Velocity-Aware Geo-Indistinguishability

Ricardo Mendes
rscmendes@dei.uc.pt
CISUC and Department of Informatics Engineering, University of Coimbra
Coimbra, Portugal

Mariana Cunha
mariana.cunha@fc.uc.pt
CISUC, CRACS/INESCTEC, and Departamento de Ciência de Computadores, Faculdade de Ciências, Universidade do Porto, rua do Campo Alegre s/n, 4169-007 Porto, Portugal

João P. Vilela
jvilela@fc.up.pt
CISUC, CRACS/INESCTEC, and Departamento de Ciência de Computadores, Faculdade de Ciências, Universidade do Porto, rua do Campo Alegre s/n, 4169-007 Porto, Portugal

ABSTRACT

Location Privacy-Preserving Mechanisms (LPPMs) have been proposed to mitigate the risks of privacy disclosure yielded from location sharing. However, due to the nature of this type of data, spatio-temporal correlations can be leveraged by an adversary to extenuate the protections. Moreover, the application of LPPMs at collection time has been limited due to the difficulty in configuring the parameters and in understanding their impact on the privacy level by the end-user. In this work we adopt the velocity of the user and the frequency of reports as a metric for the correlation between location reports. Based on such metric we propose a generalization of Geo-Indistinguishability denoted Velocity-Aware Geo-Indistinguishability (VA-GI). We define a VA-GI LPPM that provides an automatic and dynamic trade-off between privacy and utility accordingly to the velocity of the user and the frequency of reports. This adaptability can be tuned for general use, by using city or country-wide data, or for specific user profiles, thus warranting fine-grained tuning for users or environments. Our results using vehicular trajectory data show that VA-GI achieves a dynamic trade-off between privacy and utility that outperforms previous works. Additionally, by using a Gaussian distribution as estimation for the distribution of the velocities, we provide a methodology for configuring our proposed LPPM without the need for mobility data. This approach provides the required privacy-utility adaptability while also simplifying its configuration and general application in different contexts.

CCS CONCEPTS

• **Security and privacy** → **Privacy protections**; **Usability in security and privacy**; • **Human-centered computing** → *Mobile devices.*

KEYWORDS

Location Privacy, Differential Privacy, Adaptive Privacy, Privacy Budget, Location-Based Services

ACM Reference Format:
Ricardo Mendes, Mariana Cunha, and João P. Vilela. 2023. Velocity-Aware Geo-Indistinguishability. In *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy (CODASPY '23), April 24–26, 2023, Charlotte, NC, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3577923.3583644

CODASPY '23, April 24–26, 2023, Charlotte, NC, USA

ACM ISBN 979-8-4007-0067-5/23/04.
https://doi.org/10.1145/3577923.3583644

1 INTRODUCTION

Location-Based Services (LBSs) proliferate with the pervasiveness of mobile devices and connectivity. While useful to the user, sharing location data with service providers raises privacy concerns that are beyond physical safety. Specifically, location data may reveal identity, habits, health conditions and social connections, even if data is anonymized [18, 31].

Privacy protection has originally and predominantly been employed by the service providers after the data has been collected. However, this scenario requires trust from the users that their data is handled properly, as after the data is collected, the user has no (or limited) control over it [25]. More recently, mechanisms that protect privacy at data collection, that is, in an online fashion before the data is sent to the provider, have been raising research interest due to empowering users with control over their privacy. This is specially true for Location Privacy-Preserving Mechanisms (LPPMs), where a great portion of the recent studies are mechanisms for online privacy protection [31].

Geo-Indistinguishability has been proposed to design online LPPMs with provable and rigorous privacy guarantees [3], a property inherited from differential privacy [11]. Geo-Indistinguishability guarantees that an obfuscated report is generated with (almost) the same probability regardless of the user-position within a certain radius. This approach conceals the exact location while allowing for the same data to be released, and is thus suitable for LBSs.

Geo-Indistinguishability is only effective for the sporadic use of an LBS as the privacy degrades linearly with the number of queries (c.f. [3, 7]) and due to the fact that continuous location reports are highly correlated [7, 38]. This correlation can be used by an adversary to track users over time and even predict future locations [18, 21, 40].

Adaptations of Geo-Indistinguishability have been proposed to the scenario of online continuous release of location data [2, 7, 10]. Such approaches resort to estimations and distance metrics to evaluate the correlation and subsequently apply obfuscation accordingly. However, using simple estimators such as linear regressions result in a non-negligible amount of outliers due to time-gaps in reports, which occur due to failures in the GPS/communications [24]. Additionally, dynamically adapting the obfuscation requires additional parameters that a user must configure. This is often challenging [16] and potentially misleading [8, 20, 29], specially since users are typically unaware of the privacy risks and privacy-utility trade-offs [1].

Table 1: Desired behavior of a velocity-aware LPPM as a function of the velocity of the user (v_u) and of the velocity of reports (v_r). The symbols ↑ and ↓ denote a high and a low value, respectively.

Velocity		Desired Result
$\uparrow v_u$	$\uparrow v_r$	Balance Privacy and Utility
$\uparrow v_u$	$\downarrow v_r$	Favor Utility
$\downarrow v_u$	$\uparrow v_r$	Favor Privacy
$\downarrow v_u$	$\downarrow v_r$	Balance Privacy and Utility

Moreover, a misconfigured parameter can result in no relevant privacy protection [22].

In this work we argue that the correlation between reports can be estimated by the velocity of the user and the frequency of reports. Consider the following example as an illustration of this argument: a user reporting his location every 30 seconds while walking (∼5 km/h) will have a point every ∼ 42 meters. If the same user was driving in an highway at 120km/h, a point every 1000m would be reported instead. Even though the frequency of updates is the same, the correlation between points might be lower in the case of the highway, as the speed of the user is higher and therefore, the points are sparser. A similar (yet inverse) effect is observed for a constant user speed and varying frequency of reports. If the same user in the highway at 120km/h would instead report every 5 minutes, the distance between reports would increase to 10km. In conclusion, the reports become sparser as the velocity of the user increases or the frequency of reports decreases. Inversely, the reports become denser as the user velocity decreases or the frequency increases.

The previous example paired with the degradation of privacy with the increase in the correlation [18, 21, 40] and frequency of reports [22] lead us to the following conclusion. From the point-of-view of a privacy-preserving mechanism, high frequency of reports or a low user velocity should be met with an increase in obfuscation to increase privacy, and a low frequency of reports or a high user velocity should be met with a decrease in obfuscation as to increase utility. Following these desired properties, which are summarized in Table 1, this work makes the following contributions:

- We generalize Geo-Indistinguishability for effective privacy preservation under online continuous reports. In this proposal, termed Velocity-Aware Geo-Indistinguishability (VA-GI), the velocity of the user and the frequency of reports are used to dynamically adapt the privacy and utility level.
- We devise a VA-GI LPPM that according to our empirically evaluation with real trajectories, outperforms previous literature LPPMs regarding the dynamic adaptability between privacy and utility under different scenarios. Moreover, by using data in its formulation, the proposed LPPM requires only two user-set parameters, thus facilitating usability and mitigating misconfigurations that can lead to no effective privacy [22]. Furthermore, the considered data can be from a specific region or from a single person, thus providing an adaptability to the environment in which it is applied or personalized to the user.
- We evaluate the feasibility of generalizing the VA-GI LPPM for wide deployment through an approximation of the formula using publicly available data. We show empirical evidence on the feasibility and effectiveness of doing so. Specifically, by using data from one location to formulate the LPPM, and evaluating such formulation on another dataset from a different location results in relative differences of the configured privacy parameters inferior to 10%.

The remainder of this paper is structured as follows. Section 2 provides a background on location privacy. Section 3 formally details our proposed mechanism. Section 4 describes the experimental setup, whose results are presented and discussed in Section 5. Section 6 presents and evaluates a generalization of the proposed mechanisms for wide deployment. Section 7 discusses limitations and future remarks, and Section 8 concludes this work.

2 BACKGROUND

This section provides an overview on location privacy and details the location privacy-preserving mechanisms (LPPMs), attacks and metrics used in this work. The reader is referred to [18, 21, 31] for detailed surveys.

A typical framework to evaluate an LPPM consists of a (or multiple) user(s) and his real and obfuscated locations, the LPPM, an adversary which is characterized by its attacks and background knowledge, and metrics [32]. The following subsections address each of these elements while detailing the selected LPPMs, attacks and metrics used in this work to evaluate and compare the effectiveness of VA-GI.

2.1 Location Privacy-Preserving Mechanisms

In this work we focus on online LPPMs suitable to be ran in mobile devices, as this is also a requirement for our VA-GI proposal. Consequently we focused on the following LPPMs: the Planar Laplace [3], which was the first proposed geo-indistinguishable mechanism, the Adaptive [2] and the Clustering [10] Geo-Indistinguishability. The latter two mechanisms were proposed as an extension of Geo-Indistinguishability for the continuous scenario. The following sections detail each of these LPPMs.

2.1.1 Geo-Indistinguishability and Planar Laplace. Geo-Indistinguishability [3] (Geo-Ind) has been proposed as a formal notion based on differential privacy [11] to design user-centric LPPMs. Geo-Ind guarantees that the user location is indistinguishable to any other *nearby* location based on the observed (obfuscated) report independently of an attacker's background information. In other words, the obfuscated report could have been generated with (almost) the same probability from any location around the exact user location.

