Hiding the start of Brownian motion: towards a Bayesian analysis of privacy for GPS trajectories

SIRIO LERAMANTI
Bocconi University, Milan
sirio.legramanti@phd.unibocconi.it
June 26, 2018

Abstract
The diffusion of GPS sensors and the success of applications for sharing GPS trajectories raise serious privacy concerns. In this paper, we show that a Bayesian approach is natural for a rigorous analysis of both home identification attacks and their countermeasures. Our Bayesian framework allows to naturally incorporate the adversary’s background knowledge and quantify the bias and level of uncertainty after the attack. We propose measures for both utility and privacy: while the first is by definition application-specific, the second extends beyond the present application and can be regarded as a Bayesian measure of privacy. Based on our utility measure, we restrict to “privacy region cut strategies”, a family of countermeasures consisting in publishing the trajectories from the first exit to the last entrance from/into a privacy region. We run experiments on Brownian motion trajectories for two of these strategies, showing that our generalization of the previously proposed “two balls strategy” performs better than “random radius strategy”, which in turn generalizes a strategy currently employed in industry. Beyond the location privacy application, the problem of hiding the start of Brownian motion is of interest in itself, with possibly many other applications.

KEYWORDS: Bayesian methods, Evaluation framework, Harmonic measure, Inference attack, Location privacy, Privacy metric

I. Introduction
The integration of GPS sensors in smartphones, smartwatches, bike computers and sport watches has made easy for everybody to record GPS trajectories of fitness activities, for example running and cycling. Sport social networks such as Strava, Mapmyride, Runtastic, etc. allow people to share these GPS trajectories with followers. Unfortunately, sharing this kind of data with strangers poses serious privacy concerns.

In fact, whoever has access to a user’s GPS trajectories can retrieve his home address [12, 6, 10]. Moreover, some sport social networks encourage users to add information about their equipment, for example make and model of their bike. According to UK Police, thieves are exploiting these data to locate garages that are worth a visit [16]. Sharing GPS trajectories endangers not only expensive sport equipment but also people, especially if we consider that GPS trajectories contain temporal information. In fact, each latitude-longitude couple is paired with a timestamp of the date and time it was recorded. Hence, anybody who has access to a user’s GPS trajectories, including burglars and stalkers, could learn a user’s habits and infer when the house is empty, when somebody is home alone or jogging by himself in an isolate place, etc. These concerns get
even more serious if we consider the possibility of sharing activities in real time. Such risks are not limited to private citizens. Recently, a huge outcry has been caused by the discovery that aggregated data (heatmaps) from fitness apps can reveal military sites [19]. The problem can be tackled from two sides: on one hand, each user should select carefully who can access his GPS trajectories; on the other hand, sport social networks should limit the amount of potentially dangerous information that is given away while sharing GPS trajectories. The first approach doesn’t need any special algorithm, it just requires that each user shares his activities only with people that he trusts, but this is somehow against the nature of online social networks and relies on the users’ privacy awareness, which has been shown to be not much reliable [11]. The second approach is then preferable and necessary. It requires algorithms that take as input an original trajectory and give as output a modification of it from which is harder to infer sensible information. These algorithms, that we will call obfuscation strategies, range from publishing nothing to publishing everything as it is. These two extremes are both undesirable, in opposite ways that highlight the tradeoff between privacy and utility, which is the value of data for the considered application. In fact, publishing nothing guarantees perfect privacy but null utility, while full disclosure provides maximum utility but usually this comes at the cost of insufficient privacy.

In the present work we will focus on protection from home identification attacks, which consist in the attempt of an adversary to localize the house of a user, in this case based on his GPS trajectories.

The remainder of this paper is organized as follows. In Section II we review some relevant literature. In Section III we model home identification attacks as Bayesian inference problems and define the house hiding problem as an optimization problem. In Section IV we assess the performance of two obfuscation strategies on Brownian motion trajectories. Finally in Section V we draw conclusions.

i. Relevant literature

Home identification attacks based on GPS trajectories have already been treated in location privacy literature, but none of the previous works has provided at the same time a modelling of the adversary’s background knowledge, a modelling of the attack as rigorous statistical inference, uncertainty quantification for the adversary’s estimate of the user’s house location, an updating mechanism for such estimate as more trajectories are released, a definition of utility for the considered application, the tuning of obfuscation strategy parameters as a well defined optimization problem.

