

Analysing Fairness of Privacy-Utility Mobility Models

YUTING ZHAN, HAMED HADDADI, Imperial College London, UK

AFRA MASHHADI, University of Washington, USA

Preserving the individuals' privacy in sharing spatial-temporal datasets is critical to prevent re-identification attacks based on unique trajectories. Existing privacy techniques tend to propose ideal privacy-utility tradeoffs, however, largely ignore the fairness implications of mobility models and whether such techniques perform equally for different groups of users. The quantification between fairness and privacy-aware models is still unclear and there barely exists any defined sets of metrics for measuring fairness in the spatial-temporal context. In this work, we define a set of fairness metrics designed explicitly for human mobility, based on structural similarity and entropy of the trajectories. Under these definitions, we examine the fairness of two state-of-the-art privacy-preserving models that rely on GAN and representation learning to reduce the re-identification rate of users for data sharing. Our results show that while both models guarantee group fairness in terms of demographic parity, they violate individual fairness criteria, indicating that users with highly similar trajectories receive disparate privacy gain. We conclude that the tension between the re-identification task and individual fairness needs to be considered for future spatial-temporal data analysis and modelling to achieve a privacy-preserving fairness-aware setting.

1 INTRODUCTION

Understanding human mobility based on collected locations from mobile devices has become a fundamental part of urban and environmental planning in cities [28]. These GPS traces enable the scientific community and policymakers to model citizens' daily mobility patterns (*e.g.*, crowd-sensed car sharing, ride sharing, city bicycles sharing, and RFID-card-based public transportation, or build predictive algorithms to estimate people's flows and community structure [13]. However, location-based traces corresponding to human mobility, even at an aggregate level, have raised numerous privacy concerns [8, 38], mainly when the data contains sensitive and revealing insights about people's identity, behaviour, associations, religion, and others [23].

In the past decades, the research community has examined various ways of ensuring the privacy of mobility traces. Previous work, ranging from k -anonymity [1], differential privacy (*i.e.*, DP) [35, 41], to information-theoretic metrics [32, 45], explore scientific guarantees that data subjects cannot be re-identified while the data remain practically useful. More recently, *privacy-utility trade-off (PUT)* models based on machine learning or deep learning techniques that aim to optimize both privacy and utility (*i.e.*, inference accuracy) have been studied and shown to be superior to the previous approaches [11]. These techniques can be summarized as representation learning [22], generative adversarial network (GAN) [31, 33], reinforcement learning [10, 11], etc. In these works, researchers have shown that it is possible to design and implement frameworks that enhance the privacy protection of individual trajectories without a significant reduction of the trace's utility.

A dimension that has been vastly overlooked is whether privacy-preserving algorithms work equally for all users or whether they could lead to unexpected consequences of protecting the privacy of only a group of people. Indeed, as recent evidence from the broader machine learning domain has shown, the systematic discrimination in making decisions against different groups has been shifted from people to autonomous algorithms [19, 24]. In many applications, discrimination may be defined by different protected attributes, such as race, gender, ethnicity, and religion, that directly prevent favourable outcomes for a minority group in societal resource allocation, education equality, employment opportunity, etc [36]. Similarly, in the context of spatial-temporal data, mobility demand prediction algorithms have been

shown to offer higher service quality to neighbourhoods with more white people [5]. However, in such contexts, only a handful of recent studies exist that examine the fairness of location-based systems [16, 42, 43], with little consensus on how fairness should be defined and measured for spatial-temporal applications.

In this work, we aim to measure and evaluate the fairness of the location privacy-preserving algorithms applied to mobility traces. We seek to answer the research question as to *whether the outcome of the PUT models satisfies fairness*. Extended from the notion of fairness in broader machine learning literature, fairness in location privacy-preserving mechanisms could also be concluded in two categories: *individual fairness* and *group fairness*. *Individual fairness* ensures that similar users receive similar outcomes with respect to the specific privacy-aware inference tasks [4]. That is, whether these privacy-aware models preserve the privacy and service quality of similar users equally. In order to do so, we first posit a set of similarity metrics to mathematically denote a notion of user similarity grounded on the human mobility literature [15], in terms of both the structural similarity of their heatmap images and the entropy of their trajectories. On the other hand, *group fairness* ensures the independence between the model outcome and a sensitive attribute (i.e., gender, age, ethnicity, etc) of interest. That is, it ensures equal privacy gain and utility loss over multiple groups.

We examine two machine learning-based privacy-preserving approaches (i.e., TrajGAN [33] and Mo-PAE [44]), compared to the original inference tasks that optimize only for privacy or for utility. We evaluate their fairness on two real-life mobility datasets: Geolife [46] and MDC [25]. Our results indicate that both TrajGAN and Mo-PAE do not guarantee *individual fairness*; users with similar trajectories might receive different privacy gain outcomes where the *individual fairness* criteria are violated in these location privacy-aware settings. More specifically, we observe that for the users with similar traces, even when the outcome of the prediction task is identical, the privacy gains amongst those users are highly different, leading to some users not advantaging from obfuscation as others do. Different to the highlights of individual fairness, the results of *group fairness* of privacy-aware models show that there is no demographic disparity in the privacy and prediction outcome. However, as we discuss, this observation highly reflects the socio-cultural settings where these traces have been collected and are less of a by-product of the privacy-preserving models. The contributions of our paper are as follows:

- We theoretically denote the notion of *similarity* for tackling the measurements of individual fairness of spatial-temporal datasets.
- We offer a set of *individual fairness* metrics specifically defined based on mobility characteristics that can help the broader research community measure fairness for spatial-temporal applications.
- We examine the privacy-preserving algorithms in terms of both individual fairness and group fairness on two representative mobility datasets, and show their deficiencies in accounting for fairness can lead to undesired consequences.
- We systematically discuss why individual fairness and group fairness are competing in the privacy-aware setting.

2 RELATED WORK

2.1 Fairness in Machine Learning

Literature on fairness in machine learning (Fair-ML) tends to focus on *the absence of any prejudices or favoritism toward an individual or group based on their inherent or acquired characteristics* [30]. The majority of fairness research strives to avoid the decision made by automated systems skewed toward the advantaged groups or individuals. In [15], authors proposed a framework for understanding different definitions of fairness through two views of the world: i) *we are*

all equal (WAE, mostly ensuring the *group fairness*), and ii) *what you see is what you get* (WYSIWYG, mostly ensuring the *individual fairness*). The framework shows that the fairness definitions and their implementations correspond to different axiomatic beliefs about the world, described as two fundamentally incompatible worldviews. A single algorithm cannot satisfy either definition of fairness under both worldviews [15].

The most adopted metrics for fairness in machine learning are widely based on the WAE assumption and denoted as *group fairness*, which is also known as *statistical parity* and *demographic parity* [9]. These metrics aim to ensure that there is independence between the predicted outcome of a model and sensitive attributes of age, gender, and race. If variations of *statistical parity* exist, Fair-ML will concentrate on relaxation of this measure by ensuring that groups from sensitive attributes and non-sensitive attributes meet the same misclassification rate (*false negative rate*, also known as *equalized odds* [17]), or *equal true positive rate* (also known as *equal opportunity* [17]).

In the context of mobility data and its applications, such as equitable transportation, research attention also mainly devoted to group fairness. Transportation equity heavily employs statistical tests for equity analysis, which is appropriate for discovering unfairness [43]. The such metric is often defined based on census tract information, which offers an aggregate demographic characteristic of the residing population. (author?) [42] defined fairness in terms of the region-based fairness gap and assesses the gap between mean per capita ride-sharing demand across groups over time. The two metrics differ from each other. One is based on a binary label associated with the majority of the sub-population (e.g., white), and the other is based on a continuous distribution of the demographic attributes. Similarly, (author?) [18] proposed a graph-based approach for integrating group-based (census) into e-scooter demand prediction. Through the integration of an optimization regularizer, they showed that it is possible for their model to jointly learn the flow patterns and socio-economic factors, and returns socially-equitable flow predictions. Hosford et al. [20] investigated the equity of access to bike sharing in multiple cities in Canada. Ge et al. [16] studied racial and gender discrimination in the expanding transportation network companies. These handfults of recent works all focus on group-based fairness metrics and collective methods (e.g., demand or flow prediction).

