Time Distortion Anonymization for the Publication of Mobility Data with High Utility
Vincent Primault, Sonia Ben Mokhtar, Cédric Lauradoux, Lionel Brunie

To cite this version:

HAL Id: hal-01170060
https://hal.archives-ouvertes.fr/hal-01170060
Submitted on 1 Jul 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Time Distortion Anonymization for the Publication of Mobility Data with High Utility

Vincent Primault*, Sonia Ben Mokhtar*, Cédric Lauradoux† and Lionel Brunie*

*Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205, F-69621, France
{vincent.primault,sonia.ben-mokhtar,lionel.brunie}@liris.cnrs.fr
†INRIA
cedric.lauradoux@inria.fr

Abstract—An increasing amount of mobility data is being collected every day by different means, such as mobile applications or crowd-sensing campaigns. This data is sometimes published after the application of simple anonymization techniques (e.g., putting an identifier instead of the users’ names), which might lead to severe threats to the privacy of the participating users. Literature contains more sophisticated anonymization techniques, often based on adding noise to the spatial data. However, these techniques either compromise the privacy if the added noise is too little or the utility of the data if the added noise is too strong. We investigate in this paper an alternative solution, which builds on time distortion instead of spatial distortion. Specifically, our contribution lies in (1) the introduction of the concept of time distortion to anonymize mobility datasets (2) Promesse, a protection mechanism implementing this concept (3) a practical study of Promesse compared to two representative spatial distortion mechanisms, namely Wait For Me, which enforces k-anonymity, and Geo-Indistinguishability, which enforces differential privacy. We evaluate our mechanism practically using three real-life datasets. Our results show that time distortion reduces the number of points of interest that can be retrieved by an adversary to under 3%, while the introduced spatial error is almost null and the distortion introduced on the results of range queries is kept under 13% on average.

Keywords—location privacy; data publication; time distortion

I. INTRODUCTION

The widespread adoption of location-aware devices such as smartphones, GPS navigation systems or GPS-enabled watches has dramatically increased the quantity of mobility data that is being continuously collected. This data is either sent online by the users when querying location-based services (e.g., finding nearby restaurants, friends) or is collected by applications and then published/sold offline to third parties (e.g., GPS traces sold for marketing purposes) or as part of crowd-sensing campaigns (e.g., publication of jogging/cycling traces, publication of pictures with location metadata on Flickr).

However, collecting and sharing mobility data raises serious privacy concerns. Among the known threats is the extraction of users’ points of interest [1] (or POIs), which can be defined as places where individuals regularly spend some time, e.g., home, work, a cinema or a mall. By studying the semantics of these places, it is possible to infer sensitive knowledge like religious or political preferences, for example if an individual often goes to a worship place or a political party’s headquarter. Learning users’ POIs can ultimately lead to learn about their real identity with a good accuracy [2]. Mobility data can also be used as an input of next place prediction algorithms, enabling to guess where users are likely to be in the next hour [3] as well as in four years [4], which opens the door to harmful threats such as house robbing.

Nevertheless, mobility data is still very valuable. Publishing such information allows researchers to perform real-time traffic predictions, to find out interesting mobility patterns or to discover social and economic tendencies. A number of research efforts have thus been carried out in the last few years to protect users’ location privacy. While effective mechanisms exist for the protection of privacy when a user opportunistically sends her current location to a location-based service (e.g., by using two non-colluding servers [5]), it is still an open issue to publish mobility traces of a set of users in a privacy-preserving manner. The major challenge in the latter case is due to the regularity of users’ mobility habits. Indeed, as users go to the same places at similar times of the day/days of the week, simple clustering techniques applied on published traces allow to infer users’ POIs.