In [10], for example, home identification attacks are based on four different heuristics. The first three output a point estimate: the median of final points, the centroid of the largest cluster of final points, the weighted median of all points with dwell time as weight. The fourth heuristic, instead, outputs a probability distribution over the map, with the underlying idea that if you are somewhere at a time that you are usually home, then that location is likely to be home. The efficacy of these attacks is evaluated in terms of distance between estimated and true house location. The first three heuristics exhibit median errors of 60.7, 66.6 and 66.6 meters respectively, which are close to the median distance between two consecutive points in the trajectories from this dataset (about 65 meters). The last heuristic is the least successful, with median error exceeding 2 kilometers. This may be due to the fact that it relies on the timestamps of points instead of their position within each trajectory, and is based on the assumption that subjects’ schedule is very stable. The study also experiments with a few countermeasures to home identification attacks, among which only spatial cloaking can be applied to our case, because the others would correspond
Spatial cloaking is a data suppression technique, consisting in hiding the part of the trajectory within a privacy region [4]. As privacy region, the author rules out a ball of fixed radius centered in the user’s house and this is our very same starting point. He instead proposes a ball of radius $R$ and random center $c$ uniformly distributed within a ball of radius $r < R$ centered in the user’s house location. We will generalize this obfuscation strategy in Section III.ii. The author then experiments on the dependence of privacy, measured through the number of incorrect inferences, on the strategy parameters $r$ and $R$. These experiments, though, do not take into account that the adversary may have some background knowledge about the user’s house location and that he may know or infer how the adopted countermeasure works, and this may result in underestimating the effectiveness of the attacks.

In [6] authors experiment with another countermeasure to home identification attacks: reducing the sampling frequency. Like spatial cloaking, this a data suppression technique, but the suppression rule is temporal instead of spatial and affects the whole trajectory. The simulated attack is based on clustering of low-speed points and incorporates some prior knowledge in the form of a target region selected by the adversary. The output is a list of likely home locations, with no ordering in probability. The performance of the attack is evaluated in terms of precision/recall. While at 1 minute frequency the recall is about 85%, it drops to about 40% for the 4 minutes frequency. Precisions are not so far apart, at about 60% and 50% respectively. A proper probability distribution as output of the attack is rare but not completely new to privacy literature. In [21], for example, the output of the attack is the probability distribution of the user’s position at a fixed time, over a map discretized in regions, given one obfuscated trajectory. What is missing is a prior distribution, an updating mechanism when multiple trajectories from the same user are available and a definition of utility of published trajectories. A Bayesian approach is rare but not completely new in privacy literature. For example, [8, 13, 1] take advantage of the possibility of modeling the background knowledge of the adversary through the prior in order to give new privacy definitions. The Bayesian approach naturally provides uncertainty quantification for the adversary’s inference and handles the dynamic release of trajectories through the posterior updating mechanism. This is crucial for our application, where GPS trajectories are shared one by one, after each activity is completed. In privacy literature, instead, datasets of trajectories are usually static, in the sense that no new trajectory is to be added, and they are meant to be released as a whole. Hence, utility is typically defined, when it is defined at all, as the preservation of aggregate properties of the dataset [5, 17, 14]. For our application, instead, utility must depend on the preservation of the relevant features of each single trajectory.

II. THE HOUSE HIDDING PROBLEM

We assume and try to model the following scenario:

• after a fitness activity is completed, the user uploads the corresponding GPS trajectory on a sport social network;
• before the publication, the trajectory can be altered through an obfuscation strategy to hide the user’s house location;
• all trajectories from the same user are published under the same permanent pseudonym and are visible to any of his followers;
• one or more of the user’s followers may act as an adversary and make an home identification attack based on the published trajectories;
• the adversary knows or has inferred the obfuscation algorithm and may have some background knowledge about the user’s house location, but has no access to the original trajectories.
We model both original and published trajectories as continuous time stochastic processes taking values in \( \mathbb{R}^2 \). GPS devices actually sample at discrete time, but the sampling frequency is sufficiently high to make this choice sensible (1 Hz is a popular option on GPS bike computers). We denote:

- the user’s house location with \( \theta \);
- his original trajectories with \( \{ x_i^{(i)} \}_{i \in [0,T]} \), \( i = 1, \ldots, n \);
- the corresponding published trajectories with \( \{ y_i^{(i)} \}_{i \in [0,T]} \), \( i = 1, \ldots, n \).