On the other hand, individual fairness claims that similar individuals should be treated similarly concerning the target task [9]. For example, in making hiring decisions, the algorithm has to possess perfect knowledge of comparing the “qualification” of two individuals. In most cases, the difficulty with individual fairness lies in the notion of measuring *similarity*. For example, (author?) [43] used the population and employment density of each city area for achieving individual fairness in bike-sharing demand prediction. The difficulty, again, lies in the fact that there is often a lack of perfect knowledge to determine the *similarity* in demand between two areas. In broader spatial-temporal data and application, the definitions of mobility similarity are almost non-existence, so as individual fairness of spatial-temporal data. Although previous work in fairness literature [6] has examined the boundary of fairness and privacy, these works have been applied to low dimensional datasets (e.g., COMPAS) that differ greatly from complex mobility data of people. In this work, we offer a new perspective on how to measure individual fairness metrics defined based on the literature on mobility and examine its application in assessing the fairness of the privacy-preserving algorithms applied to mobility traces.

2.2 Privacy Methods for Spatial-Temporal Data

Large-scale human mobility data contain crucial insights into understanding human behaviour but are hard to share in non-aggregated form due to their highly sensitive nature. Decades of research on privacy examined various anonymous mechanisms on human trajectories [1, 35, 41]. A mobility privacy study conducted by De Montjoye et al [8] illustrates that four spatial-temporal points are enough to identify 95% of the individuals in a certain granularity, demonstrating

the necessity of the anonymous mechanism against the re-identification attack. Previous work, ranging from k-anonymity [1], differential privacy [35, 41], to information theoretic metrics [32, 45], explore scientific guarantees that the subjects of the data cannot be re-identified while the data remain practically useful. More recently, PUT models based on machine learning, which simultaneously aim to optimize for data privacy protection and utility, are emerging. In these lines of work, researchers have focused on the objective of training neural network models that optimize for reducing privacy leakage risk of individual trajectories while at the same time minimizing the depreciation in the mobility utility. These models have been shown to be superior to differential privacy techniques. In this paper, we selected two machine learning-based PUT models based on two different strategies of GAN and Representation Learning, but both with promising high performance in terms of both utility and privacy. These two PUT models mainly focus on temporal correlations in time-series data and aim to reduce the user re-identification risk (i.e., privacy) while minimizing the downgrade in the accuracy of mobility prediction task (i.e., utility). We describe the details of these two privacy-aware spatial-temporal models:

TrajGAN [33]: it is an end-to-end deep learning model to generate synthetic data that preserves the real trajectory data’s essential spatial, temporal, and thematic characteristics. Compared with other standard geo-masking methods, TrajGAN can better prevent users from being re-identified. TrajGAN claims to preserve essential spatial and temporal characteristics of the original data, verified through statistical analysis of the generated synthetic data distributions, which is in a line with the data utility assessment based on the mobility prediction task in our work. Hence, we train a TrajGAN-based PUT model to evaluate the mobility predictability and privacy protection of synthetic data generated by TrajGAN.

Mo-PAE [44]: it is a **privacy-preserving adversarial feature encoder**. In contrast to the TrajGAN that aims to generate synthetic data, Mo-PAE trains an encoder Enc_L that forces the extracted representations f to convey maximal information about data utility while minimizing private information about user identities via adversarial learning. It consists of a multi-task adversarial network to learn an LSTM-based encoder Enc_L , which can generate the optimized feature representations $f = Enc_L(X)$ via lowering the privacy disclosure risk of user identification information (i.e., privacy) and improving the mobility prediction accuracy (i.e., utility) concurrently.

3 FAIRNESS DEFINITION AND METRICS

In this section, we first define the mathematical representation of fairness in spatial-temporal applications before we incorporate it into our analysis.

3.1 Formulation of the Problem

In this work, we aim to measure and evaluate the fairness of the privacy-preserving algorithms applied to mobility traces. We seek to figure out whether these models equally preserve the user privacy and inference accuracy of similar users. We try to determine whether fairness metrics benefit from a privacy-preserving model simultaneously, laying a theoretical foundation for further research on the privacy-preserving fairness-aware mechanism for human mobility. Both individual- and group-based fairness are discussed.

We first introduce some basic notations and abbreviations utilized in this work: individuals are labelled as u , if individuals u_i and u_j are similar, that is $u_i \sim u_j$; sensitive or protected attributes are denoted as S ; raw data without sensitive attributes is denoted as X ; Y is the ground-truth labels for a specific inference task and Y' is the predicted one, which is the variant that depends on S and X . The true positive rate (i.e., TPR, recall, or sensitivity) is utilized to judge the performance of the multi-categorical classifiers, which refers to the proportion of who should be predicted

accurately that received a positive result. TPR is also utilized in the inference tasks’ quality of the examined models and is denoted as *task accuracy*.

3.2 Individual Fairness

Individual fairness [9] states that individuals who are similar, with respect to a specific task, should be treated similarly (i.e., $P_{u_i} \sim P_{u_j}$ when $u_i \sim u_j$) [34]:

$$P(Y'|u_i, S, X) = P(Y'|u_j, S, X) \quad (1)$$

As we have mentioned in Section 2.1, the difficulty with individual fairness lies in the notion of measuring *similarity*. To measure individual fairness in the context of spatial-temporal data, we need two sets of definitions corresponding to i) the similarity between users’ *trajectories* (SIM_t); and ii) the similarity of the *outcome* of the PUT models (SIM_o), as well as their generalizability for different mobility datasets and PUT models. We define each next:

3.2.1 Similarity of Trajectories. Grounded on the literature on mobility [13, 29, 38], we mathematically denote the notion of trajectory similarity (SIM_t) based on i) the *structural similarity index* of mobility heatmap images; and ii) the *entropy* of trajectories.

Structural Similarity Index Measure (SSIM): SSIM was initially designed to quantify image quality degradation caused by processing, such as data compression or losses in data transmission, which leverages the differences between the reference image and the processed image [40]. To apply SSIM metrics in this work, we construct *heatmap* images from the raw geo-located data with the methodology proposed by [13]. Figure 1 shows some sample heatmap images with spatial granularity coarsening from 50 meters to 900 meters by the left to right. These heatmap images structurally represent mobility features extracted from mobility traces, which use pixel intensity to encode the *frequency* of the visit spent in a given area; hence, the brighter pixels denote the more frequently visited locations of the user. SSIM has been shown to be a well-suited metric to compute the image similarity of the heatmap images specifically when applied to mobility heatmap images [13, 29]. Unlike Mean Square Error, the SSIM metric has been shown not to be significantly impacted by the changes in luminosity and contrast.

In this work, we formulate the SSIM measure as the perceptual difference of two similar users’ heatmap images, H_i and H_j . See the Appendix for full definitions. We then leverage the integrated heatmap image, which combines all user trajectories, to calculate the effective SSIM index ($SSIM_{eff}$) that indicates the overall trajectory similarity of users. The SSIMs between individual ($SSIM_{one}$) and integrated trajectory ($SSIM_{eff}$) are denoted by calculating the SSIM *maps* (i.e., local values of the SSIM, $SSIM_{maps} = abs(SSIM_{eff} - SSIM_{one})$). $SSIM_{maps}$ is utilized to lower the impact of the unreached area, that is, only the swept area in the integrated heatmap image was selected for further analysis. Hence, the average SSIM value of the selected points is what we define as $SSIM_{eff}$. Additionally, as this metric relies on heatmap images, it is highly influenced by spatial granularity, where each pixel in the image corresponds to the spatial boundary of the data. Intuitively, in Figure 1, as the granularity coarsens, the trajectories become blurry and, thus, more similar. The impact of the spatial granularity on the SSIM index will discuss in Section 5.1.1.

