individual users under a given LPPM and its configuration parameter set of values. The profiler takes as input a user's mobility trace and loops on all LPPMs and on a set of their possible configurations. The outputs are the resulting list of privacy and utility metrics values for all cases. Specifically, the profiler considers two LPPMs, GEO-I with $\epsilon = [10^{-4}, 1]$ in $meter^{-1}$ and PROMESSE with $\epsilon = [50, 10^4]$ in $meter$ where range values are chosen according to LPPMs' authors recommendations. The number of configuration values needed is driven by the fitting accuracy of the modeler. One shall choose the set of configuration values to run and its size such that a certain accuracy of the model is reached. The number of values required depends on the accuracy target as well as the functional form of models. Suggestions on how to choose them are given in Section VI-B3.

### C. Modeler

The aim of the modeler is to derive the functional relation between privacy/utility metrics and the configuration parameter of a given LPPM, i.e., $Pr(i,j) = F^i_{pr}(\epsilon_j)$ and $Ut(i,j) = F^i_{ut}(\epsilon_j)$.

To search for the most suitable and general function, we conduct numerous data fitting schemes on our datasets. Figure 5 depicts commonly seen dependency between privacy/utility and $\epsilon$, via an example of applying GEO-I and PROMESSE on a CABS user (continuous lines). Experimental conditions are further detailed in Section VI-A. The curves' shape can be explained by the limited ranges of privacy and utility metrics in $[0, 1]$ and insensitiveness of metrics to extreme values of $\epsilon$. These observations lead to choose arctan function as our base model, instead of general polynomial functions. The general shape of our observations makes us to use $ln(\epsilon)$ to fit the arctan model of $F_{pr}$ and $F_{ut}$, instead of $\epsilon$ directly.

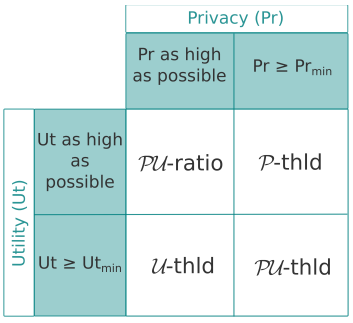
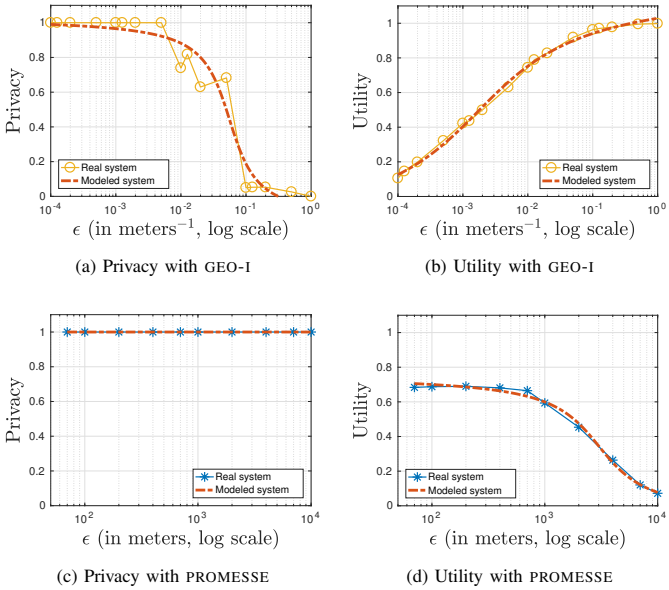



Fig. 4. Automatic configuration laws in *PULP*




(a) Privacy with GEO-I (b) Utility with GEO-I

(c) Privacy with PROMESSE (d) Utility with PROMESSE

Fig. 5. Impact of LPPMs configuration on a user's privacy and utility metrics – Real system vs. modeled system (CABS user)

Now, we formally introduce the utility and privacy models with four coefficients each. Index $i$ of user and $j$ of LPPM are not used in the following notation even if there is one privacy model and one utility model per user and per LPPM.

$$F_{pr}(\epsilon) = a_{pr}.tan^{-1}\left(b_{pr}(ln(\epsilon) - c_{pr})\right) + d_{pr}, \quad (3)$$

$$F_{ut}(\epsilon) = a_{ut}.tan^{-1}\left(b_{ut}(ln(\epsilon) - c_{ut})\right) + d_{ut}. \quad (4)$$

An illustration of model shapes is given in Figure 5, in dashed line.

The physical meaning of model parameters in both $F_{pr}$ and $F_{ut}$ are: $a$ and $d$ represent the two saturation levels, $a$ models their amplitude and $d$ their offset. $b$ characterizes the transition speed between the saturation levels while parameter $c$ corresponds to the $\epsilon$ value that results into the median privacy (or utility) value. Specific values of parameters in $F_{ut}$ and $F_{pr}$ need to be learned from each combination of user $i$ and LPPM $j$. The proposed models have the computational advantage that there are only four coefficients to be learned.

### D. Configurator

The aim of *PULP* configurator is to select and configure a LPPM from the available LPPM set so as to satisfy the

user defined objectives. These objectives are related to the user privacy (the proportion of her POIs to be hidden) and to the data utility (the proportion of correct map cells coverage). We consider four types of user's objective formulation, that combines privacy and utility differently (see also Figure 4):

- $\mathcal{PU}$-ratio: keeping both privacy and utility as high as possible, with a given ratio between privacy and utility, e.g., privacy is twice more important than utility;
- $\mathcal{P}$-thld: guaranteeing that privacy is above a given threshold, while keeping utility as high as possible;
- $\mathcal{U}$-thld: guaranteeing that utility is above a threshold, while keeping privacy as high as possible;
- $\mathcal{PU}$-thld: keeping both privacy and utility as high as possible, above given thresholds.

Using the models that link each LPPM configuration parameter to privacy and utility values, one can reformulate the objectives on privacy and utility as requirements on LPPMs' configuration. Then, in the case where several LPPMs are able to fulfill the objectives, *PULP* selects the most efficient one to achieve the specified objectives. *PULP*'s output is then the recommended $LPPM^*$ and its configuration $\epsilon^*$.

In the following, we first present *PULP*'s ratio-based configuration law $\mathcal{PU}$-ratio (Section IV), and then describe *PULP*'s threshold-based configuration laws $\mathcal{P}$-thld, $\mathcal{U}$-thld and $\mathcal{PU}$-thld (Section V).

## IV. *PULP*'s Ratio-Based Configuration Law

A first configuration law proposed by *PULP* is $\mathcal{PU}$-ratio. Its objectives are as follows:

- $(O1)$ Privacy-to-utility ratio is fixed: $Pr = w_{pr/ut}.Ut$, and
- $(O2)$ Privacy and utility are as high as possible.

For example, when a user specifies $w_{pr/ut} = 0.5$, that means that utility is twice more important for her than privacy. On the contrary, $w_{pr/ut} = 2$ implies that a user thinks preserving the privacy is twice more important than contributing to the LBS accuracy. We now detail the solving procedure, in two steps, to find $LPPM^*$ and its configuration parameter $\epsilon^*$, based on a relative trade-off $w_{pr/ut}$ provided by the user.

