A Survey and Experimental Study on Privacy-Preserving Trajectory Data Publishing

A Survey and Experimental Study on Privacy-Preserving Trajectory Data Publishing

Fengmei Jin, Wen Hua✉, Matteo Francia, Pingfu Chao, Maria E Orlowska, Xiaofang Zhou, *Fellow, IEEE*

Abstract—Trajectory data has become ubiquitous nowadays, which can benefit various real-world applications such as traffic management and location-based services. However, trajectories may disclose highly sensitive information of an individual including mobility patterns, personal profiles and gazetteers, social relationships, etc, making it indispensable to consider privacy protection when releasing trajectory data. Ensuring privacy on trajectories demands more than hiding single locations, since trajectories are intrinsically sparse and high-dimensional, and require to protect multi-scale correlations. To this end, extensive research has been conducted to design effective techniques for privacy-preserving trajectory data publishing. Furthermore, protecting privacy requires carefully balance two metrics: privacy and utility. In other words, it needs to protect as much privacy as possible and meanwhile guarantee the usefulness of the released trajectories for data analysis. In this survey, we provide a comprehensive study and a systematic summarization of existing protection models, privacy and utility metrics for trajectories developed in the literature. We also conduct extensive experiments on two real-life public trajectory datasets to evaluate the performance of several representative privacy protection models, demonstrate the trade-off between privacy and utility, and guide the choice of the right privacy model for trajectory publishing given certain privacy and utility desiderata.

Index Terms—Trajectory data publishing, attack models, privacy protection models, privacy metrics, utility metrics

1 Introduction

Privacy is usually referred to as the *"ability of an individual to control the terms under which personal information is acquired and used"* [1]. Privacy entails the protection of several data aspects such as collection [2], mining [3], querying [4], and publication [5]. Each of these aspects involves its own privacy protection models as well as measures to evaluate privacy level. We focus on data publication in this work, i.e., releasing datasets without leaking any sensitive information. Privacy-preserving data publishing has been extensively studied in the database community, and well-known techniques have been proposed to anonymize tabular records stored in the database including k-anonymity [6], [7], [8] (1998), l-diversity [9], [10] (2006), t-closeness [11] (2007), and differential privacy [12] (2006).

With the increasing popularity of GPS-enabled devices, a wide range of location-based services keep track of moving objects, resulting in massive available spatial trajectory data. Nowadays, trajectory data analysis has become ubiquitous, as evidenced by a huge amount of trajectory-related techniques, which can benefit various real-world applications including urban planning, traffic management, personalized recommendation. However, the analysis of trajectory data can disclose sensitive information of an individual, making it essential to design techniques for privacy protection. In general, the protection of trajectory privacy is based on two major directions: *location-based services* (LBSs) and *privacy preserving trajectory publication* (PPTD). On one hand, privacy protection in LBSs requires that a sufficient quality-of-service is ensured while preventing an adversary from learning the exact locations of an individual [13], [14]. On the other hand, privacy concerns hinder data-holders in the publication of private trajectories which, thus, has spawned extensive research on privacy-preserving trajectory data publishing. These directions are orthogonal and can be distinguished according to the amount of adversary's knowledge (i.e., a sequence of real-time locations for LBSs; the entire movement history for PPTD) and the protection scope (i.e., the current location for LBSs; the entire trajectory for PPTD). We focus on PPTD in this paper, considering the proliferation of applications relying on the availability of trajectory data. Formally, a trajectory of an individual is recorded as a sequence of (geo-position, time) ordered chronologically. Although trajectory data is representable in a tabular format (e.g., organizing each historical trace as a record), trajectories cannot be easily anonymized as "classic" tabular data due to the following reasons:

- Trajectory data fulfills spatial constraints (e.g., mobility in an urban area).
- Trajectory locations are not independent (e.g., there is spatiotemporal continuity between adjacent locations; it is impossible to jump from a road to another).
- Although trajectory data is highly sparse, only a few locations can link 95% of individuals [15]. The longer the trajectory, the easier to break individuals' privacy.