Entropy of Trajectories (EOTs): Mobility literature defines the highest potential accuracy of predictability of any individual, termed as “maximum predictability” (Π_{max}) [27]. Maximum predictability is determined by the *entropy* of a person’s trajectory information (e.g., frequency, sequence of location visits, etc.). Hence, some similar characteristics of user spatial-temporal patterns are able to be captured by leveraging the entropy of trajectory. In this paper, we conclude and define four types of entropy to measure trajectory similarity for spatial-temporal applications, denoted as *Shannon Entropy* (SE), *LonLat Entropy* (LE), *Heatmap Entropy* (HE), *Actual Entropy* (AE). The integrated entropy of these

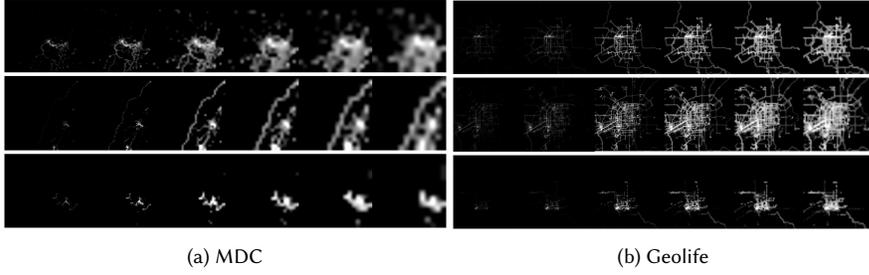


Fig. 1. Sample mobility heatmap images with various spatial granularities of MDC and Geolife. Three different trajectories are shown with different granularities (50 m, 100 m, 300 m, 500 m, 700 m, and 900 m).

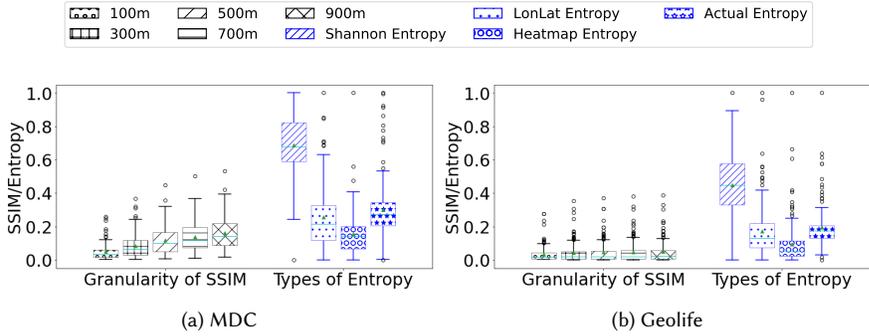


Fig. 2. Overview of SSIM and entropy distribution of trajectories of MDC and Geolife datasets. Different granularities of SSIM are compared in a row, where the granularity are ranging from 100-meter to 900-meter.

four different types of entropy is denoted as EOTs. The details of them are as followed, and see the Appendix for full definitions.

i) *Shannon Entropy (SE)*: the entropy of probabilities of visited location distribution. Leveraging the common definition of Shannon entropy (E_h), a classic notion of data uncertainty, we first calculate E_h of the trajectory to characterize visited location distribution and their probabilities. A larger E_h indicates greater disorder and consequently reduces the predictability of an individual's movements.

ii) *LonLat Entropy (LE)*: the entropy of the geo-located locations in a time-series format. Considering the spatial-temporal pattern of the mobility data, the entropy of visited locations in terms of longitudes and latitudes are separately estimated by using the fuzzy entropy E_f . This entropy reflects the probability of a new sub-string and quantifies the irregularity or complexity of the time-series data.

iii) *Heatmap Entropy (HE)*: the entropy of the users' heatmap images. In contrast to the aforementioned entropy models, we define a two-dimensional entropy (E_{2D}) to quantify the irregularity (i.e., unpredictable dynamics) of the user's heatmap image. The entropy of trajectory heatmap images is calculated using the two-dimensional sample entropy method (*SampEn_{2D}*) [37]. In a trajectory heatmap image, the features are extracted by accounting for the spatial distribution of pixels in different m -length square windows.

iv) *Actual Entropy (AE)*: the entropy of capturing entire spatial-temporal order present in user's mobility pattern. To capture AE, (author?) [38] proposed an actual entropy model using the Lempel-Ziv algorithm. Different to other types

of entropy, AE depends not only on the frequency of visited locations but also on the order in which the nodes were visited and the time spent at each location [38]. In this work, the given area is segmented using structured grids, where each grid is initialized as 0. Then the visited locations and whether the person reached the cell previously are tracked. If the person visits an unreached cell, the location is marked as 1, generating time-series binary data to characterize the trajectory.

See the Appendix for full definitions and related equations of these four different entropies.

3.2.2 Similarity of Users. With the aforementioned definition of trajectory similarity, we mathematically define the users with similar trajectories as the *similar users* by two techniques:

i) ϵ -thresholding: setting the threshold ϵ to filter similar users based on their trajectories' similarity. To be specific, if the trajectory similarity of u_i and u_j is greater than a threshold ϵ , that is $SIM_t(u_i, u_j) > \epsilon$, this pair of users will be selected out as the *similar users*, that is $u_i \sim u_j$.

ii) clustering: grouping similar users together via *clustering* techniques. We use k-means clustering to cluster users based on their SSIM and EOTs features. We apply the Elbow and Silhouette method [26] to determine the number of clusters (k values). The resulting clusters present a group of highly similar users together.

3.2.3 Similarity of Outcome. To understand whether users with similar trajectories receive similar outcomes from the models, we first need to define what it means to receive a *similar outcome* mathematically. As the objective of the PUT models is to optimize privacy gain and minimize utility loss, we consider privacy gain and utility gain as positive outcomes. After selecting out the similar users, we then measure the difference of them in privacy gain outcome, $\Delta D_{pri} = 1 - D_{pri}(u_i)/D_{pri}(u_j)$, and utility gain outcome, $\Delta D_{uti} = 1 - D_{uti}(u_i)/D_{uti}(u_j)$. Both ΔD_{pri} and ΔD_{uti} contribute to the evaluation of SIM_o . When with the *clustering* approach, the average pairwise differences of ΔD_{pri} and ΔD_{uti} for all the members of each cluster are assessed.

Regardless of the grouping technique in similar users, we argue that ΔD_{pri} or ΔD_{uti} satisfies fairness if it is within $1 - \epsilon$, otherwise, the PUT model is considered to be *violating individual fairness* for user pair u_i and u_j . The threshold of different combinations of SSIM and EOTs are utilized to distinguish similar users and map all users into a list of *pairs* with trajectory similarity and performance discrepancy. To measure the fairness of systems as a whole for each model and outcome, we report the percentage of user pairs for whom fairness was violated (i.e., *violation%* or *V%*). As we will show, in our experiments, we set $\epsilon = 0.8$ to correspond to users with at least 80% similarity of trajectory which imposes the model's outcome to be within 20% difference between the similar users. The choice of $\epsilon = 0.8$ is based on the various literature in fairness and literature [2, 12]. We discuss the impact of this threshold on policy making in the Discussion section of this article.

3.3 Group Fairness

Different to individual fairness lies heavy on the similarity definition, group fairness has been vastly discussed and shares a systematic analysis approach in broader Fair-ML study. In this work, we bridge the gap between the standard group fairness metrics and the specific privacy-preserving mechanism of spatial-temporal data.