To address this issue, a classical solution is to alter user locations (e.g., [6], [7]). However, according to the amount of added noise, this may also alter the utility of the published data. Indeed, the introduced spatial distortion may prevent the exploitation of protected traces for some well-known use cases such as transportation mode detection (e.g., [8]). An alternative solution that one may think of is the adoption of time distortion instead of spatial distortion. In this paper, we investigate this alternative and introduce our main contribution, Promesse, which is, at the best of our knowledge, the first mechanism that aims at hiding users’ POIs by distorting time. Specifically, Promesse hides users’ POIs by: (1) smoothing the users’ speed along their trajectories and (2) blurring the start and end points of these trajectories to make them less easily identifiable.

A preliminary version of this work has been presented in [9]. We present in this paper a detailed version along with a thorough evaluation. We practically study the effectiveness of Promesse compared to two representative mechanisms relying on spatial distortion, namely Wait For Me [10], which guarantees k-anonymity, and Geo-Indistinguishability [6], which guarantees differential privacy. Our evaluation, performed using three real-life datasets, shows that the number of retrieved POIs with Promesse is under 3%, which is comparable to what the other mechanisms can achieve. In the same time, Promesse provides no spatial error, while the other mechanisms’ error
Differential privacy has been used by Jiang et al. [15] to protect ships’ trajectories. Endpoints of trajectories are preserved while intermediate locations are altered. This results in a large distortion of trajectories and consequently in a difficult trade-off between privacy and utility. Andres et al. [6] introduced the concept of Geo-Indistinguishability (or Geo-I) which is a generalization of differential privacy specifically suited for location privacy. They propose to achieve Geo-I by adding a calibrated noise drawn from a two-dimensional Laplace distribution. Two use cases are studied in their paper, one of them being the anonymization of statistical geolocated datasets. We have previously studied this mechanism in [16] and shown that it is not suitable to anonymize entire mobility datasets, because POIs can still be retrieved. Chen et al. [17] anonymized public transportation usage data, which can be seen for each user as a sequence of places (metro/bus stations) she went to. They built a method to anonymize such data in a differentially private way and evaluated their mechanism by studying the impact of anonymization on range queries and sequential pattern mining. Like other k-anonymity works, their mechanism is focused on the anonymization of sequences of records. Acs et al. [18] proposed a mechanism to anonymize spatio-temporal densities datasets, which reports counts of active users within small areas for given time windows. Such data can be obtained for instance from call data records that are gathered by mobile phone operators. The authors proposed an approach that adapts to the original data in order to guarantee differential privacy with the highest possible utility. Counting users is one interesting thing to do with mobility data, but we want to publish entire trajectories and allow more mining tasks to be performed.

III. Problem Statement

We present in this section the privacy and utility objectives we target in this paper.

A. Utility concerns

We aim at the publication of fine-grained and spatially accurate mobility datasets. By fine-grained datasets we mean a set of records, each one being a location (i.e., a point on the Earth) associated with a (virtual) user identifier and a timestamp. Indeed, publishing fine-grained datasets allows analysts to freely implement their data mining tasks with the tools and languages they want and run them on the published data. Moreover, there are use cases where such datasets are required. For example, researchers working on delay-tolerant networks test their algorithms with real-life datasets (e.g., [19]). Another example is the case of transportation mode detection. In this context, state-of-the-art algorithms need to extract a number of information from mobility traces such as speed, acceleration [8] or proximity to rail lines/bus stops [20], which is not possible using alternative solutions such as interactive querying, where data analysts are restricted to a pre-defined query language provided by the data owner (e.g., [21]), or the publication of pre-aggregated datasets (e.g., [18]).

B. Privacy challenges

In terms of privacy, we focus on hiding the users’ POIs that could be inferred by an attacker. As defined in [1], POIs
are important places that a user regularly visits and in which she spends a given amount of time. POIs have been shown to be very sensitive because they allow to infer knowledge such as where one lives, her work occupation, her hobbies, her political or religious preferences, etc. Thus, it has been shown in [22] that even the simple pair home/work can lead to a re-identification of a large number of users. We hence focus in this paper on the protection of users’ POIs that we assess using a state-of-the-art POI extraction attack presented in [23].