The first step consists in finding, for each $LPPM_j$, its configuration $\epsilon_j^*$ that satisfies objectives $(O1)$ and $(O2)$. To achieve the trade-off ratio of $w_{pr/ut}$ between the privacy and utility of objective $(O1)$, we need to find configuration $\epsilon_j^*$ such that $Pr = w_{pr/ut} \cdot Ut$. Applying the model of Eq. (3) and (4), we obtain $\epsilon_j^*$ by solving

$$F_{pr}(\epsilon_j^*) = w_{pr/ut} \cdot F_{ut}(\epsilon_j^*).$$

Due to the complexity of this equation, we do not derive closed-form solution for $\epsilon_j^*$. Instead, we numerically solve it as the minimization problem of the absolute difference between $F_{ut}$ and $F_{pr}$

$$\epsilon_j^* = arg \min_{\epsilon_j} |F_{pr}(\epsilon_j) - w_{pr/ut} \cdot F_{ut}(\epsilon_j)| \quad (5)$$

The convergence of the solution is ensured by the convexity of the function to minimize in Eq. (5). However, when the resulting configuration parameter value does not fall into its legitimate range (which depends on the LPPM), we then consider $LPPM_j$ as an infeasible LPPM to provide the target trade-off between privacy and utility. Thus, this first step results in the set of values $\{\epsilon_j^*$ s.t. Eq. (5) is minimized for a feasible $LPPM_j\}$ that fulfill objective $(O1)$ .

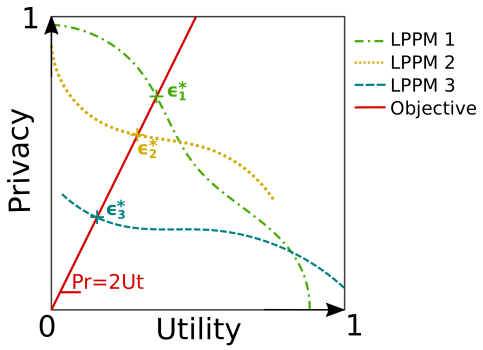



Fig. 6. Illustration of $\mathcal{PU}$-ratio configuration law for three schematic LPPMs with $w_{pr/ut} = 2$.

To better understand this step, we schematically illustrate in Figure 6 three LPPMs's behavioral models. Here, the model equations of each $LPPM_j$ are represented, each point of the curve of a $LPPM_j$ represents one of $LPPM_j$'s configuration. In this example, objective $(O1)$ specifies a privacy twice more important than utility, i.e., $w_{pr/ut} = 2$. Thus, the result of the first step of *PULP* ratio-based Configurator $\mathcal{PU}$-ratio is the set of values of $\{\epsilon_1^*, \epsilon_2^*, \epsilon_3^*\}$, the configuration of each feasible LPPM that fulfills objective $(O1)$.

Among the subset of LPPMs that can achieve the target trade-off with a valid configuration parameter, the $\mathcal{PU}$-ratio Configurator then selects the LPPM that maximizes the weighted sum of the resulting privacy and utility, to keep privacy and utility as high as possible, c.f., objective $(O2)$. Thus, the resulting $LPPM^*$ and its configuration $\epsilon^*$ for a user are

$$LPPM^* = arg \max_j (F_{pr}(\epsilon_j^*) + w_{pr/ut} \cdot F_{ut}(\epsilon_j^*)). \quad (6)$$

From the example in Figure 6, the LPPM that best achieves objective $(O2)$ which aims at maximizing privacy and utility is the one crossing the objective $(O1)$ line at the upper point. Here, *PULP* ratio-based Configurator $\mathcal{PU}$-ratio returns $< LPPM_1, \epsilon_1^* >$.

## V. *PULP*'s Threshold-Based Configuration Laws

In addition to the ratio-based configuration law, *PULP* provides three threshold-based laws, namely $\mathcal{P}$-thld, $\mathcal{U}$-thld and $\mathcal{PU}$-thld that are presented in the following section.

### A. $\mathcal{P}$-thld Law: Privacy Above a Minimum Threshold

Another possible set of objectives is to guarantee a minimum privacy level while keeping utility as high as possible:

- $(O3)$ Privacy is higher or equal to a minimum privacy value: $Pr \geq Pr_{min}$, and
- $(O4)$ Utility $Ut$ is as high as possible.

For each LPPM, we define $\epsilon_{pr}$ as the configuration parameter satisfying the equation $F_{pr}(\epsilon_{pr}) = Pr_{min}$. Using eq. (3), we can express $\epsilon_{pr}$ as

$$\epsilon_{pr} = \exp\left(\frac{1}{b_{pr}} tan\left(\frac{Pr_{min} - d_{pr}}{a_{pr}}\right) + c_{pr}\right). \quad (7)$$

Due to the trade-off between utility and privacy, the higher the utility is, the lower the privacy will be. Then for objective $(O4)$, utility can be increased until the privacy reaches its lower bound specified in objective $(O3)$. Thus for each $LPPM_j$, the configuration $\epsilon_j^*$ that achieves objectives $(O3)$ and $(O4)$ for that LPPM is:

$$\epsilon_j^* = \epsilon_{pr} \quad for \quad LPPM_j \quad (8)$$

Finally, as the privacy level is achieved for each combination $< LPPM_j, \epsilon_j^* >$, the overall $LPPM^*$ is the one that maximizes utility:

$$LPPM^* = arg\max_j(F_{ut}(\epsilon_j^*)). \quad (9)$$

### B. $\mathcal{U}$-thld Law: Utility Above a Minimum Threshold

Similarly, one can set the constraint on a minimal level of utility while keeping privacy as high as possible:

- $(O5)$ Utility is not below a given minimum utility threshold $Ut \geq Ut_{min}$,
- $(O6)$ With the highest data privacy $Pr$.

For each LPPM, we define $\epsilon_{ut}$ such that $F_{ut}(\epsilon_{ut}) = Ut_{min}$:

$$\epsilon_{ut} = \exp\left(\frac{1}{b_{ut}} tan\left(\frac{Ut_{min} - d_{ut}}{a_{ut}}\right) + c_{ut}\right). \quad (10)$$

Here, objective $(O6)$ is ensured given objective $(O5)$ iff $\epsilon$ converges to the highest value that guarantees $F_{ut}(\epsilon) \geq Ut_{min}$, due to the trade-off between the two. Thus, the configuration $\epsilon_j^*$ for $LPPM_j$ is

$$\epsilon_j^* = \epsilon_{ut} \quad for \quad LPPM_j \quad (11)$$

In order to elect the protection mechanism $LPPM^*$, we compare the values of privacy of the obfuscated data and choose the following:

$$LPPM^* = arg\max_j(F_{pr}(\epsilon_j^*)). \quad (12)$$

### C. $\mathcal{PU}$-thld: Privacy and Utility Above Minimum Thresholds

This configuration law aims at guaranteeing that the level of privacy *and* the level of utility of the obfuscated data are above given thresholds:

- $(O7)$ Privacy is higher than or equal to a given minimum threshold: $Pr \geq Pr_{min}$, and
- $(O8)$ Utility is higher than or equal to a given minimum threshold: $Ut \geq Ut_{min}$.

The trade-off between privacy and utility described in Section II shows that the utility function $F_{ut}(\epsilon)$ and privacy function $F_{pr}(\epsilon)$ have opposite directions of variation, both functions being monotonous. Let us first make the hypothesis that privacy function is decreasing and utility function increasing (which is the case for GEO-I for example). Then the objective $(O7)$ of a threshold value on privacy $F_{pr}(\epsilon) \geq Pr_{min}$ can be written as:

$$\epsilon \leq \epsilon_{pr}$$

And $(O8)$ $F_{ut}(\epsilon) \geq Ut_{min}$ as

$$\epsilon \geq \epsilon_{ut}$$

Then the two objectives can be combined in one condition regarding to the value of the configuration parameter:

$$\epsilon_{ut} \leq \epsilon \leq \epsilon_{pr}$$

Then for users for whom $\epsilon_{ut} \leq \epsilon_{pr}$, all the parameters in the range $[\epsilon_{ut}; \epsilon_{pr}]$ satisfy the objectives. *PULP* returns the mean value of the range:

$$\epsilon^* = \frac{\epsilon_{pr} + \epsilon_{ut}}{2}. \quad (13)$$

Otherwise if $\epsilon_{ut} \geq \epsilon_{pr}$, there is no solution to both objectives $(O7)$ and $(O8)$ for this particular combination of LPPM and user.