Group fairness as also referred to as Demographic Parity [15] states that demographic groups should receive similar decisions, inspired by civil rights laws in different countries [3]. To be specific, group fairness argues that a disadvantaged group (in terms of the sensitive attributes) should receive similar treatment to the advantaged group, that is:

$$P(Y' = 1|S = 0, Y = 1) = P(Y' = 1|S = 1, Y = 1) \quad (2)$$

It is worth nothing that PUT spatial-temporal models are by definition group unaware that is S indicating a sensitive attribute (e.g., race, or gender) is not an explicit feature into these models. However specific demographic groups of users may exhibit certain properties in their mobility behaviour (e.g., students) that could still impact the outcome of the PUT models. For instance, age and employment status can highly influence peoples’ day-to-day trajectory. A user whose trajectory data is limited to his home and office location could be highly predictable by the PUT model, however, also highly re-identifiable (with low privacy gain). This means the notion of group fairness in the context of this study is highly dependent on the examined *dataset*. We elaborate more on this discussion in Section 6.

In order to quantify the group fairness in a more statistical approach, *group fairness score* (i.e., *GFS*) for spatial-temporal data are calculated by disparate impact for disadvantaged groups:

$$GFS = \frac{P(Y' = 1|S = advantaged, Y = 1)}{P(Y' = 1|S = disadvantaged, Y = 1)} \quad (3)$$

4 EXPERIMENT SETUP

In this section, we describe the datasets we used to evaluate the fairness of PUT models and the steps we took to set up the PUT models for examination.

4.1 Datasets

In order to evaluate the fairness of the examined models, we use two datasets that the original papers used to assess the privacy level of their models.

4.1.1 MDC. This dataset is recorded from 2009 to 2011, contains a large amount of continuous mobility data for 184 volunteers with smartphones running a data collection software, in the Lausanne/Geneva area. Each record of the *gps-wlan* dataset represents a phone call or an observation of a WLAN access point collected during the campaign [25]. In addition to the trajectory data, MDC includes individual user demographic information: categorical age groups, gender, and employment status. To the best of our knowledge, MDC is the only dataset that has published users’ demographic information along with their trajectories.

4.1.2 Geolife. This dataset is collected by Microsoft Research Asia from 182 users in the four-and-a-half-year period from April 2007 to October 2011 and contains 17,621 trajectories [46]. As the Geolife dataset does not include demographic attributes of individuals, we are unable to measure the group fairness for this dataset and our analysis suffices for the individual fairness dimension.

As mentioned in Section 3.2.1, in Figure 1, with the granularity coarsens, the trajectories become blurry and thus more similar to each other. Figure 2 confirms this observation by illustrating the SSIM- and EOTs-based similarity of all the users for varying spatial granularity for both datasets. As the spatial granularity coarsens, we observe an increase in the SSIM values, with users becoming more similar to each other. Furthermore, as different types of entropy are considering different features of the spatial-temporal data, Figure 2 presents the expected similarity of users for various EOTs-based measures. In addition to the distribution of the entropy values presented in the Figure 2 for each dataset, we observe that across both datasets, SSIM along with SE and AE correspond to the most relaxed measure of similarity, LE and HE correspond to stricter measures of similarity. The corresponding percentage of user pairs that meet each similarity criterion is described in Table 1.

Metrics	% of pairs	Original, V% of (DIFF>0.2)		Mo-PAE, V% of (DIFF>0.2)		TrajGAN, V% of (DIFF>0.2)		
		Trajectory	Mobility	Privacy	Utility	Privacy	Utility	
		Uniqueness	Predictability	Gain	Decline	Gain	Decline	
MDC	SE	36.17%	10.50%	11.11%	87.69%	39.75%	41.65%	27.32%
	LE	12.85%	8.31%	7.90%	88.81%	36.95%	41.32%	25.10%
	HE	14.11%	12.89%	9.60%	86.88%	41.30%	38.23%	27.14%
	AE	33.05%	12.64%	10.28%	87.10%	35.95%	45.26%	29.42%
	SSIM	65.06%	14.57%	13.17%	88.98%	42.02%	39.50%	27.50%
	EOTs	1.73%	6.10%	1.22%	84.76%	30.49%	44.51%	27.44%
	EOTs+SSIM	1.64%	6.45%	1.29%	83.87%	31.61%	43.23%	28.39%
Geolife	SE	33.16%	57.91%	61.09%	94.14%	71.84%	67.85%	58.58%
	LE	9.11%	57.41%	61.20%	94.32%	71.50%	65.09%	56.05%
	HE	7.29%	61.37%	63.47%	94.09%	70.43%	69.78%	58.87%
	AE	27.03%	57.44%	59.36%	93.23%	72.96%	71.96%	58.54%
	SSIM	63.52%	59.88%	61.49%	94.13%	74.77%	63.53%	53.05%
	EOTs	0.62%	61.54%	58.46%	89.23%	66.15%	78.46%	72.31%
	EOTs+SSIM	0.61%	62.50%	57.81%	89.06%	65.63%	78.13%	71.88%

Table 1. Individual fairness among diverse models and datasets with SSIM and EOTs. % of pairs represents the ratio of the pairs that meet the thresholding requirements. The maximum/minimum instances of each column are highlighted in **bold font**.

4.2 Original Properties of the Trajectory

Before describing the privacy and utility trade-off for mobility trajectories of the PUT models, we first give brief definitions of two popular inference tasks (i.e., *user re-identification and mobility prediction*), which are also applied to assess the privacy gain and utility decline in the PUT models we discussed. These two popular inference tasks are named *original tasks* in this paper, where the *original* demonstrates the nature of the data before being processed by any privacy-aware model. These *original* tasks are leveraged to assess the native data characteristics in terms of *user re-identification (UR)* and *mobility predictability (MP)*, respectively. See the Appendix for full definitions.

5 FAIRNESS ANALYSIS

In this section, we present our analysis in studying whether the PUT models can be considered fair. To do so, we analyze these models in terms of individual fairness and group fairness. The similarity SIM_t applied in the individual fairness is defined by SSIM and EOTs, and group fairness is grouping users based on demographic attributes such as gender, age, and employment status.

5.1 Individual Fairness

The metrics of trajectories' similarity SIM_t are crucial for quantifying individual fairness. As definitions in Section 3.2, the SIM_t can be quantified by SSIM and EOTs. In this section, we discuss individual fairness with two different similarity quantification approaches. First, the SIM_t discriminated based on ϵ -thresholding metrics of SSIM and EOTs directly. Second, the k-means clustering approach, based on the characteristics of SSIM and EOTs aforementioned, is leveraged to classify similar users.

5.1.1 Similarity Based on ϵ -Thresholding. Table 1 presents the individual fairness of different models by the ϵ -thresholding metrics based on SSIM and EOTs. The threshold ϵ of different combinations of SSIM and EOTs are utilized to distinguish similar users ($u_i \sim u_j$) and map all users into a list of *pairs* with trajectory similarity and performance discrepancy. Based on fairness thresholding criteria defined in Section 3.2.3, *similar users* (i.e., *user pairs*)

imply at least 80% pairwise similarity of their trajectories. "% of pairs" in the table represents the percentage of the user pairs that meet the corresponding metric threshold requirements. For instance, with the MDC dataset, 36.17% of *user pairs* have a more than 80% similarity when under the *SE* metric. That is, under the *SE* metric, 36.17% user pairs are qualified for further analysis of outcome similarity.

The *user pair* is defined to achieve individual fairness when the outcome difference (ΔD_{pri} or ΔD_{uti}) between u_i and u_j is within 20%. Table 1 shows the percentage of *user pairs* that commit fairness violation (i.e., $V\% = \% \text{ of } (\Delta D > 0.2)$). For instance, in Table 1, with the MDC dataset under the *SE* metric, there are only 10.50% and 11.11% of the *qualified user pairs* violate the fairness criteria in two original tasks, which implies that the individual fairness is achieved, as both $V\%$ are within 20%. Different from the original tasks, two PUT models have $V\%$ that are all higher than 20%, hence, they violate individual fairness. The higher $V\%$ indicates that the model causes more disparities in performance. The values in the *italic format* present the cases where the outcome to meet individual fairness (i.e., $V\% \leq 20\%$) in the Table 1.