We model the home identification attack as a parametric Bayesian inference problem described by the following hierarchical model:

\[
\theta \sim \pi(\theta), \tag{1}
\]

\[
\{ x_i^{(i)} \}_{i \in [0,T]} \mid \theta \sim p(\{ x_i^{(i)} \}_{i \in [0,T]} \mid \theta), \quad i = 1, \ldots, n \tag{2}
\]

\[
\{ y_i^{(i)} \}_{i \in [0,T]} \mid \theta, \{ x_i^{(i)} \}_{i \in [0,T]} \sim p(\{ y_i^{(i)} \}_{i \in [0,T]} \mid \theta, \{ x_i^{(i)} \}_{i \in [0,T]}), \quad i = 1, \ldots, n \tag{3}
\]

where:

- \( \pi \) is a probability distribution over \( \mathbb{R}^2 \) acting as the prior;
- \( p(\{ x_i^{(i)} \}_{i \in [0,T]} \mid \theta) \) is a model for original trajectories, that act as latent variables;
- \( p(\{ y_i^{(i)} \}_{i \in [0,T]} \mid \theta, \{ x_i^{(i)} \}_{i \in [0,T]} \) models the obfuscation strategy and acts as the likelihood function;

The posterior distribution of \( \theta \) given the first \( n \) published trajectories is denoted by \( \pi_n \):

\[
\theta \mid \{ y_i^{(1)} \}_{i \in [0,T]}, \ldots, \{ y_i^{(n)} \}_{i \in [0,T]} \sim \pi_n(\theta). \tag{4}
\]

The house hiding problem consists in identifying the obfuscation strategy that maximizes privacy among all the strategies ensuring a given level of utility. In the present work, we limit ourselves to comparing two obfuscation strategies with respect to privacy under a constraint on utility.

### i. Privacy measure

Since we model a privacy attack as a point estimation problem, privacy can be measured by the quality of the adversary’s estimate. In the present work, we adopt a quadratic loss function and we then employ the Mean Square Error (MSE) as a measure of the quality of the adversary’s estimate and hence as a measure of privacy. The MSE incorporates the concepts of incorrectness and uncertainty through the well known bias/variance decomposition:

\[
MSE = E \left[ \| \theta - \theta_{true} \|^2 \right] = E \left[ \| \theta - E [\theta] \|^2 \right] + \| E [\theta] - \theta_{true} \|^2 = Var(\theta) + Bias(\theta)^2, \tag{5}
\]

where \( \| \cdot \| \) is the Euclidean norm in \( \mathbb{R}^2 \) and all the expectations are taken with respect to the posterior \( \pi_n \) defined in \( 4 \). Since the posterior will not be analytically derived but rather sampled from, expectations in \( 5 \) will be estimated through posterior sample means.

Note that incorrectness and uncertainty have been previously discussed as privacy measures, for example in \( 21 \), where incorrectness corresponds to bias while uncertainty is measured by the entropy of the estimated posterior.
ii. Utility measure

Utility is the value of data for the considered application, hence our utility measure depends on what we regard as valuable for our application. On social networks, fake times or locations cannot be allowed, since they would alter rankings. Hence, trajectories can be cut, but what is published must be authentic. In particular, users may accept cuts to the beginning and the end of their trajectories, since these segments likely consist of how they get out of their neighborhood and back in. Still, we assume that users want these cuts to be as small as possible.

Based on these assumptions, we define the utility \( U \) of a perturbed trajectory \( \{y_t\}_{t \in [0,T]} \) given the corresponding original trajectory \( \{x_t\}_{t \in [0,T]} \) as:

\[
U \left( \{y_t\}_{t \in [0,T]} \mid \{x_t\}_{t \in [0,T]} \right) = f \left( \|y_0 - x_0\|^2 + \|y_T - x_T\|^2 \right) \cdot 1 \{ \{y_t\}_{t \in [0,T]} = \{x_t\}_{t \in [t_1,t_2]} \} \tag{6}
\]

with \( f : \mathbb{R}^+ \rightarrow \mathbb{R} \) monotonically decreasing, \( 0 \leq t_1 \leq t_2 \leq T \) and \( T = t_2 - t_1 \). When positive, utility depends on the original and published trajectory only through the squared perturbation \( (SP) \), where

\[
SP = SP \left( \{y_t\}_{t \in [0,T]} \mid \{x_t\}_{t \in [0,T]} \right) = \|y_0 - x_0\|^2 + \|y_T - x_T\|^2. \tag{7}
\]

The variability of original trajectories and the stochasticity of the obfuscation strategy make the \( SP \) random, with distribution \( p^{(s)}_{SP} \) under strategy \( s \). We then define the utility of a generally non-deterministic obfuscation strategy \( s \) as

\[
U (s) = \phi \left( p^{(s)}_{SP} \right) \cdot 1 \{ \{y_t\}_{t \in [0,T]} = \{x_t\}_{t \in [t_1,t_2]}, s \text{ a.s.} \}, \tag{8}
\]

where \( \phi \) is a real-valued functional on the space of probability distributions on \( \mathbb{R}^+ \), \( 0 \leq t_1 \leq t_2 \leq T \) and \( T = t_2 - t_1 \). The expression \( s \text{ a.s.} \) refers to the probability distribution \( (3) \) induced by strategy \( s \). Since we will not quantify utility but just set a level for it, we don’t need to choose a specific functional \( \phi \): we can set the level of utility by fixing the distribution of the \( SP \).

iii. Privacy region cut strategies

The only perturbed trajectories with positive utility are those obtained cutting the initial and/or final part of the original trajectory. If original trajectories start and finish at the user’s house location \( \theta \), this can be achieved by publishing only the part of the trajectory between the first exit and the last entrance from/into a privacy region containing \( \theta \). We will focus on this class of strategies, that we will call privacy region cut strategies.