Similarly, if the utility function on a LPPM decreases while the privacy function increases (as for PROMESSE for instance), the objectives can be written as:

$$\epsilon_{pr} \leq \epsilon \leq \epsilon_{ut}$$

and the condition of the existence of a solution is thus $\epsilon_{ut} \geq \epsilon_{pr}$. However, in the case where a solution exists, *PULP* returns the same solution of eq. (13).

Once the configuration $\epsilon_j^*$ of each $LPPM\ j$ is found, if any, the protection mechanism $LPPM^*$ is selected as follows. The values of privacy and utility are compared by computing their weighted sum, after using each $LPPM_j$ in its previously calculated configuration :

$$LPPM^* = arg\max_j(Pr_{min} \cdot F_{pr}(\epsilon_j^*) + Ut_{min} \cdot F_{ut}(\epsilon_j^*)). \quad (14)$$

## VI. *PULP* Evaluation

*PULP*'s validation is carried out in three strokes: first, an analysis of the modeler with an emphasis on the accuracy of the derived models and on their robustness; second, the configurator evaluation that illustrates its effectiveness in choosing a suitable LPPM to achieve different user's objectives; and eventually a comparison with the state of the art. Prior to those core results, the experimental setup is depicted.

### A. Experimental Setup

For the experimental validation of *PULP*, two different machines were used. The profiler was executed on a machine running Ubuntu 14.04 and equipped with 50Gb of RAM and 12 cores cloaked at 1,2 GHz. We run the profiler using the 30-days datasets. The modeler and the configurator use Matlab R2016b on a Ubuntu 14.04 and equipped with 3.7Gb of RAM and 4 cores cloaked at 2,5 GHz.

The number of configuration of each LPPM to be tested by the profiler has been set at first to 17 for GEO-I and 10 for PROMESSE, corresponding to 4 values per decade of the

definition range, uniformly distributed. The modeler searches for each user's model by fitting the experimental data using *fminunc*, and $\mathcal{PU}$-ratio configuration law uses *min* (both are Matlab functions).

The metrics of privacy and utility used have first been parametrized to correspond to our datasets collected in dense-cities. For measuring privacy, we consider POIs of a maximum diameter of $d = 200$ m and a minimal stay time of $t = 15$ min. In order to calculate intersections between sets of POIs, we consider that two POIs matched if their centroids are within $d_{max} = 100$ m from each other. Google's S2 geometry library [27] is used for cell extraction when computing utility. The size of the cells is highly related to the nature of the LBS. Indeed, a navigation application needs a spatial accuracy at a really fine level while a recommendation system needs accuracy at a neighborhood level. We consider cells at level 15, which corresponds to areas having the size of around 300 meters (city block or neighborhood).

As initial values for the models parameters of eq. (3) and (4), we choose the following:

a The metric amplitude. The arctan function is varying between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$. Our metrics have been defined to vary between 0 and 1. Moreover, we expect GEO-I utility function and PROMESSE privacy function to be increasing and GEO-I privacy function and PROMESSE utility function to be decreasing. Consequently, we set $a = \frac{1}{\pi}$ for GEO-I utility and PROMESSE privacy and $a = -\frac{1}{\pi}$ for GEO-I privacy and PROMESSE utility.

b The transition speed between the saturations. It shall be non-null and positive, the value $b = 1$ was chosen.

c The offset - configuration parameter value. This parameter should be the default value of the configuration parameter defined by the authors of the LPPM: $c = ln(10^{-2})$ for GEO-I and $c = ln(200)$ for PROMESSE.

d The offset - metric value. As metrics vary between 0 and 1 (or 1 and 0), the offset was set to $d = 0.5$.

When considering a new LPPM, all needed for configuring the modeler is a standard value of its parameter (update of $c$ initial value only).

### B. Evaluation of PULP Modeler

This section evaluates the ability of *PULP* modeler to capture the behavior of privacy and utility metrics when the LPPM configuration varies. We focus particularly on the accuracy of the modeling, its robustness regarding the amount of input data and its adaptability to model any privacy and utility metric.

*1) Modeler's theoretic guarantees:* The working hypothesis regarding LPPMs are their configuration by a single parameter, influencing both privacy and utility metrics. Thus, the variation of those metrics will be stable or monotonous (at least in average in case of stochastic LPPM), varying at most between 0 and 1 (by definition of the metrics), with eventual saturated levels for high and low values of the parameter. The arctan shape of the model enables to capture this behavior. The accuracy of the modeling is thus given by the relevance of the parameters of the model, output by the *fminunc* function. The tolerance for stopping the iterative search for the best parameters have been set to $10^{-6}$, nonetheless with a maximum number of iteration set to 400.

*2) Modeler's performance:* As a preliminary analysis, one can take a look at Figure 5. Experimental data (continuous line with circles for GEO-I and stars for PROMESSE) is compared to their model (dotted lines) for both utility and privacy metrics. The closer the model curves are to the real data, the better the model fitting is for that user. For our CABS user example of Figure 5, the modeler accuracy is good for GEO-I and PROMESSE utility and PROMESSE privacy; however for GEO-I privacy the modeler is less accurate but still relevant as it avoid overfitting the experimental data.

In order to ensure that *PULP* modeler is accurate *for every user*, we compute the variance of the fitting error (difference between experimental data and model prediction), which is a relevant indicator for non-linear modeling. Results are shown in Figure 7, in form of a cumulative distribution function where low values of error variance show a high accuracy of the modeling. For all metrics and all LPPMs, the median modeling accuracy is less than $5.10^{-2}$, which, when put into the perspective of our metrics varying between 0 and 1, is a really good fitting. PROMESSE privacy is by far the better modeled data, 95% of user have an accuracy of less than $10^{-4}$. This can be easily explained as many users have a privacy of 100% no matter the configuration of PROMESSE, as it is the case for the user illustrated in Figure 5. From Figure 7, one can also notice that the modeler has still a high accuracy when dealing with outliers, as the $99^{th}$ percentile of the error variance is smaller than $2.10^{-1}$ for all metrics and all LPPMs.

*3) Modeler's robustness and adaptability:* In the next paragraphs, we comment on the robustness of the modeler regarding both its sensibility to input data and its adaptation to metrics parametrization. Illustrative figures of the kind of Figure 7 are eluded due to space limitation, the results are directly commented in the text.

First, we study the impact of the amount of profile data needed for accurate modeling. We vary the number of values of $\epsilon$ taken for the profiling phase, from 1 value per decade up to 4. Results show that the modeling accuracy varies depending on the LPPM. In all cases, the more data are used, the better the modeling is. However the improvement is negligible when modeling GEO-I, which makes us recommend to use few

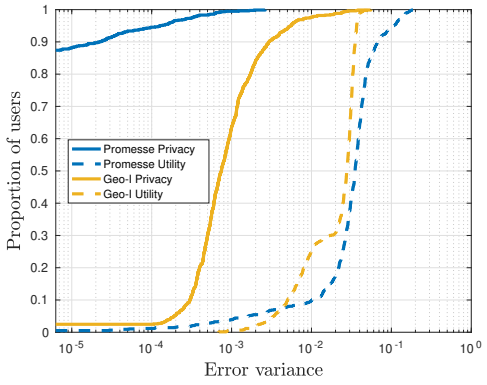



Fig. 7. *PULP* modeler accuracy. Cumulative distribution function (cdf) among all users.

