The Impact of Mobility on Location Privacy: A Perspective on Smart Mobility

Ekler P. de Mattos[1,2], Augusto C. S. A. Domingues[1], Bruno P. Santos[1,3], Heitor S. Ramos[1] and Antonio A. F. Loureiro[1]

***Abstract*—People use smart transportation systems to move around in smart cities, producing a massive amount of valuable mobility data. Although this characteristic enables the development of many intelligent applications, it can expose users to privacy threats. Location privacy is an issue addressed in many mobility contexts, in which there is a privacy concern. Currently, there are some proposals to tackle this problem, and some questions naturally arise: are these proposals suitable for a dynamic environment, such as smart mobility? What are the impacts of mobility on privacy? In this work, we answer these questions to explore location privacy in smart mobility considering open and online data, one of the fundamental pillars of smart city platforms. We have evidenced the hypothesis that mobility can impact privacy approaches of anonymization (mix-zones) and obfuscation (GEO-I) in the context of smart mobility. For this, we performed experiments to characterize and find similarities in the statistical distributions extracted from two stay points metrics, which operate as substrates to build location privacy protection mechanisms. We use an accuracy metric to quantify the datasets' distributions that matched each other. We conducted a comprehensive evaluation of seven real datasets of mono or multimodal mobility. The results showed that the stay point count metric reached 100% and 83.3% accuracy for coarse-grained (person and vehicle) and fine-grained (bus, taxi, and person) data. Additionally, we show a similarity between distributions for the same vehicle type for mono and multimodal datasets. Results suggest that privacy has a high dependence on mobility in different granularity levels.**

***Index Terms*—Smart Cities, Smart Mobility, Location Privacy, Mix-zone, Geo-indistinguishability, Statistical Analysis.**

I. Introduction

Human mobility refers to the movement of human beings in space and time [1]. The study of human mobility has a fundamental role in developing smart cities, such as urban planning, estimating migratory flows, and developing traffic forecasting applications to help people, vehicles, and things move safely and efficiently [1], [2]. In this context, Smart Mobility emerged as an essential feature associated with smart cities [3]. For example, Smart Mobility can foster a smart transportation system that improves traffic safety and efficiency, reduces citizens' time commuting, and enhances the quality of life [4]. In such a smart transportation scenario, users can combine different transportation modes (e.g., bike, bus, car, and walking) to reduce travel times, traffic, and air pollution. Moreover, people, vehicles, and things act as sensors and produce geo-tagged mobility data, valuable to help the management of assets and efficiently interact with resources and services in smart mobility scenarios. As a result, we typically find three datasets classes according to the transport mode prevalence and its granularity level:

1) Unimodal traces (UT): all trajectories contained in a dataset have a single transportation mode, i.e., there is only one vehicle type in the dataset (Fig. 1a I)).
2) Multimodal traces with unimodal trajectories (MT-UT): in a single dataset, there may be at least more than one transport mode, but the trajectories are associated with a single transport mode (Fig. 1a II)).
3) Multimodal traces with multimodal trajectories (MT-MT): in a single trajectory, there may be more than one mobility type. In this scenario, a user can take a taxi, then walk, and, finally, get a bus to reach his/her destination (Fig. 1a III)).

The benefits of smart mobility are clear, but there are also many privacy concerns. For example, mobility datasets contain not only a set of positions on a map or sensitive places such as home and workplace, among other Points of Interest (POIs). The contextual information attached to a trace tells much about the individuals' habits, interests, activities, and relationships [5]. Thus malicious entities can be mining latent information on these datasets to identify and track users without their consent, which aggravates the privacy threats related to sharing multimodal mobility data, voluntarily or not [6].

Location privacy is a longitudinal issue in mobility. It is a particular type of information privacy for avoiding other entities from learning one's current or past location [7]. Notably, it has gained attention when the Internet of Things (IoT) and the Internet of Vehicles (IoV) contributed to smart mobility connecting objects sharing location information, but unrestricted [8]. Some traditional strategies for anonymization and obfuscation, such as Mix-zones [9] and GEO-I [10], respectively, are being applied to provide location privacy. However, when this happens, we have to ask whether these traditional location privacy approaches are suitable for such a new smart mobility environment, as Scenarios 2 and 3 described above. Location privacy solutions based on obfuscation or anonymization are generally static regarding the setup of its parameters. They are not tuned for different types of datasets and their scenarios and, thus, are not resilient against het-

[1]All the authors are with the Department of Computer Science, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil. Email: {ekler.mattos, augusto.souza, bruno.ps, ramosh, loureiro}@dcc.ufmg.br.

[2]Ekler P. de Mattos is also with the Federal University of Mato Grosso do Sul, Coxim, MS, Brazil.

[2]Bruno P. Santos is also with the Department of Computer and Systems, Federal University of Ouro Preto, João Monlevade, MG, Brazil.

erogeneous mobility data containing other mobility behaviors. Furthermore, what is the degree of impact of mobility on location privacy?

In this paper, we explore the influence of mobility on location privacy. For this, we propose a framework that allows us to characterize and analyze the similarity between types of transport modes through metrics extracted from Stay Points (SPs) – regions in which an entity stays for a minimum time interval [11]. Particularly, we analyze two SPs metrics: *Stay Point Count (SPC)*, which represents the total of different locations visited by users; *Stay Point Duration (SPD)*, which refers to the time a user spends at a location. Unlike other mobility metrics, SPs can provide directions for parameter settings of Location Privacy Protection Mechanisms (LPPMs) techniques, such as the coverage radius size and noise level. Those SPs can be used as indicators of Point of Interests (POIs) and traffic intensity to choose the best place to apply LPPMs [11], [12], [13]. Thus, when studying the distributions of SPs metrics from different transport modes, we found that each transport mode can influence the LPPMs parameters tuning. In the literature, several approaches have been using equal LPPMs parameters for different transport modes [14], [15], [16]. However, this work aims to provide empirical evidence on the impact of mobility on location privacy for different transport modals, without delving into the theoretical analysis, that location data from different modes of transport must be protected by LPPMs calibrated independently.

We conducted a comprehensive evaluation applied to seven real datasets for these analyses (in which six of them refer to the case UT, and one refers to the cases MT-UT and MT-MT). The results showed that the SPC metric reached 100% and 83.3% accuracy for coarse-grained (person and vehicle) and fine-grained (bus, taxi, and person) data, respectively. Additionally, we show a similarity between distributions for the same vehicle type for mono and multimodal datasets. Results suggest that mobility has a high impact on privacy in different granularity levels, enabling us to build a resilient mobility-aware privacy solution. To our knowledge, this is the first study that analyzes stay points to observe the impacts of mobility on location privacy. An essential step in location privacy research, once stay points, is a powerful mechanism for positioning and setting privacy approaches.

II. Related work

This section presents some relevant proposals in the literature about mobility analysis and fallacies found in studies of location privacy.

A. Statistical Analysis of Mobility

The analysis of statistical properties of human mobility can reveal valuable insights for developing various services, including opportunistic networks, traffic monitors, and recommender systems [18], [19], [25]. Thus, exist a broad study on the characterization of distributions extracted from metrics of different mobility contexts, such as banknotes in human mobility, trip displacement in mono/multimodal transport, degree distribution in Online Social Networks (OSNs), yellow intervals on intersections from monomodal transport, and distance and transfer time between POIs in-place semantics context. However, there is little research to understand location privacy from the statistical analysis of mobility aspects, such as stay point metrics. Following, we highlight some relevant literature proposals that pursue distributions characterization of the modal datasets focusing on stopping metrics, such as stopping time in semaphores' yellow light, waiting times between displacements (or trip interval), and traffic accident duration.

Brockmann et al. [17] explored traveling statistics of human mobility over a million individual displacements by analyzing banknotes' circulation in the United States. They concluded that the distribution of the traveling distances decays as a power law, indicating that trajectories of banknotes are similar to Lévy flights [26]. Also, they showed that the probability of pause time distribution (staying in a restricted region) is characterized by a long tail leading to a sub-diffusive process.

