PULP: Achieving Privacy and Utility Trade-off in User Mobility Data

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

PULP: Achieving Privacy and Utility Trade-off in User Mobility Data

Sophie Cerf[•], Vincent Primault[⋆], Antoine Boutet[⋆], Sonia Ben Mokhtar[⋆], Robert Birke[⋄],
Sara Bouchenak[⋆], Lydia Y. Chen[⋄], Nicolas Marchand[•], Bogdan Robu[•]
[•] Univ. Grenoble Alpes – GIPSA-Lab – CNRS, Control Theory Research Group, Grenoble, France
[⋆] INSA Lyon – LIRIS – CNRS, Distributed Systems Research Group, Lyon, France
[⋄] IBM Zurich Research Lab, Zurich, Switzerland
[⋆]{firstname.lastname}@insa-lyon.fr,[•]{firstname.lastname}@gipsa-lab.fr,[⋄]{bir, yic}@zurich.ibm.com

***Abstract*—Leveraging location information in location-based services leads to improving service utility through geo-contextualization. However, this raises privacy concerns as new knowledge can be inferred from location records, such as user's home and work places, or personal habits. Although Location Privacy Protection Mechanisms (LPPMs) provide a means to tackle this problem, they often require manual configuration posing significant challenges to service providers and users. Moreover, their impact on data privacy and utility is seldom assessed. In this paper, we present *PULP*, a model-driven system which automatically provides user-specific privacy protection and contributes to service utility via choosing adequate LPPM and configuring it. At the heart of *PULP* is nonlinear models that can capture the complex dependency of data privacy and utility for each individual user under given LPPM considered, i.e., Geo-Indistinguishability and Promesse. According to users' preferences on privacy and utility, *PULP* efficiently recommends suitable LPPM and corresponding configuration. We evaluate the accuracy of *PULP*'s models and its effectiveness to achieve the privacy-utility trade-off per user, using four real-world mobility traces of 770 users in total. Our extensive experimentation shows that *PULP* ensures the contribution to location service while adhering to privacy constraints for a great percentage of users, and is orders of magnitude faster than non-model based alternatives.**

I. Introduction

Location Based Services (LBSs) such as navigation applications, social networks, or on-line games have been widely adopted by people carrying mobile devices. Although location-aware systems have greatly improved the quality of many services by introducing geo-contextualization, such systems raise important privacy concerns. Indeed, these services generate mobility traces with timestamped locations reflecting users' moving activities. This can reveal sensitive information about users such as their home and work places [9], hobbies, religious or political leanings [8], or their health status.

To overcome this privacy issue, many Location Privacy Protection Mechanisms (LPPMs) have been proposed in the last decade to allow users to enjoy LBSs while protecting their privacy. These LPPMs vary according to the privacy guarantees they offer to the users as well as to the type of alteration they introduce on their location data. For instance, *Geo-Indistinguishability* (or GEO-I) adds spatial noise to user's locations [4], PROMESSE adds temporal noise to user's locations [18], while the LPPM presented in [2] performs spatial cloaking to hide the user among a set of other users in her vicinity. However, the effectiveness of these mechanisms usually rely on the tuning of a set of configuration parameters, often with a large range of possible values. For instance, GEO-I's main configuration parameter, which is named ϵ and takes its values in $\mathbb{R}^+$, has a direct impact on the amount of spatial noise added to the data. While this tuning has an impact on the privacy guarantees offered by the LPPM, it naturally affects the quality/utility of the protected data. In a nutshell, the higher the privacy guarantees, the lower the utility of the resulting data. One of the main challenges in this context is thus to provide effective means to choose appropriate LPPMs and further configure them according to a set of privacy and utility objectives that individuals specify.

To the best of our knowledge, few works have been done in this direction and mainly targeted at group of users.Specifically, in [6], authors propose to adapt the ϵ configuration of GEO-I, and hence the amount of spatial noise added to the user's location. While this solution offers privacy guarantees that are related to the chosen configuration value, it does not allow the data owner to explicitly set utility objectives. On the other hand, in [3] and [19], the authors propose a heuristic-based solution that iteratively explores a range of LPPM configurations on the spatial cloaking LPPM for the former and on GEO-I and PROMESSE LPPMs for the latter. However, there is no guarantee on the provided levels of privacy and utility as the heuristics are designed to achieve the best-effort solutions. The performance of these two approaches thus can be intrinsically limited, i.e., under-exploring the opportunities for differentiated trade-off of individual privacy and utility, because of their greedy nature. Moreover, without any guarantee regarding the performances of these solutions, their use to follow legal requirements seems quite limited. The central research questions related to individuals using LBSs still remain open challenge to a large extent: how to choose LPPMs and properly configure them with the dual objectives of achieving certain privacy *and* utility levels.

To achieve the differentiated trade-off of utility and privacy at per user level, it is necessary to understand the dependency that exists among different LPPMs between their configuration parameter and the privacy/utility metrics that one wants to maximize. In this paper we develop *PULP*, a framework which aims at efficiently capturing such a complex interplay and selecting a suitable LPPM according to users' objectives. The core of *PULP* is user-specific non-linear models that accurately describe and extrapolate how metrics of privacy and utility change with respect to different LPPMs. The *PULP* is composed of three phases: (i) off-line profiling that experiments a small number of combinations of LPPMs (ii) building of non-linear models based on those profiles and (iii) automatic choice of the LPPM and its associated configuration to correspond to users' privacy/utility objectives. The modeling and configuration in *PULP* have a computational complexity of O(1), which contrasts with state of the art solutions proposed in [3] and [19] that have quadratic and linear complexities, respectively.

The particular privacy metrics considered is the normalized percentage of POIs that are successfully hidden by LPPM, the utility metrics is defined by the normalized percentage of areas that are successfully covered after using LPPM. We evaluate *PULP*'s on two state-of-the-art LPPMs, using four real-world mobility datasets collected in San Francisco, Beijing, Lyon and Geneva. Results on 770 users show that (i) *PULP* can accurately capture the non-linear trend of privacy and utility metrics relative to LPPMs for individual user; (ii) *PULP* can strongly achieve users' privacy and utility objectives simultaneously and (iii) *PULP* is able to identify best configurations by orders of magnitude faster than its closest competitor *ALP* [19].