TABLE I
*PULP* OUTPUT FOR SELECTED USERS

| Configuration Law | $\mathcal{PU}$-ratio | | $\mathcal{PU}$-thld | | $\mathcal{P}$-thld | | $\mathcal{U}$-thld | |
|---|---|---|---|---|---|---|---|---|
| Objectives | $w_{pr/ut} = 2$ | | $Pr_{min} = 0.6$ $Ut_{min} = 0.7$ | | $Pr_{min} = 0.7$ | | $Ut_{min} = 0.5$ | |
| *PULP* Output | LPPM* | $\epsilon^*$ | LPPM* | $\epsilon^*$ | LPPM* | $\epsilon^*$ | LPPM* | $\epsilon^*$ |
| User 1 | PROMESSE | 694 | GEO-I | 0.014 | PROMESSE | 69 | GEO-I | 0.004 |
| User 2 | GEO-I | 0.001 | PROMESSE | 244 | PROMESSE | 197 | GEO-I | 0.0034 |
| User 3 | PROMESSE | 173 | NaN | NaN | GEO-I | 0.0097 | GEO-I | 0.0074 |

experiments for the profiling phase in order to limit computing. When modeling PROMESSE, only 4 values per decade enable to properly capture the behavior of users.

Metrics described in Section II are parametrized. The privacy metric depends on the diameter and duration of a POI as well as on the maximum distance between two POIs to consider they match. As for the utility metric, one can vary the size of the cells that discretize the map. When varying these parameters, the metrics reflect several notions of privacy and utility. To ensure the performances of the modeler even with these other notions of privacy and utility, we varied the four metrics' parameters and computed again the modeler accuracy. In a general way, the modeler is able to keep a good accuracy: around $10^{-3}$ for the median value and $10^{-2}$ for the $99^{th}$ percentile (excluding extreme cases).

We now detail the impact of each metric parameter on modeling performance.We varied the duration of a POI between 5 and 120 minutes. For PROMESSE, the longer the POI, the better the modeling. For GEO-I however, medium duration POIs (around 15 to 30 minutes) are well modeled while extreme ones have error variance close to $10^{-1}$. When looking at the impact of POI diameter (from 100m to 1000m) on the modeling accuracy, we found none on GEO-I (all metrics are well modeled), while for PROMESSE the smaller the POI is, the better the modeling is. As for the maximum distance between two matched POIs (ranging from 25% of the POI diameter to 250%), we obtained similar results: no impact for GEO-I modeling and the smaller the distance is, the better the modeling is for PROMESSE. When looking at the size of the cells in the utility metric computation, we found that the larger the cells are, the better PROMESSE behavior on users' traces is captured. However, GEO-I modeling is more accurate for large or small cells, and a sightly worst for medium-size cells.

### C. Evaluation of PULP Configurator

We now analyze the configurator's ability to fulfill users' objectives. To do so, we ran the four versions of *PULP* configuration laws, each of them with several objectives.

*PULP* outputs for a set of users (selected to show diversity) and a few objectives can be found on Table I. Results show that for each user, a LPPM is recommended with a configuration value, except when no LPPM can fulfill the objectives, which is the case of user 3 with $\mathcal{PU}$-thld objective. As a preliminary analysis, we can see that users have different recommendations even when having the same objectives. Moreover, a single user gets various recommendation depending of her objectives.

The next sections and Figures 8 to 11 detail the results for each law, by looking at four indicators: the LPPM selected for each user (subfigure (a)) and its associated configuration parameter (subfigures (b) and (c)), the privacy and utility of users' data when obfuscating with *PULP* recommendation (subfigures (d) and (e)), and the corresponding trade-off between privacy and utility (subfigure (f)).

*1) Evaluation of $\mathcal{PU}$-ratio Configuration Law:* In this variant of the configurator, the objective is to achieve a given trade-off between privacy and utility. For its evaluation, we run *PULP* on all users with various objective ratio $w_{pr/ut}$ ranging from 0.5 (utility is twice more important than privacy) to 3 (privacy is three times more important than utility). Results are shown in Figure 8. We computed the actual privacy to utility ratio after applying the LPPM selected with its right configuration. Results illustrated in Figure 8 (f) show that at least 95% of the users have a resulting ratio in a range of +/- 1% of user specified values.

From Figure 8 (a) we can see that all users ended with a recommended LPPM. For a given objective, the LPPM chosen by *PULP* varies depending on the user, and the distribution changes according to the objective, meaning that the adequate LPPM of a single user may vary depending on the objective. There is no a priori relation between the objective $w_{pr/ut}$ and the distribution of selected LPPM.

When analyzing *PULP* choice of LPPM configuration parameters from Figure 8 (b) and (c), we make two observations: (i) users need different configurations to fulfill the same objective and (ii) different objectives lead to various configurations distribution. When looking at users for whom GEO-I is chosen, the higher the objective ratio is, the lower the recommended $\epsilon$ value is. Whereas for users with PROMESSE recommended, the higher the objective ratio is, the larger the suitable $\epsilon$ value is and the higher diversity there is in the recommended value. For instance with $w_{pr/ut} = 2$, users have $\epsilon$ from 100m to 2km.

When using the appropriate LPPM configured in a suitable way, users can maintain privacy and utility levels that jointly respect to the objective trade-off. In terms of absolute values, when looking at the utility, one can notice that most users have the same level of utility (Figure 8 (e)). The lower the objective ratio is (i.e. the more utility matters), the higher the utility is and the more diversity there is in the utility values. For privacy, the same trend is observed: most users have the same privacy but the lower the objective ratio is, the more there is diversity in the privacy values (see Figure 8 (d)).

*With an objective expressed as a trade-off between privacy and utility, PULP enables to find a suitable LPPM for all users and guarantee a high utility and privacy to almost all of them.*

*2) Evaluation of $\mathcal{PU}$-thld Configuration Law:* In this section, we evaluate *PULP* $\mathcal{PU}$-thld configuration law aiming at achieving privacy and utility at levels higher than some minimal thresholds. The evaluation, reported in Figure 9, has been carried out with five couples of objectives.

First, it is important to notice that some users do not have any LPPM recommended, as can be seen in Figure 9 (a). High utility constraints seem to hamper the feasibility to recommend suitable LPPMs. Recommendations range from less than 10% of all users ($Pr_{min} = 0.5, Ut_{min} = 0.9$) to more than 95% ($Pr_{min} = 0.3, Ut_{min} = 0.5$). Most of the recommendations are GEO-I. The higher the utility constraint is, the more GEO-I is recommended. When looking at values of the LPPM configuration parameter, for GEO-I the general trend is that $\epsilon$ is lower when privacy constraint in high, except in the extreme case where $Ut_{min} = 0.9$. For PROMESSE, the higher the utility constraint, the smaller $\epsilon$.

The privacy criteria is always satisfied, and almost no user gets the limit privacy $Pr_{min}$, see Figure 9 (d). However when the utility constraint is high, most users tend to have a privacy close to its bound. All users have their utility criteria fulfilled (see Figure 9 (e)). Utility of most users is really high: 80% of them have a utility above 0.8 (0.6 for $Pr_{min} = 0.9, Ut_{min} = 0.4$). As for the privacy to utility ratio, within a set of objectives most users (70-90%) have the same privacy to utility ratio (Figure 9 (f)).

*PULP is not able to find a suitable LPPM for all users using this objective formulation; however, when it can, the privacy and utility are most of the time way above the minimum values required.*

*3) Evaluation of $\mathcal{P}$-thld Configuration Law:* Here the objective is to guarantee that the privacy is above a given level. Results are given in Figure 10. *PULP* can recommend a suitable LPPM and fulfill all users' objectives considered here. Around 10 to 20% of users can even achieve higher privacy levels than requested threshold (even 80% for the objective $Pr_{min} = 0.9$), see Figure 10 (d). These proportions correspond to users to which PROMESSE is recommended, see correspondence with Figure 10 (a).

For those PROMESSE users, the utility is quite low but for the other 80% of users, utility is more than 0.5 (Figure 10 (e)). Moreover, the higher the privacy limit is, the lower the utility is. Looking at the privacy to utility ratio (Figure 10 (f)), for 80% of users (those with GEO-I recommended), the lower the privacy limit is the higher the ratio is (allowing more utility to the data) and all users have the same privacy to utility ratio. However, for 20% of them (PROMESSE users) the ratio is always the same no matter the objective.