Geo-Ind is formally defined as follows [24]. Consider a location privacy mechanism as a probabilistic function $K(\cdot)$ that assigns to each location $x \in \mathcal{X}$ a probability distribution on $\mathcal{Z}$, the set of all possible obfuscated locations, where $\mathcal{X}$ and $\mathcal{Z}$ are assumed to be discrete to simplify notation. A mechanism K satisfies ϵ-Geo-Indistinguishability iff:

$$d_{\mathcal{P}}\left(K(x), K(x')\right) \leq \epsilon d_{\mathrm{x}}(x, x') \quad \forall x, x' \in \mathcal{X} \tag{1}$$

where $d_x(\cdot)$ is any distance function and $d_{\mathcal{P}}(\cdot)$ is the multiplicative distance between two distributions, defined as $d_{\mathcal{P}}(\sigma_1, \sigma_2) = \sup_{S \in \mathcal{S}} \left|\log \frac{\sigma_1(S)}{\sigma_2(S)}\right|$, where σ_1 and σ_2 are two distributions on some set S, with the convention that $\mathcal{L} = \left|\log \frac{\sigma_1(S)}{\sigma_2(S)}\right| = 0$ if $\sigma_1(S) = \sigma_2(S) = 0$ and $\mathcal{L} = \infty$ if one of the two is 0.

Intuitively, equation (1) states that the probability of reporting location z while standing in location x is similar to that of standing in any location x'. In fact, both probabilities differ at most by the distance between x and x' factored by a small constant ϵ, where ϵ may be used to tune Geo-Indistinguishability. Commonly, and as specified in the seminal work [3], this constant is set to $\epsilon = l/r$, such that for any x, x' s.t. $d_x(x, x') \leq r$, $d_{\mathcal{P}}(K(x), K(x')) \leq l$, where d_x is an arbitrary metric and l is a user defined parameter termed *privacy loss*. This enforces that any x' within r distance of x discloses at most l information. Consequently, the true location x is better concealed for closer x' locations, while allowing higher dissimilarity for distant locations, thus preserving some degree of utility.

The Planar Laplace (PL) mechanism was the first proposed mechanism to achieve the notion of Geo-Indistinguishability [3] and consists of adding 2-dimensional Laplacian noise centered at the exact user location x and with PDF [3]:

$$p(z|x) = \frac{\epsilon^2}{2\pi} e^{-\epsilon d_x(x,z)} \tag{2}$$

Obtaining z from x using equation (2) can be efficiently done using polar coordinates [3].

2.1.2 Clustering Geo-Indistinguishability. The composability property of differential privacy states that the privacy loss is linear with the number of reports. Specifically, reporting n locations under Geo-Ind results in a privacy loss of $n.\epsilon$ [3, 7]. Under continuous reports, this privacy loss becomes prohibitive and correlations between subsequent reports can be used to improve the efficiency of potential attacks [21, 40]. Clustering Geo-Indistinguishability [10] tackles this problem by reducing the number of obfuscations as a function of the traveled distance. Let x_c and r be the center and radius of an area, denoted cluster, and x_i and z_i the user position and obfuscated report at timestamp i, respectively. Then:

$$z_i = \begin{cases} z_{i-1} & \text{if } d_2(x_c, x_i) \leq r \\ planarLaplace(x_i, \epsilon) & \text{otherwise} \end{cases} \tag{3}$$

Essentially, if the distance between the center of the cluster x_c and the current user position x_i is higher than a radius r then a new obfuscation z_i is generated using the Planar Laplace. When this happens, a new cluster is created by setting the center of the cluster to the current user position, that is, $x_c = x_i$.

In Clustering Geo-Indistinguishability, the privacy and utility level can be tuned by the radius r. Increasing the radius results in an increased privacy at the expenses of the utility, and vice-versa for a decrease of the radius. In the original paper [10], the authors have used the approach from Geo-Indistinguishability to set the value of r as $\epsilon = l/r \leftrightarrow r = l \cdot \epsilon$. Therefore, only two parameters are required, the privacy loss l and privacy budget ϵ.

2.1.3 Adaptive Geo-Indistinguishability. While Clustering Geo-Indistinguishability reduces the number of obfuscated reports to decrease the privacy loss, the privacy/utility configuration relies on the value of the radius r, which is static. In contrast, the Adaptive Geo-Indistinguishability [2] increases the privacy or utility depending on the correlation between past and current locations at each report. Specifically, this LPPM uses the Planar Laplace while dynamically adjusting the privacy budget ϵ to increase privacy if the correlation is high, and increase utility if the correlation is low. For measuring the correlation, a linear regression is used to produce an estimation $\hat{x}_i$ of the real user location x_i at each timestamp i using past locations up to i. Depending on the Euclidean distance between the estimation and real location $d_2(x_i, \hat{x}_i)$, the mechanism increases either privacy or utility as follows:

$$\epsilon_i = \begin{cases} \alpha \cdot \epsilon, & \text{for } d_2(x_i, \hat{x}_i) < \Delta_1 \\ \epsilon, & \text{for } \Delta_1 \leq d_2(x_i, \hat{x}_i) < \Delta_2 \\ \beta \cdot \epsilon, & \text{for } d_2(x_i, \hat{x}_i) \geq \Delta_2 \end{cases} \tag{4}$$

where Δ_1 and Δ_2 are thresholds and α and β two constants with the following constraints: $\Delta_2 > \Delta_1$, $0 < \alpha < 1$ and $\beta > 1$. Fundamentally, if the distance between the estimation and the user location is lower than a threshold Δ_1, then the correlation between past and current locations is high. Therefore, the mechanism decreases the privacy budget ϵ_i to increase privacy. If instead the correlation is low, signaled by a distance between the real and estimated locations higher than a threshold Δ_2, then the mechanism adjusts for increasing utility.

As defined in equation (4) and in contrast with previous LPPMs, Adaptive Geo-Indistinguishability provides a dynamic adjustment of the privacy and utility by taking into account previous reports. Note however, that this adjustment comes at the expense of usability. Namely, in addition to setting ϵ, the user must also define four extra parameters: Δ_1, Δ_2, α and β. This is a crucial drawback on the usability of the LPPM, as a misconfiguration may lead to an ineffective privacy/utility adjustment, as we further discuss in Section 4.2.

2.2 Attacks

Location data is extremely sensitive, not only because it reveals whereabouts, but also because it can disclose identity, habits, health conditions and social connections [18, 31]. In this regard, an adversary can have different objectives and therefore employ different methods [39]. In this work we focus on the problem of tracking, that is, finding the location of the user throughout time. This can be considered the most general type of attack as having exact geolocation data then allows for more specific inferences [12, 14, 17], such as the extraction of sensitive locations.

For the tracking problem we consider two different approaches: a Map-Matching (MM) technique from [15] and the optimal localization attack given an LPPM and a mobility profile [33]. The following two sections provide a high level overview of each of the techniques. The reader is referred to the original work for details.

2.2.1 Map-Matching. Map-Matching (MM) is the process of continuously positioning a vehicle on a road network given noisy location readings [19]. In traditional map-matching, the noise in the readings come from the positioning system, such as the GPS.

However, MM can also be used as a tracking attack by considering the noisy readings to be originated in the use of an LPPM over the exact locations [22]. In fact, MM as an attack has the advantage of using the road network, which can reduce the potential area of locations where the target can be, and some techniques resort to Hidden Markov Models (HMM), that have been shown effective in modeling the temporal correlations of location traces [40, 41].

As LPPMs can generate obfuscations significantly distanced from the user location, an attacker should use a MM technique that is robust to noisy and sparse data. Therefore, in this work we considered the method proposed by Jagadeesh and Srikanthan [15]. This proposal is based on a seminal MM work [27], with adaptations to consider extremely noisy readings, as the cellular network positioning was used instead of the GPS. We refer the reader to the original work for the full implementation details.

2.2.2 Optimal Localization Attack. Localization attacks have the objective to locate a target at a certain point in time [32]. Therefore, in contrast with tracking, this type of attacks do not reconstruct a continuous trajectory, but instead an estimated location for a given location report. Nevertheless, these techniques can reconstruct a discrete trajectory, where each location is a cell grid, instead of the exact location [33].

For this work we considered the optimal localization attack given a mobility profile and an LPPM [33]. Formally, the adversary wants to compute the estimation $\hat{x}_i$ that minimizes the expected distance to the real user location x_i. Knowing that the user is using an LPPM, the expectancy is taken over the probability of the user being at x_i while reporting z_i:

$$\hat{x}_i = \underset{\hat{x}_i}{\operatorname{argmin}} \sum_{x_i \in \mathcal{X}} p(z_i|x_i) \cdot p(x_i) \cdot d(x_i, \hat{x}_i) \tag{5}$$

The probability of the user being at x_i, $p(x_i)$, is computed from the adversary prior knowledge. In practice, this corresponds to using available data to train a mobility profile [6].