Instances of this class differ for the choice of the distribution of the privacy region \( D_i \) and are specified by the following hierarchical structure:

\[
D_i \mid \theta, \{x_i^{(j)}\}_{t \in [0,T]} \sim p(D_i \mid \theta, \{x_i^{(j)}\}_{t \in [0,T]}), \quad i = 1, \ldots, n \tag{9}
\]

\[
t_i^{(j)} \mid D_i, \theta, \{x_i^{(j)}\}_{t \in [0,T]} = \inf\{t \in [0,T] : x_i^{(j)} \notin D_i\}, \quad i = 1, \ldots, n \tag{10}
\]
If the infimum in (10) is over an empty set, i.e., the original trajectory never leaves the privacy region, no trajectory is published, otherwise:

$$t^{(i)}_2 | D_i, \theta, \{x^{(i)}_t\}_{t \in [0,T^{(i)}]} = \sup \{ t \in [0,T^{(i)}] : x^{(i)}_t \notin D_i \}, \quad i = 1, \ldots, n$$

(11)

$$\tilde{T}^{(i)} | t^{(i)}_1, t^{(i)}_2, D_i, \theta, \{x^{(i)}_t\}_{t \in [0,T^{(i)}]} = t^{(i)}_2 - t^{(i)}_1, \quad i = 1, \ldots, n$$

(12)

$$\{y^{(i)}_t\}_{t \in [0,\tilde{T}^{(i)}]} | \tilde{T}^{(i)}, t^{(i)}_1, t^{(i)}_2, D_i, \theta, \{x^{(i)}_t\}_{t \in [0,T^{(i)}]} = \{x^{(i)}_{t^{(i)}_1+t}\}_{t \in [0,\tilde{T}^{(i)}]}, \quad i = 1, \ldots, n$$

(13)

Note that $D_i$ may be different for each trajectory and its distribution may depend on the user’s house location $\theta$ and on the original trajectory $\{x^{(i)}_t\}_{t \in [0,T^{(i)}]}$: The dependence on $\theta$ may be useful to have a larger privacy radius in sparsely populated areas, in order to avoid that the privacy ball contains too few houses other than the user’s one. The dependence on the original trajectory may allow to design interactive obfuscation strategies. In the present work we will not implement any of these two dependences, leaving this research direction to future work, but our framework is totally ready to accommodate it.

### III. Simulations on Brownian motion

Analyzing a home identification attack within the framework of Section II requires to specify a model for original trajectories in (2). Our framework poses no limitation to the complexity of such model, but we start with one of the simplest possible models, assuming that, given the user’s house location $\theta$, original trajectories are i.i.d. as a Brownian motion (BM) starting in $\theta$. Original trajectories actually have finite duration, as stated in (2), but in this section, for simplicity, we will consider the time index $t$ ranging in $\mathbb{R}^+$.

$$\{x^{(i)}_t\}_{t \in \mathbb{R}^+} | \theta \sim BM(\theta), \quad i = 1, \ldots, n.$$

(14)

Even if the behavior of vehicles along roads is not a BM, the results obtained under this simplified model can give useful insights about obfuscation strategies. Moreover, the problem of hiding the start of Brownian motion has mathematical and probabilistic interest in its own and may have applications other than the present.

Under (14), thanks to the memoryless property of BM, we can consider a simplified version of privacy region cut strategies, consisting in cutting just the part of the trajectory before the time $t^{(i)}$ of first exit from the privacy region $D_i$. A privacy region cut strategy for BM is then defined as:

$$D_i | \theta, \{x^{(i)}_t\}_{t \in \mathbb{R}^+} \sim p(D_i | \theta, \{x^{(i)}_t\}_{t \in \mathbb{R}^+}), \quad i = 1, \ldots, n$$

(15)

$$t^{(i)} | D_i, \theta, \{x^{(i)}_t\}_{t \in \mathbb{R}^+} = \inf \{ t \in \mathbb{R}^+ : x^{(i)}_t \notin D_i \}, \quad i = 1, \ldots, n$$