As for the parameter values, for users with GEO-I recommended, the higher the objective is, the smaller $\epsilon$ is. Users with PROMESSE recommended seem to always have the same $\epsilon$ recommended no matter the value of the objective, except for $Pr_{min} = 0.9$ where $\epsilon$ is at its upper bound for most users.

*When PULP guarantees a minimal privacy level, two distinctive types of users are observed. Some have* GEO-I *recommended, a limited privacy and a high utility; while the others use* PROMESSE *with a high privacy but a low utility.*

*4) Evaluation of $\mathcal{U}$-thld Configuration Law:* The results of the configuration law guaranteeing a minimum level of utility are illustrated in Figure 11. They show similar patterns than those of $\mathcal{P}$-thld configuration law.

All users had a LPPM recommended, and the general trend is that the lower the utility limit is, the more PROMESSE is recommended. Hence, PROMESSE tends to protect better than GEO-I but result into lower utility, see Figure 11 (a).

All users have exactly the minimum utility they wanted, no matter the value of the limit (see Figure 11 (e)). The lower the utility limit is, the higher the privacy is. Most users (almost 70%) have good privacy levels, i.e. more than 0.7, except with $Ut_{min} = 0.9$ (see Figure 11 (d)). Therefore, the lower the utility limit is, the higher the ratio is, allowing more privacy preservation in the data. Within a set of objectives, most users have the same privacy to utility ratio (Figure 11 (f)).

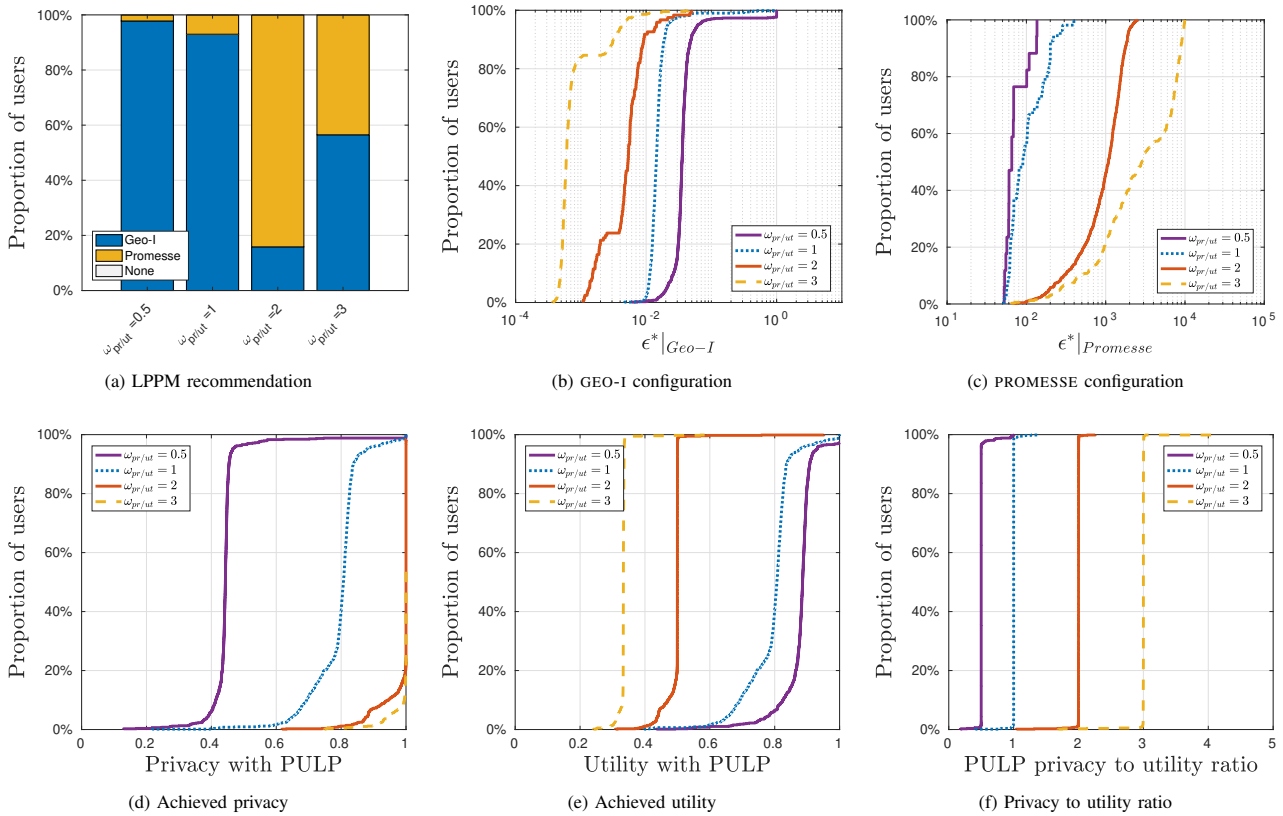
For users with GEO-I recommended, the higher the objective is, the higher $\epsilon$ is. However, for users with PROMESSE recommended, the higher the limit is, the smaller $\epsilon$ is.

*In this objective formulation, PULP always sets utility to its minimum value, ensuring a good privacy to users especially for those with* PROMESSE *recommended.*

### *D. Comparison with Competitor*

As *PULP* works with few profiling experiments, its execution time is significantly shorter compared to state of the art. Indeed, all LPPMs configuration mechanisms that we are aware of use greedy processes that need to run many experiments in order to converge to a suitable configuration (if ever they converge). We compare our framework *PULP* to the closest work from the state of the art, the configurator ALP from [26]. ALP is a framework that iteratively look for a LPPM configuration that satisfies high level objectives objectives such as *maximizing privacy and utility*. We consider only one LPPM in *PULP* that is GEO-I and set our objective to $w_{pr/ut} = 1$ to be as close as possible to the ALP working conditions. The execution time of *PULP* in theses conditions is of the order of the minute for GEOLIFE dataset while ALP requires around ten hours to converge. This makes a difference of 3 orders of magnitude. The execution time of *PULP* is barely all spent on the profiling phase. Indeed modeler and configurator execution time are of few milliseconds. This enables a user to change her objective and easily found again the new adequate LPPM and its configuration.