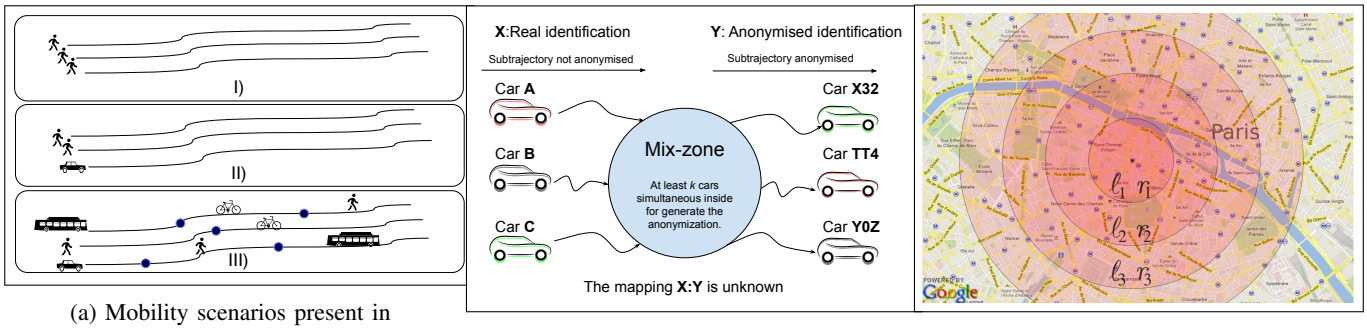
Zhao et al. [18] explored the Lévy walk behavior of human mobility [26]. They decomposed mobility patterns of multimodal datasets into different classes according to transport modes such as, Walk/Run, Bike, Train/Subway, and Car/Taxi/Bus [18]. They concluded that human mobility could be modeled as a mixture of different transport modes. Moreover, single movement patterns can be approximated by a log-normal distribution rather than a power-law distribution.

Xia et al. [19] also analyzed human mobility in both subway and taxi with three metrics: trip displacement (TD), trip duration (TD), and trip interval (TI). The results showed that TD patterns by subway and taxi are similar and follow log-normal distribution rather than an exponential model. Additionally, TD on weekends is different from that on weekdays, no matter the modal. The TD metric is fitted to Weibull distribution for subway and log-normal distribution for taxis. For TI, they concluded that the Weibull distribution can fit the probability curve by taxi rather than log-normal distribution. For the subway, the TI obeys the distribution composed of Weibull and log-normal distributions.

Li et al. [21] explored the stopping behavior during yellow intervals on the semaphores. Notably, they evidenced that the survival curves extracted from stopping time confirm the existence of group-specific effects on drivers based on two metrics: stopping time and drivers' age. The results showed that the log-logistic-based frailty model with age as a grouping variable presents the best goodness of fit and prediction accuracy.

Zhang et al. [25] investigated the prediction curves of traffic accident duration, which provide an important basis for traffic mitigation measures after accidents. They applied AIC and BIC to fit the probability distribution of the accident duration, and the results showed that the log-normal distribution fitted best.

Alessandretti et al. [22] verified the relationship between spatial and temporal properties of human mobility using trajectories of 850 individuals of Copenhagen Network Study composed of GPS and Wi-Fi data. They showed that a log-normal distribution best describes displacements' distribution and waiting times between displacements. They also noticed



(a) Mobility scenarios present in smart mobility datasets.

(b) Mix-zone with $k = 3$.

(c) GEO-I: privacy levels in relation to r.

Fig. 1: (a) item I) UT; II) MT-UT; III) MT-MT. (b) Mix-zone where three cars with pseudonyms A, B, and C enter a mix-zone and attend the minimal *k*=3 and at the exit, receive new pseudonyms (TT4, Y0Z and X32, respectively) without any association with previous ones, cloaking their identities. (c) GEO-I scenario where the privacy level is proportional to the radius.

TABLE I: Related work about statistical analysis of mobility.

Ref	Metric	Accuracy Model[1]	#Datasets	Modal	Similarity Analysis	Similarity Method[1]	Loc. Privacy
[17]	Banknotes	AIC	1	people	No	-	No
[18]	Trip Displacement	AIC	2	walk/run,bike, train/subway, car/taxi/bus	No	-	No
[2]	Trip Displacement	AIC	3	bus, taxi, subway	partially	-	No
[19]	Trip Displacement, Trip Duration, Trip Interval	MLE, BIC	2	taxi, subway	No	-	No
[20]	Degree Distribution, Friendship Node, Hashtag	AIC, BIC, SSE Approach proposed	2	OSN	No	-	No
[21]	Yellow Intervals, Driver's Age, Gender, Phone Status, Maximum Decel./Accel., Vehicle's approaching Speed, Distance between Vehicle's Position, Stopping line when yellow light go up.	AIC, BIC	1	car	No	-	No
[22]	Distr. of Displacements, Waiting Time	AIC	3	people	No		No
[23]	Distance and Transfer Time between POIs	AIC	4	bus, taxi, subway	partially	Pearson Corr.	No
[24]	Srrive-stay-leave, Trajectory Entropy, Number of Trips, Average Velocity, Trip Length, Total Driving Days, and Average Mileage per Day	MLR, DCNN, Approach proposed	3	private car, taxi	yes	MSE, RMSE, MAE, KL, R^2	No
Our work	Stay Point Count (SPC), Stay Point Duration (SPD)	AIC, SSE	7	people, bus, taxi, private car, multimodal	Yes	Wasserstein	Yes

[1] AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion; MLE: Maximum Likelihood Estimation; SSE: Sum of Squared Estimate of Errors; MLR: Multiple Linear Regression; DCNN: Deep Convolutional Neural Network; MSE: Mean-square Error; RMSE: Root-mean-square Error; MAE: Mean Absolute Error, KL: Kullback–Leibler divergence; R^2: degree-of-fit test and the closer the R^2.

correlations between displacements length and the waiting time at destination.

There are also studies about understanding place semantics in mobility [23] and hot zones evolution [24]. Papandrea et al. [23] noticed that POIs have some statistically similar properties among individuals. They classified POIs in terms of their relevance on a per-user basis. Farther, they applied travel metrics (spatial and temporal distances) to four datasets: trajectory, continuous mobility datasets, and two CDR datasets. The trajectory dataset is defined as a unique trip with origin and destiny. In contrast, after starting in the continuous mobility dataset, the user sampling never stops unless the sample collector gets switched off, yielding many trips in a trajectory. The results showed a correlation between these metrics in trajectory datasets.

Xiao et al. [24] investigated the spatiotemporal evolution of urban hot zones (a kind of POIs) from stay points behavior on private cars dataset. They noticed that the hot zones' formation is intricately related to the spatiotemporal coupling correlation of stay points, and its spatiotemporal variation shows certain predictability. Farther, they analyzed mobility patterns between taxis and private cars with trajectory entropy, number of trips, average velocity, trip length, total driving days, and average mileage per day. They concluded that taxis trip, unliked private cars, are different, irregular, and has a high degree of randomness.

Despite the vast literature on the statistical analysis of mobility, few proposals analyze privacy from the perspective of mobility metrics. Further, there are few studies about the similarity between different data sources and transport modes. Alessandretti et al. [22] made an initial analysis of Pearson's correlation between the two datasets. Although the proposal of Papandrea et al. [23] and Xiao et al [24] presented relevant contributions for human mobility analysis, they did not compare similarity levels between datasets. Further, to the best of our knowledge, no previous proposal analyzed these metrics when characterizing location privacy.

Unlike previous studies, here we advance the state of the art w.r.t. SPs. We explore the stay points metrics for characterizing and evaluating the impacts of mobility data on location privacy. In location privacy, the stay point metrics stand out over the mobility metrics discussed above. Once mining the SPs, it is possible to get valuable information about the users' mobility profile (e.g., whereabouts and diary routines) and then define the best placement and configuration of LPPM's instances, such as radius, noise level (ϵ), and k[1]. For this, we propose an analytical framework to analyze two SPs metrics extracted

[1]Minimum of entities into a region for anonymizing.

from different transport modal datasets (see Section IV). Table I summarizes the related work discussion.

B. Mobility Effects on Security and Location Privacy

In a smart mobility scenario, people equipped with smartphones, vehicles, and things can be seen as mobile devices, allowing users to access a rich set of mobile services. Nevertheless, these resource-constrained devices need fast and easy access to mobile services without multiple credentials of the users. In this way, password-based single-sign-on authentication has been widely applied in mobile environments. An authentication token is generated on an identity server, and one can request mobile services from related service providers without multiple registrations [27]. However, this model introduces privacy and security threats. For instance, if an adversary accesses the identity server, one can retrieve users' passwords by performing Dictionary guessing attacks (DGA) and overissue authentication tokens to break the security [27], [28].

Other securities and privacy issues occur when cloud services provide data deduplication to users equipped with mobile devices to save storage space. For instance, the user's data must be cyphered by a symmetric encryption method like Message-locked encryption (MLE) to avoid leaking private information. However, MLE is vulnerable to brute-force DGA. Additionally, MLE schemes are subject to key management problems, mainly if users access different devices. Thus, to mitigate DGA and key management problems, some proposals have focused on applying secure distributed secret sharing protocols [28], [29].

In different location privacy studies [14], [15], [16], we can see evidence about the impact of mobility on privacy by analyzing the performance discrepancy of an LPPM applied to different transport modes. For instance, in vehicular mobility, some proposals showed significant differences in the re-identification rate between the datasets of buses and cars of the same city [14]. Additionally, other studies also found divergences in the re-identification rate in datasets of different cities, such as datasets of cabs in Rome and buses in Shanghai [15]. In both studies, datasets were submitted to LPPMs and configured with the same parameters.