(16)

$$\{y^{(i)}_t\}_{t \in \mathbb{R}^+} | t^{(i)}, D_i, \theta, \{x^{(i)}_t\}_{t \in \mathbb{R}^+} = \{x^{(i)}_{t^{(i)}+t}\}_{t \in \mathbb{R}^+}, \quad i = 1, \ldots, n$$

(17)

The starting points of published trajectories coincide with the exit points of the corresponding original trajectories from the privacy region. We just denote them by $y_i$:

$$y_i \doteq y^{(i)}_0 = x^{(i)}_{t^{(i)}}, \quad i = 1, \ldots, n.$$

(18)
Under (14), if we assume that privacy regions $D_i$’s are a.s. bounded domains containing $\theta$, we have that:

(i) $t(i)$ is a stopping time for $\{x^{(i)}_t\}_{t \in \mathbb{R}^+}$ [15] Remark 2.14 and is a.s. finite [3] Chapter 4.3;

(ii) $\{y^{(i)}_t\}_{t \in \mathbb{R}^+} \mid t(i), D_i, \theta, \{x^{(i)}_t\}_{t \in \mathbb{R}^+} \sim \text{BM}(y_i)$ independent of $\{x^{(i)}_t\}_{t \in [0,t(i)]}$ [15] Theorem 2.16;

(iii) exit points $\{y_i\}_{i=1}^n$ are sufficient statistics for $\theta$;

(iv) by Kakutani’s theorem [7], the distribution of exit points is the harmonic measure on the boundary of $D$ parametrized by $\theta$. We will denote such distribution with $\mathcal{H}_\theta^D$.

In conclusion, assuming Brownian motion as model for original trajectories, home identification attacks against privacy region cut strategies can be modelled as parametric Bayesian inference problems with the following hierarchical structure, which is a special case of the general framework given in Section [11]:

\[ \theta \sim \pi(\theta), \]  
\[ \{x^{(i)}_t\}_{t \in \mathbb{R}^+} \mid \theta \sim \text{BM}(\theta), \]  
\[ i = 1, \ldots, n \]  
\[ D_i \mid \theta, \{x^{(i)}_t\}_{t \in \mathbb{R}^+} \sim p(D_i \mid \theta, \{x^{(i)}_t\}_{t \in \mathbb{R}^+}), \]  
\[ i = 1, \ldots, n \]  
\[ y_i \mid D_i, \theta, \{x^{(i)}_t\}_{t \in \mathbb{R}^+} \sim \mathcal{H}_\theta^D, \]  
\[ i = 1, \ldots, n \]  
\[ \theta \mid y_1, \ldots, y_n \sim \pi_n(\theta). \]

We now analyze two privacy region cut strategies, corresponding to different choices for the distribution of privacy regions in (21).

i. Random radius strategy

In random radius strategy, privacy regions are balls of random radius centered in the user’s house:

\[ D_i = B(\theta, r_i), \quad i = 1, \ldots, n. \]  

Hence, this strategy is completely specified by the distribution of the privacy radius $r_i$, that may in general depend on the user’s house location $\theta$ and on the original trajectory $\{x^{(i)}_t\}_{t \in \mathbb{R}^+}$. As anticipated in Section [II], in the present work we will not implement such dependences. A particular case of this strategy is the fixed radius strategy, in which $r_i \sim \delta_{r^*}$, for some fixed $r^* \in \mathbb{R}^+$. We will not analyze this strategy, since it allows to locate $\theta$ with just three distinct trajectories. Instead, we let $r_i$ be random. To make transformations easier, we equivalently consider the distribution of $r^2_i$. If we choose $r^2_i$ to be distributed according to a $\chi^2(2)$, through the Box-Muller transform [2] we get:

\[ r^2_i \mid \theta, \{x^{(i)}_t\}_{t \in \mathbb{R}^+} \sim \chi^2(2) \iff y_i \mid \theta \sim \mathcal{N}_2(\theta, \mathbb{I}_2). \]  

By closure of the exponential distribution under scaling by a positive factor:

\[ r^2_i \mid \theta, \{x^{(i)}_t\}_{t \in \mathbb{R}^+} \sim \text{Exp}(\text{mean} = 2\sigma^2) \iff y_i \mid \theta \sim \mathcal{N}_2(\theta, \sigma^2\mathbb{I}_2). \]

Inference in this normal noise setting is easy and potentially conjugate, but our goal is to make inference as hard as possible for the adversary. Moreover, we may want the mode of the distribution
of $r_i^2$ to be further from zero, in order to avoid that the majority of the trajectories remain almost unaltered. This can be obtained generalizing the exponential distribution in (26) to a Gamma, which corresponds to

$$r_i^2 | \theta, \{x_i^{(t)}\} \sim \text{Gamma}(\alpha, \beta) \iff f(y_i | \theta) = \frac{\beta^\alpha}{\pi \Gamma(\alpha)} ||y_i - \theta||_2^{2(\alpha - 1)} e^{-\beta ||y_i - \theta||_2^2}, \quad y_i \in \mathbb{R}^2.$$  (27)