- *Fengmei Jin and Wen Hua are with The University of Queensland, Brisbane, QLD 4072, Australia. E-Mail: {fengmei.jin, w.hua}@uq.edu.au*
- *Matteo Francia is with The University of Bologna, Via Dell'Università, 50, 47522 Cesena FC, Italy. E-Mail: m.francia@unibo.it*
- *Pingfu Chao is with Soochow University, Suzhou, Jiangsu, China. Email: pfchao@suda.edu.cn*
- *Maria Orlowska is with Polish-Japanese Academy of Information Technology, Koszykowa 86, 02-008 Warszawa, Poland. E-Mail: omaria@pjwstk.edu.pl*
- *Xiaofang Zhou is with Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. E-Mail: zxf@cse.ust.hk*

- Trajectory locations represent geographical features mappable into semantics (e.g., POIs) that can directly reflect individuals' interests and demographics.
- Trajectories do not have fixed quasi-identifiers [16], [13]. Sensitivity depends on both single locations and arbitrary spatiotemporal patterns (e.g., day and nighttime mobility).

The sensitivity, uniqueness, and low anonymizability of trajectory data raise many issues and concerns, and hence extensive research has been conducted to develop effective techniques for privacy-preserving trajectory data publishing. In the 2000s, two main approaches *ad-hoc* for spatiotemporal data were introduced to protect individual locations either by producing dummy locations indistinguishable from the real ones [17] (2005) or by mixing identifiers of individuals entering/leaving mix-zones [18] (2008). Due to the need for publishing trajectories, both dummy and mix-zone models have been adapted to trajectory data. Additionally, since 2008, *generic* privacy models for sequential patterns [8], [10], [11], [12], [19] have also been specialized to protect trajectory data, with trajectory k-anonymity being implemented first [20] by making a trajectory indistinguishable in an anonymity group including k-1 other trajectories. Differential privacy has been introduced for trajectory data in [21] (2012) where, rather than generalizing/suppressing locations to achieve k-anonymity, authors release synthetic trajectories resembling the original ones. Recently, l-diversity and t-closeness [22] have also been applied to trajectories to protect semantic locations (e.g., residence and workplace).

1.1 Positioning and Contributions

In this survey, we analyze and organize the articulated spectrum of threat and anonymization models on the publication of trajectory data. Although lots of trajectory privacy papers have been recently published in top-tier venues, understanding which anonymization models fulfill the publication requirements is hard especially because ad-hoc privacy and utility metrics are usually leveraged in these papers, requiring exhaustive comparison which is missing in these works. Our goal is to provide a comprehensive and clear overview of the privacy issues related to trajectory data as well as the privacy models countering these issues. We target readers approaching trajectory privacy problems or with only partial knowledge, and organize the content of the survey at an increasing level of details to drive readers from a general perspective to technical and empirical details.

While other surveys on trajectory privacy have been published already, they are either vertical (e.g., focusing on wireless sensor networks [23], opportunistic mobile networks [24], and automotive applications [25]), or lacking of a systematic categorization and evaluation of utility and privacy metrics (e.g., [26], [27], [28]), or have been published before well-known recent results (e.g., [29]). Overall, the main contributions of this paper are as follows:

- We provide a detailed overview of trajectory sensitivity and attacks, to highlight the privacy issues related to the publication of trajectory data.
- We conduct a systematic analysis of privacy models applied to trajectory data publishing and the ways to quantify their privacy level and utility preservation.
- We provide an open-source library integrating implementations of the most representative trajectory anonymization models developed in the literature, and systematically evaluate these models using two publicly-available trajectory datasets.
- Through extensive empirical evaluation of the privacy models with respect to different utility and privacy metrics, we guide the logical meaning and the choice of algorithms for the release of trajectories given certain privacy and utility desiderata.

The remaining of the paper is organized as follows: We summarize the privacy threat of trajectories in Section 2 and state-of-the-art privacy protection models for trajectory publishing in Section 3; Quantitative utility and privacy metrics are introduced in detail in Section 4; Our experimental results and analysis are reported in Section 5; We conclude this survey in Section 6 with a summary of some insightful findings and promising future work.