Some investigations have identified privacy divergences of users registers [16] in datasets, which are not all equal in the face of re-identification attacks. This means that some users' profiles can never be re-identified even in the absence of LPPMs, while others can be easily re-identified. The authors argued that this difference in users' protection level is that no generic LPPM provides the same protection level for different users' profiles. Moreover, the resilience of an LPPM against re-identification attacks depends on the underlying data. For instance, the LPPM settings to protect the location data of a user walking and using his/her smartphone may differ from those of a user driving a private car due to mobility characteristics such as speed, direction, and frequency of visits to places.

One of the reasons for these accuracy differences of re-identification attacks is that possibly these datasets were protected by inappropriate or misconfigured LPPMs or with the static setting, which did not consider the dynamic scenarios with different transport modes. As a result, we have datasets with low protection and utility. Next, we emphasize some essential privacy issues when using classical LPPMs to smart mobility.

III. Location Privacy Issues in Smart Mobility

This section shows the classical LPPMs and collection of issues concerning privacy and mobility aspects addressed in smart mobility, organized in three scenarios: general, anonymization, and obfuscation. Nevertheless, we need first to understand what privacy threats are through an adversary model.

A. Adversary Model

Defining a consistent adversary model is important to outline a location privacy attack's limits. This model makes it possible to have a panoramic view of privacy threats and define more appropriate mitigation actions. Therefore, we present an adversary model capable of carrying out both anonymized and obfuscated data attacks.

The adversary model can be defined as follows. Let be $\mathcal{F}$ and $\mathcal{G}$ be two functions that represent the LPPMs anonymization and obfuscation, respectively. Also, let be $\mathcal{D}'$ an open dataset $\mathcal{D}$, but protected by $\mathcal{F}$ or $\mathcal{G}$, being $\mathcal{D}' \leftarrow \mathcal{F}(\mathcal{D})$ or $\mathcal{D}' \leftarrow \mathcal{G}(\mathcal{D})$. The adversary may also have access to some training traces (possibly noisy or incomplete) of users and other public contextual information, represented by a profile B_u for each user u. The above information applied to $\mathcal{D}$ represents the adversary's background knowledge about the users $\mathcal{B} = (B_1, \ldots, B_b, \ldots, B_m)$, where $[1 \leq b \leq m]$ and b represents the number of elements in $\mathcal{B}$ known by the adversary, enabling the adversary to execute a Tracking Attack $\mathcal{Z}$ or Points of Interest Attack $\mathcal{W}$, for example.

In a Tracking Attack (TA), the adversary's objective is to determine the whole sequence (or a partial subsequence) of events in a user's trace. Given an anonymized dataset $\mathcal{D}'$ composed of users and background B_u, a tracking attack is defined as $T_u \leftarrow \mathcal{Z}(\mathcal{D}', B_u)$, where T_u represents the reconstructed trajectory of user u.

A Points of Interest Attack (POIA) uses location points or regions where people commonly stay at a given instant, such as home or workplace, to characterize users' profiles. Given a set of regions on map $R \in \mathcal{B}$, a period of time t, and an obfuscated dataset $\mathcal{D}'$ composed of users. An adversary can be interested in discovering the most visited locations L_s of users s in $\mathcal{D}'$ at time t. That is, $L_s \leftarrow \mathcal{W}(\mathcal{D}', R, t)$. It is not needed the exact location, but the region on the map.

We can observe privacy threats through an adversary model, even if an LPPM protects data $\mathcal{D}$, from adversary's background knowledge $\mathcal{B}$, which is the type of LPPM (anonymization $\mathcal{F}$, or obfuscation $\mathcal{G}$) applied in $\mathcal{D}$. For example, if the adversary has $\mathcal{B}$ that the $\mathcal{D}$ was protected by $\mathcal{G}$. The adversary knows that users' sensitive locations have been obfuscated and may have low success with a POI attack $\mathcal{W}$, but the users' identity was not protected. Thus, the adversary can be highly accurate in identifying the identity of users with tracking attack $\mathcal{Z}$.

Likewise, if the adversary has $\mathcal{B}$, which $\mathcal{D}$ was protected by $\mathcal{F}$, a POI attack $\mathcal{W}$ will have a high accuracy in identifying the location, as the POIs was not protected. A way to protect open data is to apply a hybrid LPPM based on anonymity and obfuscation. However, we encountered issues between privacy and utility, as defined below.

B. General Scenario

In smart mobility open data, there are issues related to the decision about the type of privacy to achieve, the order of applying these approaches, and how to set up the LPPMs to get an optimal trade-off between obfuscation and anonymization. However, applying these techniques does not mean achieving location privacy fully but can be effective for datasets used for a specific purpose that requires one type of protection than another. Depending on the systems that use the protected dataset (called consumers), these issues can affect/change privacy and utility goals.

Suppose a congestion reduction scenario, where it is necessary to test a system that monitors traffic flow. In the test dataset, the fine-grained geolocation points are needed for measuring the traffic efficiently, and a high distortion on this data may lose its utility. In this case, the test dataset should consider obfuscation since it is more sensitive than anonymization. Thus, this dataset should have a high level of anonymization and a low level of obfuscation. These questions are described in Fig. 2.

Questions **Q1** and **Q2**, hybrid mechanisms (obfuscation and anonymization), deal with a new trend for the LPPM design but, at the same time, pose new challenges. For instance, hybrid mechanisms are intricately dependent on the context in which each customer's family will use the protected dataset.

Question **Q3** addresses the issue that there is no one-size-fits-all LPPM for many privacy scenarios without losing privacy and utility, such as in datasets MT-UT and MT-MT. We detail this issue and directions for a possible solution in the following.

C. Anonymization Scenario

Although there are studies on mix-zones to optimize anonymization effectiveness, whether in silence period strategies [30], [31], cryptography [32], positioning [33], or modeling of its geometric region [34], few efforts have been conducted to investigate mix-zones in the context of smart mobility open data. Specifically, the side effects regarding privacy and data utility when the LPPM parameters are not calibrated in an environment with different modals. Here we have identified some issues about these concerns.

Mix-zones is a technique to anonymize a dataset and, thus, its protection. It selects urban regions where the simultaneous anonymization of vehicles (or people) occurs by changing their current pseudonym [9]. We need to have at least k entities within the mix-zone (see Fig. 1b) to anonymize them. The mix-zone parameters are the radius (r), the minimum number of entities in the mix-zone to change the pseudonym (k), and the geo-position.

From a spatial point of view, we need to calibrate the mix-zone radius (r) for different entities considering the trade-off between privacy and utility. We can gain anonymity coverage with a larger radius since more entities are likely to be inside the zone simultaneously. However, within a mix-zone, there is a period of silence in which location records are discarded. Therefore, the larger the mix-zones radius, the greater the data gaps, which might compromise the dataset's utility. Additionally, people's mix-zone radius may be smaller, as people tend to have a lower speed than vehicles. But, what are the radii for two different types of entities (like MT-UT and MT-MT scenarios) within mix-zones?

Other spatial issues are the number and location of mix-zones on a map, depending on that area's mobility characteristics. If high data privacy is desirable or needed, the mix-zones must be positioned in higher traffic regions [35]. The positioning of mix-zones over time is a temporal issue that we need to consider. Once the best positioning for mix-zones is defined, these points may lose their effectiveness to anonymize over time. With the flow variation of entities, some mix-zones may not make sense anymore, whereas other regions may need them. Thus, the problem is how to define the lifetime of a mix-zone.

Tuning the parameter k of a mix-zone is also affected by a given area's mobility characteristics. In high vehicle traffic or crowd, it is desirable to have a high value of k. In contrast, a low-traffic region requires a small value of k, which must be defined to cover data anonymization and a higher privacy level significantly. Also, k is decisory for pseudonyms changing. A k low may yield excessively pseudonyms changing that implies a high anonymity level. However, this harms open data once it produces a significant number of sliced trajectories, losing its utility. Further, in an online context, the excess on pseudonyms changing can negatively affect communication protocols (e.g., routing task) and applications that need long-term communication relationships (e.g., file transfer or interactive chat sessions) [30], [31]. Fig. 2 presents a summary of these anonymization issues.

D. Obfuscation Scenario

An obfuscation scenario also presents many issues concerning privacy and utility. For example, Geo-indistinguishability (GEO-I) [10] is an obfuscation technique based on data perturbation. It protects the user's location by adding spatial noise extracted from a Laplace distribution to the actual user's location in the mobility trace [10]. GEO-I considers the privacy level l to be proportional to the radius r, and defines an ϵ-geo-indistinguishability as $\epsilon = l/r$. The value of ϵ represents a level of privacy for l within r, and proportionally selects a privacy level for all other radii observing that the lower the ϵ, the higher the noise (see Fig. 1c). The GEO-I approach is necessary to consider the noise level applied to data. If it is high, one may risk distorting the data by losing its utility. Nevertheless, if the noise level is low, the data may not be properly protected to ensure its privacy. Therefore, the noise level must be adjusted to handle this trade-off.