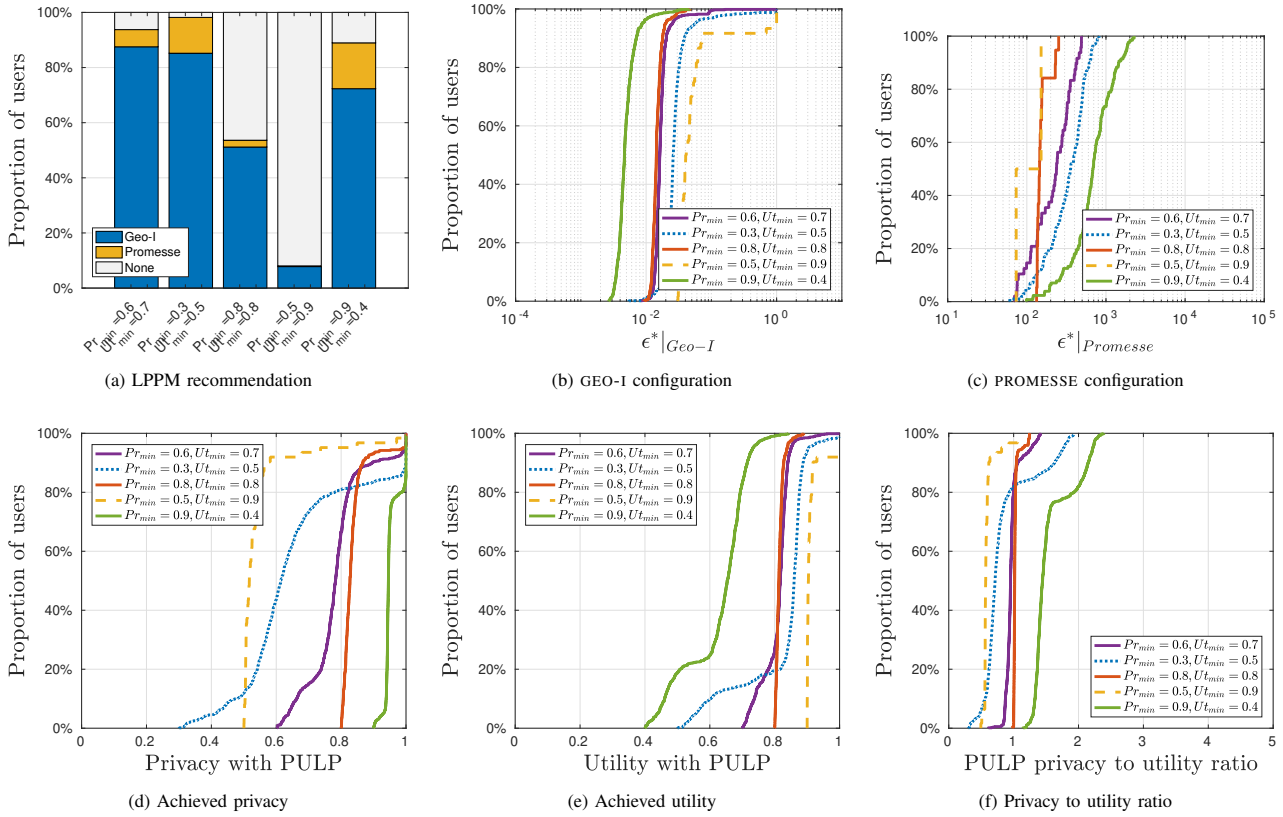
ALP only considers the configuration challenge and does not allow to choose between several LPPM. To still compare the accuracy of *PULP* regarding to the state of the art, the focus will be done on the users' privacy and utility preservation after using the frameworks. While the objective given to ALP is to maximize both utility and privacy (no preference is given to one or the other), the ratio between the two after running ALP is almost always greater than 1, meaning that more importance is given to privacy than to utility [26]. With *PULP*, the ratio is almost always 1, see Figure 8 (f). Moreover, with *PULP*, 80% of the users have a utility and privacy higher than 0.7, while with ALP 90% of the users have a privacy higher than 0.8, while 80% of them have their utility only between 0.4 and 0.7 [26]. The low utility with ALP comes from a small configuration parameter (less that $5.10^{-2}$ for 70% of the users). Hence, the objective are more evenly treated when using *PULP*, and enable to achieve better utility.




(a) LPPM recommendation (b) GEO-I configuration (c) PROMESSE configuration

(d) Achieved privacy (e) Achieved utility (f) Privacy to utility ratio

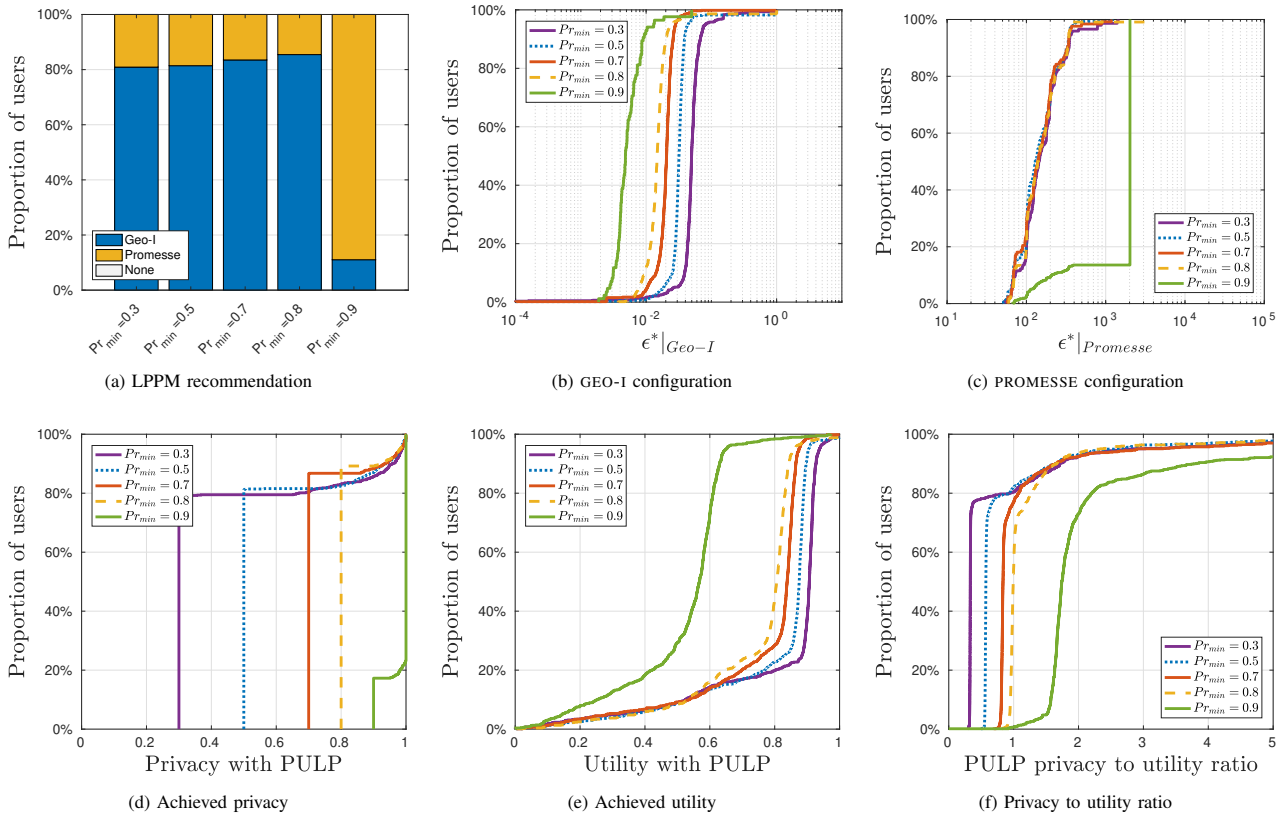
Fig. 8. $\mathcal{PU}$-ratio configuration law evaluation. (a) Recommended LPPM and its configuration (b) for GEO-I, (c) for PROMESSE. Achieved (d) level of privacy and (e) utility when users are protected according to *PULP* recommendations, and the corresponding (f) privacy to utility ratio. Four objective ratio $w_{pr/ut}$ are illustrated: 0.5, 1, 2 and 3.