C. The need of a new mechanism

State-of-the-art protection mechanisms suitable to publish fine-grained mobility datasets in a privacy-preserving way focus on spatial distortion, i.e., they alter the spatial component of each record to protect the privacy of users. We present in the next section an alternative solution based on temporal distortion and analyse the pros and cons of each approach in the rest of the paper.

IV. A TIME DISTORTION BASED MECHANISM

We begin by proposing definitions of concepts that will be used in the remaining of this paper (Section IV-A), before presenting our mechanism (Section IV-B).

A. Definitions

Physical and logical user. An individual, also called a physical user, is a person with an identity (first name, last name, social security number, etc.). A logical user (or user for short), is a consistent source of mobility data coming from a single physical user. A physical user can be associated with many logical users (e.g., multiple anonymized traces can be originated from a single physical user).

Location. A location is a point at the surface of the Earth. It can be represented in many ways, like by a latitude and a longitude or by a projection in cartesian coordinates. Locations are elements of \( \mathcal{P} \) and we consider that there is a suitable function \( d_X : \mathcal{P}^2 \to \mathbb{R} \) to compute the distance between two locations.

Record. A record is the smallest indivisible piece of data that can be collected. It represents the location of a user at a specific time. More specifically, it is a triplet composed of (1) a user identifier, linking a record to a specific user. Whereas user identifiers could be anything, we suppose they do not contain anything that could lead to the physical user’s identity. This is why user identifiers are generally sequential or random integers; (2) a timestamp, storing the date and time at which the record was generated; and (3) a location, encoding the spatial representation of the place at which the record was generated. The set of all possible records being \( \mathcal{R} \), we use a dotted notation to ease access to records’ members. For any record \( r \in \mathcal{R} \), \( r.user \) refers to the user identifier, \( r.loc \) refers to the location of the record and \( r.time \) refers to the timestamp of the record.

Dataset. A consistent set of records forms a dataset. All records inside a dataset usually come from a single collection campaign and feature similar characteristics (e.g., the sampling rate or the geographical area). The set of all possible datasets being \( \mathcal{D} \), for any \( D \in \mathcal{D} \), \( D \subseteq \mathcal{R} \).

Mobility trace. A subset of a dataset that contains all records belonging to a same user is called a mobility trace. Each user of a dataset is associated with one single mobility trace containing all her records. The set of all possible mobility traces associated with user \( u \) is noted \( \mathcal{T}_u \), and hence for any \( T \in \mathcal{T}_u \), \( T \subseteq \mathcal{R} \). There is a total order on \( \mathcal{T}_u \) using the chronological order, i.e., more formally, \( \forall a, b \in \mathcal{T}_u \ a \preceq b \Leftrightarrow a.time \leq b.time \).

Location-privacy protection mechanism. Datasets can be modified to offer privacy guarantees by using a location-privacy protection mechanism (or simply a protection mechanism). More formally it is defined as a function \( \mathcal{D} \rightarrow \mathcal{D} \) which produces a new dataset offering privacy guarantees from an input dataset. Protected datasets are also referred to as published datasets.

B. Speed smoothing

As pointed out in Section III, our objective in terms of privacy protection is to hide users’ POIs. These correspond to places where users stop and spend some time, before starting again to move to another place. A trace can be viewed as a list of POIs, that appear on visualized traces as clusters of locations (as shown on Figure 1a), with transitions in between. Our counter-measure to hide POIs is thus to enforce a constant speed in the whole trace of a user, i.e., with speed smoothing. If we can guarantee that speed is constant throughout a trace, it becomes difficult for an adversary to spot where a user stopped because there is no point at which she appears to be stationary. Clues can still be obtained from background knowledge (e.g. the probability is higher to stop in a park that in the middle of a highway) but there will be no certainty for an attacker (e.g. a user can either have just crossed the park or had a picnic there). Moreover, we want to guarantee that there is a constant duration and distance between two successive records in a trace. This prevents an attacker from inferring information by studying spatio-temporal intervals at which traces have been sampled. Figure 1b shows the result of speed smoothing applied to two mobility traces. From this figure, we can see that the POIs of the users have been removed and that records on each trace are regularly spaced. After this step, two records remain unchanged in each trajectory: the first and the last record. Because they are likely to be POIs (e.g., a home), they need to be protected. Our solution is to blur endpoints to reduce the precision around them and hence protect users’ privacy around these places. Figure 1c shows the effect of blurring endpoints.