Another parameter to consider is the GEO-I radius to identify the user's POIs and introduce noise. The higher the radius is, the greater the chance of identifying the POIs of a mobile entity, but also the greater the noise to be applied to protect

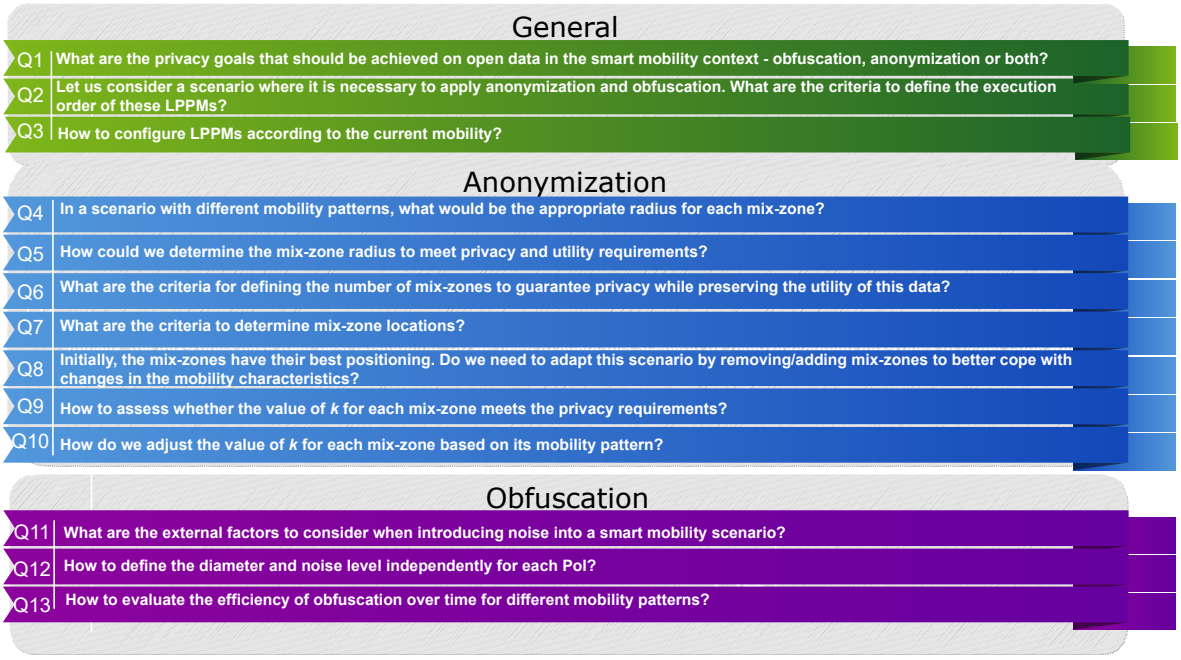


Fig. 2: Location privacy issues in smart mobility.

the region. Thus, for datasets with multimodal trajectories, the radii may have different values. The radius to identify the vehicle's POIs may differ from that of people's. People's POIs usually have smaller perimeters, such as home, workplace, tourist point, campus building. In contrast, vehicles' POIs have a larger perimeter, such as the airport area for POI of taxis or a parking yard for car rental. However, what would be the ideal radius for MT-UT and MT-MT scenarios? Must these POIs radius/noise levels be varied with time? Fig. 2 presents a summary of these obfuscation issues.

Based on these facts, we can state the following hypotheses:

> ***Hypothesis 1****: Mobility reflects directly on location privacy.*
>
> If we consider Hypothesis 1 valid, which is quite reasonable, then we have:
>
> ***Hypothesis 2****: Different types of mobility need different profiles of privacy and utility, independent of the applied privacy model.*

These questions have no trivial answers. However, a possible direction is to understand how privacy behaves in different mobility patterns to address Hypotheses 1 and 2. In this work, we propose analyzing the impacts of mobility on location privacy by analyzing SPs, as a substrate used in many privacy algorithms, to identify specific behaviors in different transport modes. Specifically, we intend to characterize and analyze distributions extracted from SPs of different datasets and compare them. The goal is to verify the similarity between distributions of the same and different transport modes. The similarity level is expected to be low between different transport modes and high for the same modes of transportation. In this way, we highlight the hypothesis. More details are presented below.

[2]This algorithm is also used for the *stay point duration* metric. To do this, replace Line 7 with the function that calculates the stay point duration metric.

Algorithm 1: Characterization and Similarity Analysis of Stay Points Count metric[2].

Data: $\mathfrak{D} = \{\mathcal{D}_1, \mathcal{D}_1, ..., \mathcal{D}_n\}$ set of distinct datasets.
Result: Υ: best fit distribution for each $D_i \in \mathfrak{D}$; S: similarity matrix; Φ: Accuracy matching of mobility group;

1 **for** $\mathcal{D}_i \in \mathfrak{D}$ **do**
2 $\quad P_i \leftarrow Extract\_SP\_Param(\mathcal{D}_i)$
3 $\quad SP_i \leftarrow Extract\_SP(\mathcal{D}_i, P_i)$
4 $\quad \mathcal{SP} \leftarrow \mathcal{SP} \cup SP_i$
5 **end**
6 **for** $SP_i \in \mathcal{SP}$ **do**
7 $\quad SPU_i \leftarrow Extract\_SP\_by\_User(SP_i)$
8 $\quad \mathcal{SPU} \leftarrow \mathcal{SPU} \cup SPU_i$
9 **end**
10 **for** $SPU_i \in \mathcal{SPU}$ **do**
11 $\quad distr_i \leftarrow Best\_Fit\_Distr(SPU_i)$
12 $\quad \Upsilon \leftarrow \Upsilon \cup distr_i$
13 **end**
14 **for** $SPU_i \in \mathcal{SPU}$ **do**
15 $\quad$ **for** $SPU_j \in \mathcal{SPU}$ **do**
16 $\quad\quad S[i,j] \leftarrow WM(SPU_i, SPU_j)$
17 $\quad$ **end**
18 **end**
19 $\Phi \leftarrow ACC(S)$
20 **return** Υ, S, Φ

IV. Methodology

This section presents an analytical framework for analyzing location privacy with SPs. Firstly, we define the SPs and their relationship to the LPPMs. Next, we detail the steps for extraction, characterization, and similarity analysis of stay points metrics.

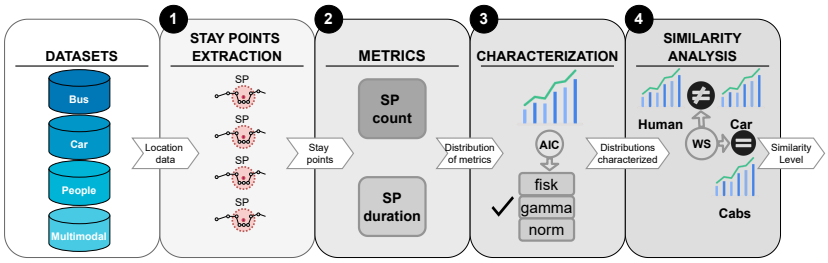


Fig. 3: The framework for analyzing location privacy with stay points.

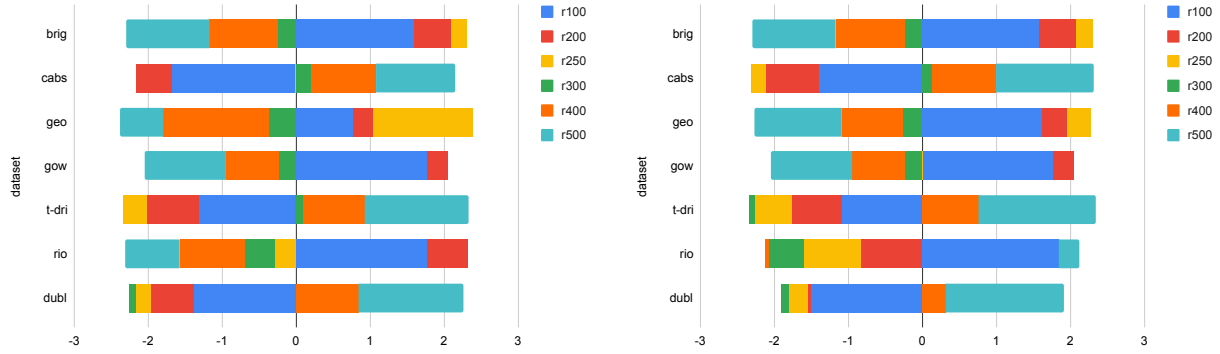
TABLE II: Datasets details.