(a) LPPM recommendation (b) GEO-I configuration (c) PROMESSE configuration

(d) Achieved privacy (e) Achieved utility (f) Privacy to utility ratio

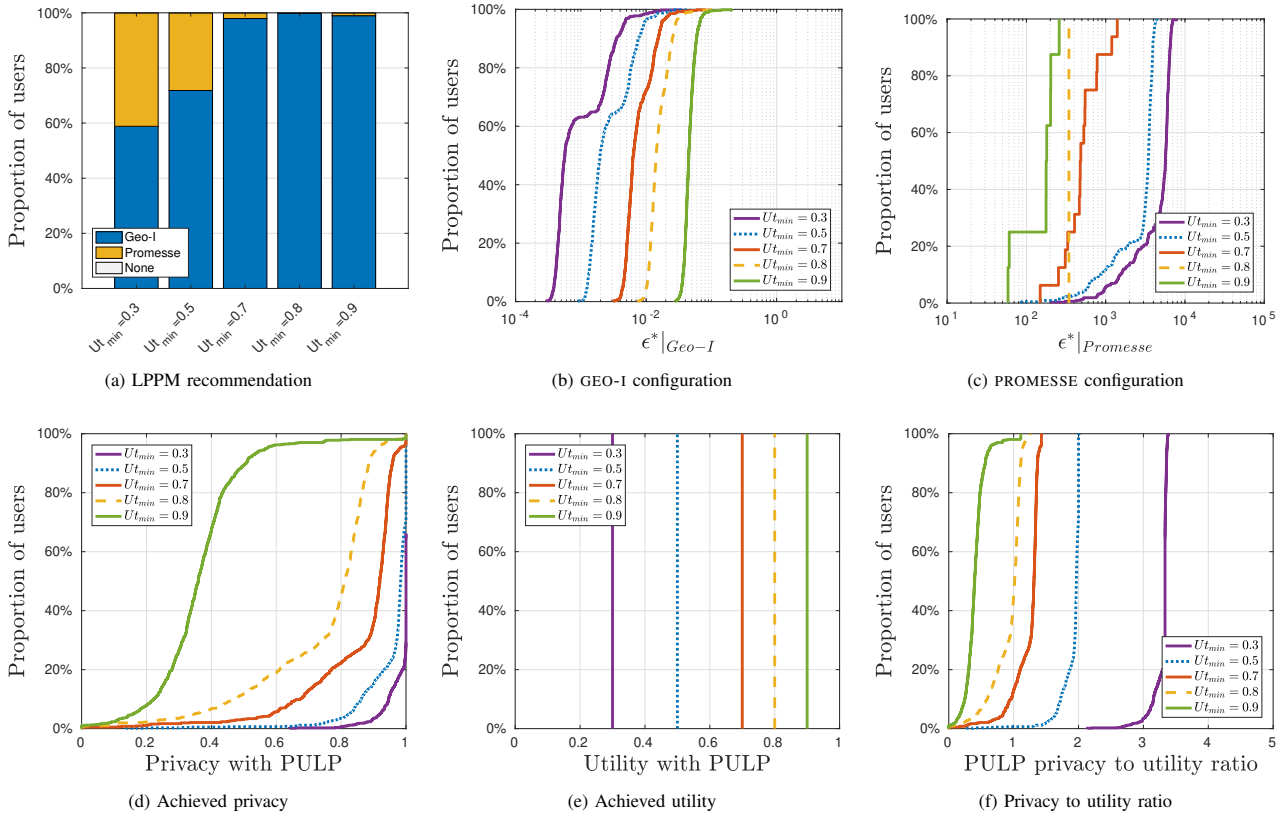
Fig. 9. $\mathcal{PU}$-thld configuration law evaluation. (a) Recommended LPPM and its configuration (b) for GEO-I, (c) for PROMESSE. Achieved (d) level of privacy and (e) utility when users are protected according to *PULP* recommendations, and the corresponding (f) privacy to utility ratio. Five objectives couple of constraints on privacy and utility are illustrated.




(a) LPPM recommendation (b) GEO-I configuration (c) PROMESSE configuration

(d) Achieved privacy (e) Achieved utility (f) Privacy to utility ratio

Fig. 10. $\mathcal{P}$-thld configuration law evaluation. (a) Recommended LPPM and its configuration (b) for GEO-I, (c) for PROMESSE. Achieved (d) level of privacy and (e) utility when users are protected according to *PULP* recommendations, and the corresponding (f) privacy to utility ratio. Five objectives constraints on privacy $Pr_{min}$ are illustrated: 0.3, 0.5, 0.7, 0.8, 0.9.




(a) LPPM recommendation (b) GEO-I configuration (c) PROMESSE configuration

(d) Achieved privacy (e) Achieved utility (f) Privacy to utility ratio

Fig. 11. $\mathcal{U}$-thld configuration law evaluation. (a) Recommended LPPM and its configuration (b) for GEO-I, (c) for PROMESSE. Achieved (d) level of privacy and (e) utility when users are protected according to *PULP* recommendations, and the corresponding (f) privacy to utility ratio. Five objectives constraints on utility $Ut_{min}$ are illustrated: 0.3, 0.5, 0.7, 0.8, 0.9.

## VII. Related Work

### A. Location Privacy Protection Mechanisms

LPPMs attempt to enhance location privacy of users willing to interact with location-based services. Although our work is not concerned in designing a new LPPM, we quickly present here some prominent privacy protection schemes. Generally speaking, LPPMs can be classified according to the privacy guarantees they offer to the users. A well-known privacy guarantee is $k$-anonymity [29], which states that a user is $k$-anonymous if she is hidden among $k-1$ other users sharing similar properties. In the context of location privacy, it means that, instead of reporting their exact location, users report to be inside cloaking areas containing at least $k$ users. This has been successfully implemented using a trusted third party to compute cloaking areas (see for instance [21]) as well as in distributed systems relying on peer-to-peer communication between users (e.g., PRIVÉ [16]).

Another popular privacy guarantee is differential privacy [9], which ensures that the presence or absence of a single user from a dataset should not significantly affect the outcome of any query on this dataset. Geo-Indistinguishability [3] is an extension of differential privacy designed specifically to be used on mobility traces. Differential privacy is guaranteed by adding noise, drawn from a two-dimensional Laplace distribution. Further version of Geo-I have been developed in [5] and [22].

Eventually, some application-specific protection mechanisms have been developed; such as private decision making for smart cities [2], private classification of human activities [15] or private task recommendation for crowdsourcing [28].

### B. LPPM Configuration

What makes LPPMs difficult to use in practice is that they rely on a set of configuration parameters. For instance, the $\epsilon$ parameter of differentially-private protection mechanisms is a sensitive parameter that has a great impact on the resulting data privacy and utility. With the inherent trade-off between privacy and utility, it is a difficult task to set LPPM configuration parameters to an appropriate value.

In [18], the author showed that defeating a well-performing privacy attack would require adding so much noise that it would make the resulting data unusable by any LBS, and hence useless. This means that we do have to consider the right balance between privacy and utility in order to satisfy a system designer objectives.

A few works have been proposed to help a user choose a LPPM configuration that fits his actual needs. Agir et. al proposed an adaptive mechanism that dynamically computes the size of the cloaking area the user will be hidden within [1]. However, their privacy estimation routine has a complexity of $O(L^2)$, $L$ being the maximum number of locations that a cloaked area can be formed of. Chatzikokolakis et. al introduced an extension of GEO-I that uses contextual information to adapt the effective privacy level to the density of the area [8]. However, this approach still requires some parametrization from the user side and is not objective-driven. Primault et al. presented *ALP*, a system that configures a LPPM depending users objectives [26]. This solution relies on a greedy approach that iteratively evaluates the privacy and utility for refining configuration parameters. Their evaluation of the metrics has a complexity varying between $O(L)$ and $O(L^2)$. Moreover, the convergence is not ensured and consequently there is no guarantee that the objectives are actually met. *PULP* is first presented in [7]. In this paper, we extend this previous work by developing three more configuration laws enabling a dataset owner to specify several combinations of privacy and utility objectives, and we present extensive experimental evaluations for both the modeling and the configuration parts of *PULP*.

## VIII. Conclusion

In this paper we present *PULP*, a framework that ensures users' objectives regarding privacy and utility for mobility databases by automatically choosing and configuring LPPMs. Our notion of privacy relies on the hiding of users' points of interest, and the utility of the services is measured by looking at the spatial proximity of obfuscated data to the original ones. *PULP* realizes an in-depth analysis of the considered LPPMs applied at a user scale in order to capture the formal relationship between the configuration parameters of the LPPMs and both privacy and utility metrics. Then *PULP* leverages the models derived to identify the adequate LPPM and its configuration that enables to fulfill the objectives. The considered objectives aim at maximizing privacy and utility with various constraints regarding minimal levels or the ratio between the two metrics.