Overall, individual fairness is **not achieved** in the two selected PUT models, especially for the unfairness of the privacy gain, which is generally higher than the utility decline. When comparing two different privacy models in a row, TrajGAN achieves less fairness violation rate than Mo-PAE in both privacy gain and utility decline outcomes. For instance, in the MDC dataset, when 45.26% and 29.42% of user pairs commit fairness violations in privacy gain and utility decline, respectively, the Mo-PAE reports twice as many fairness violations for both outcomes. While both the Geolife and MDC data exhibit individual unfairness, the Geolife is worse in both the PUT models and the accuracy of the *original tasks*. In both original tasks, Geolife’s unfairness rate is as high as 60%, and this inequity is exacerbated when with PUT models. In contrast to Geolife, the performance of the MDC in the original tasks conforms to the definition of individual fairness, that is, the performance difference of task accuracy in MDC is within 20% in both user re-identification tasks and mobility prediction tasks.

Impact of Spatial Granularity on Similarity: After the overall comparison of threshold metrics, we discuss the model discrepancy when trajectory similarity is based on the SSIM index under varying granularity. As a crucial metric in distinguishing the trajectory similarity, the SSIM index could be affected by different parameters, which will result in subtle performance disparities in the quantification of individual fairness. The spatial granularity of trajectory is the most important one among these parameters. These disparities could be intuitively observed in the heatmaps (Figure 1). In contrast to the SSIM, the spatial granularity has less impact on different types of entropy, hence, they are not discussed here.

The Figure 3 then shows the impact of varying spatial granularity on the model discrepancy. The model that achieves individual fairness should perform less discrepancy with higher SSIM. The accuracy of original tasks and two PUT models are compared in granularity at 100 meters, 300 meters, 500 meters, and 900 meters. In conclusion, different models have diverse sensitivities of varying granularities. Both original tasks (UR and MP) in the two datasets have an increasing difference with a higher SSIM index, which means they violate individual fairness. For the Mo-PAE, individual fairness is met on MDC data but not on Geolife. The Mo-PAE is also the most sensitive model for varying granularities. For instance, when granularity changes from 100-meter (Figure 3a) to 900-meter (Figure 3d), Mo-PAE has the most obvious change in its line trend on the UR (i.e., privacy gain), and the decreasing trend at 100-meter granularity is lost at 900-meter. Overall, the selection of SSIM granularity has a significant impact on the judgement of the individual fairness of a model. However, these impacts become subtle when the SSIM is applied to the trajectory similarity distinction, as the user pairs table reduced the granularity impact to some extent. For the remaining of the analysis, the granularity of the SSIM is chosen as 100-meter.

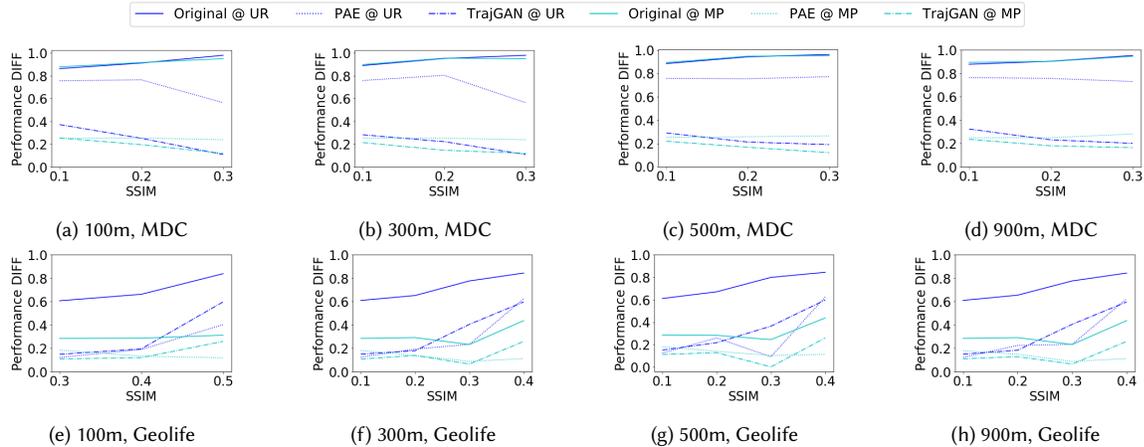


Fig. 3. The model performance discrepancy when trajectory similarity is based on the SSIM in different granularities. Figure (a) to Figure (d) are the results of MDC dataset, Figure (e) to Figure (h) are of Geolife. The performance discrepancy (i.e., Performance DIFF) of each model in different granularities compares in each sub-figure.

Metrics	Cluster Size	Original, V% of (DIFF>0.2)		Mo-PAE, V% of (DIFF>0.2)		TrajGAN, V% of (DIFF>0.2)		
		Trajectory Uniqueness	Mobility Predictability	Privacy Gain	Utility Decline	Privacy Gain	Utility Decline	
MDC	Cluster 1	26	14.77%	15.38%	83.69%	48.92%	50.77%	33.85%
	Cluster 2	5	0.00%	0.00%	90.00%	0.00%	40.00%	0.00%
	Cluster 3	43	17.39%	17.17%	88.15%	48.84%	55.92%	43.63%
	Cluster 4	24	0.00%	0.00%	86.23%	16.67%	35.87%	17.75%
	Clusters Average	-	12.99%	12.18%	88.86%	39.36%	51.26%	34.49%
Geolife	Cluster 1	21	55.71%	13.81%	60.00%	1.43%	18.57%	0.00%
	Cluster 2	17	46.32%	8.09%	49.26%	13.97%	16.91%	0.00%
	Cluster 3	9	13.89%	11.11%	38.89%	16.67%	13.89%	11.11%
	Cluster 4	10	31.11%	0.00%	40.00%	0.00%	44.44%	4.44%
	Cluster 5	36	44.92%	8.57%	29.84%	5.56%	23.02%	0.16%
Clusters Average	-	43.91%	9.15%	47.11%	7.55%	26.53%	2.02%	

Table 2. K-means-clustering-based individual fairness among diverse models and datasets. The numbers present the percentage of users for whom individual fairness was violated based on their difference in the outcome being greater than 0.2. The fair instances are highlighted in *italic font*. The maximum/minimum instances of each column are highlighted in **bold font**.

5.1.2 *Similarity Based on K-means Clustering.* Alternative to the results presented based on the similarity thresholding, Table 2 demonstrates the results of individual fairness based on the clustering technique described in Section 3.2.3. Applying the Elbow and Silhouette methods, we decide the number of clusters (k) to be 4 and 5 for MDC and Geolife, respectively. For each cluster, the table reports the percentage of users whose individual fairness was violated for a given outcome and under various models. More precisely, the results presented here indicate that the original model that objectifies a single task (prediction or privacy) is able to meet the individual fairness criteria for the MDC dataset. We can observe that in the case of the Mo-PAE model, the privacy gain exhibits high variations across users in the same clusters. Even in the cases where the model satisfies individual fairness by performing similarly in terms of utility

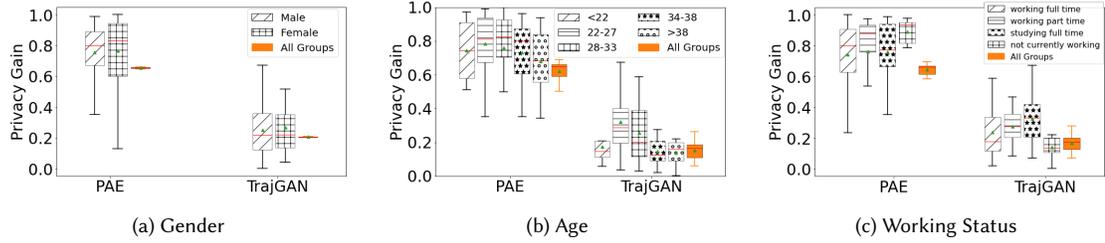


Fig. 4. The privacy protection outcome of PUT models across different demographic groups for the MDC dataset.