Id	Name	Location	Transp. type	#users	#reg.	#Staypoints[1]
brig	Brightkite [36]	LBSN	human	51,406	4,747,287	2,679,758
gow	Gowalla [36]	LBSN	human	107,092	6,442,892	3,733,344
cabs	Cabspotting [37]	San Francisco, USA	taxi	532	6,837,027	28,589
t-dri	T-Drive [38]	Beijing, China	taxi	10,320	17,652,648	187,183
dubl	Dublin Bus [39]	Dublin, Ireland	bus	911	43,851,182	52,961
rio	Rio Bus [37]	Rio de Janeiro, Brazil	bus	13,954	51,845,217	570,894
geo	Geolife [40]	Beijing, and 36 cities in China, USA, South Korea, and Japan.	multimodal	182	24,876,978	27,862
geo-car	Geolife cars [40]	—	cars	36	512,807	770
geo-cabs	Geolife cabs [40]	—	taxi	29	242,018	449
geo-bus	Geolife bus [40]	—	bus	43	1,276,632	1679
geo-human	Geolife human [40]	—	human	61	2,535,433	3320

[1] Stay points setup $SP_p\langle R = 250, T = 30\rangle$ for brig and gow datasets, $SP_v\langle R = 500, T = 30\rangle$ for cabs, t-dri, dubl, rio, geo, and geo-* datasets.

TABLE III: Staypoints values from analysis radius and time to stay.

	t15						t30					
Dataset	**r100**	**r200**	**r250**	**r300**	**r400**	**r500**	**r100**	**r200**	**r250**	**r300**	**r400**	**r500**
brig	**3050695**	2874132	2826407	2753173	2642387	2611167	**2882499**	2722982	2679758	2612758	2509555	2481030
cabs	31437	31467	31479	31484	31501	**31506**	28453	28487	28513	28530	28566	**28589**
geo	37349	37295	**37409**	37229	37117	37206	**28433**	27864	27862	27603	27360	27198
gow	**4340465**	3967355	3894363	3838986	3718058	3620534	**4146983**	3799599	3733344	3683107	3571545	3480181
t-dri	319472	356156	380332	405721	450533	**486246**	162470	166066	167868	171871	179502	**187183**
rio	**893318**	839561	803565	797655	777507	783632	**582907**	562466	563038	565332	568527	570894
dubl	93308	96572	98176	98668	102506	**104893**	51879	52390	52308	52371	52511	**52961**



(a) Radius [100-500] and time to stay at least 15 mins.

(b) Radius [100-500] and time to stay at least 30 mins.

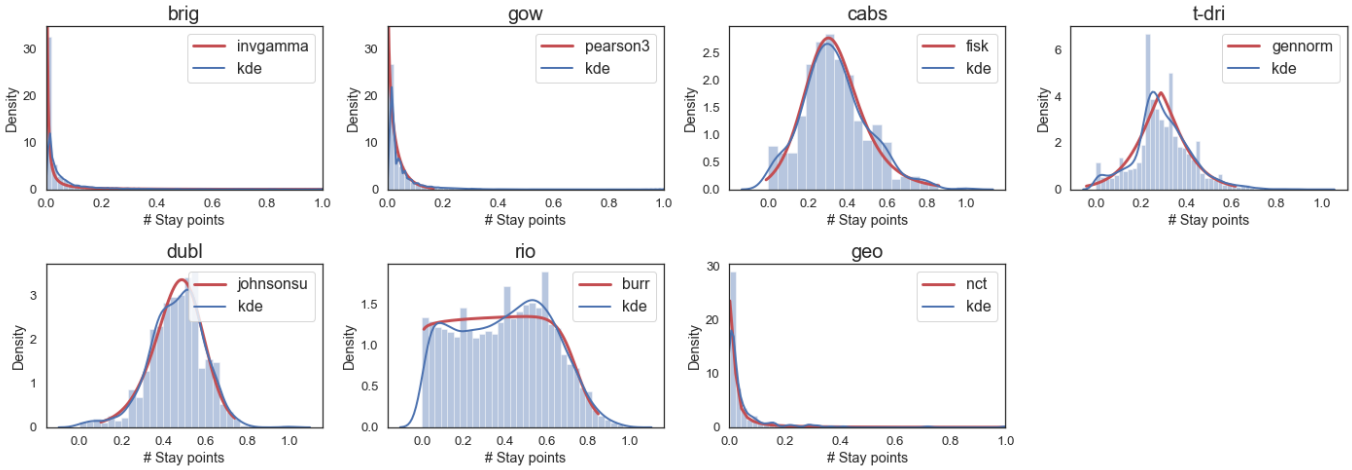
Fig. 4: Stay points extraction with many radius and time to stay time threshold.

A. Analyzing Location Privacy with Stay Points

Stay Point (SP) is a region where an entity stays for a minimum time interval [11]. The parameters of an SP are the radius r in meters of the region and the minimum time to stay there t in minutes. These points are relevant for detecting many mobility characteristics, such as traffic lights and even traffic jams. Stay points are commonly used as a substrate for many privacy mechanisms in the location privacy context. In LPPM design, stay points are typically used to detect POIs and apply obfuscation methods. Additionally, stay points can be used for mix-zones placement [6]. In location attacks, stay point mining enables to identify and characterize behaviors in the

Best Distr.	brig	gow	cabs	t-dri	dubl	rio	geo
invgamma	−**207,956**	−392,287	−3,218	−54,411	−4,966	−111,489	−430
pearson3	−78,572	−**452,154**	−3,189	−54,387	−5,406	−112,594	72
fisk	−136,175	−416,304	−**3,377**	−55,407	−5,332	−110,197	122
gennorm	−73,731	−321,543	−3,171	−**55,626**	−5,276	−123,696	−180
johnsonsu	−83,889	−421,870	−3,214	−55,570	−**5,496**	−112,781	−207
burr	−114,167	−437,034	−3,167	−55,216	−5,431	−**134,550**	223
nct	−205,646	−387,561	−3,228	−55,435	−5,112	−112,661	−**451**

(a) Best distribution and AICc(SSE) value for the SPC metric.



(b) Distribution of each dataset and best fit distribution (in red color: probability density function (pdf)), calculated with AICc(SSE).

Fig. 5: Best fit distribution for SPC metric.

users' trajectory, revealing sensitive information, such as social preferences, since victims regularly go to those places [6], [41]. In this way, stay points bring valuable information w.r.t. location privacy.

Fig. 3 details the steps for analyzing location privacy with stay points. Step 1, we extract SP from seven different real mono and multimodal mobility datasets (in which six of them refer to the case UT and one referring to the cases MT-UT and MT-MT). Step 2, from each SP set, we extract the distributions of two SP metrics: SPC and SPD. Step 3, we characterize the distributions according to the best fit statistic model. Step 4, for each SP metric, we compare the similarity between datasets' distributions. Specifically, we verify if there is a divergence between the distributions, in terms of SPs, for vehicles and people. Next, we refine the vehicle category for cars, cabs, and buses. We use an accuracy metric to quantify the number of distributions that matched each other. The steps are defined in Algorithm 1 and detailed in the following sections.

B. Extraction of Stay Points

The number of collected SPs. on stay points extraction is related to their radius and time to stay parameters. Many empirical studies in the literature set up the SP parameters [11], [42]. For instance, Zheng et al. [11] argued that radius and time to stay parameters enable finding significant places, such as restaurants and shopping malls. At once, it is possible to ignore geo-regions without semantic meaning, like places where people wait for traffic lights. They extracted 10,354 stay points from the dataset with 107 users using mobile devices in Beijing, including 36 cities in China and a few cities in the USA, South Korea, and Japan. Chen et al. [42] used spatial density clustering and temporal Gaussian Kernel Density to extract spatiotemporal features from sparse and non-stationary stay behavior data. Concerning SP set up parameters, questions naturally arise: what is these parameters' setup to obtain optimal stay points extraction? Is there a setup pattern for datasets of different natures, such as transport modals?

Here, we analyzed the parameter tuning for the SPs extraction considering various transport modals to answer these questions (Line 2 of Algorithm 1). The goal is to identify how tuning the radius and time parameters affects the result set of SPs in different transport modals collected from seven datasets (see Section V). We defined a testing scenario of ranging the radius threshold r from 100 meters to 500 meters ($S_R = \{100, 200, 250, 300, 400, 500\}$), and the time threshold S_T from 15 minutes to 30 minutes ($S_T = \{15, 30\}$). For each combination of radius $r \in S_R$ and time value $t \in S_T$, we extracted the set of SPs for all users in each mobility trace (see Table III). We can see that the settings to extract stay points for people and vehicles are different.