2 Trajectory Sensitivity and Attacks

Sensitive data is personal data (i.e., any information related to an identifiable person) which, by its nature, is particularly sensitive and might cause forms of discrimination or undesired profiling. In this section, we categorize *what* sensitive data could be exposed by trajectories, and *how* (technically) attack models expose those sensitive information from published trajectory data, as summarized in Table 1.

2.1 Sensitive Data

Inspired by GDPR [54], we distinguish three categories of sensitive data: *identity* (i.e., any data that directly identifies an individual; e.g., fiscal code and social security number), *personal profile* (i.e., any information related to an identifiable person; e.g., religion and ethnicity), and *social relationship* (i.e., any relationship between individuals; e.g., friendship or partnership). Although the value of trajectory data is out of question, its peculiar spatiotemporal, sequential, and recurrent natures threaten the protection of sensitive data.

Identity: Since human mobility is highly unique [15], individual trajectories act as fingerprints, making individuals in trajectory datasets likely to be re-identified using only a few *known locations*. For instance, a trajectory in a rural area generates outlier locations that are easily exposed [55], and the identity of an individual might be uncovered by linking *shared paths* (i.e. connecting individuals with high trajectory similarity). Additionally to single trajectory locations, individual moving history unveils personal routines and idiosyncratic behaviors that are easily linkable to individual identities. For instance, the *personal gazetteer* identifies recurrent locations in everyday life, such as home, work, and favorite restaurants. Similarly, the *location probability distribution* identifies how likely an individual is in a given location at a given time. Although some *spatiotemporal patterns* are extracted for the good purposes such as destination prediction [41], point-of-interest (POI) recommendation [56], [57], and personalized navigation [58], [59], acquiring these

TABLE 1
Categorization of sensitive information, sensitive spatiotemporal patterns, and attack models.

Sensitive Data	Attack Model	Exploited Spatiotemporal Pattern	Reference
Identity	Record linkage	Known locations	[30], [31], [32]
		Location probability distribution	[33], [34], [35], [36], [37]
		POI / Personal gazetteer	[38]
		Shared path	[39]
Personal profile	Attribute linkage	Recurrent mobility pattern	[40], [41], [42]
		POI / Personal gazetteer	[40], [43], [44], [45], [46], [47]
	Probabilistic attack	Known locations / Subtrajectories / Outliers	[48], [49] [50], [51]
	Table linkage	Aggregated location statistics	[52]
Social relationship	Group linkage	Encounter / Proximity	[53]

distinguishable knowledge dramatically enhances attackers' capability of identifying a specific individual.

Personal Profile: Besides identities, personal gazetteers (e.g., frequent locations, check-ins, POIs) and individual mobility also unveil personal profiles. The semantic information on locations contained in *personal gazetteers* expose individual habits (e.g., religion, wage) to user profiling [47]. Similarly, *mobility preferences* or *recurrent mobility patterns* (e.g., how likely an individual rides a bicycle instead of driving a car, or knowing her preferred routes or frequent stops) vary from person to person [60], [61], exposing even religion [42]. Indeed, by the analysis and prediction of individual trajectories, it is possible to infer demographics, lifestyle, and previously-unknown locations [40], [49]. Interestingly, also from *aggregated location statistics* (e.g., the number of individuals covered by a GSM cell) it is possible to infer the presence of an individual in certain dataset, allowing the inference of her personal data related to the dataset (e.g., her health condition if the dataset is about the movement of hospitalized people).

Social Relationship: Social relationships affect user mobility [62]. Following the ever-increasing amount of geo-tagged contents (e.g., check-ins or geo-localized games), individuals not only expose themselves through personal gazetteer, but also give the chance of inferring their social relationships [63]. Additionally, the wide-spreading of positioning systems (e.g., GPS and wireless access points) exposes aggregated patterns such as the *encounter* of people in area of interest (i.e., a continuous time interval in which individuals are close in space; e.g., concerts and manifestations). For instance, as individuals tend to group in communities (e.g., family and colleagues), the encounter and proximity of people in restricted areas unveils social ties based on co-located trajectories [53].