2.3 Metrics

In location privacy, numerous metrics can be used to evaluate the privacy, utility and performance [21, 31]. In this work we focus on the differential privacy budget ϵ and the F_1-score as proposed by the authors of the MM method [15]. This latter metric is computed as follows:

$$\begin{gathered} precision = \frac{L_{correct}}{L_{matched}} \qquad recall = \frac{L_{correct}}{L_{truth}} \\ F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \end{gathered} \tag{6}$$

where $L_{matched}$ is the length of the output path, L_{truth} is the length of the corresponding ground truth and $L_{correct}$ is the length of the portions of the output path that overlap with the ground truth path. Intuitively, the precision and recall measure the length of the segments that were correctly matched as a fraction of the map-matching output and the true path, respectively. The F_1-score is then the harmonic mean between both metrics.

Given the definition in equation (6), the F_1-score can be seen as a privacy metric from the point-of-view of an adversary reconstructing the original trajectory . It can additionally be seen as an utility metric from the point of view of a service that requires access to the full trajectories, such as a navigation service or for traffic statistics. In this context, a low F_1-score corresponds to a high privacy and low utility and a higher F_1-score corresponds to a lower privacy and higher utility.

To complement the F_1-score, we additionally consider the adversary error as privacy metric and the quality loss as utility metric. Empirically, the adversary error at each report is computed as the distance from the exact user location x_i and the respective adversary estimation $\hat{x}_i$. The quality loss is the distance from the exact user location x_i and the obfuscated reported location z_i. It is important to note that these two metrics of adversary error and quality loss only consider discrete trajectories, as the error/loss is only measured at the time of the report. That is, they do not take into consideration the continuous trajectory that connects the positions at each timestamp. Therefore, in this work we strongly focus on the F_1-score as it effectively captures the reconstruction of the full continuous trajectory.

3 VELOCITY-AWARE GEO-INDISTINGUISHABILITY

In differential privacy, setting the value of ϵ is challenging as it highly depends on the data, specially in the presence of correlations [8, 20]. In fact, in the context of location privacy it has been shown that there is an upper bound on the value of the privacy budget necessary to guarantee relevant privacy protection [22]. Additionally, from the composability properties of Geo-Ind, the privacy loss increases linearly with the number of reports. Therefore, this notion is only suitable for sporadic reports.

To solve the privacy budgeting problem under continuous reports, we propose a generalization of Geo-Ind termed Velocity-Aware Geo-Indistinguishability (VA-GI). VA-GI adjusts the privacy and utility as a function of the user's velocity and the frequency of reports in accordance with the desired behavior of a velocity-aware LPPM as described in Table 1. For this dynamic adaptability, we set the privacy budget ϵ as a function of both velocities. Formally, for each timestamp i, ϵ is set dynamically as:

$$\epsilon_i \coloneqq \epsilon_i(v_{u,i}, v_{r,i}) \tag{7}$$

where $v_{u,i}$ and $v_{r,i}$ are the velocity of the user and the velocity (or frequency) of the reports at timestamp i, respectively. This formulation leads us to definition 1.

Definition 1. *An obfuscation mechanism $K(\cdot)$ is Velocity-Aware Geo-Indistinguishable iff for any timestamp i:*

$$d_{\mathcal{P}}\left(K(x_i), K(x_i')\right) \leq \epsilon_i(v_{u,i}, v_{r,i}) \cdot d(x_i, x_i'), \quad \forall x_i, x_i' \in \mathcal{X}$$

Definition 1 states that the difference in the output of a VA-GI mechanism with input location x_i or x_i' at timestamp i differs at most by the distance between both locations multiplied by a variable privacy budget that is function of the user and report velocities at the same timestamp i. Note, however that contrary to Geo-Ind, the privacy bound depends on the bounds of the function $\epsilon_i(\cdot)$, which we discuss next.

In order to achieve the desired behavior for a velocity-aware LPPM as described in Table 1, $\epsilon_i(\cdot)$ must increase with an increase in the velocity of the user or a decrease in the frequency of reports,

and decrease with the decrease of the user velocity or an increase in the frequency of reports. Formally, we can describe this requirement as:

$$\epsilon_i(v_{u,i}, v_{r,i}) \propto v_{u,i} \quad \wedge \quad \epsilon_i(v_{u,i}, v_{r,i}) \propto \frac{1}{v_{r,i}} \tag{8}$$

that is, ϵ_i is directly proportional to the user velocity, and inversely proportional to the frequency of reports. Towards this goal, we depart from the standard Geo-Ind [3] where $\epsilon = l/r$, and set the privacy budget as:

$$\epsilon_i = \frac{\epsilon}{m} \cdot m^{\left(2 \cdot f(v_{u,i}, v_{r,i})\right)} \tag{9}$$

where m is a privacy and utility multiplier (as further discussed below) with $m \in [1, \infty[$, $v_{u,i}$ and $v_{r,i}$ are the user and report velocities at timestamp i, and $f(\cdot)$ is any function of $v_{u,i}$ and $v_{r,i}$, that holds the proportionalities from equation (8) and with $f(\cdot) \in [0, 1]$. Equation (9) corresponds to the exponential regression on $f(\cdot)$ such that the following bounds for ϵ_i are achieved:

$$\frac{\epsilon}{m} \leq \epsilon_i \leq m \cdot \epsilon \;\Leftrightarrow\; \frac{l}{m} \leq r \cdot \epsilon_i \leq m \cdot l, \quad \forall i \tag{10}$$

where the multiplier m is used to adjust the privacy and utility bounds. Equations (9) and (10) provide a dynamic balance in where the privacy and utility levels can be increased or decreased up to m times the initial ϵ value, depending on the velocity of the user and frequency of reports. As long as $f(\cdot)$ provides the proportionality from equation (8), an increase in the user velocity ($v_{u,i}$) and/or a decrease in the frequency of updates ($v_{r,i}$) is met with an increase in the privacy budget ϵ_i, and vice-versa for a decrease in ϵ_i. Therefore, we refer to this VA-GI formulation as (m, ϵ)-VA-GI. Finally note that, if $m = 1$, $(1, \epsilon)$-VA-GI becomes Geo-Ind as $\epsilon_i = \epsilon$, $\forall i$. Therefore, (m, ϵ)-VA-GI can be seen as a generalization of Geo-Ind. This result leads us to Theorem 1.

Theorem 1. *(m, ϵ)-VA-GI satisfies $m\epsilon$-Geo-Indistinguishability and guarantees a maximum privacy loss of $m \cdot l$ within a radius r. Namely, for any timestamp i:*

$$d_{\mathcal{P}}\left(K(x_i), K(x_i')\right) \leq m \cdot l, \; \forall x_i, x_i' \in \mathcal{X} \quad \textbf{s.t.} \quad d_{\mathrm{X}}(x_i, x_i') \leq r$$

Proof. The Planar Laplace has been proven to provide Geo-Indistinguishability by abiding to equation (1) [3]:

$$d_{\mathcal{P}}\left(K(x), K(x')\right) \leq \epsilon d_{\mathrm{X}}(x, x') \quad \forall x, x' \in \mathcal{X}$$

In (m, ϵ)-VA-GI, the privacy budget ϵ varies for each timestamp i. From equation (10) we have that $\epsilon_i \leq m \cdot l$, $\forall i$. Consequently, from equation (1) and for any two x_i, x_i' such that $d_{\mathrm{X}}(x_i, x_i') \leq r$:

$$d_{\mathcal{P}}\left(K(x_i), K(x_i')\right) \leq m \cdot l, \quad \forall i \tag{11}$$

□

3.1 An (m, ϵ)-VA-GI LPPM

Equation (9) defines the generic formula to achieve (m, ϵ)-VA-GI, where any definition of the function $f(\cdot) \in [0, 1]$ that respects the desired properties from Table 1 can be used. However, for an effective privacy and utility balance, the function should respect the nature of the velocities, specifically their distributions. Unfortunately, the velocities do not follow any unimodal distribution, and in fact depend on the underlying road features and drivers [35]. Therefore, to design an (m, ϵ)-VA-GI LPPM one can approximate or estimate the distributions by using available data. This section describes such methodology.

Since there is no a priori best choice, we leave the comparison between VA-GI LPPMs for future work and instead choose a simple velocity function $f(\cdot)$ defined as the average between a function of the user velocity $f_u(v_{u,i})$ and a function of the report velocities $f_r(v_{r,i})$:

$$f(v_{u,i}, v_{r,i}) = \frac{1}{2} \cdot \left(f_u(v_{u,i}) + f_r(v_{r,i})\right) \tag{12}$$

Where $f_u(\cdot),\ f_r(\cdot) \in [0, 1]$. To take into consideration the distributions of $v_{u,i}$ and $v_{r,i}$ and to favor the variance of f_u and f_v near the typical values of $v_{u,i}$ and $v_{r,i}$, we set:

$$\begin{aligned} f_u(v_{u,i}) &= cdf(v_{u,i}) \\ f_r(v_{r,i}) &= 1 - cdf(v_{r,i}) \end{aligned} \tag{13}$$

where $cdf(\cdot)$ stands for the Cumulative Density Function (CDF). With equation (13) we guarantee that the codomain of $f(v_{u,i}, v_{r,i})$ as defined in equation (12) is in the interval $[0, 1]$ and that the proportionalities from equation (8) are respected. Additionally, because the slope of the CDF is higher in the typical values, smaller deviations from these will have a steeper privacy/utility adjustment. Combining equation (12) and (13) in equation (9), we reach:

$$\epsilon_i = \epsilon \cdot m^{\left(cdf(v_{u,i}) - cdf(v_{r,i})\right)} \tag{14}$$

The advantage of using the $cdf(\cdot)$ functions of the user velocities and frequency of reports relates to the minimization of the required parameters. By using equations (13) in the (m, ϵ)-VA-GI privacy budget equation (9), we limit the LPPM to 2 parameters: the initial ϵ value and the multiplier m. This is in contrast with other LPPMs for continuous report, that either require several parameters to provide the dynamic adaptability, such as the Adaptive Geo-Ind, or that have few parameters but do not adapt to the dynamics of the movement (e.g. Clustering Geo-Ind and the Planar Laplace). Sections 4.2 and 5 demonstrate these disadvantages of previous works.