Finally note that, for the random radius strategy, the square perturbation (SP, see Section II.ii) coincides with $r_i^2$, hence its distribution is directly available.

ii. Two balls strategy

In two balls strategy, the privacy region is, for all trajectories of the same user, the same ball of fixed radius $R$ and random center $c$ with distribution $p_c$ within a smaller ball of radius $r < R$ centered in the user’s house location $\theta$.

Note that $r < R$ ensures that $\theta \in B(c, R)$, since $d(\theta, c) < r < R$. The distribution $p_c$ may in general depend on the user’s house location $\theta$ and on the original trajectory $\{x_i^{(t)}\}$, but, as anticipated in Section II.iii, in the present work we will not implement such dependences.

This strategy generalizes the one proposed in [10], where $c$ was uniformly distributed within the smaller ball, while we will consider the following distribution for $c$:

$$c = \theta + r \rho \begin{bmatrix} \cos \tau \\ \sin \tau \end{bmatrix};$$  (28)

$$\tau \sim \mathcal{U}(-\pi, \pi);$$  (29)

$$\rho^2 \sim \text{Beta}(\alpha, \beta).$$  (30)

Distributions of $\rho^2$ concentrated near 0 or 1, corresponding to $p_c$ concentrated near $\theta$ and $\partial B(\theta, r)$ respectively, should be avoided. In fact, a $p_c$ concentrated near $\theta$ leads almost back to the fixed radius strategy, that we have shown to be inadequate. If instead $p_c$ is concentrated near $\partial B(\theta, r)$ and moreover $r \approx R$, then $\theta$ is with high probability close to the boundary of the privacy ball, leading to a very concentrated distribution of the exit points. Both phenomena are captured by the distribution of the square perturbation (SP, see Section II.ii), which concentrates near $R^2$ in the first case and near zero in the second case.

iii. Experiments and results

In this section, we simulate home identification attacks against the obfuscation strategies described above, with i.i.d. realizations of a Brownian motion starting from the home location as original trajectories, following equations (19)-(23).

Since we did not implement the dependence of the strategies on the house location $\theta$, in this work strategies are translation invariant. Hence we will just fix $\theta$ in the origin. We do not generate the whole original and perturbed trajectories but just the exit points from the privacy region, that we have shown to be sufficient statistics for $\theta$. Since for these simulations there is no background information from a map, we will employ a uniform prior over the plane, coherently with Laplace’s principle [9]. The posterior is not computed exactly but rather sampled from using PyStan [20]. On such sample from the posterior, we compute a Monte Carlo estimate of variance and squared bias as defined in (5). Since both quantities depend on the generated data, we perform multiple experiments for each setting and then consider the median together with the 2.5% and 97.5% percentiles. We expect the median of both quantities to decay to zero as the number of published
Figure 1: Squared Perturbation (SP) theoretical pdf, posterior variance and squared bias for a random radius strategy with \( SP = r^2 \sim \text{Gamma}(\alpha, \beta) \) with different choices for \( \alpha \) and \( \beta \) such that: (a) the SP variance is constant, (b) the SP expectation is constant. For each couple \((\alpha, \beta)\) we have run 50 experiments and plotted the median (solid) together with 2.5% and 97.5% quantiles (dashed).

trajectories grows, even if they are not monotonically decreasing for each single realization. We will focus on the speed of such decay for different choices of the parameters in each strategy.

We first experiment with a random radius strategy with \( SP = r^2 \sim \text{Gamma}(\alpha, \beta) \). The strategy parameters \( \alpha \) and \( \beta \) are in a bijection with mean and variance of the SP distribution, which are more interpretable, hence we act on them. In Figure 1a we can observe that, keeping the SP variance fixed, the larger the SP mean the faster the decay of posterior variance, surprisingly in contrast with the usual tradeoff between privacy and utility, while the squared bias decays similarly. More predictably, when the SP mean is kept fixed, the smaller the SP variance, the faster the decay of both posterior variance and squared bias (see Figure 1b). This is natural, since an SP variance equal to zero corresponds to a fixed radius strategy, which brings to null posterior variance after just 3 published trajectories.