2.2 Attack Models

Due to the high sensitivity of trajectory data, an adversary can gather sensitive information of individuals within or across the datasets. We classify existing attack models on trajectories into two orthogonal categories: *linkage* and *probabilistic*. Linkage attack models refer to *what* sensitive data is inferred, and are categorized depending on such information, while the probabilistic attack models quantify *how much* knowledge is revealed by accessing the dataset. As for sensitive data, the spatiotemporal nature of trajectories opens new opportunities to specialize these generic attacks to the spatiotemporal domain.

2.2.1 Linkage Models

Depending on the attack target, linkage models are categorized into *record linkage* (i.e., inferring individual identity), *attribute linkage* (i.e., inferring personal profile such as health condition), *table linkage* (i.e., inferring personal data through the presence of a known individual in the dataset), and *group linkage* (i.e., inferring social relationships).

Record Linkage: Record linkage is the mainstream attack addressed by the state-of-the-art contributions. An adversary with some background knowledge (e.g., exposed locations [30], [31], origin and destination locations [32], and social relationships [64]) can attempt to identify the record of a known victim (i.e., run a re-identification attack). In [37], linkage is formalized as a k-nearest-neighbor search (i.e., finding the most similar k individuals to the query). While in [34], authors model a linkage attack as a bipartite graph in which individuals are modeled as two disjoint vertex sets connected by edges weighted by the similarity between the two individuals (e.g., the number of co-occurrences at a certain spatiotemporal bin). The maximal match within the bipartite graph [65] identifies the optimal linkage.

Existing record linkage attacks differentiate by how *individual similarity* is computed (i.e., what spatiotemporal patterns are exploited to link two individuals). In [35], authors discretize a map into a uniform grid, define the similarity between two individuals as the Jensen-Shannon divergence between their two location probability distributions, and finally link users minimizing the divergence. In [33], authors link datasets through a spatiotemporal join on co-occurring locations and time periods, leveraging known locations to prune the join space. In [39], authors model linkability in terms of spatiotemporal closeness between two trajectories. Additionally, when a location is missing from a trajectory at a certain time, authors interpolate such location by leveraging the distribution of historical locations. In [37], [66], authors map trajectories into road network locations, build compressed spatial signatures of trajectories by selecting the locations with the highest TF-IDF scores, and formalize linkage as k-nearest neighbor problem. While these attacks are based on trajectory micro-data (i.e., raw trajectory locations), aggregated trajectory data (e.g., the number of users within an area) also poses privacy issues. In [36], authors exploit the uniqueness and regularity of human mobility [67] (e.g.,

night and daytime mobility behaviors) to recover individual trajectories from aggregated mobility data without any prior knowledge. Given a dataset representing the number of trajectories in a cell at a given time, authors iteratively estimate the probability for an individual to move from a cell to another in its neighborhood and link adjacent locations by maximizing such probability.

Attribute Linkage: If sensitive values frequently occur within similar trajectories, an adversary can uncover sensitive information even though cannot unequivocally isolate single trajectories (i.e., perform an attribute linkage attack but not a record linkage attack). Despite value diversity can be ensured through l-diversity, if distinct sensitive values sharing a semantic similarity occur frequently within trajectories, an adversary can still cause a privacy breach (i.e., perform an attack based on similarity).

POIs and personal gazetteer easily expose personal data, since they characterize the individual interests [46]. Examples of POIs are home, work, religion or political parties' locations [40]. Revealing the POIs can cause a privacy breach as such data may be sensitive (e.g., frequent visits to a hospital suggest potential diseases). In [40], authors introduce a Markov model that represents the mobility behavior of an individual. POIs are states and transitions correspond to movements from one POI to another. Then, authors leverage such model to infer home locations (i.e., where individuals usually spend their night) and regular patterns emerging from circles in the mobility models. In [43], for each individual in a dataset of call records, authors extract her top-N locations (i.e., locations with high frequency) and join them with census data. In [44], authors introduce an algorithm to classify the POI semantic. Given two government diary studies (i.e., logs of two-day individual locations), a multi-class classifier [68] is trained to assign semantic labels based on individual demographics, time of visits, and nearby businesses. Furthermore, by extracting and predicting individual movement patterns (either short-term [41] or long-term [69]), it is possible to infer sensitive information such as the mode of transport, demographics and lifestyle [40]. In [45], given a dataset of location check-ins, authors use spatiotemporal knowledge and the regularity of human mobility to classify demographics attributes such as gender, age, education, and marital status based on the individual's POIs extracted from check-in dataset. In [42], a Reddit user identifies Muslim taxi drivers in New York City by integrating anonymized taxi trips to the daily praying time. By uncovering which taxi drivers are inactive at such time, it is possible to infer sensitive information such as religion. In [47], authors collect and integrate GPS locations with open data to profile the income, home and working locations of individuals frequenting a specific mall by summarizing frequent location patterns.