The rest of the paper is organized as follow. Section II presents background on geolocation services, LPPMs and motivates the problem addressed in the paper. Section III presents our proposed *PULP* framework. Section IV presents experimental evaluation of *PULP* and discussion. Finally, related work is reviewed in Section V, before we draw our conclusions in Section VI.

II. Background and Problem Statement

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

A. Geolocation Services and Mobility Traces

Mobility Traces. The base of this work is mobility datasets collected in the wild: the Cabspotting (CABS), the PRIVAMOV, the GEOLIFE, and the Mobile Data Challenge (MDC) datasets, amounting to a total of 770 users. Datasets are constituted of a set of recorded locations (points on the surface of the earth at a precise time) called *records*. The set of all records corresponding to a user is called user's *trace.* Table I details statistics of each dataset. These datasets contain mobile information about users during their daily life. We make no assumption regarding the shape or patterns of the traces. The CABS dataset [17] contains the GPS traces of almost 550 taxi cabs (referred as users here) in San Francisco, USA, collected in 2008. The PRIVAMOV dataset [5], collected in 2015, involves 48 students and staff from various campuses in the city of Lyon equipped with smartphones running a data collection software. The GEOLIFE dataset [23] gathers the GPS trajectory of 42 users collected from April 2007 to August 2012 in Beijing, China. Finally, the MDC dataset [16], [14] involves around 142 volunteers in the Lake Geneva region, Switzerland, collected in 2014. To have homogeneous datasets, we align the length period of the four datasets to the one of the smallest one (i.e., CABS which has 30 days of mobility data). Hence, we extracted the most active period of 30 days from PRIVAMOV, GEOLIFE, and MDC data collections.

TABLE I
30-DAYS MOBILITY DATASETS

Dataset	Location	#users	#records
CABS	San Francisco, USA	548	11 219 955
GEOLIFE	Beijing, CN	42	1 574 338
MDC	Geneva Region, CH	142	904 422
PRIVAMOV	Lyon, FR	48	973 684

We index each user by the subscript of i, and each LPPM by j. We denote the mobility trace of user 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 i her location on the surface of the Earth defined by latitude-longitude coordinates (i.e., lat, lng), at a given time t. Whereas T_i reflects the actual location of user i, T'_{ij} contains a modified version T_i that depends on the $LPPM$ used for obfuscation and its configuration.

B. Location Privacy Protection Mechanisms (LPPMs)

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

GEO-I Geo-Indistinguishability protects user's location data by adding spatial noise drawn from a Laplace distribution to the actual user's location of each record in the mobility trace (see [4] for the algorithm details). GEO-I has a configuration parameter ϵ, expressed in meters^{-1} varying in $\mathbb{R}^+$, which quantifies the amount of noise to add to raw data. The lower the ϵ is, the more noise is added. GEO-I is a state of the art LPPM that follows the differential privacy model [7]; 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 [19] is a LPPM that has been developed in order to prevent the extraction of Points-Of-Interest (users' stop places) while maintaining a good spatial accuracy. 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 (parametrized by ϵ in meters). One can see its behavior as adding temporal noise to a trace instead of spatial noise as in GEO-I.

Although we specifically consider GEO-I and PROMESSE, the proposed methodology in the next section is general for any LPPM working for every user independently and having one configuration parameter. For some LPPMs, the computation of obfuscated trace is done accordingly to the obfuscation of other users, in cloaking solutions for instance. *PULP* works only for LPPM for which the obfuscation for one user only depends on this user. *PULP* is not designed for LPPM with several configuration parameters, however they can still be integrated in the framework by fixing all the LPPM parameters except the one having the most influence on privacy and utility. In the following, we define privacy and utility metrics for user mobility data, before illustrating them when applying GEO-I and PROMESSE LPPMs.

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. To our knowledge, there is no standard way of assessing these two complementary dimensions associated to LPPMs at a user level. Part of the literature focus on evaluating the privacy of a user by comparing her to others using re-identification attacks, see [20] or [10] for more details. However, we chose to use metrics that treat each user independently in order to be able to work at the user level, which is a key point as will be shown in section II-E. We choose to define privacy by looking at a user's POI (i.e. significant stops) protection [9] and utility by evaluating the accuracy of revealed locations [6]. Both metrics evaluate the gain of privacy and the loss of utility of the obfuscated data compared to the raw data. The next two sections define privacy and utility metrics while the third one gives illustrated example of metrics computation.

1) Data Privacy Metric: To evaluate data privacy from mobility traces, we first consider the retrieval of POIs from some location data. A POI (point of interest) is a meaningful geographical point around which where a user made a significant stop. A POI is defined by the position of a centroid of a given diameter d 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's privacy level by 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 set of $poi(T_i)$ and $poi(T'_{ij})$. 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 its completeness. We advocate to use Fscore to reconcile the precision and recall.

We formally write the privacy metric, showing the normalized percentage of successfully (non-matched) hidden 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.