We now present one implementation of the speed smoothing protection mechanism, i.e., Promesse, depicted in Figure 2. This algorithm takes as input a mobility trace and an \( \epsilon \) parameter. The first step of this algorithm, is to extract regularly spaced locations, each being at a distance \( \epsilon \) from the following one (lines 2-16). The larger \( \epsilon \), the better the privacy guarantee, but the higher the information loss along the trajectory. To perform this sampling, locations are interpolated along segments joining known locations. This means that our method is more suited for traces with a high sampling rate (e.g., ten to thirty seconds between consecutive records). The Interpolate function is not given here because its implementation depends on how locations are represented. After this, we remove the first and last locations from the list (lines 22-23), in
order to help hiding endpoints that would otherwise be easily guessed. More precisely, we reduce the precision by ϵ around there areas. If after this step there is not enough remaining locations left to recreate a valid trace (i.e., two or less), we simply discard the trace (lines 18-20). The second step is then to assign to each of the previously sampled locations a timestamp by uniformly distributing the duration of the original trace (lines 25-32). The implementation of the full Promesse protection mechanism is shown in lines 36-44. The speed smoothing algorithm is independently applied to each mobility trace of a dataset. The algorithm depicted here is sequential but operations can be parallelized if necessary to anonymize large datasets containing a large number of users.

From the above it appears that a variable that might have an important impact on the results of Promesse is ϵ. This value should be chosen according to the granularity of POIs that the data owner wants to hide. Intuitively, we expect POIs whose diameter is smaller than ϵ to be hidden, where the diameter of a POI refers to the diameter of the circular area where all the records related to this POI fall. This parametrization question is not unique to our mechanism. Indeed, differentially-private mechanisms [14] or k-anonymous ones [24] must also be parametrized according to the level of privacy to achieve.

V. EXPERIMENTAL STUDY

We start by describing in Section V-A our experimental settings (i.e., datasets, Promesse implementation and configuration as well as a description of the other studied mechanisms). Then, we describe the evaluation of Promesse in terms of privacy, utility and performance in Section V-B, V-C and V-D, respectively.

A. Experimental settings

Datasets. We study the protection mechanisms using three real-life datasets: the Cabspotting dataset [25], the Geolife dataset [26] and the MDC dataset [27], [28]. Cabspotting has been collected over a month by 536 taxis in the San Francisco Bay Area. The entire dataset is composed of 11M records. Geolife has been collected by Microsoft Research Asia over four years and by 182 users. It is not restricted to people during working hours but instead follows people during their daily life. This dataset includes 25M records, but is quite inconsistent: some people have been tracked during the whole four years whereas others have only contributed for a few hours. MDC has been collected between 2009 and 2011 around Lausanne, Switzerland and involves a total of 185 users, followed during their daily life. The entire dataset is composed of 11M records containing location information.

We pre-processed our datasets to make all traces begin the same day and removed entire days with no data. Then we only kept the first 20 days of data, to have a dataset with a similar duration for all users. We further performed another type of pre-processing on the three datasets. We divided each trace into individual parts, each one being a set of records with no temporal gap between two consecutive records. Specifically, a trace is divided into two parts when no record is logged during four consecutive hours. Each part is then considered as an independent trace associated with a new virtual user identifier, no matter to which logical user it really belongs. This pre-processing helps to preserve privacy, as it breaks the correlation between multiple journeys of a same logical user. We applied this pre-processing for all studied mechanisms to allow a fair comparison. Final figures about datasets after the pre-processing are summarized in Table I.