For the people datasets, the best configuration was the smallest radius ($r100$), whereas for the vehicle datasets were the largest radius ($r500$). This fact is best seen in Figs. 4a and 4b that represent the z-score standardization of the stay

points count in the datasets for all the radius and time values $t15$ and $t30$. For both time configurations, the vehicle datasets – cabs, t-dri, rio and dubl – got a positive score above value 2, i.e., collected more SPs with $r500$. The only exception was the rio dataset, which for $t15$ prevailed $r200$. The people datasets gow, brig and geo prevailed radius $r250$, $r200$ and $r100$, respectively. Geolife is a multimodal dataset, but most of its mobility records refer to people, which leads to bias, with lightning configurations below $r250$. The SP extraction results suggest that the transport modal significantly influences the parameters' choice to collect a more significant amount of stay points. Therefore, we adopted two parameters set, $SP_p\langle R = 250, T = 30\rangle$, $SP_v\langle R = 500, T = 30\rangle$, applied to human and vehicular datasets, respectively. These stay points setups are already used in [11] for people datasets and in [42] for vehicles datasets, as they have contexts close to ours. Optimal tuning of SPs parameters is context-dependent and has several open issues [11], [42].

C. Stay Points Metrics

We use the metrics related to the stay points aspects: *SPC per user* and *SPD* (Lines 6–9 of Algorithm 1).

Stay Point Count (SPC) per user refers to the total of different locations one visits. This metric can be used to understand the different mobility characteristics of users. Its distribution contains the locations visited by users in which some may have only a few places while others may have a large collection. That can be used for POI extraction, routing algorithms, and contagion models.

Stay Point Duration (SPD), also called the *duration of stay at a stay point*, refers to the time a user spends at a location. The lower bound is defined by the stay time algorithm's parameter, and the upper bound has no limit. Understanding the time users (or population) spend on average at a location can be a good indicator of its capacity regarding data offloading, helping to design handoff solutions.

D. Distributions characterization

The fundamental step for identifying divergences between transport mode datasets is to characterize the distributions (*dist*) obtained from SP metrics (Lines 11–12 of Algorithm 1). That is, identify a representative model of *dist* from a set of candidate models. For this, we calculated the Akaike Information Criterion (AIC) from the Sum of squared estimate of errors (SSE) of each distribution candidate (AIC (SSE))[3]. AIC is an estimator of out-of-sample prediction error, widely used in statistical analysis, that evaluates a collection of models for the data and estimates the quality of each model w.r.t. the other models [18], [2], [19], [25]. The lowest value of AIC will be the best fit distribution for *dist*. Thus, AIC provides a means for model selection. Specifically, we use an AIC correction (AICc) to address potential overfitting for small sample sizes. SSE is a measure of the discrepancy between the data and an estimation model. It is used as an optimal criterion in parameter selection and model selection. A small SSE indicates a tight fit of the model to the data. AIC works for samples of different sizes, including non-normal distributions.

E. Analysis of Similarities between Distributions

Statistical distance is the approach we use to identify the distance between two probability distributions. We applied the Wasserstein metric (WM), which measures the difference between two distributions by the optimal cost of rearranging one distribution into the other[4] (Lines 14–18 of Algorithm 1). The smaller the WM value is, the less effort to transform one distribution into another, and, consequently, the two distributions show high similarity (Line 19 of Algorithm 1). The Wasserstein distance is asymmetric, (weakly) continuous, and ideal for analyzing corrupted data, in contrast to common distributions divergence approaches, such as Kullback-Leibler or Jensen-Shannon [43]. For the WM, we carry out three types of analysis to identify:

WMA1 similarities of distributions between the two transport generic types: vehicular and human.

WMA2 similarities between distributions of the same category of vehicles. For example, if a taxi distribution is more similar to the distribution of another taxi than a distribution of buses or people.

WMA3 similarities between distributions of the same generic type but extracted from monomodal and multimodal transport datasets. For instance, if the taxicab dataset matches a vehicle category from a multimodal dataset, such as car, cabs, or bus.

V. Experimental Results

We conducted an experimental similarity analysis of different transport modes. To do so, we used distributions extracted from stay points in datasets of different transport modes. We use SP metrics to extract the distributions, as previously defined.

We used seven datasets with different transport modes (see Table II). We analyzed datasets of human and vehicular mobility and multimodal transport. Additionally, we selected at least two datasets for each type of vehicle to check for possible similarities between them. Also, for Geolife, we analyzed the trajectories of humans, cars, buses and, cabs categories[5] separately.

We evaluated the performance of the WM, in the task of associating the distributions generated by the stay points metrics, using the accuracy metric, i.e., $Accuracy = \frac{distributions\ match\ correctly}{total\ distributions} \times 100$.

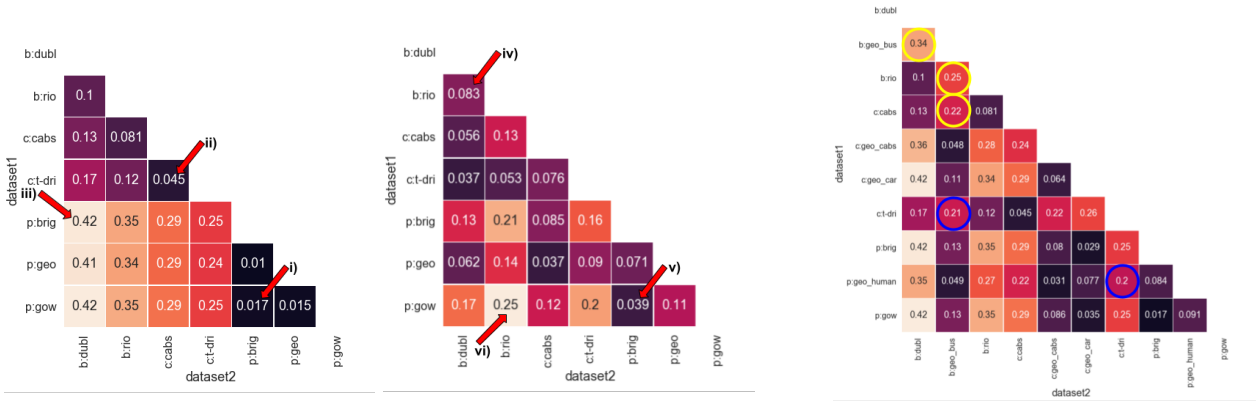
For instance, the WMA2 analysis with two datasets' distributions for each type of vehicle distinct: cabs, people, and buses; Totalize six distributions. During the WM matching, there was a more significant similarity between distributions of the same type, such as taxis-taxis, and buses-buses, matching a total of four datasets and providing an accuracy of 67%.

For the characterization of the distributions produced by the metric *SPC* for each dataset, we used 89 types of distributions

[3]For more details please refer to K. P. Murphy, Machine Learning: A Probabilistic Perspective. MIT Press, 2012.

[4]For more details, please refer to C. Villani, "Topics in Optimal Transportation". American Mathematical Soc., 2003, no. 58

[5]We used all datasets, except the Rio de Janeiro bus dataset, in which we used a sampling of 10 days corresponding to 33% of all dataset



(a) SP count.

(b) SP duration.

(c) SP count for the categories inside Geolife: human, bus, car, and cabs.

Fig. 6: Wasserstein distance for SPC and SPD metrics. Analysis with people=r250 and vehicles=r500, both t=30 minutes. The metrics are grouped by vehicle category: people (p), car (c), and bus (b).

available in the library in Scipy[6]. Table 5a shows the results of AIC(SSE), in which the best-fit distribution to the SPC metric for each dataset[7]. For example, for the stay points count of the Cabspotting and Brightkite datasets, the distributions that obtained the best fit were the fisk and invgamma distributions, respectively, among the set of distributions.

Fig. 5b shows the distribution of the number of stay points by users for each dataset. We can see the similarity between datasets of the same category. For instance, human datasets, such as Gowalla and Brightkite, tend to be similar. Both distributions indicate that many users have only one stay point. The vehicular category, like cabs datasets Cabspotting and T-Drive, also have similar distributions. Still, they tend to a normal distribution, in which we observed that many vehicles show more than one stay point. However, datasets of different categories, such as human and vehicular, tend to present divergences, as Brightkite and Rio de Janeiro Bus. Differences between stay points distributions can affect LPPMs regarding the number of instances, parameters (*radius*, ϵ, and k), lifetime, and placement. For example, when applying user-level obfuscation with GEO-I to people and using SP as a preliminary step, we will have one GEO-I instance per user, due to the nature of the SP distribution for people, with few places to obfuscate. In contrast, for vehicles with many SPs per user, we will have more than one GEO-I instance per user.