One of the disadvantages of this approach is that the distribution of the velocities is unknown. To solve this issue, one can use data to estimate the CDFs using non-parametric density estimation. From the point of view of the LPPM, this data can belong to all drivers in a specific city, global or even be personalized to the user, by using their past data. Regardless of the data used, a better CDF fit will favor the privacy and utility trade-off. Nevertheless, to faithfully fit a CDF, a decent amount of data is required. In Section 6 we show empirical evidence on the effectiveness of generalizing the CDFs to Gaussian distributions in the context of (m, ϵ)-VA-GI.

In summary, the (m, ϵ)-VA-GI LPPM consists of using the PL mechanism from equation (2) with the epsilon from equation (14).

3.2 Setting LPPM Parameters

One of the challenges in the wide deployment of privacy mechanisms is the configuration parameters. In differential privacy, for instance, setting the value of ϵ depends on the dataset and must take into account the presence of correlations [8, 20]. This is specially true for mechanisms that act at collection time, where the responsibility to properly tune the mechanism lies on the user.

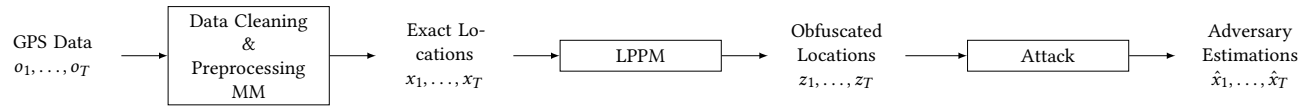


Figure 1: Diagram of the followed methodology. The LPPM step is repeated for each of the LPPMs and the Attack step is repeated for each combination of LPPM/Attack.

Statically set parameters such as in Geo-Ind and Clustering fail at providing effective privacy against continuous, and hence, correlated reports. The Adaptive Geo-Ind provides the dynamic adaptability but introduces four additional parameters. Unfortunately, misconfigured parameters can result in ineffective privacy/utility trade-offs (c.f. Section 4.2), or even in no effective privacy [22]. In contrast, the (m, ϵ)-VA-GI LPPM only requires the initial ϵ value and a straightforward privacy/utility multiplier m. Setting these values can take into consideration personal preferences or application requirements, such as a minimum required utility. For instance, one can see the typical range of values from the literature and set ϵ to the mid value of such range and then set m such that ϵ_i automatically adjusts within the range, so as to either favor privacy or favor utility, as required by the context at hand. Specifically, with $\epsilon = 16$ km^{-1}, a value that is commonly used in the context of continuous reports [2, 22], and $m = 10$ we obtain $1.6 \leq \epsilon_i \leq 160$, a range that contains values that are used in sporadic scenarios [3] and values used in the continuous scenario [2, 22].

4 EXPERIMENTAL SETUP

This section describes the conducted simulations by detailing the datasets and experimental setup. To evaluate the effectiveness of (m, ϵ)-VA-GI, our methodology consisted in applying the LPPMs detailed in Section 2 to the data, followed by each of the attacks. The results are compared between the output of the attacks and the original dataset, along with the comparison between the different LPPMs and respective configurations. The diagram in Figure 1 summarizes the methodology, which is repeated (except the preprocessing MM) for each pair of LPPM/Attack. The following subsections detail the dataset and respective preprocessing, and the configurations/parameters of the LPPMs and attacks.

4.1 Dataset Characterization And Preprocessing

The dataset used in our experiments was the Cabspotting, a dataset of taxi trajectories over the city of San Francisco, California, USA. The trajectories belong to 536 taxis and were collected over a period of 30 days, containing not only the GPS position and timestamp, but also whether the cab had a customer at each time [30].

To preprocess the dataset, we first filtered out trajectories with points outside the bounding-box defined from South and North by the latitudes 37.600, 37.811, and from West and East by longitudes −122.517, −122.354 . We additionally removed trajectories without occupancy, as to avoid trajectories where the cab is waiting for a client. Finally, public datasets present a significant number of location/time outliers (c.f. [23]), resulting in spurious trajectories. Consequently, we applied the data cleaning procedure from [37] by discarding: 1) trips with duration lower than 1 minute and higher than 3 hours; 2) trips with total displacement over 100 km; 3) trips with average velocity lower than 5 km/h or over 120 km/h; 4) non-smooth trips. Non-smooth trips were removed by using a filter with a sliding window that detects whether the average velocity between points is within normal intervals. If more than a defined percentage of points in each trajectory has abnormal average velocities, then the trajectory is rejected. The original default parameters were used for this filter [37]. After this preprocessing, 307983 trajectories remained from the original dataset.

Because we apply four LPPMs under different configurations and multiple attacks, we further subsampled the dataset as to reduce the number of trajectories. In order to evaluate the adaptability of the LPPMs under continuous reports, we divide the trajectories in four different sets depending on the average user velocity v_u and average report velocity v_r:

(1) Balance Privacy/Utility 1 ($\downarrow v_u \downarrow v_r$): trajectories with average user velocity $v_u \leq 20$ km/h and velocity of reports $v_r \leq 45$ reports/h;
(2) Favor Privacy ($\downarrow v_u \uparrow v_r$): trajectories with average user velocity $v_u \leq 20$ km/h and velocity of reports $v_r \geq 100$ reports/h. This is the worst case with respect to privacy, as it has the largest density of reports per distance traveled. Therefore, and according with the desirable properties from Table 1, LPPMs should ideally adjust for privacy to account for the higher correlation between reports;
(3) Favor Utility ($\uparrow v_u \downarrow v_r$): trajectories with average user velocity $v_u \geq 100$ km/h and velocity of reports $v_r \leq 45$ reports/h. This scenario has the lowest density of reports per distance traveled and hence, the lowest correlation between reports. Therefore, the LPPMs should ideally adjust to improve utility;
(4) Balance Privacy/Utility 2 ($\uparrow v_u \uparrow v_r$): trajectories with average user velocity $v_u \geq 100$ km/h and velocity of reports $v_r \geq 100$ reports/h. This scenario is similar to the "Balance Privacy/Utility 1 ($\downarrow v_u \downarrow v_r$)" with respect to the density of reports and therefore to the desired response.

The threshold values were chosen by looking at the speed limits[1] and empirical distributions, which we omit due to space constraints. Specifically, speed limits in alleys and residential areas are 24 and 40 km/h, respectively, and therefore, a vast number of trajectories will have an average speed lower to 20. The high user velocity trajectories, with average over 100 km/h, will correspond to trajectories in highways, where the speed limit is 105 km/h. For the frequency of reports, we picked intervals directly from the empirical distribution.

From the four data set divisions, we picked the 100 trajectories from each partition with lowest standard deviation, to select trajectories where the instant velocities are closest to the filtered mean

[1] https://data.sfgov.org/Transportation/Map-of-Speed-Limits/ttcm-fwt2

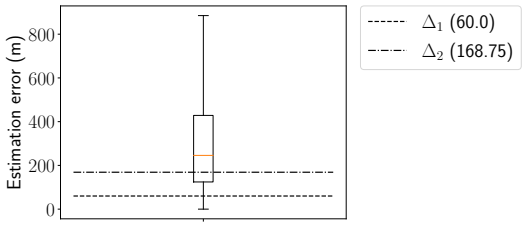
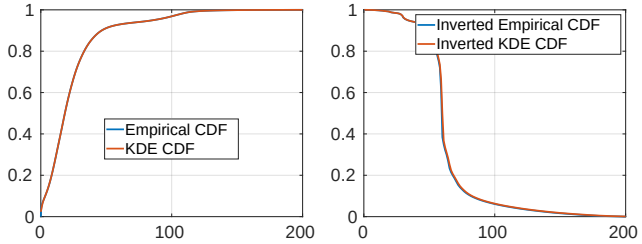


Figure 2: Boxplot of the Adaptive estimation errors and the Δ_1 and Δ_2 original thresholds.



(a) Empirical and kernel CDF of the users' velocities. **(b) Inverted empirical and kernel CDF of the reports' velocities.**

Figure 3: Empirical and kernel density estimation cumulative density functions of the velocities.