### Table I: Figures about datasets used in experiments

<table>
<thead>
<tr>
<th>Metric</th>
<th>Cabspotting</th>
<th>Geolife</th>
<th>MDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records</td>
<td>8,986,419</td>
<td>3,804,788</td>
<td>1,124,454</td>
</tr>
<tr>
<td>Traces</td>
<td>5,503</td>
<td>2,475</td>
<td>4,693</td>
</tr>
<tr>
<td>Avg trace duration</td>
<td>32h 44 min</td>
<td>3h 4 min</td>
<td>3h 15 min</td>
</tr>
<tr>
<td>Max trace duration</td>
<td>20 days</td>
<td>1 day 22h</td>
<td>20 days</td>
</tr>
<tr>
<td>Avg sampling rate</td>
<td>72s</td>
<td>7s</td>
<td>32s</td>
</tr>
</tbody>
</table>

Implementation. Our mechanism is implemented in Java 8. To enable the reproduction of our experiments, the source code of Promesse and datasets we use are available on our website.

1. We ran our experiments on a single Debian virtual machine having access to 8 Gb of RAM and 8 cores clocked at 1.8 GHz each.

Configuration. We tested Promesse with various values of the ϵ parameter: 50 (only for Geolife and MDC, because this value is impracticable for Cabspotting), 100, 200 and 500

function SMOOTH_SPEED(T ∈ T_u, ϵ)
points ← \{\}, prev ← ∅
for curr ∈ T do
if prev = ∅ then
points ← points ∪ \{point\}
prev ← curr.loc
else
d ← d_X(curr.loc, prev)
while d ≥ δ do
loc ← INTERPOLATE(prev, curr.loc, ϵ)
points ← points ∪ \{(loc, curr.time)\}
prev ← loc
d ← d_X(curr.loc, prev)
end while
end if
end for
if |points| ≤ 2 then
return \{\} \triangleright Non protectable trace
end if
remove first element of points
remove last element of points
t_min ← min_r∈points r.time
t_max ← max_r∈points r.time
interval ← (t_max − t_min)/(|points| − 1)
output ← \{\}, time ← t_min
for point ∈ points do
output ← output ∪ \{(u, point.loc, time)\}
time ← time + interval
end for
return output
end function

definition 1: The matching operation associates to each location of a given set, the closest location in another set iff its distance w.r.t. d_X is lower than a threshold ℓ:
match_ℓ(P, P') = \bigcup_{p,p' \in P \cap P'} \{p \mid d_X(p, p') ≤ ℓ\}.
definition 2: The recall is the ratio between the number of matching POIs retrieved from a protected dataset and the real number of POIs:
recall_ℓ(P, P') = \frac{|P \cap \text{match}_ℓ(P, P')|}{|P|}.
definition 3: The precision is the ratio between the number of matching POIs retrieved from a protected dataset and the total number of POIs retrieved:
precision_ℓ(P, P') = \frac{|P \cap \text{match}_ℓ(P, P')|}{|P'|}.
definition 4: The F-score is the harmonic mean of precision and recall:
f_\text{score}_ℓ(P, P') = \frac{2 \times \text{precision}_ℓ(P, P') \times \text{recall}_ℓ(P, P')}{\text{precision}_ℓ(P, P') + \text{recall}_ℓ(P, P')}.