2) Data Utility Metric: To evaluate data utility from a single user's trace, we resort to the comparison between the area coverage of the original mobility trace and of the obfuscated one. Particularly, we define the area coverage by the concept of cells. The size of the cell reflects the granularity of the considered spatial dimension, ranging from the size of a house to a city. 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 associated to mobility trace of user i, before and after applying the LPPM. One can think of $cell(\cdot)$ as a set containing cells that are visited by a user. To enable the comparison of cell coverage across a user's trace, we use the measurement of precision and recall to describe the percentage of cells that are correctly covered by T'_{ij}, relative to the original cell sets from T'_{ij} and T_i, respectively. We formally write the precision and recall of correct recovered cells for user i as:

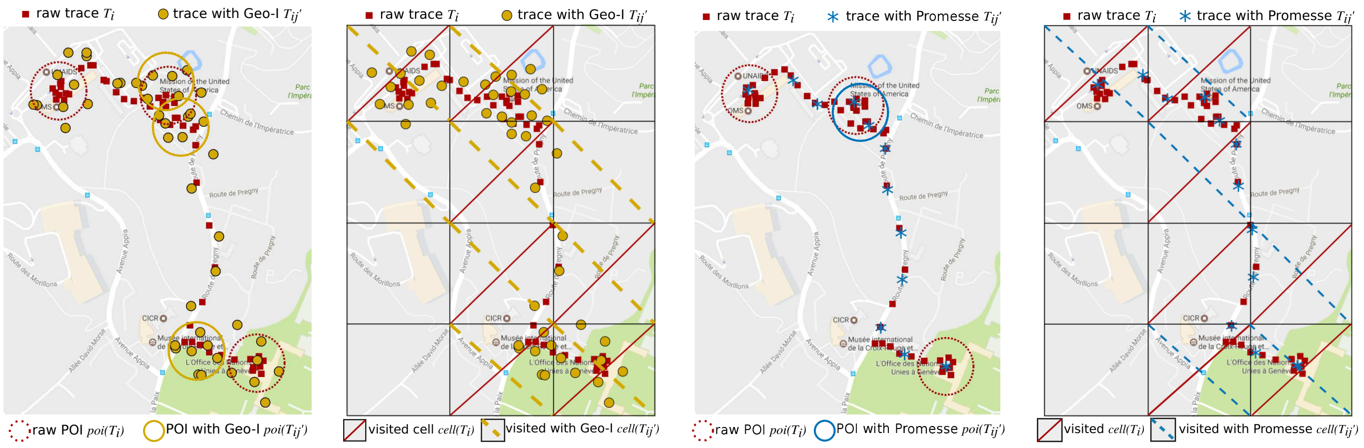
$$P_{ut}(i,j) = \frac{|cell(T_i) \cap cell(T'_{ij})|}{|cell(T'_{ij})|},$$

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

Similar to privacy metric, we finally define the utility metric of user i, $Ut(i)$, by the Fscore reconciling the 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 of $[0,1]$ where a higher value reflects a better utility, meaning a better spatial accuracy of the LBS results. The utility metric is by definition sensitive to the discretization of the map in cells. A mobility trace always moving at the boarder of two cells will produce


raw trace T_i
trace with Geo-I T_{ij}'
raw POI $poi(T_i)$
POI with Geo-I $poi(T_{ij}')$
visited cell $cell(T_i)$
visited with Geo-I $cell(T_{ij}')$
trace with Promesse T_{ij}'
POI with Promesse $poi(T_{ij}')$
visited with Promesse $cell(T_{ij}')$

(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.

a utility value different than if it were in the middle of cells. However in our application, the metric is calculated on large mobility datasets where the impact of discretization is then negligible.

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

D. 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 mobility trace, see Fig. 1.

Computing privacy metric: In Fig. 1(a), the raw mobility trace T_i of the user i is represented with the small red squares, each square being a location record. We overdraw the mobility trace of the user after using GEO-I (T'_{ij}), configured with a high ϵ (small yellow dots). We clearly see that the trace obfuscated with GEO-I corresponds to the original one but with some noise. For those two traces T_i and T'_{ij}, we illustrate the Points-of-Interest (POIs) with large circles. The set of POIs of the original trace $poi(T_i)$ are the red dashed circles, while POIs of the obfuscated trace $poi(T'_{ij})$ are the yellow continuous ones. Based on those sets, one can compute the number of obfuscated POIs that match the real ones (here the two top ones $\text{Matched}(poi(T'_{ij}), poi(T_i)) = 2$). Then our privacy metrics can then be computed using the precision of the matching of POIs and its recall, that both are $2/3$. Then, the level of privacy is $1 - 2.\dfrac{2/3 * 2/3}{2/3 + 2/3} = 0.33$.

Fig. 1(c) is similar to Fig. 1(a) but here the considered LPPM is PROMESSE. In this case, the obfuscated data T'_{ij} (the small blue stars) are spatially regularly distributed (time stamps are modified). In this illustration all obfuscated POIs correspond to the real ones, the privacy precision is 1. However, there is POIs from the raw trace that have not been retrieved, then the recall is $1/3$. The resulting privacy value is then 0.5.

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

E. Problem Statement: No Single Solution Fits All

Here, we present a motivating example showing that applying LPPM 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 $\epsilon = 100m$ on all of them. Following definitions of eq. (1) and (2), we obtain the privacy and utility values for all combinations of LPPMs, configurations, and users in Fig. 2. Let us first analyze those metrics from the perspective of individual users. Both utility and privacy metrics of user 2 (squares of all colors) differ from applying GEO-I to PROMESSE, showing the impact of LPPM and its configurations. Such an observation can also be made for users 1, 3 and 4, with varying degrees of differences. Taking the per LPPM perspective, either GEO-I or PROMESSE, one can see that it offers different levels of privacy protection and service utility to users when using only single configuration value of ϵ on all four users (symbols of the same color). These differences are due to the specificities of the users, however we will no go further into details about this as it is still a open research topic. Fig. 2 also illustrates that using one LPPM but with various configurations, can lead to totally different privacy protection and service utility. In other words, it might be impossible to find single (configuration) solution that fits all users' privacy and utility objectives. Both observations highlight the complex interplay among privacy/utility metrics, the LPPM and its configuration and the specificities of a

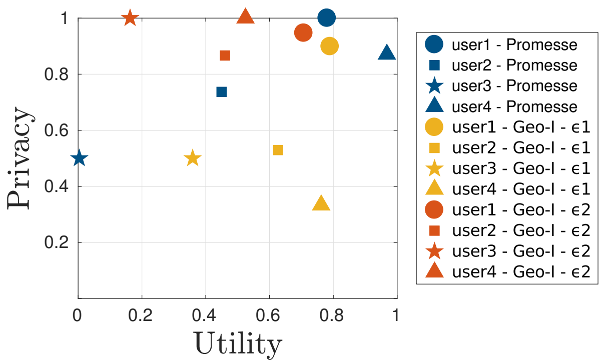


Fig. 2. Same LPPM can result into different privacy and utility values: examples from 4 users using PROMESSE with $\epsilon = 100m$ and GEO-I with two different configurations: $\epsilon_1 = 0.01m^{-1}$ and $\epsilon_2 = 0.005m^{-1}$.