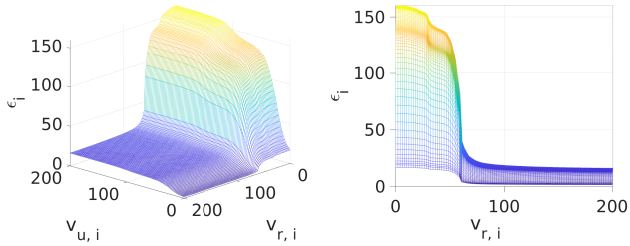


Figure 4: 3-dimensional plot of the value of ϵ_i as given by equation (14) as a function of $v_{u,i}$ and $v_{r,i}$, with $\epsilon = 16$ km^{-1} and $m = 10$.

values. This selection was made to have a strong diversity of trajectories, encompassing scenarios with a high, medium and low density of reports per trajectory. This density relates directly to the difficulty of an adversary in reconstructing the real trajectory. From now onwards, we refer to the selected 400 trajectories as test data and the remaining cleaned trajectories (307583) as training data.

To the test data, we apply the map-matching technique detailed in Section 2.2.1 as to position each location report on the road network. This preprocessing step cleans these trajectories from noisy reports, thus forming our ground-truth. For the standard deviation of the measurement error σ, we used a typical value for GPS readings of $\sigma = 6.86$ m [13]. The parameters λ_y and λ_z were estimated following the original map-matching proposal [15]. Namely, using trajectories from the training data with duration between 1 and 6 minutes with at least 2km of traveled distance (4963 trajectories). This resulted in the values $\lambda_y = 0.69$ and $\lambda_z = 13.35$.

4.2 LPPMs

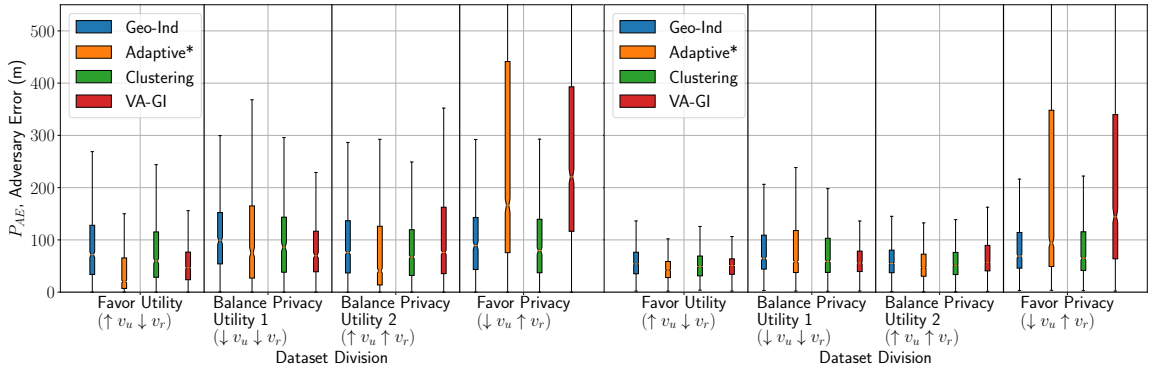
For the LPPMs, we compare the (m, ϵ)-VA-GI LPPM from Section 3.1, which we simple refer to as **VA-GI**, with the geo-indistinguishable LPPMs described in Section 2.1, that is, the Planar Laplace [3], which we refer to as **Geo-Ind**, the Clustering Geo-Ind [10], referred to as **Clustering**, and the Adaptive Geo-Ind [2], or **Adaptive**.

For the privacy budget, and for all LPPMs, we used multiple values in the typical ranges of LPPMs for continuous reports [2, 22], specifically $\epsilon = [16, 32, 64, 128]$ km^{-1}. For the Geo-Ind LPPM, this corresponds to an average obfuscation of $[125, 62.5, 31.25, 15.625]$ m, respectively. These values of obfuscation range from city block level distances to parallel streets. For the remaining parameters we attempted to use the proposed values from the original respective papers, but we found some problems in the Adaptive as follows.

In the Adaptive mechanism, the privacy budget ϵ is adjusted for privacy or utility depending on the error in estimating the current location, as described in Section 2.1.3. In accordance with equation (4), if the estimation error is smaller than Δ_1, then privacy is increased by reducing ϵ by a factor of α. If the estimation error is higher than Δ_2 the utility is increased by increasing ϵ by a factor of β. The authors heuristically proposed setting $\Delta_1 = 0.96/\epsilon$ and $\Delta_2 = 2.7/\epsilon$. However, for the ϵ values used in our work, we found that these thresholds result in a poor privacy and utility adaptability. Figure 2 illustrates this problem by plotting a boxplot of the estimation errors $(d_2(x, \hat{x}))$ for all points in the training data and the thresholds for $\epsilon = 16$ km^{-1}. From this plot it is clear that for almost 75% of location reports, the adaptive would adjust for utility, and only for less than approximately 15% of cases, it would adjust for privacy. This unbalance is even worse for higher ϵ values, as the estimation errors are the same, but the thresholds would be lower. In order to have a proper privacy/utility dynamic, we set the Δ_1 and Δ_2 thresholds to the first ($\Delta_1 \approx 124.29$) and third ($\Delta_2 \approx 428.56$) quartiles of the boxplot. We refer to this tuned LPPM to as **Adaptive*** and only present the results for this optimized variant of the original adaptive mechanism. This example illustrates the difficulty in setting the proper parameters, as discussed in Section 3.2, a problem that VA-GI solves by using the cumulative density functions, as previously discussed.

As for the VA-GI, and following the description from Section 3.1, we use a non-parametric estimation of the Cumulative Density Functions (CDF) using the training data, specifically, a Kernel Density Estimation (KDE). Figures 3a and 3b present the empirical and KDE CDF for the user and report velocities, respectively. From these images we can clearly see that the lines are mostly coincident in both cases, thus confirming that the KDE is a good estimator for the CDF. In summary, for VA-GI, at each timestamp i we compute ϵ_i as defined in equation (14) with the KDE CDFs.

One can plot equation (14) as a function of the velocities of the users and reports. Figure 4 presents this plot with $\epsilon = 16$ km^{-1} and $m = 10$. We can observe that as the velocity of the user increases, the value of ϵ_i increases as to reduce the obfuscation, and therefore increase utility, and vice-versa for a decrease in the velocity as



(a) Map-matching **(b) Optimal Localization Attack**

Figure 5: Boxplot of the adversary errors for the Map-matching and optimal localization attack for $\epsilon = 16$ km^{-1}.

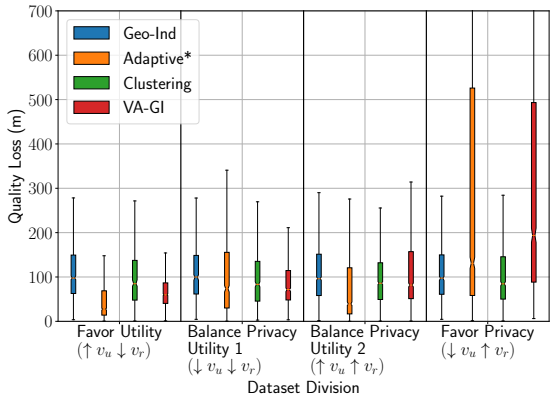


Figure 6: Boxplot of the quality loss for each dataset division and LPPM with $\epsilon = 16$ km^{-1}.

to increase privacy. The velocity of reports has the inverse effect, which is in accordance with Table 1.

For the remainder of the paper, the results for the VA-GI mechanism were obtained with $m = 10$. This value was chosen such that typical epsilon values (c.f. [2, 3, 22]) were contained within the ϵ_i bounds defined in equation (10).

4.3 Attacks

For the attacks we consider the optimal localization attack from section 2.2.2 and the map-matching attack from section 2.2.1. For the map-matching, while the original authors presented a complex route choice model, the increase in the accuracy was marginal when compared to using the shortest path [15]. Therefore, in this work we opted for the simple shortest path to reduce computational complexity. For efficiency, and similarly to [13], we only consider candidates nodes within a radius r which we calculate using the inverse cumulative distribution function of the Gaussian distribution. The radius r is computed such that the circle centered at the observation contains the exact location with 90% probability given a geo-indistinguishable obfuscation. When this circle contains no candidates, which can happen due to the use of the LPPM and selected road network, the nearest road network node is used as candidate. The road network was obtained from OpenStreetMap using the OSMnx tool [5] over the San Francisco bay area.

5 RESULTS

This section presents the results obtained following the presented methodology. Because the findings endure for all epsilon values, we present the results only for $\epsilon = 16$ km^{-1}.

Figure 5 shows the adversary error for the map-matching and optimal localization attack. From this figure we can observe that Geo-Ind and Clustering have similar adversary error (privacy level) for the different dataset divisions and for both attacks. However, the Adaptive* and VA-GI largely vary. Specifically, for the "Favor Privacy ($\downarrow v_u \uparrow v_r$)" division these two LPPMs have greater adversary errors, and for "Favor Utility ($\uparrow v_u \downarrow v_r$)" the lowest. These results indicate that both the Adaptive* and the VA-GI properly adapt in accordance with the desired properties of a velocity-aware LPPM, as described in Table 1. However, the large increase in the adversary error comes with the consequence of a high quality loss as displayed in Figure 6. This is the natural and ever present trade-off between privacy and utility [9].