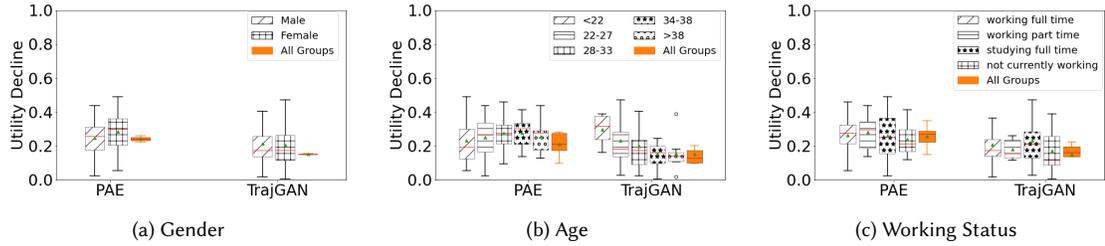


Fig. 5. The prediction accuracy outcome of PUT models across different demographic groups for the MDC dataset.

decline (clusters 2 and 4 of MDC and all clusters in Geolife), the privacy gains of those users are very different from each other.

5.2 Group Fairness

Group fairness states that groups across different sensitive attributes should receive similar outcomes. To be specific, group fairness argues that a disadvantaged group should receive similar treatment to the advantaged group. Figure 4 presents the discrepancy of the *privacy gain* from two PUT models for different demographic groups, and Figure 5 presents the *utility decline*. We observe that both Mo-PAE and TrajGAN perform equally for different gender attributes, as shown in Figure 4a, where the orange boxes (labelled as *All Groups*) on both are very small. That is while the privacy gain varies across individuals within the same gender, the model achieves group fairness when grouping individuals by gender. The same observations could be made for the age and employment status, where we see that there exist bigger differences across the classes than the gender, but they still achieve group fairness as $\Delta D < 20\%$. Similarly, in Figure 5, we can observe that both models equally meet the group fairness criteria on the utility decline.

In order to quantify the group fairness of the *disadvantaged groups* in a more statistical approach, the results of the *group fairness score (GFS)* are shown in Table 3. For instance, for different age groups, the subgroup with ages between 22 and 27 (i.e., "22 - 27") is regarded as the *advantaged group*, as it has the dominant user number for all age groups. The other age groups' GFSs are calculated based on the disparate impact between them and the *advantaged group*. Then, compare all GFSs against the fairness threshold of 0.8, which is defined in Section 3.2.3, that is, $GFS \geq 80\%$ indicates fairly treating the disadvantaged group and $GFS < 80\%$ indicates the unfairly treating. For example, the result of "28-33" group (i.e., $GFS = 98.65\%$) then indicated that the model satisfy the group fairness as $98.65\% > 80\%$.

		users #	Original, GFS		Mo-PAE, GFS		TrajGAN, GFS	
			Uniqueness	Predictability	PrivacyGain	UtilityDecline	PrivacyGain	UtilityDecline
Gender	Male	56	-	-	-	-	-	-
	Female	33	98.07%	96.35%	98.04%	90.13%	96.57%	95.00%
Age	<21	5	94.48%	99.10%	46.09%	85.86%	84.73%	87.38%
	22-27	38	-	-	-	-	-	-
	28-33	29	98.65%	94.49%	96.97%	97.36%	90.16%	93.74%
	34-38	11	97.98%	98.54%	91.13%	99.55%	81.81%	92.74%
	>39	9	95.76%	99.73%	75.51%	94.05%	83.47%	91.30%
Working	Full-time work	48	-	-	-	-	-	-
	Part-time work	8	95.80%	96.44%	82.58%	85.24%	99.67%	94.81%
	Full-time student	26	98.09%	99.23%	97.83%	88.16%	85.77%	88.37%
	Others	8	95.80%	99.33%	98.70%	93.59%	95.36%	99.07%

Table 3. Group fairness scores (*GFS*) of three models with different demographic attributes. *GFS* \geq 80% indicates the fairly treating the minority subgroup; *GFS* $<$ 80% indicates the unfairly treating.

In conclusion, except for two subgroups with age attributes (i.e., "<21" and ">39") violating the four-fifths rule, the other subgroups satisfy the group fairness. Finally, it is worth noting that the results presented here are highly dependent on the studied dataset, as we discuss in the next section.

6 DISCUSSION

In this section, we describe the limitations and implications of our work and discuss possible future directions.

6.1 Limitation

Despite our efforts, the presented work also has its limitations. Firstly, the collected mobility dataset is often biased as they only present a subset of the population who took part in data collection. In many cases, the users are limited to students or those affiliated with the research team that has collected the dataset. This limitation means the examined trajectories are not representative of everyone’s mobility behaviour. Furthermore, the demographics of the participants are also limited in terms of age and socio-economic diversity.

Secondly, in our paper, we reported that we did **not** observe any violation of *group fairness* across gender, age and employment level for the examined PUT models. However, we acknowledge that the results presented regarding group fairness are highly influenced by the city and societal structures in which the data was collected. In the case of MDC, users’ traces correspond to a level of socio-economic and cultural freedom associated with life in Switzerland. Such observations will indeed differ if we examine other cultures, such as those in the United States or Asian countries, where there is a broader socio-economic and gender inequality gap. We also believe the availability of datasets with rich demographic information could enable future work to examine the intersection of individual fairness within demographic groups. Finally, it is worth noting that, unlike online datasets, offline mobility datasets come in limited size due to the great burden the data collection imposes on participants and are handful. Although this limitation could impact the generalization of our results (e.g., that is we cannot claim that Mo-PAE is always fairer than TrajGan), the methods proposed in this study are generalizable and applicable to other PUT models and across mobility datasets. Indeed, we believe future work would focus on creating a toolkit for computing spatial-temporal fairness of datasets and models. We expand on the implications of our work next.

6.2 Implication

Our paper has multiple important implications: first, our work offers a novel methodology for defining fairness in the context of spatial-temporal datasets. We believe works such as ours will help shape the future roadmap of Fair-ML studies by offering possibilities to measure equity within different systems such as those of mobility based ones (e.g., transportation). The choice of which of the proposed similarity metrics to select for evaluating individual fairness is another critical dimension that could be highly context and application dependent. For example, for applications where there is a need for strict fairness measurement, corresponding to the WYZIWIG worldview [15], a strict similarity measure such as combined entropy (EOTs) could be chosen. In contrast, for applications where the groups are not necessarily equal, but for the purposes of the decision-making process, we would prefer to treat them as if they were, a less sensitive similarity measure such as coarse grain SSIM could be used.

Although our focus in this work was on fairness analysis of the PUT models, we believe our study can be the first step towards implementing fairness interventions embedded in these models. For example, in-processing approaches rely on adjusting the model during the training to enforce fairness goals to be met and optimized in the same manner as accuracy. This goal is often achieved through adversarial networks or fair representation learning approaches such as [21], model induction, model selection, and regularization [43]. Of course, designing such mitigation strategies requires access to the underlying architecture of the PUT models which is most of the time not possible, and is in contrast to taking these models as black-box as we did in this study.

Regarding the relationship between privacy versus fairness, location privacy-preserving mechanisms generally prevent information leakage against protected attributes, and these attributes are also essential to fairness analysis, they are used to ensure little discrimination against protected population subgroups. This dimension also explains why the PUT models achieve group fairness but not individual fairness, as these sensitive attributes considered by group fairness are in protection. The competing trend between individual- and group- fairness also implies another interesting trade-off in Fair-ML. From the individual perspective, the re-identification risk and individual fairness are in tension. We believe designing privacy-preserving models to become fairness-aware is a research direction that will receive significant attention in the future.