We illustrated the ability of our system *PULP* to efficiently protect a user while keeping utility of her service using two LPPMs from the state of the art: GEO-I and PROMESSE. Evaluation has been done for several objectives and using data from four real mobility datasets of 770 users in total. *PULP* can accurately model the behavior of LPPMs on individual users and thus successfully achieve privacy and utility objectives at the same time in an automated way. Moreover, when comparing with the state of the art, we proved *PULP* to be 3 order of magnitude faster (minutes versus hours) and more robust to achieve user specified privacy and utility objectives.

Future directions include further researches of *PULP* ability to work with more LPPMs, e.g. ones that have higher number configuration parameters. We also aim at exploring more metrics, corresponding to extensions of the notions of users' privacy protection and service utility. In particular, we want to include the temporal aspect of the mobility data.

## IX. Acknowledgment

This work was partly supported by the SIBIL-Lab project funded the the French National Research Agency (ANR), under grant number ANR-17-LCV2-0014.

## References

[1] B. Agir, T. G. Papaioannou, R. Narendula, K. Aberer, and J.-P. Hubaux. User-side adaptive protection of location privacy in participatory sensing. *GeoInformatica*, 18(1):165–191, 2014.

[2] A. Alabdulatif, I. Khalil, H. Kumarage, A. Y. Zomaya, and X. Yi. Privacy-preserving anomaly detection in the cloud for quality assured decision-making in smart cities. *Journal of Parallel and Distributed Computing*, 2018.

[3] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. Geo-indistinguishability: Differential Privacy for Location-based Systems. In *CCS*, pages 901–914, 2013.

[4] I. Bilogrevic, K. Huguenin, M. Jadliwala, F. Lopez, J.-P. Hubaux, P. Ginzboorg, and V. Niemi. Inferring Social Ties in Academic Networks Using Short-Range Wireless Communications. *Wpes*, 2013.

[5] N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi. Optimal geo-indistinguishable mechanisms for location privacy. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 251–262. ACM, 2014.

[6] A. Boutet, S. Ben Mokhtar, L. Bouzouina, P. Bonnel, O. Brette, L. Brunie, M. Cunche, S. D'alu, V. Primault, P. Raveneau, H. Rivano, and R. Stanica. PRIVA'MOV: Analysing Human Mobility Through Multi-Sensor Datasets. In *NetMob*, 2017.

[7] S. Cerf, V. Primault, A. Boutet, S. Ben Mokhtar, R. Birke, S. Bouchenak, L. Y. Chen, N. Marchand, and B. Robu. PULP: Achieving Privacy and Utility Trade-off in User Mobility Data. In *SRDS*, 2017.

[8] K. Chatzikokolakis, C. Palamidessi, and M. Stronati. Constructing elastic distinguishability metrics for location privacy. In *PETS*, volume 2015, pages 156–170, 2015.

[9] C. Dwork. Differential Privacy. In *Automata, Languages and Programming*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin Heidelberg, 2006.

[10] L. Franceschi-Bicchierai. Redditor cracks anonymous data trove to pinpoint muslim cab drivers. http://mashable.com/2015/01/28/redditor-muslim-cab-drivers/, Jan. 2015.

[11] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez. De-anonymization Attack on Geolocated Data. *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*.

[12] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez. Show Me How You Move and I Will Tell You Who You Are. *Transactions on Data Privacy*, 4(2):103–126, Aug. 2011.

[13] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez. Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, page 3. ACM, 2012.

[14] GDPR. reform of eu data protection rules, May 2018.

[15] Z. Gheid, Y. Challal, X. Yi, and A. Derhab. Efficient and privacy-aware multi-party classification protocol for human activity recognition. *Journal of Network and Computer Applications*, 98:84–96, 2017.

[16] G. Ghinita, P. Kalnis, and S. Skiadopoulos. PRIVE: Anonymous Location-based Queries in Distributed Mobile Systems. In *WWW*, pages 371–380, 2007.

[17] N. Kiukkonen, B. J., O. Dousse, D. Gatica-Perez, and L. J. Towards rich mobile phone datasets: Lausanne data collection campaign. In *ICPS*, 2010.

[18] J. Krumm. Inference Attacks on Location Tracks. In *PerCom*, 2007.

[19] M. Maouche, S. Ben Mokhtar, and S. Bouchenak. Ap-attack: A novel user re-identification attack on mobility datasets. In *MobiQuitous*, 2017.

[20] K. Micinski, P. Phelps, and J. S. Foster. An Empirical Study of Location Truncation on Android. *Most'13*, 2013.

[21] H. Ngo and J. Kim. Location privacy via differential private perturbation of cloaking area. In *Computer Security Foundations Symposium (CSF), 2015 IEEE 28th*, pages 63–74. IEEE, 2015.

[22] S. Oya, C. Troncoso, and F. Pérez-González. Is geo-indistinguishability what you are looking for? In *Proceedings of the 2017 on Workshop on Privacy in the Electronic Society*, pages 137–140. ACM, 2017.

[23] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. CRAWDAD dataset epfl/mobility (v. 2009-02-24). Downloaded from http://crawdad.org/epfl/mobility/20090224, Feb. 2009.

[24] V. Primault, S. Ben Mokhtar, C. Lauradoux, and L. Brunie. Differentially Private Location Privacy in Practice. In *MoST'14*, 2014.

[25] V. Primault, S. Ben Mokhtar, C. Lauradoux, and L. Brunie. Time distortion anonymization for the publication of mobility data with high utility. In *TrustCom*, pages 539–546, 2015.

[26] V. Primault, A. Boutet, S. Ben Mokhtar, and L. Brunie. Adaptive location privacy with ALP. In *SRDS*, 2016.

[27] S2, a spherical geometry library. Available online at https://github.com/google/s2-geometry-library-java.

[28] J. Shu, X. Jia, K. YANG, and H. Wang. Privacy-preserving task recommendation services for crowdsourcing. *IEEE Transactions on Services Computing*, 2018.

[29] L. Sweeney. k-Anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[30] theguardian.com. Fitness tracking app strava gives away location of secret us army bases, January 2018.

[31] G. Yavaş, D. Katsaros, Ö. Ulusoy, and Y. Manolopoulos. A data mining approach for location prediction in mobile environments. *Data & Knowledge Engineering*, 54(2):121–146, 2005.

[32] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma. Mining interesting locations and travel sequences from gps trajectories. In *WWW*, 2009.


**Sophie Cerf** is a PhD candidate working in the GIPSA-lab of the Univ. of Grenoble Alps and the LIRIS laboratory of INSA Lyon. Her research focuses on the use control theory for computer sciences, with specific application to performance and dependability of cloud services; and utility-aware solutions for privacy preserving mobile applications.

**Sara Bouchenak** is Professor of Computer Science at INSA Lyon. She is a member of the LIRIS laboratory, DRIM research group, where she conducts research on dependable, trustworthy and highly available distributed computer systems.

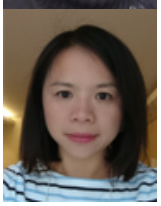
**Bogdan Robu** is associate professor at the Grenoble Alps University (UGA) and a researcher in GIPSA-lab laboratory, Grenoble, France since September 2011. He received his PhD in 2010 from the University of Toulouse.

**Nicolas Marchand** is researcher and deputy director of GIPSA-lab since 2015. His researches focus on Event-based control, Control and stabilization of flying robots and Control theory for computer sciences.

**Vincent Primault** is a research associate at the University College of London. He received his PhD in 2018 from the LIRIS laboratory of INSA Lyon. He works on privacy-preserving data collection.

**Sonia Ben Mokhtar** is a CNRS researcher at the LIRIS lab since October 2009. She is at the head of the distributed systems and information retrieval group (DRIM) since 2017.

**Antoine Boutet** is assistant professor at INSA Lyon working on privacy and security in the Inria Privatics research group of the Citi laboratory, France.



**Lydia Y. Chen** is a research staff member at IBM Zurich Research Lab. Her researches focus on approximate big data discovery, privacy, reliability, and dependability.