According to Figures 5 and 6, the VA-GI had the strongest privacy (highest adversary error) for the "Favor Privacy ($\downarrow v_u \uparrow v_r$)" scenario, while the Adaptive* had the best utility (lowest quality loss) for the "Favor Utility ($\uparrow v_u \downarrow v_r$)" division. However, due to the fact that the quality loss and adversary error metrics do not take into consideration the continuous nature of the trajectories, these results can be inconclusive. Therefore, in the following discussion we focus on the F_1-score metric as it considers not only the user individual reports but every traversed segment between each location.

Figure 7 shows the F_1-score for each dataset division and each LPPM. From the plot it is clear that the Adaptive* and VA-GI adapt to both the user and report velocities in accordance with the desirable properties of a velocity-aware LPPM, thus confirming previous results. This is observable from the fact that for the "Favor Utility ($\uparrow v_u \downarrow v_r$)" division these two LPPMs present the highest F_1-scores,

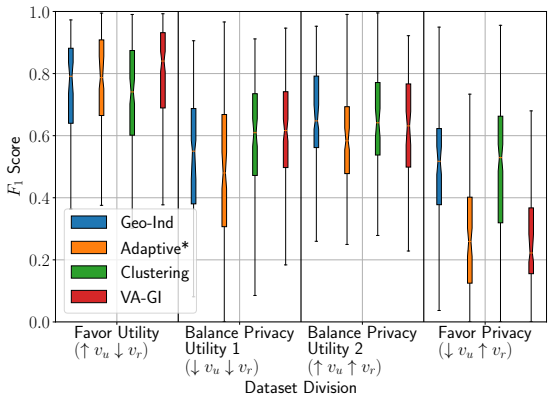


Figure 7: Boxplot of the F_1-score for each dataset division and LPPM with $\epsilon = 16\ \text{km}^{-1}$.

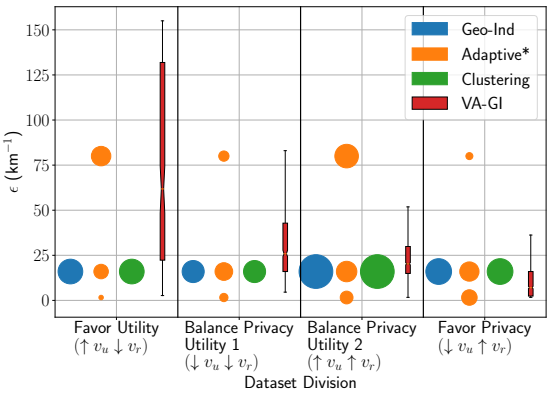


Figure 8: Distribution of the ϵ_i values for each dataset division and LPPM with $\epsilon = 16\ \text{km}^{-1}$ in the form of a scatter plot for Geo-Ind, Adaptive* and Clustering, where the size of the points is the absolute frequency of the corresponding ϵ value, and a boxplot for VA-GI, due to the continuous nature of the epsilon values in this latter LPPM.

meaning that both LPPMs adjusted for utility, and the lowest F_1-scores for "Favor Privacy ($\downarrow v_u \uparrow v_r$)", signaling an adjustment for privacy. However, the VA-GI outperformed the Adaptive* in both cases, presenting higher score for the "Favor Utility ($\uparrow v_u \downarrow v_r$)" and lower for the "Favor Privacy ($\downarrow v_u \uparrow v_r$)", as desired. Furthermore, it should be noted that the displayed results for the adaptive mechanism were obtained with an improved selection of parameters, the Adaptive* – the default values would have lead to an ineffective privacy-utility adaptability, as depicted in Figure 2. Figure 7 also shows that Geo-Ind and Clustering present similar yet smaller fluctuations in the F_1-score for the different dataset divisions. This is due to the underlying selection of trajectories for the division. Specifically, the selected traces with high user velocity correspond to movements in highways, where there is less entropy in finding the right trajectory with the map-matching, thus resulting in a higher F_1-score. As for lower v_u trajectories, these correspond to alleys and residential areas, where the density of the road network is higher and therefore resulting in a lower F_1-score. Nevertheless, these fluctuations in the scores for the different divisions are inferior to the ones obtained with the Adaptive* and VA-GI, signaling that the latter two LPPMs effectively adapt to the velocities.

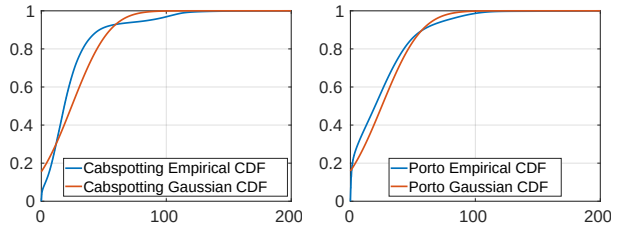
The variations in the F_1-score for the Adaptive* and VA-GI originate from the dynamic adaptability of the ϵ_i value according to the velocities as in equation (14). Therefore, it is useful to look at the distribution of these values to confirm the aforementioned findings. Figure 8 presents the distributions of the ϵ_i values for each dataset and LPPM with $\epsilon = 16\ \text{km}^{-1}$. Note that Geo-Ind and Clustering are a single scatter point as the ϵ is constant, while the Adaptive* presents three possible values as per equation (4). Results for VA-GI are presented as a boxplot, due to the continuous nature of the epsilon values obtained. These plots firmly agree with F_1-score results. Namely, both the Adaptive* and VA-GI adapt for privacy for the "Favor Privacy ($\downarrow v_u \uparrow v_r$)" by decreasing ϵ_i and for utility for the "Favor Utility ($\uparrow v_u \downarrow v_r$)" by increasing ϵ_i. Notice however, that while the VA-GI has continuous spectrum of values for ϵ_i, the Adaptive* mechanism considers only three values, resulting from the application of formula (4). Therefore, the VA-GI is able to provide a more fine grained privacy/utility adaptability. This is relevant as the Adaptive* might erroneously not adapt for privacy or utility in cases where it should, which is further aggravated by possibly misconfigured threshold values Δ_1 and Δ_2 (c.f. Figure 2). The VA-GI mitigates this problem by using the CDF of the velocities for defining the system parameters, as previously discussed.

In summary, both the VA-GI and Adaptive* adapt in accordance to the desired properties of a velocity-aware LPPM. For this dynamic adjustment, the VA-GI outperfoms all other tested LPPMs from both privacy and utility metrics. Although the performance advantage of VA-GI is more limited with respect to the Adaptive mechanism, the latter requires setting parameters on demand for each new context/scenario (see Section 4.2), whereas VA-GI provides a mechanism for automatic adjustment of parameters and a finer grained continuous adaptability, while requiring fewer parameters, thus mitigating misconfiguration issues that can lead to no privacy [22].

6 GENERALIZING THE VA-GI LPPM

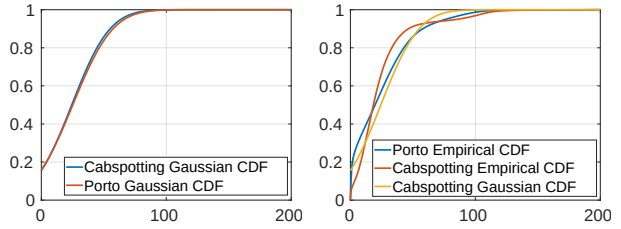
The VA-GI LPPM requires data to estimate the CDFs of the velocities, which can limit the wide deployment of such mechanism. This section addresses this limitation by proposing a methodology to generalize the VA-GI LPPM for wide deployment by approximating the CDFs to a Gaussian distribution modelled using publicly available data.

In order to use the CDF of the velocities for the definition of function $f(\cdot)$ in (12), real mobility data is needed, thus posing a limitation on the practicability of VA-GI. However, from the CDF plots for the Cabspotting data illustrated in Figure 3, we can observe that an approximation to a Gaussian distribution might fit as an estimation. In this section we do such evaluation, specifically, we use publicly available data to generate a CDF and measure the fitness of the approximation to a new dataset. Without loss of generality, we focus on fitting the CDF of user velocity, since the same methodology could be used to approximate the CDF of the



(a) Cabspotting Empirical and Gaussian CDFs.

(b) Porto Empirical and Gaussian CDFs.



(c) Cabspotting and Porto Gaussian CDFs.

(d) Porto empirical CDF and Cabspotting Empirical and Gaussian CDFs.

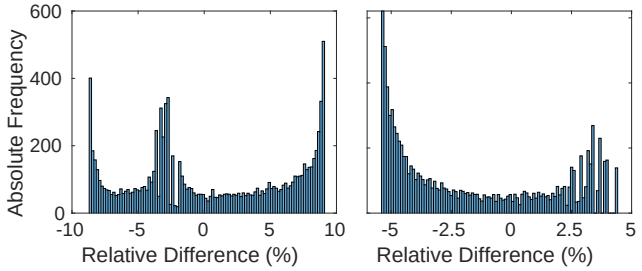
Figure 9: CDF and PDF for the user velocities of the Cabspotting and Porto datasets.

report velocity, with the difference that different applications might use different sampling frequencies. To solve such dissimilarity, a normalization of the distribution would suffice.