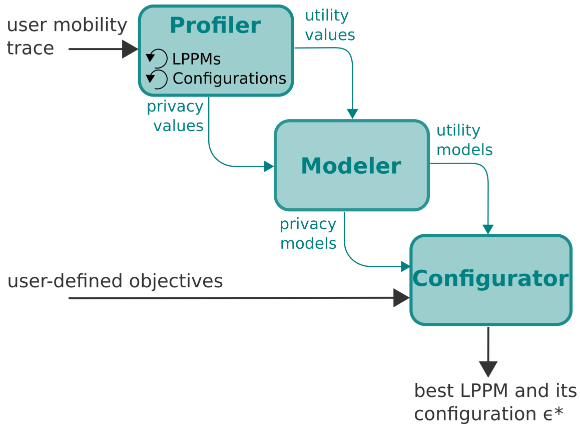


Fig. 3. System schematics of *PULP*

user. 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 configuration at the level of individual users.

III. Design Principles of *PULP* Framework

In this section, we describe the methodology and design of *PULP*, a framework that can efficiently select and configure LPPMs according to each user's privacy and utility objectives. Particularly, a user specifies the area coverage of her mobility to be used to improve the LBS and also the percentage of her POIs to be hidden. To such an end, *PULP* leverages a non-linear modeling approach and is composed of three key components: profiler, modeler and configurator shown in Fig. 3.

The profiler conducts off-line experiments to build users' privacy and utility profiles, with respect to LPPMs considered and a set of their configuration parameters. For each user, the modeler uses the 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 (for instance a ratio between privacy and utility) and the models learned by the modeler, the configurator suggests the suitable LPPM and its configuration. We explain the details of each components in the following subsections.

A. 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. 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 privacy and utility metrics values for all cases. Specifically, the profiler considers two LPPMs, GEO-I with $\epsilon = [10^{-4}, 1]$ in $meters^{-1}$ and PROMESSE with $\epsilon = [50, 10^4]$ in $meters$ where range values were taken from LPPM authors recommendations. The number of configuration values needed is driven by the fitting accuracy of the proposed model in the following subsection. 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. As a default setting, we propose to select 4 values of the configuration parameter per log-decade of its definition range, the picked values being equally distributed along the range.

B. Modeler

The aim of the modeler is to derive the functional relationship between privacy/utility metrics and the configuration parameter of a given LPPM, i.e., $Pr(i,j) = F_{pr}(\epsilon|_{LPPM_j})$ and $Ut(i,j) = F_{ut}(\epsilon|_{LPPM_j})$.

To search for the most suitable and general function, we conduct numerous data fitting schemes on our datasets. Fig. 4 depicts commonly seen dependency between privacy/utility and ϵ, via an example of applying GEO-I and PROMESSE on a CABS user (continuous line). Experimental conditions of these experiments are further detailed in Section IV-A. The shape of curves can be explained by the limited ranges of privacy and utility metrics in $[0,1]$ and insensitiveness of privacy and utility metrics to extreme values of ϵ. The first two observations lead us to choose arctan function as our base model, instead of general polynomial functions, that could fit the experimental data but using more parameters. The observation of experimental data makes us to use $ln(\epsilon)$ to fit the arctan model of F_{pr} and F_{ut}, instead of ϵ directly.

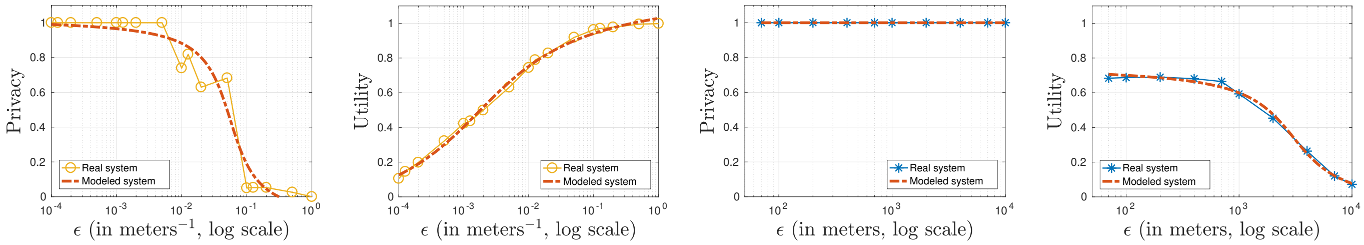
Now, we formally introduce the utility and privacy models with four coefficients, i.e., a, b, c, and d,

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

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

An illustration of model shapes are given in Fig. 4.

The physical meanings of model parameters in both F_{pr} and F_{ut} are: a and d representing the two saturation levels, and b characterizing the transition speed between saturation levels. Parameter c corresponds to ϵ 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



(a) Privacy with GEO-I (b) Utility with GEO-I (c) Privacy with PROMESSE (d) Utility with PROMESSE

Fig. 4. Impact of LPPMs' configurations on a user's privacy and utility metrics – Real system vs. modeled system (cabspotting user)

advantage that there are only four coefficients to be learned, meaning a minimum of 4 profiling runs is needed. Hence, one can choose a set of configuration parameters logarithmically distributed in its definition range, with a minimum of four values in it. An insight of maximal useful values are given in Section IV-D.