We then consider a two balls strategy with the center of the privacy ball distributed according to (28)-(30). Unlike for random radius strategy, in this case the SP distribution is not directly available, hence we work on the Gaussian KDE of the SP pdf obtained from our experiments through \texttt{sklearn.neighbors.KernelDensity} class within \texttt{scikit-learn} [18]. Since the center \( c \) of the privacy ball is fully revealed after just 3 exit points are released and we are mostly interested in the decay of privacy as the number of published trajectories grows (larger than 3), we pass \( c \) to \texttt{PyStan} as data. Inference is then carried out on the starting point \( \theta \) only. The strategy parameters are now \( r, R, \alpha \) and \( \beta \), with \( 0 < r < R \), \( \alpha \) and \( \beta \) both positive. In Figure 2a we can observe that the larger the radius \( r \) within which the center of the privacy ball is distributed, with the other parameters fixed, the smaller the SP and the faster the decay of both posterior variance and squared bias, with the usual utility-privacy tradeoff. Instead, the larger the radius \( R \) of the privacy ball, with the other parameters fixed, the larger the SP and the slower the decay of both posterior variance and bias, with the usual utility-privacy tradeoff (see Figure 2b). We finally act on \( \alpha \) and \( \beta \), keeping \( r \) and \( R \) fixed. Note that \( \alpha < \beta \) implies a distribution of \( c \) more concentrated near \( \theta \), while \( \alpha > \beta \) implies a distribution of \( c \) more concentrated near the boundary of \( B(\theta, r) \). This latter case, when moreover \( r \approx R \) corresponds to \( \theta \) likely close to the boundary of the privacy ball, resulting in a very concentrated distribution of exit points. This explains why for \((\alpha, \beta) = (5, 2)\) we observe a smaller SP and a faster decay of the posterior variance, with the usual utility-privacy tradeoff (see Figure 2c).
Figure 2: KDE of Squared Perturbation (SP) pdf, posterior variance and squared bias for a two balls strategy with different choices for: (a) $r$, (b) $R$, (c) $(\alpha, \beta)$. For each parameter combination we have run 10 experiments and plotted the median (solid) together with 2.5% and 97.5% quantiles (dashed).

Figure 3: Squared Perturbation (SP) pdf, posterior variance and squared bias for two balls strategy and random radius strategy. Strategy parameters have been chosen to match the first two moments of the SP distributions. For each strategy we have run 50 experiments and plotted the median (solid) together with 2.5% and 97.5% quantiles (dashed).
We finally compare the two abovementioned strategies. For the comparison to be fair, the utility should be the same and hence the SP distribution should match. Since the theoretical SP distribution is directly available only for the random radius strategy, we set the parameters $\alpha$ and $\beta$ of the random radius strategy such that the first two moments of the theoretical SP distribution for this strategy match the first two sample moments for a two balls strategy with parameters set to $r = 1.5$, $R = 3$, $\alpha = 2$, $\beta = 2$. These may not be the parameter values guaranteeing the absolute best performance for the two balls strategy, but they avoid its main drawbacks. In Figure 3 we can observe that with two balls strategy the posterior variance decays slower than with random radius strategy. As for the squared bias, there is a large overlap of confidence bands, but still the median decays slower with two balls strategy than with random radius strategy. We can then conclude that, at least for the distributional and parameter choices made in the present work, the two balls strategy preserves privacy longer than the random radius strategy.

Even if running time is not the focus of the present work and hence our code was not optimized for it, we were able to make a home identification attack based on 50 trajectories in about 1 second on a regular laptop (Intel Core i7-3632QM CPU @ 2.20GHz x 8, 7.7 GB of RAM) with running time scaling linearly in the number of trajectories (see Figure 4). The modest amount of time and computational resources needed for these attacks confirms that the risk for users’ privacy is real.

IV. Conclusions

Home identification attacks based on GPS trajectories have already been treated in location privacy literature, but none of the previous works has provided at the same time a modelling of the adversary’s background knowledge, a modelling of the attack as rigorous statistical inference, uncertainty quantification for the adversary’s estimate of the user’s house location, an updating mechanism for such estimate as more trajectories are released, a definition of utility for the considered application, the tuning of obfuscation strategy parameters as a well defined optimization problem. By modelling home identification attacks as Bayesian inference problems, our framework allows to naturally address these issues and to simulate home identification attacks against non-deterministic obfuscation strategies. In order to compare different obfuscation strategies, we have defined a measure of utility for the present application and a measure of privacy, that instead extends beyond this application and can be regarded as a Bayesian measure of privacy. Based on our utility measure, we have restricted to a class of obfuscation strategies, privacy region cut strategies, consisting in publishing the trajectories from the first exit to the last entrance from/into