Vehicular velocities have been previously shown to follow Gaussian distributions in highways [4]. However, urban traffic is more complex due to intersections, traffic signals, congestions and other factors [35]. Therefore, in order to visually compare the goodness of fit of the Gaussian distribution, we use the Cabspotting dataset as publicly available data to form the CDFs and a second dataset from a different geographic location to evaluate the goodness of fit. This second dataset is also composed of vehicular trajectories belonging to 441 taxis in the city of Porto, Portugal, with a sampling rate of 15 seconds and collected over a full year [26].

Figure 9 shows the obtained CDFs for a subsample of 10000 velocities from the Cabspotting and Porto datasets. From Figure 9a it can be seen that for the Cabspotting dataset, the distributions differ considerably. However, for the Porto dataset, Figure 9b reveals a high similarity between the empirical and Gaussian distributions. A Kolmogorov-Smirnov normality test confirms that both velocity sets do not follow a Gaussian distribution for any confidence level (p-value is 0). Additionally, a two-sample Kolmogorov-Smirnov goodness-of-fit hypothesis test also discards the possibility of both velocities following the same distribution (p-value is also 0). Finally, a Wilcoxon rank sum test and a Mood's median test reject the hypothesis that both velocities have the same mean and median, respectively.

Despite the fact that the velocities follow an unknown seemingly multimodal distribution, the use of the Gaussian distribution as an approximation might suffice for the purpose of the function of the user velocities $f(v_{u,i})$. We can see this effect by observing Figure 9c, where the Gaussian CDF for both datasets is similar, even though



(a) Cabspotting KDE and Porto KDE epsilon values

(b) Cabspotting Gaussian and Porto KDE epsilon values

Figure 10: Relative differences for the ϵ_i values between each pair of distributions, with initial $\epsilon = 16$.

both cities have different speed limits, and hence different velocity distributions. In fact, in Figure 9d, the Gaussian CDF obtained from the Cabspotting dataset is similar to the Porto empirical CDF. This suggests that using publicly available data, even from different geolocations might result in effective approximations. Note however, that the training data must be diverse, as training with data from a rural area and applying it to a metropolitan area might result in poor results.

From the point of view of VA-GI, an approximation of the definition of $f(\cdot)$ will result in sub-optimal velocity awareness, which can cause the LPPM to overshoot the privacy or utility in certain cases. In order to measure this unbalance, we use a subsample of the velocities from the Porto dataset and plot the differences between the ϵ_i values obtained using the Cabspotting approximations and the Porto KDE CDF, which is the baseline reference. In other words, we use the Cabspotting dataset as public available data, from which we extract the KDE and Gaussian CDFs, and then apply these distributions in the form of equation (14) to a subsample of the Porto dataset. To focus on user velocity, without loss of generality, we set the value of $cdf(v_{r,i}) = k,\ \forall i$ with $k = 0.5$, such that equation (14) becomes $\epsilon_i = \epsilon.m^{(cdf(v_{u,i})-0.5)}$, with the bounds $\epsilon \cdot m^{-0.5} \leq \epsilon_i \leq \epsilon \cdot m^{0.5}, \quad \forall i$ and $m = 10$, as defined previously.

Figure 10a presents the differences between the epsilons obtained with the Cabspotting KDE and the Porto KDE, while Figure 10b shows the differences between the epsilons obtained with the Cabspotting Gaussian and the Porto KDE, with an initial ϵ value of 16km^{-1}. A negative value in these plots corresponds to an overshoot in the privacy adjustment, as the ϵ_i in the Porto KDE is higher than the same ϵ_i for the Cabspotting estimation, and vice-versa for a positive value, thus corresponding to an overshoot in the utility adjustment. From these plots we can clearly see that even though the velocity distributions might differ, the ϵ_i obtained using equation (14) are similar. Specifically, the ϵ_i values differed less than 10% when using the Cabspotting KDE and up to approximately 5% when using the Gaussian approximation, for both privacy and utility adjustments. Nevertheless, these results confirm that it is possible to use publicly available data, even from a different geolocation with different speed limits, to approximate the function $f(\cdot)$ of VA-GI. Furthermore, from a practitioner perspective, the Gaussian distribution as an estimator can be used in production by

simply setting the mean and standard deviation parameters, thus giving a practical benefit over the KDE. Under these configurations, the user just has to provide the parameters ϵ and m as previously discussed, where m can be used to adjust VA-GI for more privacy or utility, as desired. By configuring two parameters alone, our proposed scheme enables adaptation of the privacy and utility levels according to the user and report velocities.

7 LIMITATIONS AND FUTURE WORK

The use of correlations between subsequent requests is an important venue for LPPMs. In this paper we have used the velocity of the user and the frequency of reports as a metric for the correlation between reports, and proposed a particular instance of a VA-GI LPPM. While our results have shown an effective adaptability we leave for future work the comparison between different VA-GI formulations. Furthermore, the use of the velocity of the user and the frequency of reports might not be an efficient proxy for the correlation between reports in some cases. For instance, in a highway, the density of reports can be low, but the correlation might be high due to the lack of intersections/exits. As future work, we intend to consider the underlying map in the design of the LPPM, similarly to the work in [38] but in the context of differential privacy. For example, the consideration of the road-network in vehicular trajectories or the density of buildings, can allow for a metric on the adversary confusion, thus potentially resulting in a more effective privacy and utility adjustment.

Our scheme requires access to user velocity data that may not always be available. We have shown in Section 6 how to generalize VA-GI through approximation to known CDFs, whereby the Gaussian CDF generalized better than the KDE CDF. However, further validation using other datasets is required to confirm these results.

Inherited from differential privacy, the repeated use of any geo-indistinguishable LPPM, including VA-GI, results in increasing and unbounded information leakage, which can be measured through the composition properties [11]. In practice, one can define a maximum privacy budget, such that after exhausting it, no more data is sent to the service provider. Unfortunately, this would result in not having access to the service. Practical implementations of differential privacy incur in a trade-off where the privacy budget is reset after a certain amount of time, thus limiting exposure within a given time frame, while (misleadingly) considering contributions between periods independent [34]. This is an active line of research that could spark future work in LPPMs at data collection.

Finally, location data has been considered personal data under privacy laws such as the General Data Protection Regulation (GDPR) [36] and the California Consumer Privacy Act (CCPA) [28]. VA-GI preserves location privacy even against the service provider, thus providing some degree of anonymity. In practice, however, location-based services require an account, thus identifying even obfuscated reports. Regardless, the legal requirement to anonymize the data lies on the service provider, which becomes the data owner/controller. Therefore, VA-GI provides location privacy, but not necessarily anonymity. Further anonymization might be required from the service provider before sharing/storing the data.

8 CONCLUSION

The widespread of mobile and connected devices has lead to the pervasiveness of Location-Based Services. While vast, the research on location privacy has fallen behind this development, specially in Location Privacy-Preserving Mechanisms (LPPMs) that act at collection time. In this context, for effective privacy protection, LPPMs must take into consideration the potential threat that arises from the correlation of reports. In this paper we adopted the velocity of the user and the frequency of reports as metric for the correlation and proposed a generalization of Geo-Indistinguishability termed Velocity-Aware Geo-Indistinguishability (VA-GI). Under such notion, we design a VA-GI LPPM that outperforms previous LPPMs in adapting the privacy and utility under different dynamic scenarios. The proposed VA-GI LPPM provides a mechanism for automatic adjustment of privacy parameters, while requiring fewer parameters than previous adaptive mechanisms, thus mitigating misconfiguration issues that can lead to no effective privacy protection. This adaptability of VA-GI can be tuned for general use, by using city or country-wide data for configuration, or for specific user profiles, thus warranting fine-grained tuning for specific users/profiles. Finally, we generalized our proposed VA-GI LPPM by using publicly available data for defining system parameters, thus facilitating effective wide deployment using real data to automatically configure VA-GI.

ACKNOWLEDGMENTS

The authors would like to thank Guilherme Duarte for his contributions porting the code to python and running some of the experiments. This work was funded by European Structural and Investment Funds in the FEDER component, through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) and Portugal 2020 [Project nº 047264 - Theia; Funding Reference: POCI-01-0247-FEDER-047264]. Ricardo Mendes and Mariana Cunha wish to acknowledge the Portuguese funding institution FCT - Foundation for Science and Technology for supporting their research under the Ph.D. grant SFRH/BD/128599/2017 and 2020.04714.BD, respectively.

REFERENCES

[1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* 347, 6221 (2015), 509–514.
[2] Raed Al-Dhubhani and Jonathan M. Cazalas. 2018. An adaptive geo-indistinguishability mechanism for continuous LBS queries. *Wireless Networks* 24, 8 (01 Nov 2018), 3221–3239. https://doi.org/10.1007/s11276-017-1534-x
[3] Miguel E. Andrés, Nicolás E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. 2013. Geo-Indistinguishability: Differential Privacy for Location-Based Systems. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security* (Berlin, Germany) *(CCS '13)*. Association for Computing Machinery, New York, NY, USA, 901–914.
[4] M. Boban, T. T. V. Vinhoza, M. Ferreira, J. Barros, and O. K. Tonguz. 2011. Impact of Vehicles as Obstacles in Vehicular Ad Hoc Networks. *IEEE Journal on Selected Areas in Communications* 29, 1 (2011), 15–28.
[5] Geoff Boeing. 2017. OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems* 65 (2017), 126–139.
[6] Konstantinos Chatzikokolakis, Ehab Elsalamouny, and Catuscia Palamidessi. 2017. Efficient utility improvement for location privacy. *Proceedings on Privacy Enhancing Technologies* 2017, 4 (2017), 308–328.
[7] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. 2014. A Predictive Differentially-Private Mechanism for Mobility Traces. In *Privacy Enhancing Technologies*, Emiliano De Cristofaro and Steven J. Murdoch (Eds.). Springer International Publishing, Cham, 21–41.