Table Linkage: The inference of an individual's presence in a private dataset can also leak sensitive information. For instance, knowing that a victim is part of a dataset of hospital patients implies that she suffers from some disease [52]. Membership disclosure attacks determine the presence of target individuals within a dataset. In [52], authors train a classification model to infer whether an individual is part of the aggregated released data. Although differential privacy reduces the attack success ratio, it yields a significant utility loss. Authors consider an adversary with different knowledge (e.g., locations or how aggregates were previously computed). Given a trajectory dataset, authors extract features for each region of interest (e.g., variance and sum of values of each location over time), then split the dataset into training and testing sets, and train the classifier mentioned above. A peculiar case of disclosure (not directly related to individual privacy) is the identification of military bases from the publication of a visual map representing sport activities using the Strava mobile application [70].

Group Linkage: The analysis of trajectory data can leak social relationships between individuals in the published dataset. For instance, individuals in the vicinity of each other on a frequent basis can share home or work places, or share the same religious and political orientation [40]. In [62], authors investigate the influence of social relationships on human mobility, showing that social relationships can explain about 10% to 30% of all human movement. In other words, individuals tend to group in communities (e.g., family and colleagues) where community members share some traits with other members stronger than with non-members [71]. Such phenomenon motivates group linkage attack. In [53], authors exploit the ubiquity of Wi-Fi access points to infer social ties based on co-located trajectories. Relationships are represented by an undirected weighted graph where vertices are individuals, edges are relationships, and edge weights quantify the relationship intensity. Communities are represented as sub-graphs. Authors characterize three relationship types: friends, classmates, and others. To construct the ground truth data, each relationship is assigned with one (or more) labels based on survey questionnaires. Then, they define an *encounter* as a continuous time interval in which individuals are close in space, and extract spatiotemporal features to train a classifier to label social relationships.

2.2.2 Probabilistic Models

A probabilistic attack quantifies *how much* information an adversary can gather by accessing the dataset rather than focusing on exactly what records, attributes, or tables the adversary can link to a target victim [72]. Intuitively, the access to a trajectory dataset should not reveal too much additional information to what is already known by the adversary. Probabilistic attacks can be considered as a generalization of attribute linkage [28], since their goal is not to infer a specific sensitive attribute, but rather to increase the *generic* knowledge of an adversary. For instance, given some locations known by an adversary, while linkage attacks focus on specific sensitive data, a successful probabilistic attack can reveal the entire trajectory of an individual (as in record linkage) as well as the sensitive attributes related to that trajectory (as in attribute linkage).

Recently, probabilistic attack to the trajectory dataset has been formalized in [48], where given τ known locations, an adversary is limited to learn only additional ϵ locations. The adversary knowledge can be any continued sequence of spatiotemporal samples, and the maximum additional knowledge that she can learn is called leakage. Similarly, [49] formalizes a probabilistic attack as the probability to learn a location previously unknown, and produces a privacy model to remove all the privacy breaches given some

known locations. Intuitively, such probability is related to the uniqueness of unknown locations belonging to the trajectories containing the known locations.

Normally, differential privacy providing strong and rigorous promises can handle these inference-based attacks like inferring whether an individual is included in a database. However, [73] observes that, even under differential privacy guarantee, the attack which focuses on learning properties of a population rather than directly learning attributes of an individual can be quite accurate and effective. Later, [74] formally distinguishes the fundamental difference between syntactic anonymity (targeting privacy-preserving data publishing) and differential privacy (targeting privacy-preserving data mining). Following these, [50] argues the importance of *syntactic attacks* in trajectory data privacy and formally classifies them into three types of threats including: 1) *Bayesian Inference Threat*, in which a malicious posterior belief is formulated after observing the sanitized trajectories and then is compared with the informed priors. A privacy leakage takes place if the gap is remarkable; 2) *Partial Sniffing Threat*, in which the locations exposed in sniffed regions cause the leakage of a subtrajectory of the user's full trace; and 3) *Outlier Leakage Threat*, in which outlier trajectories with highly unique features such as travel time and origin/destination locations can be easily singled out and a specific user might be identified with high confidence.