Similarity analysis in Fig. 6 shows the Wasserstein distance of SPC and SPD metrics for the datasets. The WM values in squares are similarity levels between the pairs of datasets analyzed. Low WM near zero (dark-colored squares) represents more similarity between the pairs of datasets distributions than high WM near 0.43 (light-colored squares). The WMA1 analysis for the SPC metric (see Fig. 6a) reaches 100% match accuracy between distributions of the same mobility group. Distributions of datasets of the same transport mode tend to be similar, while distributions of different transport modes tend to be distant. For example, the dataset pairs (Gowalla, Brightkite) and (Cabspotting, T-Drive) have WM values of 0.017 and 0.045 (see red arrows i and ii in Fig. 6a). In the meantime, the WM value for (Brightkite, Dublin) is 0.42 (red arrow iii). This behavior is also similar to the SPD metric, where it shows 100% accuracy in matching for distributions of people and vehicles (Fig 6b). For instance, the WM value of bus distributions (Dublin, Rio de Janeiro) and people distributions (Brightkite, Gowalla) is 0.083 and 0.039 (red arrows iv and v in the figure). In contrast, the WM value for (Rio de Janeiro, Gowalla) is 0.25 (red arrow vi).

In the WMA2 analysis, for the SPC metric, the distributions of the same transport modes tend to be similar, while the distributions of different transport modes tend to distance themselves. The matching accuracy for distributions of the same transport mode is 83.3%. Considering taxicabs (Cabspotting, T-Drive) distribution, the WM reaches 0.045, while the people (Brightkite, Gowalla) distribution reaches 0.017 (Fig. 6a). Although Geolife is a multimodal transport dataset, it is close to people's transport mode, with a WM value 0.01 for the pairs (Geolife, Gowalla). However, the SPD metric achieves a matching accuracy of 34% for the same transport mode. Although there is no direct matching for this metric, the WM values for distributions of the same transport mode category are very close, such as (T-Drive, Dublin), (T-Drive, Rio de Janeiro) pairs, in which both belong to vehicular mobility (Fig. 6b). There are two reasons why Geolife, a multimodal dataset, resembles people's mobility. First, the data was sensed by users carrying cell phones, different from vehicles with fixed sensors. Second, 55% of the dataset records labeled are assigned as people.

Further, we analyzed the association between monomodal datasets and categories of the multimodal dataset. To do so, we extracted four datasets from Geolife that represent the labeled data of people (p:geo human) and vehicles as private cars, taxis, and buses (c:geo car, c:geo cabs, and b:geo bus, respectively), totalizing ten datasets. We extracted their SPs,

[6]For more details, please refer to https://docs.scipy.org/doc/scipy/reference/stats.html

[7]We omitted the *SPD* AICc(SSE) results due to space limitations in this work.

then the SPC metric to analyze their similarities with the WS distance and verify the accuracy as depicted in Fig. 6c. In this analysis, for WMA1, the accuracy reaches 70% in classifying the distributions according to the transport modes (vehicular and human). For the WMA2, the accuracy is 50%. The Geolife categories had a more significant similarity, except c:geo car, which had a considerable similarity with the p:brig and p:gow datasets.

Concerning WMA3 is 50% accuracy of association between monomodal and Geolife categories. Additionally, among the four distributions of monomodal transport (Cabspotting, T-Drive, Rio de Janeiro, and Dublin datasets), three had a match with the Geolife vehicle categories, highlighted by circle yellow in the figure. The bus distributions, such as Rio de Janeiro and Dublin, were associated with the Geo-Bus category. Although the T-Drive distribution is associated with Geo-Human, we can see that it is close to the vehicle distribution Geo-Bus (highlighted with a blue circle in the figure), with a difference of 0.01 between WM values.

Based on the analysis of different types of datasets with the SPC and SPD metrics, we have the following insight:

> Independently of granularity level, whether in coarse (person and vehicle) or fine granularity, such as type of vehicle (taxi, bus, and person), distributions of the same mobility type tend to converge to each other. However, distributions of different types of mobility tend to diverge from each other.

Indeed, the SPs metrics applied to different transport mode datasets can be considered a fingerprint for each type of mobility, containing their inherent characteristics. Therefore, these fingerprints reflect the context of location privacy approaches, evidencing Hypotheses 1 and 2. This is especially true in smart mobility, in which multimodal trajectories can be joined in a unique trace having an identity from a set of signatures.

VI. A Light at the End of the Tunnel

The previous results show strong evidence of the hypothesis: mobility affects location privacy. The use of classic LPPMs approaches, such as mix-zone and GEO-I, reveal promising perspectives for the future of location privacy in smart mobility. Remarkably, we can model the privacy issues about mix-zone and GEO-I as optimization problems given smart mobility's dynamic nature. Hence, it encourages designing adaptive LPPMs aware of mobility to preserve its subtleties of privacy and utility. Each LPPM instance must be independent, including a variation of its parameters. For instance, for mix-zones, the radius can vary according to the type of vehicle in it as well as the LPPM instances' lifetime and positioning also can be modeled as optimization problems.

A proposal of a privacy framework based on optimization can be an iterative process composed of LPPM's tuning, protection, attack, and testing steps. In the tuning step, the LPPM's parameters are tuned according to the transport mode. In the protection step, the dataset is protected with anonymization and/or obfuscation. Attack step, the protected dataset is submitted to re-identification attacks. The testing step verifies the privacy and utility levels. If the attack step succeeds, then the defined parameters are not satisfactory, and thus need to be tuned again and resubmitted to the attack process. This cycle is repeated until reaching an acceptable trade-off of privacy and utility.

VII. Conclusion

In this work, we have explored location privacy in smart mobility. We identified gaps in the privacy approaches, the anonymization (mix-zones), and obfuscation (GEO-I) in the context of smart mobility. We have evidenced the hypothesis that mobility can affect privacy. For this purpose, we carried out experiments to find similarities in the distributions extracted from two SP metrics, applied to seven datasets of mono and multimodal mobility. Specifically, an accuracy metric was used to quantify the datasets' distributions that matched each other. The results showed that the SPC metric reached 100% and 83.3% accuracies for coarser-grained (person and vehicle) and finer-grained (bus, taxi, and person), respectively. Additionally, we showed a similarity between distributions of the same vehicle type for mono and multimodal datasets. We also showed that vehicles and people have specific patterns about stay points extraction. The results suggest that privacy has a high dependence on mobility in different levels of granularity. To our knowledge, this is the first work that uses statistical analysis of stay points to observe the mobility influence on location privacy.

As future work, we plan to extend our work with theoretical analysis. Also, we intend to design a mobility-aware privacy solution based on an optimization technique that considers both privacy and utility in smart mobility open data. Furthermore, we plan to add more mobility datasets, such as private cars, contact patterns, and call details record datasets. Finally, we intend to explore more stay points with complex networking and collective mobility metrics to pave the hypothesis addressed in this work.

References

[1] H. Barbosa *et al.*, "Human mobility: Models and applications," *Physics Reports*, vol. 734, pp. 1–74, 2018.

[2] S. Jiang, W. Guan, W. Zhang, X. Chen, and L. Yang, "Human mobility in space from three modes of public transportation," *Physica A: Statistical Mechanics and its Applications*, vol. 483, pp. 227–238, 2017.

[3] D. Eckhoff and I. Wagner, "Privacy in the smart city?applications, technologies, challenges, and solutions," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 1, pp. 489–516, 2017.

[4] Z. Ning, F. Xia, N. Ullah, X. Kong, and X. Hu, "Vehicular social networks: Enabling smart mobility," *IEEE ComMag*, vol. 55, no. 5, pp. 16–55, 2017.

[5] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in *2011 IEEE symposium on security and privacy*. IEEE, 2011, pp. 247–262.

[6] V. Primault, A. Boutet, S. B. Mokhtar, and L. Brunie, "The long road to computational location privacy: A survey," *IEEE Communications Surveys & Tutorials*, 2018.

[7] A. R. Beresford and F. Stajano, "Location privacy in pervasive computing," *IEEE Pervasive computing*, no. 1, pp. 46–55, 2003.

[8] S. Paiva, M. A. Ahad, S. Zafar, G. Tripathi, A. Khalique, and I. Hussain, "Privacy and security challenges in smart and sustainable mobility," *SN Applied Sciences*, vol. 2, pp. 1–10, 2020.

[9] E. P. Mattos, A. C. Domingues, and A. A. F. Loureiro, "Give me two points and i'll tell you who you are," in *Proceedings of the IEEE Intelligent Vehicles Symposium (IV'19)*. IEEE, 2019.

[10] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," *arXiv preprint arXiv:1212.1984*, 2012.

[11] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *Proceedings of the 18th international conference on World wide web*, 2009, pp. 791–800.

[12] Q. Yu, Y. Luo, C. Chen, and X. Zheng, "Road congestion detection based on trajectory stay-place clustering," *ISPRS International Journal of Geo-Information*, vol. 8, no. 6, p. 264, 2019.

[13] D. Minatel, V. Ferreira, and A. A. Lopes, "A multilevel approach for building location-based social network by using stay points," in *8th Brazilian Conf. on Intelligent Systems*. IEEE, 2019, pp. 359–364.