[8] C. Clifton and T. Tassa. 2013. On syntactic anonymity and differential privacy. In *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*. IEEE, IEEE, 88–93. https://doi.org/10.1109/ICDEW.2013.6547433

[9] Lorrie Cranor, Tal Rabin, Vitaly Shmatikov, Salil Vadhan, and Daniel Weitzner. 2015. *Towards a Privacy Research Roadmap for the Computing Community*: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association.

[10] M. Cunha, R. Mendes, and J. P. Vilela. 2019. Clustering Geo-Indistinguishability for Privacy of Continuous Location Traces. In *4th International Conference on Computing, Communications and Security (ICCCS)*. IEEE, 1–8.

[11] Cynthia Dwork. 2008. Differential Privacy: A Survey of Results. In *Theory and Applications of Models of Computation*, Manindra Agrawal, Dingzhu Du, Zhenhua Duan, and Angsheng Li (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–19.

[12] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. 2010. Show Me How You Move and I Will Tell You Who You Are. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS* (San Jose, California) *(SPRINGL '10)*. Association for Computing Machinery, New York, NY, USA, 34–41. https://doi.org/10.1145/1868470.1868479

[13] C. Y. Goh, J. Dauwels, N. Mitrovic, M. T. Asif, A. Oran, and P. Jaillet. 2012. Online map-matching based on Hidden Markov model for real-time traffic sensing applications. In *2012 15th International IEEE Conference on Intelligent Transportation Systems*. IEEE, 776–781. https://doi.org/10.1109/ITSC.2012.6338627

[14] Eelco Herder, Patrick Siehndel, and Ricardo Kawase. 2014. Predicting User Locations and Trajectories. In *User Modeling, Adaptation, and Personalization*, Vania Dimitrova, Tsvi Kuflik, David Chin, Francesco Ricci, Peter Dolog, and Geert-Jan Houben (Eds.). Springer International Publishing, Cham, 86–97.

[15] George R Jagadeesh and Thambipillai Srikanthan. 2017. Online map-matching of noisy and sparse location data with hidden Markov and route choice models. *IEEE Transactions on Intelligent Transportation Systems* 18, 9 (2017), 2423–2434.

[16] Nesrine Kaaniche, Maryline Laurent, and Sana Belguith. 2020. Privacy enhancing technologies for solving the privacy-personalization paradox: Taxonomy and survey. *Journal of Network and Computer Applications* 171 (2020), 102807. https://doi.org/10.1016/j.jnca.2020.102807

[17] John Krumm. 2007. Inference Attacks on Location Tracks. In *Pervasive Computing*, Anthony LaMarca, Marc Langheinrich, and Khai N. Truong (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 127–143.

[18] John Krumm. 2009. A survey of computational location privacy. *Personal and Ubiquitous Computing* 13, 6 (2009), 391–399.

[19] Matej Kubicka, Arben Cela, Hugues Mounier, and Silviu-Iulian Niculescu. 2018. Comparative Study and Application-Oriented Classification of Vehicular Map-Matching Methods. *IEEE Intelligent Transportation Systems Magazine* 10, 2 (2018), 150–166.

[20] Jaewoo Lee and Chris Clifton. 2011. How Much Is Enough? Choosing ϵ for Differential Privacy. In *Information Security*, Xuejia Lai, Jianying Zhou, and Hui Li (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 325–340.

[21] Bo Liu, Wanlei Zhou, Tianqing Zhu, Longxiang Gao, and Yong Xiang. 2018. Location Privacy and Its Applications: A Systematic Study. *IEEE Access* 6 (2018), 17606–17624.

[22] Ricardo Mendes, Mariana Cunha, and João P. Vilela. 2020. Impact of Frequency of Location Reports on the Privacy Level of Geo-indistinguishability. *Proceedings on Privacy Enhancing Technologies* 2020, 2 (2020), 379 – 396.

[23] Ricardo Mendes and João Vilela. 2018. On the Effect of Update Frequency on Geo-Indistinguishability of Mobility Traces. In *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. ACM, 271–276.

[24] Ricardo Mendes and João Vilela. 2018. On the Effect of Update Frequency on Geo-Indistinguishability of Mobility Traces. In *Proceedings of the 11th ACM Conference on Security & Privacy in Wireless and Mobile Networks* (Stockholm, Sweden) *(WiSec '18)*. Association for Computing Machinery, New York, NY, USA, 271–276.

[25] Ricardo Mendes and João P Vilela. 2017. Privacy-Preserving Data Mining: Methods, Metrics, and Applications. *IEEE Access* 5 (2017), 10562–10582.

[26] Luis Moreira-Matias, Joao Gama, Michel Ferreira, Joao Mendes-Moreira, and Luis Damas. 2013. Predicting taxi–passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems* 14, 3 (2013), 1393–1402.

[27] Paul Newson and John Krumm. 2009. Hidden Markov Map Matching through Noise and Sparseness. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Seattle, Washington) *(GIS '09)*. Association for Computing Machinery, New York, NY, USA, 336–343. https://doi.org/10.1145/1653771.1653818

[28] State of California. 2018. The California Consumer Privacy Act of 2018. *California Civil Code* Assembly Bill No. 375 (2018), 1–24.

[29] Simon Oya, Carmela Troncoso, and Fernando Pérez-González. 2017. Is Geo-Indistinguishability What You Are Looking For?. In *Proceedings of the 2017 on Workshop on Privacy in the Electronic Society* (Dallas, Texas, USA) *(WPES '17)*. Association for Computing Machinery, New York, NY, USA, 137–140. https://doi.org/10.1145/3139550.3139555

[30] M. Piorkowski, N. Sarafijanovic-Djukic, and M. Grossglauser. 2009. A parsimonious model of mobile partitioned networks with clustering. In *2009 First International Communication Systems and Networks and Workshops*. IEEE, 1–10. https://doi.org/10.1109/COMSNETS.2009.4808865

[31] V. Primault, A. Boutet, S. B. Mokhtar, and L. Brunie. 2019. The Long Road to Computational Location Privacy: A Survey. *IEEE Communications Surveys Tutorials* 21, 3 (2019), 2772–2793. https://doi.org/10.1109/COMST.2018.2873950

[32] R. Shokri, G. Theodorakopoulos, J. Le Boudec, and J. Hubaux. 2011. Quantifying Location Privacy. In *2011 IEEE Symposium on Security and Privacy*. IEEE, 247–262. https://doi.org/10.1109/SP.2011.18

[33] Reza Shokri, George Theodorakopoulos, Carmela Troncoso, Jean-Pierre Hubaux, and Jean-Yves Le Boudec. 2012. Protecting Location Privacy: Optimal Strategy against Localization Attacks. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security* (Raleigh, North Carolina, USA) *(CCS '12)*. Association for Computing Machinery, New York, NY, USA, 617–627. https://doi.org/10.1145/2382196.2382261

[34] Jun Tang, Aleksandra Korolova, Xiaolong Bai, Xueqiang Wang, and Xiaofeng Wang. 2017. Privacy loss in apple's implementation of differential privacy on macos 10.12. *arXiv preprint arXiv:1709.02753* (2017).

[35] O. K. Tonguz, W. Viriyasitavat, and F. Bai. 2009. Modeling urban traffic: A cellular automata approach. *IEEE Communications Magazine* 47, 5 (2009), 142–150.

[36] European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal L110* 59 (2016), 1–88.

[37] Wenjun Wang, Lin Pan, Ning Yuan, Sen Zhang, and Dong Liu. 2015. A comparative analysis of intra-city human mobility by taxi. *Physica A: Statistical Mechanics and its Applications* 420 (2015), 134–147.

[38] Yong Wang, Yun Xia, Jie Hou, Shi meng Gao, Xiao Nie, and Qi Wang. 2015. A fast privacy-preserving framework for continuous location-based queries in road networks. *Journal of Network and Computer Applications* 53 (2015), 57–73. https://doi.org/10.1016/j.jnca.2015.01.004

[39] Marius Wernke, Pavel Skvortsov, Frank Dürr, and Kurt Rothermel. 2014. A classification of location privacy attacks and approaches. *Personal and Ubiquitous Computing* 18, 1 (2014), 163–175.

[40] Yonghui Xiao and Li Xiong. 2015. Protecting Locations with Differential Privacy under Temporal Correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Denver, Colorado, USA) *(CCS '15)*. Association for Computing Machinery, New York, NY, USA, 1298–1309. https://doi.org/10.1145/2810103.2813640

[41] Shaobo Zhang, Xiong Li, Zhiyuan Tan, Tao Peng, and Guojun Wang. 2019. A caching and spatial K-anonymity driven privacy enhancement scheme in continuous location-based services. *Future Generation Computer Systems* 94 (2019), 40–50. https://doi.org/10.1016/j.future.2018.10.053