![Figure 4: Running time (in seconds) as a function of the number of published trajectories. For each strategy we have run 50 experiments and plotted the median (solid) together with 2.5% and 97.5% quantiles (dashed).](image)
a privacy region. We have proposed two instances of this class of strategies, random radius strategy and two balls strategy, with the latter generalizing a strategy proposed in [10]. Our framework poses no limitation to the complexity of the model for original trajectories, but in the present work we have employed one of the simplest possible models, a Brownian motion starting at the user’s house. Even if the behavior of vehicles along roads is not a BM, the results obtained under this simplified model can give useful insights about obfuscation strategies. On Brownian motion trajectories, the two balls strategy outperformed the random radius strategy. Since the latter is a generalization of a strategy currently employed in industry, the impact of the present work on the motivating application could be immediate and effective. Beyond this application, the problem of hiding the start of Brownian motion has mathematical and probabilistic interest in its own and may have applications other than the present.

**Acknowledgments**

I would like to thank Nelly Litvak for encouraging me to address this problem and Giacomo Aletti, Daniele Durante and Antonio Lijoi for their helpful comments on a first version of this work.

**References**

[1] R. Bassily, A. Groce, J. Katz, and A. Smith. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *Foundations of Computer Science (FOCS), IEEE 54th Annual Symposium on*, pages 439–448. IEEE, 2013.

[2] G. E. Box, M. E. Muller, et al. A note on the generation of random normal deviates. *The annals of mathematical statistics*, 29(2):610–611, 1958.

[3] K. L. Chung. *Lectures from Markov processes to Brownian motion*, volume 249. Springer Science & Business Media, 2013.

[4] M. Gruteser and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. In *Proceedings of the 1st international conference on Mobile systems, applications and services*, pages 31–42. ACM, 2003.

[5] X. He, G. Cormode, A. Machanavajjhala, C. M. Procopiuc, and D. Srivastava. DPT: differentially private trajectory synthesis using hierarchical reference systems. *Proceedings of the VLDB Endowment*, 8(11):1154–1165, 2015.

[6] B. Hoh, M. Gruteser, H. Xiong, and A. Alrabady. Enhancing security and privacy in traffic-monitoring systems. *IEEE Pervasive Computing*, 5(4):38–46, 2006.

[7] S. Kakutani. Two-dimensional Brownian motion and harmonic functions. *Proceedings of the Imperial Academy*, 20(10):706–714, 1944.

[8] S. P. Kasiviswanathan and A. Smith. On the ‘semantics’ of differential privacy: A Bayesian formulation. *Journal of Privacy and Confidentiality*, 6(1):1–16, 2014.

[9] R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

[10] J. Krumm. Inference attacks on location tracks. *Pervasive computing*, pages 127–143, 2007.
[11] J. Krumm. A survey of computational location privacy. *Personal and Ubiquitous Computing*, 13(6):391–399, 2009.

[12] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition. In *Advances in Neural Information Processing Systems*, pages 787–794, 2006.

[13] A. Machanavajjhala, J. Gehrke, and M. Götz. Data publishing against realistic adversaries. *Proceedings of the VLDB Endowment*, 2(1):790–801, 2009.

[14] A. Monreale, G. L. Andrienko, N. V. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, and S. Wrobel. Movement data anonymity through generalization. *Trans. Data Privacy*, 3(2):91–121, 2010.

[15] P. Mörters and Y. Peres. *Brownian motion*. Cambridge University Press, 2010.

[16] G. Mullin. Thieves using apps Strava and MapMyRide to spy on cyclists and steal their expensive bikes after finding out where they live. [http://www.dailymail.co.uk/news/article-2928129/Warning-thieves-using-cyclists-apps-Strava-MapMyRide-live-steal-expensive-bikes.html](http://www.dailymail.co.uk/news/article-2928129/Warning-thieves-using-cyclists-apps-Strava-MapMyRide-live-steal-expensive-bikes.html) 2015. Online: accessed June 22, 2018.

[17] M. E. Nergiz, M. Atzori, and Y. Saygin. Perturbation-driven anonymization of trajectories. Technical report, ISTI-CNR, Pisa, 2007.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[19] R. Pérez-Peña and M. Rosenberg. Strava fitness app can reveal military sites, analysts say. [https://www.nytimes.com/2018/01/29/world/middleeast/strava-heat-map.html](https://www.nytimes.com/2018/01/29/world/middleeast/strava-heat-map.html) 2018. Online: accessed June 22, 2018.

[20] Stan Development Team. Pystan: the Python interface to Stan, Version 2.16.0.0. [http://mc-stan.org](http://mc-stan.org), 2017. Online: accessed June 22, 2018.

[21] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux. Quantifying location privacy. In *2011 IEEE Symposium on Security and Privacy*, pages 247–262. IEEE, 2011.