For conciseness, we only report on the F-score metric with ℓ = Δℓ/2 = 100m, which takes into account both precision and recall. We report on the average F-score across all traces in each dataset in Table II. From this table, we can see that in the Cabspotting dataset, POIs are always hidden with Promesse, no matter the value of ϵ. With Geolife and MDC, we closely reach our goal with ϵ = 200m. This value of ϵ was indeed
expected to be its optimal parametrization, given that with this value we have $\epsilon = \Delta f$. Still, the F-Score is close but not equal to zero, most likely due to some edge cases. It is worth noting that behind this low F-Score, we have in reality both a low precision and a low recall. This means that very few POIs are retrieved, and that they are lost inside many false positives. Furthermore, we observe that $W4M$ hides all POIs in the three datasets, but this is a consequence of the great quantity of noise that has to be added to enforce $k$-anonymity. Here the privacy comes at the cost of a very degraded utility, as it will be shown in Section V-C. Finally, retrieving POIs from the Geo-I protected datasets is very dependent on the quantity of noise that has been added. With lower values of $\epsilon$, and hence more privacy, retrieving POIs becomes very challenging. With $\epsilon = \ln(2)/200$, almost no POI is found, as with Promesse. Weaker values of $\epsilon$ allow many POIs to be retrieved, up to an F-Score of 55%. We stayed with a simple POIs extraction routine, but risk is even higher if we enlarge the clustering radius $\Delta f$ to adapt to Geo-I, as it has been shown in [16].

### C. Utility evaluation

We evaluate the utility of the protected datasets with two metrics: the spatial error and range queries as described below.

#### Spatial error

We compute the difference between two traces by measuring the distance between each record in the protected trace and its projection onto the original trace. We consider original traces as polylines whose vertices are records. This allows us to fill the gaps between records and project points onto these polylines. The distance between a record in the protected trace and its projection onto the original trace represents the spatial error, i.e., the spatial inaccuracy of a protection mechanism. Please note that we do not take into account the error due to the interpolation between two consecutive locations, which is what our mechanism actually does. We consider the original trace to be precise enough to allow interpolation without losing accuracy.

<table>
<thead>
<tr>
<th>Protection mechanism</th>
<th>Cabspotting</th>
<th>Geolife</th>
<th>MDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promesse, $\epsilon = 50m$</td>
<td>-</td>
<td>17.22 %</td>
<td>13.91 %</td>
</tr>
<tr>
<td>Promesse, $\epsilon = 100m$</td>
<td>0 %</td>
<td>11.06 %</td>
<td>7.94 %</td>
</tr>
<tr>
<td>Promesse, $\epsilon = 200m$</td>
<td>0 %</td>
<td>2.27 %</td>
<td>1.01 %</td>
</tr>
<tr>
<td>Promesse, $\epsilon = 500m$</td>
<td>0 %</td>
<td>0 %</td>
<td>0 %</td>
</tr>
<tr>
<td>Geo-I, $\epsilon = \ln(10)/100$</td>
<td>30.72 %</td>
<td>42.38 %</td>
<td>55.95 %</td>
</tr>
<tr>
<td>Geo-I, $\epsilon = \ln(6)/200$</td>
<td>6.33 %</td>
<td>9.41 %</td>
<td>16.31 %</td>
</tr>
<tr>
<td>Geo-I, $\epsilon = \ln(4)/200$</td>
<td>3.51 %</td>
<td>4.50 %</td>
<td>9 %</td>
</tr>
<tr>
<td>Geo-I, $\epsilon = \ln(2)/200$</td>
<td>1.49 %</td>
<td>0.85 %</td>
<td>0.54 %</td>
</tr>
</tbody>
</table>

| $W4M, k = 2, \delta = 200m$ | 0 % | 0 % | 0 % |

We report about the average spatial error for all records in each dataset in Table III. From this table, we observe that the spatial error of Promesse is equal to zero for the three datasets. Indeed, by construction, the only inaccuracy introduced by Promesse is due to the interpolation between sampled records, which shows to be negligible in this experiment. Geo-I instead adds noise to locations, depending on its $\epsilon$ parameter, which results in an average error ranging from 24 to 378 meters on the three datasets. This has to be compared with the average error due to GPS measurements which is about 15-20 meters. This means that at its weakest level of privacy, Geo-I is just a little bit less precise than the error that could come from the normal usage of a GPS. However, when the level of privacy increases, the error can go as high as 378 meters, which is enough to highly disturb data mining tasks, especially in a dense urban environment. Finally, among the three tested mechanisms, $W4M$ is the one with the worst spatial error, which is at least equal to 13,046 meters in our experiments. This is due to the large amount of noise $W4M$ introduces to enforce k-anonymity. These results highlight the benefit of a time distortion protection mechanism for use cases where a high spatial accuracy is needed, which cannot be achieved with the other mechanisms building on spatial distortion.