C. Configurator

The aim of the configurator is to select and configure a LPPM from the available LPPM set so as to satisfy the user defined objectives that are related to the privacy (the proportion of POIs to be hidden) and utility metrics (proportion of correct map coverage to be communicated to the LBS). To such an end, the configurator takes inputs from the modeler and users as shown in Fig. 3. Inverting the user-specific models derived by the modeler, the configurator can then choose the configuration of all LPPMs considered such that the resulting utility and privacy values can fulfill individual user's objective. We consider four types of user's objective that combine privacy and utility differently:

- keeping both privacy and utility above given levels, i.e., $Pr \geq Pr_{min}$ and $Ut \geq Ut_{min}$;
- keeping the privacy above a given level then maximize the utility, i.e., $\max Ut$ with a constraint on $Pr \geq Pr_{min}$,
- keeping the utility above a given level then maximize the privacy, i.e., $\max Pr$ with a constraint on $Ut \geq Ut_{min}$,
- guaranteeing a given ratio between privacy and utility: $Pr = w_{pr/ut} \cdot Ut$.

While the first three objectives aim to achieve absolute values of privacy and utility metrics, the last objective addresses the relative trade-off between privacy and utility. For example, when a user specifies $w_{pr/ut} = 2$, she prefers the privacy to the utility in a two to one ratio, meaning that every contribution to the LBS (in the percentage of area coverage) is at the cost of half unit of privacy loss (in the percentage of exposed POIs). On the contrary, $w_{pr/ut} = 0.5$ implies that a user thinks contributing to the LBS is twice more important than preserving her privacy. Detailing all configuration laws would take a lot of space and be relatively repetitive. In the following, we focus only on the fourth objective specifying the relative trade-off between the privacy and utility, $w_{pr/ut}$. This choice motivated by the user-friendliness formulation of this objective compared to the others. We now detail the solving procedure to find the LPPM, i.e., j^*, and its configuration parameter, from a set of J LPPMs based on a given relative trade-off $w_{pr/ut}$ provided by user i.

Solving procedure to achieve $Pr = w_{pr/ut}.Ut$. As there are J LPPMs available, the configurator first needs to find the best configuration parameter for a user i when applying each LPPM, i.e., ϵ_j^*, $\forall j$. There can be multiple LPPMs which best configuration parameters can be found to achieve the target relative trade-off, $w_{pr/ut}$. The configurator then compares them and recommends the best LPPM (and thus its configuration) based on the absolute values of privacy and utility metrics. Specifically, the configurator iterates through following two steps to achieve the relative trade-off for each user. To simplify the notation, here we skip the index for user i, even though all the steps depicted in the following are done at a user level.

1) Finding best configuration for LPPM $j, \forall j$
To achieve the trade-off ratio of $w_{pr/ut}$ between privacy and utility, one needs to find its configuration ϵ_j^* such that $Pr = w_{pr/ut} \cdot Ut$. Applying the model of eq. (3) and (4), we can then obtain ϵ_j^* by solving

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

Due to its complexity, we opt out deriving a closed form solution for ϵ_j^*. Instead, we resort to numerically solve it as the minimization problem of the absolute weighted difference between F_{ut} and F_{pr},

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

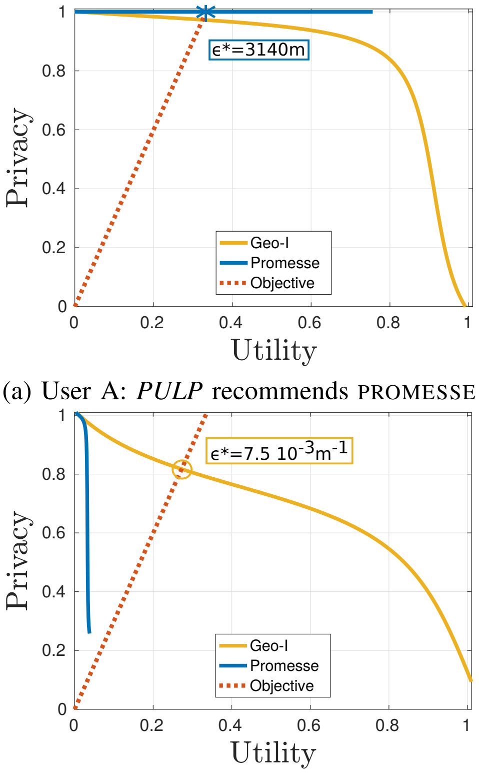
The convergence of the solution is ensured by the convexity of the function to minimize in eq (5). However, when the resulting configuration parameter value does not fall into legitimate range (which depends on the LPPM), we then consider LPPM j as an infeasible LPPM to provide the target trade-off between privacy and utility.

2) Selecting the best LPPM.
Among a subset of LPPMs that can achieve the target trade-off with valid configuration parameters, the configurator then selects the LPPM that can maximize the weighted sum of the resulting privacy and utility metrics. We then can obtain the best LPPM j^* and its configuration ϵ_j^* for user i by

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

D. Illustration of Configuration Law

Fig. 5 illustrates how configurator functions for two different users (exact experimental conditions are given in Section IV-A). The two figures represent the privacy versus utility plan, where each LPPM curve is composed of a set of possible couples $(Pr(i,j), Ut(i,j))$ that is achieved through different configuration parameters. We also plot the objective curve $Pr = w_{pr/ut} \cdot Ut$ (red dotted line). When the objective line crosses a LPPM curve, it gives the privacy and utility metrics values that meet the user specified trade-off ratio, as well as the configuration parameters of the LPPM to use (only represented here for the chosen configuration). For user A (Fig. 5(a)), the chosen LPPM is PROMESSE as it gives a higher weighted sum of privacy and utility metrics (1.99) compared to GEO-I (1.80). For user B (Fig. 5(b)), GEO-I is selected by *PULP* as it is the only LPPM that can fulfill the objective. The final output of *PULP* are then (user A, PROMESSE, $\epsilon^* = 3140\ m$) and (user B, GEO-I, $\epsilon^* = 7.5\ 10^{-3} m^{-1}$).



(a) User A: *PULP* recommends PROMESSE

(b) User B: *PULP* recommends GEO-I

Fig. 5. Illustration of *PULP* configurator possible behaviors, with a trade-off objective of $w_{pr/ut} = 3$. User A from CABS and User B from PRIVAMOV.