3 Protection of Trajectory Privacy

In this work, we focus on privacy protection of trajectories. We categorize privacy models for the release of anonymized trajectory data as *formal* and *ad-hoc* models. Formal models are independent from the data type, and extend the existing principles (e.g., k-anonymity, l-diversity, t-closeness, and differential privacy) to trajectories. Ad-hoc models are specific to spatiotemporal data and mobility features (e.g., road network constraints). In the following, we first briefly explain each type of privacy model, and then elaborate on well-known attempts applied to trajectories. Privacy models and their countered attacks are summarized in Table 2.

3.1 Formal models

These protection models define privacy on formal requirements which are usually expressed as parameters of the anonymization process. For instance, some models (e.g., k-anonymity, l-diversity, t-closeness) address quasi-identifier QI attributes (i.e. attributes enabling to breach identities after the anonymization process) or other sensitive attributes, while other models (e.g., differential privacy) try to guarantee an anonymized dataset leaks only controlled amount of information.

TABLE 2
Privacy models and countered attacks.

Attack Model	Privacy Model	Reference
Record link.	k-anonymity	W4M [75], GLOVE [76]
	l-diversity, t-closeness	KLT [22]
	differential privacy	DPT [77], SPLT [78]
	dummy	DTPP [31]
	mix-zone	UTMP [79]
Attribute link.	l-diversity, t-closeness	KLT [22]
Table link.	differential privacy	DPT [77], SPLT [78]
Group link.	-	-
Probabilistic	differential privacy	DPT [77], SPLT [78]
	attack resilience	AdaTrace [50], [51]

3.1.1 k-anonymity

Among the anonymity models, k-anonymity is the most extensively studied due to its intuitive anonymization process. Generally speaking, a dataset D satisfies k-anonymity if each QI value $D(QI)$ appears in at least k records. k-anonymity counters record linkage by ensuring the indistinguishability of an individual within a k-anonymous group and meanwhile minimizing information loss (intuitively, how much distortion is required to hide the individual within the group). Note that the optimal k-anonymity has been proved to be NP-hard [80].

In the context of trajectories, NWA [81] and its extension W4M [75], as well as GLOVE [76], are well-known implementations of trajectory k-anonymity and are often taken as baselines in privacy-model comparisons. However, due to the fact that the quasi-identifier (QI) in trajectories has not been formally defined yet, neither of these models follows the traditional way to achieve k-anonymity on trajectory data. Instead, two specific frameworks have been developed accordingly and widely used in the literature. On one hand, NWA and W4M share a consistent two-step *greedy* procedure: 1) building groups of at least k similar trajectories, and 2) anonymizing trajectories in each group. Apparently, the first step requires the definition of similarity/distance measures to group trajectories as well as the quantification of information loss or other utility metrics to perform locally optimal aggregation. On the other hand, GLOVE shows a different idea with two steps as well: 1) full calculation of trajectory-wise merge costs, and 2) hierarchical clustering by iteratively merging two trajectories with the smallest cost until each trajectory satisfies k-anonymity. Similarly, it is crucial to define the merge cost, since it determines not only to what extent the newly merged trajectories are protected but also how much utility will be reserved.

NWA: NWA [81] is the first implementation of (k,δ)-anonymity on trajectory data. It models trajectories as cylindrical volumes where radius δ represents the location imprecision. That is, two trajectories are indistinguishable if they move within the same cylinder (i.e., are closer than δ in the Euclidean space). In temporal dimension, NWA coarsens the start/end time of trajectories within an interval of length τ to enforce grouping trajectories with the same start/end time. In each group, NWA clusters trajectories in a greedy fashion. In brief, it selects proper centers of clusters, adds to each cluster the k-1 nearest trajectories that are closer than a given radius, and assigns the remaining trajectories to the closest cluster within the given radius. Note that clusters with less than k elements will be dropped as well as the outlier trajectories that cannot be added to any cluster. Finally, NWA ensures each cluster is (k,δ)-anonymous via space translation while minimizing distortion simultaneously.