#### Range queries

A classical operation performed on published datasets is range queries. A range query counts how many unique (i.e., different) users cross an area during a time window. Despite being simple, this operation can be employed by many useful applications (e.g., traffic prediction, finding popular places). To measure the utility related to range queries, we use the range query distortion metric defined in [10].

**Definition 5:** The distortion associated with a range query $Q$ is the relative error between its result over the original dataset $D \in \mathcal{D}$ and its result over a protected dataset $D' \in \mathcal{D}$:

$$\text{distortion}_{Q}(D, D') = \frac{|Q(D) - Q(D')|}{Q(D)}.$$

In our context, a range query is defined using two parameters: a time window and a geographical area. Because the set of all range queries is infinite, we need to generate many range queries with different time windows and areas centered around different times and locations. Additionally, many range queries are irrelevant for our experiments, because they concern areas or times for which we have no data. This is why we use a smart range query generator choosing randomly one record in the evaluated original dataset and using it to generate a query centered around this record with a random duration and a random area. This ensures that $Q(D) \neq 0$ (and that our distortion is always defined). Similarly to [10], we choose time windows ranging from 2 hours to 8 hours and square areas whose half-diagonals range from 500 to 5,000 meters.

We report about the average query distortion in Table IV, which is the average distortion over 1,000 randomly generated queries. Results show that Promesse has a query distortion ranging from 6 % to 27 % for $\epsilon = 200m$. This means that results of range queries have, on average, a relative error of at most 27 %.

<table>
<thead>
<tr>
<th>Protection mechanism</th>
<th>Cabspotting</th>
<th>Geolife</th>
<th>MDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promesse, $\forall \epsilon$</td>
<td>0 m</td>
<td>0 m</td>
<td>0 m</td>
</tr>
<tr>
<td>Geo-I, $\epsilon = \ln(10)/100$</td>
<td>24 m</td>
<td>45 m</td>
<td>52 m</td>
</tr>
<tr>
<td>Geo-I, $\epsilon = \ln(6)/200$</td>
<td>50 m</td>
<td>120 m</td>
<td>140 m</td>
</tr>
<tr>
<td>Geo-I, $\epsilon = \ln(4)/200$</td>
<td>62 m</td>
<td>156 m</td>
<td>183 m</td>
</tr>
<tr>
<td>Geo-I, $\epsilon = \ln(2)/200$</td>
<td>113 m</td>
<td>325 m</td>
<td>378 m</td>
</tr>
<tr>
<td>$W4M, k = 2, \delta = 200m$</td>
<td>13,046 m</td>
<td>69,676 m</td>
<td>19,222 m</td>
</tr>
</tbody>
</table>

Further, results show that we perform at least 71 % better than $W4M$ with all our datasets. Once again, the distortion with Geo-I is dependent on the value of $\epsilon$. The weakest value features almost no distortion (but also does not protect POIs in a satisfactory way, cf. previous section).
Table IV: Range query distortions.