IV. *PULP* Evaluation

For the validation of *PULP*, we proceed in three strokes: first analyzing the modeler's behavior with an emphasis on the accuracy of the derived models, then illustrating the effectiveness of the configurator in choosing suitable LPPM to achieve different user's objectives, finally we show the robustness of *PULP* system. Prior to presenting the core results, we first describe the experimental setup.

A. Experimental Setup

The metrics of privacy and utility used for evaluation have been parametrized to correspond to our datasets collected in dense-cities. For measuring privacy we consider a POI maximum diameter of $d = 200$ meters and a minimal stay time of $t = 15$ minutes. In order to calculate intersections between sets of POIs, we consider that two POIs are matched if their centroids are within $d_{max} = 100$ meters from each other. For measuring utility, we use Google's S2 geometry library for cell extraction [21]. 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, corresponding to a city block or a neighborhood.

For the experimental validation of *PULP* we used two different machines. The profiler is executed on a machine running Ubuntu 14.04 and equipped with 50Gb of RAM and 12 cores clocked at 1,2 GHz. We run the profiler using the 30-days datasets. The modeler and the configurator uses Matlab R2016b on a Ubuntu 16.04 and equipped with 3.7Gb of RAM and 4 cores clocked 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 search of each user's model and the configurator's configuration law uses the function *fminunc* [1].

B. Evaluation of PULP Modeler

Fig. 4 can be used to compare the model (red dotted line) to the experimental data (yellow circles for GEO-I and blue stars for PROMESSE). The closeness of the curves indicates a really good model fitting to real data for that user.

In order to ensure that *PULP* modeler is accurate for each user, we compute the variance of the fitting error, which is a relevant indicator for non-linear modeling. For all LPPM considered and for the two metrics, the median of the error variance is less than 7.10^{-4} which shows that the models have a remarkable good accuracy. They also fit properly in extreme cases, as for the 99^{th} percentile the error variance is still low, ranging from 6.10^{-4} to 4.10^{-2} for all LPPMs and all metrics.

C. Evaluation of PULP Configurator

The purpose of *PULP* configurator is to choose a LPPM and configure it in a way that ensures the fulfillment of the objective ratio between privacy and utility. When running *PULP* on all users with various objective ratio $w_{pr/ut}$, all users ended with a recommended LPPM. We computed the actual ratio after applying the LPPM selected with its right configuration. Results show that at least 97% of the users have a resulting ratio in a range of +/- 1% of user specified values. This illustrate the high efficiency of *PULP* for every user.

In the following we take a deeper look at *PULP* results and analyze their variability. All users have different mobility patterns, modeled by the adjustable parameters of eq. (3) and

(4). The variance of those parameters for all LPPMs can go up to 100% of their mean value. The impact of diversity among users on *PULP* results is illustrated in the following.

Importance of a Careful Choice of LPPM. Fig. 6 illustrates the distribution of the LPPM selected by *PULP* among users, for various objectives $w_{pr/ut}$. For a given objective, the LPPM chosen by *PULP* varies, as can be expected after seeing Fig. 5. Moreover, the distribution changes according to the objective, meaning that the adequate LPPM for every user may vary. There is no a priori relation between the objective $w_{pr/ut}$ and the repartition of selected LPPM. Hence, these results shows that it is important to adapt the LPPM according to the users as well as their objectives.

Importance of a Careful LPPM Configuration. Now we analyze *PULP* choice of LPPM configuration parameters. Fig. 7 and 8 illustrate the distribution (in a form of cumulative distribution functions) of the configuration respectively among users for whom *PULP* selected GEO-I as the suitable LPPM, and PROMESSE. Results for various objective ratios are overlaid. These figures illustrate two points (i) users need different configurations to fulfill the same objective, and (ii) different objectives lead to various configurations distribution. Once again these results enhance the importance of user-grained configuration of LPPM.

Achieved Privacy and Utility. When using the appropriate LPPM configured in a suitable way, users can maintain privacy and utility levels that jointly respect the objective trade-off. It is shown in Fig. 9 which summarizes the distribution of achieved ratios $w_{pr/ut}$ for different objectives. In terms of their absolute values, Fig. 10 and 11 show the distribution of privacy and utility among users for various objective trade-off. Achieved absolute values of privacy and utility levels are different among users even though the objective is always reached. For instance, if a user objective is to reach $Pr(i) = Ut(i)$ (equal weight on the utility and privacy), many users have $Pr(i) = Ut(i) = 0.8$ but some are only able to reach $Pr(i) = Ut(i) = 0.65$. One can notice that small proportion of the user have a low privacy and utility, as guaranteeing the ratio $w_{pr/ut}$ does not ensure absolute values. These users have specific mobility patterns that make them hard to obfuscate efficiently using GEO-I and PROMESSE, but that could show better results with other LPPMs. Moreover, we can see that there is more diversity in values of privacy and utility for the scenarios where the objective ratio $w_{pr/ut}$ is small than for large ones. Whereas, the higher diversity in configuration parameters (see Fig. 7 and 8) is found for high $w_{pr/ut}$ values.

D. Discussion

After illustrating *PULP* efficiency and accuracy in fulfilling the objective, we now focus on *PULP* behavior by presenting its performance and robustness.

The modeler *PULP* uses experimental data to derive non-linear models. The amount of experimental data needed is at least 4 to find the given the numbers of parameters. In terms of the upper limit, we conduct the following analysis. We compare a modeling phase for GEO-I when taking respectively 10 and 17 different values for its configuration parameter (i.e. 2 or 4 values per decade of the definition set). The resulting modeling errors are in the same order of magnitude for both models, meaning that 10 experiments are enough to properly derive the non-linear models. There is a high diversity in users considered for our study, particularly regarding the number of points $(lat, lng, time)$ per trace. However, *PULP* is able to model the behavior of *every* user with a good accuracy (see IV-B), independent of the number of points in user's trace. Further determination of the exact number of LPPM configuration to experiment for all LPPM are planed for future works, as well as a sensitivity analysis of *PULP* to the trace size and representativeness.