7 CONCLUSION

Intuitively, fairness has a close relationship to privacy, no matter structural data or unstructured data in machine learning. But the quantification between them is still unclear. In this paper, we proposed different metrics for measuring individual fairness in the context of spatial-temporal mobility data. We compared different location privacy-protection mechanisms (PUT models) on the defined individual- and group-based metrics. Our results on two real trajectory datasets show that the privacy-aware models **achieve** fairness at the group level but **violate** individual fairness. Our findings raise questions regarding the equity of the privacy-preserving models when individuals with similar trajectories receive a very different level of privacy gain. We leverage the empirical results of our work to make valuable suggestions for the further integration of fairness objectives into the PUT models. Especially when discussing the individual perspective, the tension between the user re-identification task and individual fairness needs to be considered for future spatial-temporal data analysis and modelling to achieve a privacy-preserving fairness-aware setting.

REFERENCES

- [1] Aristos Aristodimou, Athos Antoniadis, and Constantinos S Pattichis. Privacy preserving data publishing of categorical data through k-anonymity and feature selection. *Healthcare technology letters*, 3(1):16–21, 2016.

- [2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2017, 2017.
- [3] Solon Barocas and Andrew D Selbst. Big data’s disparate impact. *Calif. L. Rev.*, 104:671, 2016.
- [4] Reuben Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 514–524, 2020.
- [5] Anne Elizabeth Brown. *Ridehail revolution: Ridehail travel and equity in Los Angeles*. University of California, Los Angeles, 2018.
- [6] Hongyan Chang and Reza Shokri. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 292–303. IEEE, 2021.
- [7] Weiting Chen, Zhizhong Wang, Hongbo Xie, and Wangxin Yu. Characterization of surface emg signal based on fuzzy entropy. *IEEE Transactions on neural systems and rehabilitation engineering*, 15(2):266–272, 2007.
- [8] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleyesen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1):1–5, 2013.
- [9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.
- [10] Ecenaz Erdemir, Pier Luigi Dragotti, and Deniz Gündüz. Privacy-aware location sharing with deep reinforcement learning. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2019.
- [11] Ecenaz Erdemir, Pier Luigi Dragotti, and Deniz Gündüz. Privacy-aware time-series data sharing with deep reinforcement learning. *IEEE Transactions on Information Forensics and Security*, 16:389–401, 2020.
- [12] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD*, pages 259–268, 2015.
- [13] Danielle L Ferreira, Bruno AA Nunes, Carlos Alberto V Campos, and Katia Obraczka. A deep learning approach for identifying user communities based on geographical preferences and its applications to urban and environmental planning. *ACM Transactions on Spatial Algorithms and Systems*, 6(3):1–24, 2020.
- [14] Matthew W Flood and Bernd Grimm. Entropyhub: An open-source toolkit for entropic time series analysis. *Plos one*, 16(11):e0259448, 2021.
- [15] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, 64(4):136–143, 2021.
- [16] Yanbo Ge, Christopher R Knittel, Don MacKenzie, and Stephen Zoepf. Racial and gender discrimination in transportation network companies. Technical report, National Bureau of Economic Research, 2016.
- [17] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.
- [18] Suining He and Kang G Shin. Socially-equitable interactive graph information fusion-based prediction for urban dockless e-scooter sharing. In *Proceedings of the ACM Web Conference 2022*, pages 3269–3279, 2022.
- [19] Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 181–190, 2019.
- [20] Kate Hosford and Meghan Winters. Who are public bicycle share programs serving? an evaluation of the equity of spatial access to bicycle share service areas in canadian cities. *Transportation research record*, 2672(36):42–50, 2018.
- [21] Tongxin Hu, Vasileios Iosifidis, Wentong Liao, Hang Zhang, Michael Ying Yang, Eirini Ntoutsi, and Bodo Rosenhahn. Fairrn-conjoint learning of fair representations for fair decisions. In *International Conference on Discovery Science*, pages 581–595. Springer, 2020.
- [22] Dou Huang, Xuan Song, Zipei Fan, Renhe Jiang, Ryosuke Shibasaki, Yu Zhang, Haizhong Wang, and Yugo Kato. A variational autoencoder based generative model of urban human mobility. In *2019 IEEE MIPR*, pages 425–430. IEEE, 2019.
- [23] Maria Kamargianni, M Matyas, W Li, and A Schäfer. Feasibility study for “mobility as a service” concept in london. *UCL Energy Institute, Dept. Transp.*, pages 1–82, 2015.
- [24] Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 576–586, 2021.
- [25] Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen, et al. The mobile data challenge: Big data for mobile computing research. Technical report, 2012.
- [26] Rosa Lleti, M Cruz Ortiz, Luis A Sarabia, and M Sagrario Sánchez. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica Chimica Acta*, 515(1):87–100, 2004.
- [27] Xin Lu, E. Wetter, N. Bharti, A. J Tatem, and L. Bengtsson. Approaching the limit of predictability in human mobility. *Scientific reports*, 3(1):1–9, 2013.
- [28] Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. A survey on deep learning for human mobility, 2021.
- [29] Afra Mashhadi, Joshua Sterner, and Jeffery Murray. Deep embedded clustering of urban communities using federated learning. 2021.
- [30] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.
- [31] Kun O., Reza S., David S. R., and Wenzhuo Y. A non-parametric generative model for human trajectories. In *IJCAI-18*, pages 3812–3817. IJCAI, 7 2018.
- [32] Krishna PN Puttaswamy, Shiyuan Wang, Troy Steinbauer, Divyakant Agrawal, Amr El Abbadi, Christopher Kruegel, and Ben Y Zhao. Preserving location privacy in geosocial applications. *IEEE Transactions on Mobile Computing*, 13(1):159–173, 2012.
- [33] Jimmeng Rao, Song Gao, Yuhao Kang, and Qunying Huang. Lstm-trajan: A deep learning approach to trajectory privacy protection. *arXiv preprint arXiv:2006.10521*, 2020.

- [34] John E Roemer. Equality of opportunity: A progress report. *Social Choice and Welfare*, 19(2):455–471, 2002.
- [35] Nazir Saleheen, Supriyo Chakraborty, Nasir Ali, Md Mahbubur Rahman, Syed Monowar Hossain, Rummana Bari, Eugene Buder, Mani Srivastava, and Santosh Kumar. msieve: differential behavioral privacy in time series of mobile sensor data. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 706–717, 2016.
- [36] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019.
- [37] Luiz Eduardo Virgili Silva, ACS Senra Filho, VPS Fazan, JC Felipe, and LO Murta Junior. Two-dimensional sample entropy: Assessing image texture through irregularity. *Biomedical Physics & Engineering Express*, 2(4):045002, 2016.
- [38] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- [39] Yan Wang, Ali Yalcin, and Carla VandeWeerd. An entropy-based approach to the study of human mobility and behavior in private homes. *PLoS one*, 15(12):e0243503, 2020.
- [40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [41] Yonghui Xiao and Li Xiong. Protecting locations with differential privacy under temporal correlations. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1298–1309, 2015.
- [42] An Yan and Bill Howe. Fairst: Equitable spatial and temporal demand prediction for new mobility systems. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 552–555, 2019.
- [43] An Yan and Bill Howe. Fairness in practice: a survey on equity in urban mobility. *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering*, 42(3), 2020.
- [44] Yuting Zhan, Hamed Haddadi, and Afra Mashhadi. Privacy-aware adversarial network in human mobility prediction. *Proceedings on Privacy Enhancing Technologies*, 1:556–570, 2023.
- [45] Wenjing Zhang, Ming Li, Ravi Tandon, and Hui Li. Online location trace privacy: An information theoretic approach. *IEEE Transactions on Information Forensics and Security*, 14(1):235–250, 2018.
- [46] Yu Zheng, Xing Xie, and Wei-Ying Ma. GeoLife: A Collaborative Social Networking Service among User, Location and Trajectory. *IEEE Data Eng. Bull.*, 33:32–39, June 2010.