[14] C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao, "Privacy vulnerability of published anonymous mobility traces," *IEEE/ACM transactions on networking (TON)*, vol. 21, no. 3, pp. 720–733, 2013.

[15] Y. Khazbak and G. Cao, "Deanonymizing mobility traces with co-location information," in *2017 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2017, pp. 1–9.

[16] M. Maouche, S. B. Mokhtar, and S. Bouchenak, "Ap-attack: a novel user re-identification attack on mobility datasets," in *Proceedings of the 14th EAI Int. Conf. on Mobile and Ubiquitous Systems: Computing, Networking and Services*. ACM, 2017, pp. 48–57.

[17] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel," *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

[18] K. Zhao, M. Musolesi, P. Hui, W. Rao, and S. Tarkoma, "Explaining the power-law distribution of human mobility through transportation modality decomposition," *Scientific reports*, vol. 5, no. 1, pp. 1–7, 2015.

[19] F. Xia, J. Wang, X. Kong, Z. Wang, J. Li, and C. Liu, "Exploring human mobility patterns in urban scenarios: A trajectory data perspective," *IEEE ComMag*, vol. 56, no. 3, pp. 142–149, 2018.

[20] S. Bhattacharya, S. Sinha, S. Roy, and A. Gupta, "Towards finding the best-fit distribution for OSN data," *The Journal of Supercomputing*, pp. 1–19, 2020.

[21] J. Li, H. Zhang, Y. Zhang, and X. Zhang, "Modeling Drivers' Stopping Behaviors during Yellow Intervals at Intersections considering Group Heterogeneity," *Journal of advanced transportation*, vol. 2020, 2020.

[22] L. Alessandretti, P. Sapiezynski, S. Lehmann, and A. Baronchelli, "Multi-scale spatio-temporal analysis of human mobility," *PloS one*, vol. 12, no. 2, p. e0171686, 2017.

[23] M. Papandrea, K. K. Jahromi, M. Zignani, S. Gaito, S. Giordano, and G. P. Rossi, "On the properties of human mobility," *Computer Communications*, vol. 87, pp. 19–36, 2016.

[24] Z. Xiao, H. Xiao, W. Chen, H. Chen, A. Regan, and H. Jiang, "Exploring Human Mobility Patterns and Travel Behavior: A Focus on Private Cars," *IEEE Intelligent Transportation Systems Magazine*, 2021.

[25] J. Zhang, W. Junhua, and F. Shou'en, "Prediction of urban expressway total traffic accident duration based on multiple linear regression and artificial neural network," in *2019 5th Int. Conf. on Transportation Information and Safety (ICTIS)*. IEEE, 2019, pp. 503–510.

[26] P. Barthelemy, J. Bertolotti, and D. S. Wiersma, "A Lévy flight for light," *Nature*, vol. 453, no. 7194, pp. 495–498, 2008.

[27] B. C. Neuman and T. Ts'o, "Kerberos: An authentication service for computer networks," *IEEE Communications magazine*, vol. 32, no. 9, pp. 33–38, 1994.

[28] Y. Zhang, C. Xu, H. Li, K. Yang, N. Cheng, and X. Shen, "PROTECT: efficient password-based threshold single-sign-on authentication for mobile users against perpetual leakage," *IEEE Transactions on Mobile Computing*, vol. 20, no. 6, pp. 2297–2312, 2020.

[29] Y. Zhang, C. Xu, N. Cheng, and X. S. Shen, "Secure Password-Protected Encryption Key for Deduplicated Cloud Storage Systems," *IEEE Transactions on Dependable and Secure Computing*, 2021.

[30] K. Emara, W. Woerndl, and J. Schlichter, "CAPS: Context-aware privacy scheme for VANET safety applications," in *Proc. of the 8th ACM Conf. on security & privacy in wireless and mobile networks*, 2015, pp. 1–12.

[31] I. Memon, L. Chen, Q. A. Arain, H. Memon, and G. Chen, "Pseudonym changing strategy with multiple mix zones for trajectory privacy protection in road networks," *Int. journal of communication systems*, vol. 31, no. 1, p. e3437, 2018.

[32] J. Freudiger, M. Raya, M. Félegyházi, P. Papadimitratos, and J.-P. Hubaux, "Mix-zones for location privacy in vehicular networks," in *ACM Workshop on Wireless Networking for Intelligent Transportation Systems (WiN-ITS)*, no. CONF, 2007.

[33] B. Palanisamy and L. Liu, "Attack-resilient mix-zones over road networks: architecture and algorithms," *IEEE Transactions on Mobile Computing*, vol. 14, no. 3, pp. 495–508, 2014.

[34] ——, "Mobimix: Protecting location privacy with mix-zones over road networks," in *2011 IEEE 27th International Conference on Data Engineering*. IEEE, 2011, pp. 494–505.

[35] A. R. Svaigen, H. S. Ramos, L. B. Ruiz, and A. A. Loureiro, "Dynamic temporal mix-zone placement approach for location-based services privacy," in *2019 IEEE LATINCOM*. IEEE, 2019, pp. 1–6.

[36] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," http://snap.stanford.edu/data, Jun. 2014.

[37] Crawdad, "Crawdad: A Community Resource for Archiving Wireless Data At Dartmouth," https://crawdad.org/keyword-vehicular-network.html, Aug. 2020.

[38] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proceedings of the 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2011, pp. 316–324.

[39] D. C. Council. dublin bus gps sample data from dublin city council (insight project).

[40] "Microsoft Research. geolife trajectories (v. 1.3)," Downloaded from research.microsoft.com/jump/131675, Aug. 2012.

[41] X. Liu, H. Zhao, M. Pan, H. Yue, X. Li, and Y. Fang, "Traffic-aware multiple mix zone placement for protecting location privacy," in *2012 Proceedings IEEE INFOCOM*. IEEE, 2012, pp. 972–980.

[42] J. Chen, Z. Xiao, D. Wang, W. Long, J. Bai, and V. Havyarimana, "Stay time prediction for individual stay behavior," *IEEE Access*, vol. 7, pp. 130 085–130 100, 2019.

[43] J. H. Oh, M. Pouryahya, A. Iyer, A. P. Apte, A. Tannenbaum, and J. O. Deasy, "Kernel Wasserstein Distance," *arXiv preprint arXiv:1905.09314*, 2019.

Ekler Paulino de Mattos received his B.Sc. degree in Computer Science from the State University of Mato Grosso do Sul (2002) and the M.Sc. degree in Electrical Engineering from the State University of Campinas (2007). He is a Ph.D. candidate from the Department of Computer Science at the Federal University of Minas Gerais. Currently, he is an assistant professor in the Information Systems course at the Federal University of Mato Grosso do Sul - Campus Coxim. His research areas are Computer Networks, Distributed Systems, working mainly on the following topics: mobile/ubiquitous computing, network management.

Augusto C. S. A. Domingues received a B.Sc. in Computer Science from the Federal University of Viçosa, and an M.Sc. degree in Computer Science from the Federal University of Minas Gerais. Currently, he is a Ph.D. candidate at the Federal University of Minas Gerais. His research areas involve Computer Networks and Distributed Systems, working mainly on the following topics: Mobile networks, Trajectory Mining, Mobility Behavior, and Human Behavior Characterization.

Bruno Pereira dos Santos holds a B.S. in Computer Science from Universidade Estadual de Santa Cruz - UESC (2013), Master and Ph.D. degrees (2015 and 2019) in Computer Science from Universidade Federal de Minas Gerais - UFMG. Currently, he is a Professor at Universidade Federal de Ouro Preto - UFOP. His main research interest areas are Computer Networks, Internet of Things, Wireless Sensor Networks, Mobile computing, Ubiquitous computing, Software Defined Networks, 5G Networks, Distributed systems/algorithms, Parallel computing/programming, Cyber-physical systems, software, and tools.

Heitor Soares Ramos Filho received a B.Sc. in Electrical Engineering from the Federal University of Paraíba (1997), an M.Sc. in computational modeling from the Federal University of Alagoas (2006), and a Ph.D. in Computer Science from the Federal University of Minas Gerais (2012). He is currently is an associate professor in the Department of Computer Science at the Federal University of Minas Gerais (DCC / UFMG). His research interests are focused on Data Analysis for the Internet of Things, Mobility, Sensor Networks, Social Networks, Urban Computing and Vehicular Networks.

Antonio Alfredo Ferreira Loureiro received a B.Sc. in Computer Science from the Federal University of Minas Gerais (1983), an M.Sc. in Computer Science from the Federal University of Minas Gerais (1987) and a Ph.D. in Computer Science from the University of British Columbia, Canada (1995). He is currently a Full Professor of the Computer Science Department of the Federal University of Minas Gerais. He has experience in the area of Computer Science, with emphasis on distributed systems, working mainly on the following topics: distribute algorithms, mobile/ubiquitous computing, wireless communication, network management, computer networks, wireless sensor networks, and vehicular networks.