As *PULP* works well with few experiments, its execution time is significantly shorter compared to the state of the art. Indeed, all configuration mechanisms that we are aware of use greedy processes that need to run many experiments on the whole dataset in order to converge to a suitable configuration (if ever it converges). However, our proposed solution *PULP* has a complexity of O(1) for the modeling and configuration phases, as these steps only use the profiles and models that are independent on the size of the user's trace. We compare our framework *PULP* to the closest work from the state of the art, the configurator ALP from [19]. 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 condition 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 its objective and easily found again the adequate LPPM and its configuration.

In the heart of *PULP* is a user formulation of privacy and utility objectives, based on definition of POI and cell size. These metrics are parametrized by the following variable: d and t the diameter and minimal stay time defining a POI, d_{max} the matching threshold between POIs, and the cell size parameter refereed to as s in [21]. The robustness of *PULP* regarding changes of these parameters as been studied. We found that the variation of every parameter impacts the privacy and utility characteristic of users for all LPPMs, some more than others. However, they fall into the hypothesis of our modeler, knowing saturation levels and monotonous behavior between them. Hence, *PULP* is able to adapt to the changes of these metrics. Thus, we recommend the following:

- Utility: adapt the cell size to the LBS. For instance, a weather app is useful when the location is accurate at a few kilometers, thus large cells can be chosen.
- Privacy: parametrization depends on user objectives. For instance, a user willing to precisely hide her home or work places should choose low d and $d_{max}$and t to a few hours.

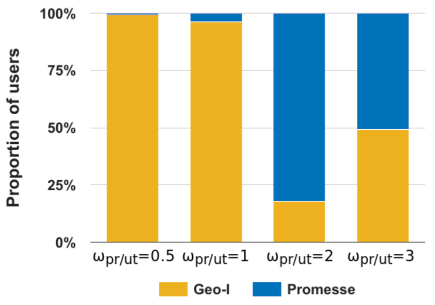


Fig. 6. Repartition of LPPM selected by *PULP* for various objectives

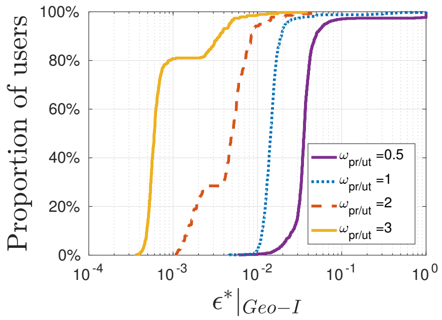


Fig. 7. Distribution (cdf) of LPPM configurations, for GEO-I-recommended users.

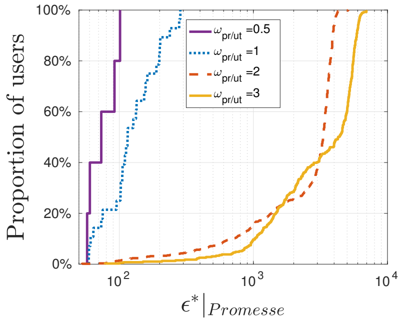


Fig. 8. Distribution (cdf) of LPPM configurations, for PROMESSE-recommended users.

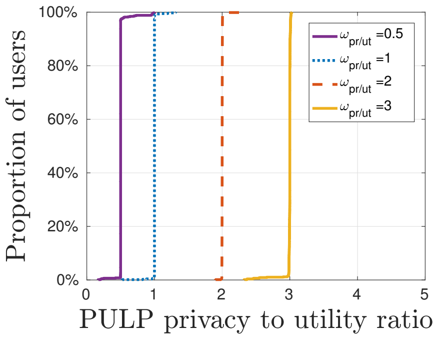


Fig. 9. Distribution of privacy to utility ratio achieved with *PULP* for varying objectives

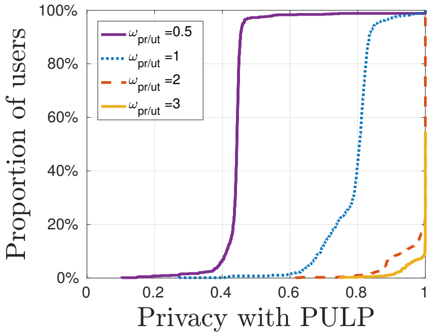


Fig. 10. Privacy of all users with *PULP* applied with varying objectives $w_{pr/ut}$

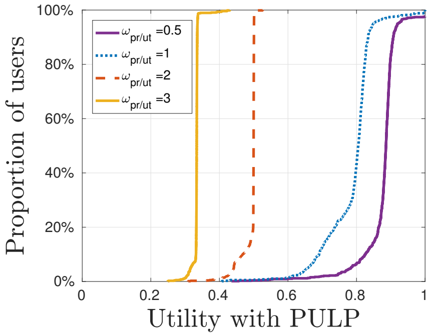


Fig. 11. Utility of all users with *PULP* applied with varying objectives $w_{pr/ut}$

V. 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 [22], which states that a user is k-anonymous if it 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 (e.g., CliqueCloak [11]) as well as in distributed systems relying on peer-to-peer communication between users (e.g., PRIVÉ [12]).

Another popular privacy guarantee is differential privacy [7], 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. Differential privacy has been applied as such in [13], where a controlled amount of noise was added to each location of a mobility trace. It has also been applied through Geo-Indistinguishability [4], which is an extension of differential privacy designed specifically to be used on mobility traces. Here again, differential privacy is guaranteed by adding noise, drawn from a two-dimensional Laplace distribution.

B. LPPM Configuration

What makes LPPMs difficult to use in practice is that they rely on a set of configuration parameters. For instance, the ϵ 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 [15], 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 objective.