<table>
<thead>
<tr>
<th>Protection mechanism</th>
<th>Cabspotting</th>
<th>Geolife</th>
<th>MDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promesse, $\epsilon = 50m$</td>
<td>-</td>
<td>15.14 %</td>
<td>25.4 %</td>
</tr>
<tr>
<td>Promesse, $\epsilon = 100m$</td>
<td>6.77 %</td>
<td>14.83 %</td>
<td>25.23 %</td>
</tr>
<tr>
<td>Promesse, $\epsilon = 200m$</td>
<td>6 %</td>
<td>15.1 %</td>
<td>27 %</td>
</tr>
<tr>
<td>Promesse, $\epsilon = 500m$</td>
<td>6.9 %</td>
<td>18.97 %</td>
<td>31.23 %</td>
</tr>
<tr>
<td>Geo-I, $\epsilon = \ln(10)/100$</td>
<td>0.7 %</td>
<td>7.58 %</td>
<td>3.08 %</td>
</tr>
<tr>
<td>Geo-I, $\epsilon = \ln(6)/200$</td>
<td>2.46 %</td>
<td>20.43 %</td>
<td>10.36 %</td>
</tr>
<tr>
<td>Geo-I, $\epsilon = \ln(4)/200$</td>
<td>3.39 %</td>
<td>27.24 %</td>
<td>13.47 %</td>
</tr>
<tr>
<td>Geo-I, $\epsilon = \ln(2)/200$</td>
<td>7.21 %</td>
<td>60.46 %</td>
<td>29.98 %</td>
</tr>
<tr>
<td>$W4M, k = 2, \delta = 200m$</td>
<td>102 %</td>
<td>102 %</td>
<td>94 %</td>
</tr>
</tbody>
</table>

Intermediary values of $\epsilon$ correspond to similar distortions than with the optimal Promesse. Note that the range query distortion metric is more sensitive with Geolife and MDC because counts are way smaller than with Cabspotting, and therefore the effect of missing one user is more important on the relative error. This explains the larger difference of distortion between Geo-I and Promesse on the Geolife and MDC datasets.

D. Performance evaluation

We also consider two non-functional metrics that can also impact the usability of the protected datasets, i.e., the execution time and the compression degree as described below.

Execution time. The execution time is the time taken by the anonymization process to generate a new protected dataset. We assume the original dataset is already stored in the file system (e.g., a database or a distributed file system) and only measure the time to generate and write the protected dataset in a text file. We report about execution times in Table V. Because execution times remain almost constant for the various configurations of Geo-I and Promesse, we do not report about subtle variations depending on parameters. Geo-I and Promesse are the fastest mechanisms because their algorithms are quite simple. Geo-I independently adds noise to each record and Promesse independently protects each trace, which enables them to be efficient. $W4M$ is the slowest mechanism because of the complexity coming from the clustering of similar traces that is the heart of this mechanism.

Table V: Execution times.

<table>
<thead>
<tr>
<th>Protection mechanism</th>
<th>Cabspotting</th>
<th>Geolife</th>
<th>MDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Promesse</td>
<td>210 s</td>
<td>25 s</td>
<td>15 s</td>
</tr>
<tr>
<td>Geo-I</td>
<td>147 s</td>
<td>64 s</td>
<td>21 s</td>
</tr>
<tr>
<td>$W4M$</td>
<td>605 min</td>
<td>70 min</td>
<td>37 min</td>
</tr>
</tbody>
</table>

Compresson. In a general context, the anonymized dataset can have a different size than the original dataset. It can be smaller, if records have been deleted, or larger, if fake records have been added. Producing datasets that are orders of magnitude larger than the original ones greatly decreases their usability, because the time needed to load and query them increases accordingly, while much smaller ones can introduce information losses (that could be quantified with the previous utility metrics).

Definition 6: We define the compression degree as the ratio between the size of a protected dataset $D' \in D$ and the size of the non-empty original dataset $D \in D$:

$$\text{compression}(D, D') = \frac{|D'|}{|D|}.$$
research in this paper used the MDC Database made available to the work supported by the LABEX IMU (ANR-10-LABX-0088) of Université de Lyon, within the program "Investissements d’Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). Portions of the research in this paper used the MDC Database made available by Idiap Research Institute, Switzerland and owned by Nokia.

ACKNOWLEDGMENT

This work was supported by the LABEX IMU (ANR-10-LABX-0088) of Université de Lyon, within the program "Investissements d’Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). Portions of the research in this paper used the MDC Database made available by Idiap Research Institute, Switzerland and owned by Nokia.

REFERENCES