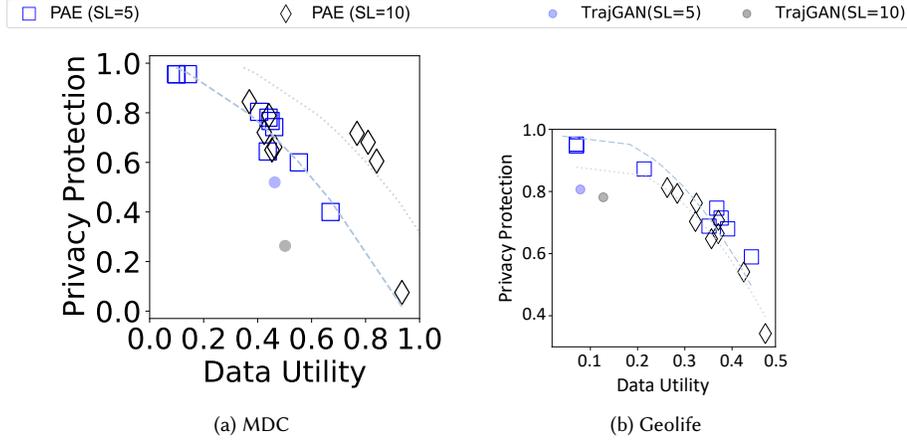


Fig. 6. Pareto Frontier trade-off of Utility and Privacy on two datasets. The hollow squares and diamonds present the results of the Mo-PAE models. The solid points present the results of the TrajGAN. Blue color presents sequence length $SL = 5$. Black color presents $SL = 10$.

APPENDIX

7.1 Inference tasks

Here we list some basic definitions of inference tasks in mobility literature.

7.1.1 User Re-identification Task (UR). The accuracy of the user re-identification task is leveraged to assess the *trajectory uniqueness* of the mobility trajectory. With more and more intelligent devices and sensors being utilized to collect information about human activities, the trajectories also expose increasing intimate details about users' lives, from their social life to their preferences. A mobility privacy study conducted by De Montjoye et al [8] illustrates that four spatial-temporal points are enough to identify 95% of the individuals in a certain granularity. As human mobility traces are highly unique, a mechanism capable of reducing the user re-identification risk can offer enhanced privacy protection in mobility data sharing. The enhanced privacy protection is referred to *privacy gain* (or *PG*) in the PUT models.

7.1.2 Mobility Prediction Task (MP). The accuracy of the mobility prediction task is leveraged to assess the *predictability* of the mobility trajectory. Mobility datasets are of great value for understanding human behaviour patterns, smart transportation, urban planning, public health issue, pandemic management, etc. Many of these applications rely on the next location forecasting of individuals, which in the broader context can provide an accurate portrayal of citizens' mobility over time. For the mobility prediction task in this work, the raw geolocated data or other mobility data commonly contain three elements: user identifier u , timestamps t , location identifiers l . Hence, each location records r could be denoted as $r_i = [u_i, t_i, l_i]$, while each location sequence S is a set of ordered location records $S_n = \{r_1, r_2, r_3, \dots, r_n\}$, namely *mobility trajectory*. Therefore, given the past mobility trajectory $S_n = \{r_1, r_2, r_3, \dots, r_n\}$, the mobility prediction task is to infer the most likely location l_{n+1} at the next timestamp t_{n+1} . The results of two PUT models indicate that a bit of mobility prediction accuracy is sacrificed in exchange for higher privacy protection. The sacrificed prediction accuracy is referred to *utility decline* in the PUT models.

7.2 Performance of the Privacy-Utility Trade-off Models

Before examining fairness, we first offer analysis and comparison of the two described PUT models that we are investigating in terms of privacy and utility. Figure 6 presents the privacy utility trade-off of Mo-PAE and TrajGAN over the two described datasets. The y-axis presents the privacy gain brought to the raw dataset by applying these models, whereas the x-axis presents the decline in privacy prediction due to this privacy gain. The data fed into the Mo-PAE [44] are a list of trajectories with specific sequence length SL , that is $\{S_{SL}^1, S_{SL}^2, S_{SL}^3, \dots, S_{SL}^j\}$. For instance, if the sequence length is 10, that indicates each trajectory contains 10 history location records r , $S_{10} = \{r_1, r_2, r_3, \dots, r_{10}\}$, and $SL = 10$.

As Mo-PAE is highly dependent on the sequence length and Lagrange multipliers that indicate to what extent privacy or utility must be optimized, each point on the corresponding plots presents experiments with one set of hyper-parameters. These results show that as the Mo-PAE achieves maximum privacy protection it comes with the cost of degrading the prediction accuracy. Similarly, TrajGAN achieves 80% privacy gain when applied on Geolife Dataset but it highly degrades the utility. For the Lagrange multipliers setting of the Mo-PAE in this work, we choose $\lambda_1 = -0.1$, $\lambda_2 = 0.8$, $\lambda_3 = -0.1$, as this combination exerts the most promising privacy-utility trade-off in the Mo-PAE model.

Related Equations

i. SSIM. In this work, we use the known SSIM measure as the perceptual difference of two similar users' heatmap images, H_i and H_j :

$$SSIM(H_i, H_j) = \frac{(2\mu_i\mu_j + c_1)(2\sigma_{ij} + c_2)}{(\mu_i^2 + \mu_j^2 + c_1)(\sigma_i^2 + \sigma_j^2 + c_2)}, \quad (4)$$

$$c_1 = (k_1L)^2, \quad c_2 = (k_2L)^2$$

where μ_i and μ_j are the averages, σ_i and σ_j are the variances, and σ_{ij} is the covariance of H_i and H_j ; L is the dynamic range of the pixel-values, $k_1 = 0.01$ and $k_2 = 0.03$ by default.

ii. Shannon Entropy (SE). SE is the entropy of the probabilities of visited location distribution. To be specific, this entropy is defined by following the notion in [27, 39] and measured as:

$$E_h = - \sum_{i=1}^n P(x_i) \log_2 [P(x_i)] \quad (5)$$

where n is the length of the probability vector, $P(x_i)$ is the probability of location x_i .

iii. LonLat Entropy (LE). LE is the entropy of the geo-located locations in a time-series format. This entropy reflects the probability of a new sub-string and quantifies the irregularity or complexity of the time-series data. The E_f of visited longitudes and latitudes are integrated as the LE:

$$E_f = \ln \Phi^m(r, n) - \ln \Phi^{m+1}(r, n) \quad (6)$$

where details and default values of the threshold r and the definition of function $\Phi^m(r, n)$ can be found in the study [7, 14].

iv. Heatmap Entropy (HE). HE is the entropy of the users' heatmap images. The entropy of trajectory heatmap images was calculated using the two-dimensional sample entropy method ($SampEn_{2D}$) [37]. In a trajectory heatmap image

(L^2), the image features were extracted by accounting for the spatial distribution of pixels in different m -length square windows with origin at $u(i, j)$.

$$\begin{aligned}
 E_{2D}(u, m, r) &= -\ln \frac{U^{m+1}(r)}{U^m(r)}, \\
 U^m(r) &= \frac{1}{N_m} \sum_{i,j,a,b=1}^{i,j,a,b=L-m} Z, \\
 Z &= P [x_m(a, b) | d [x_m(i, j), x_m(a, b)] \leq r, (a, b) \neq (i, j)]
 \end{aligned} \tag{7}$$

where r is the similarity threshold, N_m is the total number of square windows, P is the probability of pixels set $x(i, j)$ satisfying specific conditions, $U_m(r)$ is the average probability, and d is a distance function to calculate the difference of corresponding points.

v. *Actual Entropy (AE)*. AE is the entropy of capturing entire spatial-temporal order present in user's mobility pattern. In this work, the given area is segmented using structured grids, where each grid is initialized as 0. Then the visited locations and whether the person reached the cell previously are tracked. If the person visits an unreached cell, the location is marked as 1, generating time-series binary data to characterize the trajectory. The actual entropy E_a is calculated using:

$$E_a = \left(\frac{1}{n} \sum_i \Lambda_i \right)^{-1} \ln(n) \tag{8}$$

where Λ_i is the length of the shortest sub-string starting at position i which does not previously appear from position 1 to $i - 1$, and n is the length of the binary trajectory data.