A few works have been proposed to help a user choose a LPPM configuration that fits her actual needs. Agir et. al [3] proposed an adaptive mechanism that dynamically computes the size of the cloaking area the user will be hidden within. More specifically, starting a given parametrization of the LPPM, they iteratively modify the configuration in a way that strengthen the privacy until a minimum privacy level, fixed by the user, is met. 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. This routine is further repeated until required privacy level is met or at most λ times. Hence this solution is computing intensive and does not provide guarantees about its performance. Chatzikokolakis et. al [6] introduced an extension of GEO-I that uses contextual information to adapt the effective privacy level. Specifically, the amount of noise effectively added to locations depends on whether the user is located in a dense urban area or in the countryside. This qualification is done by looking at the density of venues (e.g., restaurants, monuments, amenities) in the vicinity. It is expected that the number of venues is higher in urban environments and will better hide the user's interests in the area than if located outside of a city. However, this approach still requires some parametrization from the user side and is not objective-driven, which made it difficult to use for a non-expert user. Primault et al. [19] presented *ALP*, a system that configures a LPPM depending on users objectives. This solution relies on a greedy approach that iteratively evaluates the privacy and utility for refining configuration parameters. Evaluating privacy and utility has a complexity depending on the objectives under consideration, varying between $O(n)$ and $O(n^2)$. Moreover, the convergence is not ensured, there is no guarantee that the objectives are actually met.

VI. Conclusion

In this paper we propose *PULP*, a framework that ensures privacy and utility objectives of users in the context of mobility databases. *PULP* automatically builds privacy and utility models for various LPPMs, and then select the appropriate LPPM and configuring it in order to fulfill user-defined objectives, that can be expressed as a privacy to utility ratio. *PULP* realizes an in-depth analysis of the considered LPPMs applied at a user scale, in order to provide the formal relationship between the configuration parameters of the LPPMs and both privacy and utility metrics. Then *PULP* leverages the built models to derive the adequate LPPM and its configuration that enables to fulfill the objectives.

We illustrated the ability of our system to efficiently protect a user while keeping utility to her service using two LPPM 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. *PULP* can accurately model the behavior of LPPM on users and thus successfully achieved privacy and utility objectives at the same time in an automated way. Moreover, when comparing with state of the art, we proved our system to be 3 orders of magnitude faster.

Future work will investigate *PULP*'s ability to work with new metrics and new LPPMs, including ones with more than one configuration parameter. The use of *PULP* in a real-time scenario, for instance using a navigation app, is under study.

VII. Acknowledgement

This work was supported by the LABEX IMU (ANR-10-LABX-0088) of University of Lyon in the frame of the MobiCampus-UdL project.

References

[1] Find minimum of unconstrained multivariable function - MATLAB fminunc - MathWorks Australia.

[2] Osman Abul, Francesco Bonchi, and Mirco Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *ICDE*, pages 376–385, 2008.

[3] Berker Agir, ThanasisG. Papaioannou, Rammohan Narendula, Karl Aberer, and Jean-Pierre Hubaux. User-side adaptive protection of location privacy in participatory sensing. *GeoInformatica*, 18(1):165–191, 2014.

[4] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential Privacy for Location-based Systems. In *CCS*, pages 901–914, 2013.

[5] Antoine Boutet, Sonia Ben Mokhtar, Louafi Bouzouina, Patrick Bonnel, Olivier Brette, Lionel Brunie, Mathieu Cunche, Stephane Dalu, Vincent Primault, Patrice Raveneau, Herve Rivano, and Razvan Stanica. PRIVAMOV: Analysing Human Mobility Through Multi-Sensor Datasets. In *NetMob*, 2017.

[6] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. Constructing elastic distinguishability metrics for location privacy. In *PETS*, volume 2015, pages 156–170, 2015.

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

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

[9] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. Show Me How You Move and I Will Tell You Who You Are. *Transactions on Data Privacy*, 4(2):103–126, August 2011.

[10] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. De-anonymization attack on geolocated data. *Journal of Computer and System Sciences*, 80(8):1597–1614, 2014.

[11] Bugra Gedik and Ling Liu. Location Privacy in Mobile Systems: A Personalized Anonymization Model. In *ICDCS*, pages 620–629, 2005.

[12] Gabriel Ghinita, Panos Kalnis, and Spiros Skiadopoulos. PRIVE: Anonymous Location-based Queries in Distributed Mobile Systems. In *WWW*, pages 371–380, 2007.

[13] Kaifeng Jiang, Dongxu Shao, Stéphane Bressan, Thomas Kister, and Kian-Lee Tan. Publishing Trajectories with Differential Privacy Guarantees. In *SSDBM*, pages 12:1–12:12, 2013.

[14] N. Kiukkonen, Blom J., O. Dousse, Daniel Gatica-Perez, and Laurila J. Towards rich mobile phone datasets: Lausanne data collection campaign. In *ICPS*, 2010.

[15] John Krumm. Inference Attacks on Location Tracks. In *PerCom*, pages 127–143, 2007.

[16] Juha K. Laurila, Daniel Gatica-Perez, Imad Aad, Jan Blom, Olivier Bornet, Trinh Minh Tri Do, Olivier Dousse, Julien Eberle, and Markus Miettinen. From big smartphone data to worldwide research: The mobile data challenge. *Pervasive Mob. Comput.*, 9(6):752–771, December 2013.

[17] Michal Piorkowski, Natasa Sarafijanovic-Djukic, and Matthias Grossglauser. CRAWDAD dataset epfl/mobility (v. 2009-02-24). Downloaded from http://crawdad.org/epfl/mobility/20090224, February 2009.

[18] Vincent Primault, Sonia Ben Mokhtar, Cédric Lauradoux, and Lionel Brunie. Time Distortion Anonymization for the Publication of Mobility Data with High Utility. In *TrustCom*, August 2015.

[19] Vincent Primault, Sonia Ben Mokhtar, Cédric Lauradoux, and Lionel Brunie. Time distortion anonymization for the publication of mobility data with high utility. In *TrustCom*, pages 539–546, 2015.

[20] Vincent Primault, Sonia Ben Mokhtar, Cédric Lauradoux, and Lionel Brunie. Differentially private location privacy in practice. *arXiv preprint arXiv:1410.7744*, 2014.

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

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

[23] Yu Zheng, Lizhu Zhang, Xing Xie, and Wei-Ying Ma. Mining interesting locations and travel sequences from gps trajectories. In *WWW*, pages 791–800, 2009.