W4M: Euclidean distance is employed in NWA, which makes it only applicable to trajectories with equal length.

W4M [75] extends NWA by introducing an EDR-based time-tolerant distance measurement between two trajectories. In particular, W4M adopts the greedy clustering based on the EDR distance to group trajectories in clusters having at least k elements, and then exploits the minimum space translation via spatio-temporal editing to push all the trajectories of a cluster within a cylindrical volume of radius $\delta/2$. In this way, each trajectory in a group is edited to be sufficiently similar with its center trajectory so as to make each cluster become a (k, δ)-anonymity set. Theoretically, the total computational cost of W4M is $O(|D|^2 n^2)$, where $|D|$ is the total number of trajectories to be anonymized and n is the average length of trajectories. It could be quite time-consuming due to the k-member clustering, and meanwhile the cost of measuring the EDR distance between two trajectories is proportional to the length of both trajectories.

GLOVE: GLOVE [76] represents a location as a rectangle in space with a time span rather than a cylindrical volume used in NWA and W4M. Basically, it consists of two steps: 1) computing the trajectory-wise merge cost (i.e., to what extent the two trajectories have to be stretched to produce a new one covering the others), and 2) iteratively merging two trajectories with the smallest cost until each trajectory is k-anonymous. At the point level, the stretch effort represents the smallest loss of accuracy resulted from making two spatiotemporal points indistinguishable from both spatial and temporal dimensions. During the hierarchical clustering of trajectories, the cost matrix is updated for the newly generated trajectory if it does not satisfy the k-anonymity and has to be merged further. In practice, the full calculation of trajectory-wise merge cost is also time-consuming, which leads to the time cost of GLOVE to be $O(|D|^2 n^2)$ in total.

Other Implementations: Following the framework of k-anonymity (e.g., NWA and W4M), many models attempt to further reduce information loss, such as applying minimum description length principle in a distance metric [82], coarsening begin/end timestamps to increase the number of anonymized trajectories [83], and enabling customized k for specific trajectories and time intervals considering that trajectories are not equally sensitive [84], [85], [86]. Another typical follow-up is TOPF [87], which uses a different clustering strategy by grouping trajectories with the same start/end time in k-anonymous groups, and iterates over the remaining trajectories to add sub-trajectories into existing groups with the same start/end time. In addition, [88] builds a weighted graph for each group where vertices are trajectories and trajectories overlapping in time are connected by edges weighted with their Euclidean distance. Then, the trajectory graph is partitioned into connected components until no connected component with more than k vertices exists. [89] extends [88] by including trajectory direction angle in the similarity function to achieve higher utility. Rather than directly clustering trajectories, KAM [90] groups all locations into density-based clusters, transforms each trajectory to a sequence of cluster centroids, and prunes all the trajectories whose path is shared by less than k others. [91] follows the framework of GLOVE, while achieving significant improvement on the model efficiency which is the most critical bottleneck of GLOVE. It fully utilizes the *locality* property of trajectories (i.e., individuals usually move around within certain areas) to avoid unnecessary pairwise calculation of the merge cost, with the help of hierarchical grid index and various pruning techniques. Experiments on real-life trajectory data demonstrate a model speedup by several orders of magnitude. Differently, [92] highlights the importance of semantic features hidden in trajectories. It defines sensitive areas covering various POI points and conducts trajectory ambiguity based on user motion modes, road network information for trajectory anonymization while maintaining data utility.

3.1.2 l-diversity and t-closeness

Although k-anonymity allows the release of indistinguishable data (thus counters record linkage), attribute linkage can also expose some sensitive information when individuals within an anonymity group share similar values on some sensitive attributes. Hence, l-diversity [10] is proposed to ensure that an anonymity group contains at least l *well-represented* values for each sensitive attribute. Several definitions of well-represented values exist. For instance, a dataset D satisfies *distinct* l-diversity if the number of values for the sensitive attribute in $D(QI)$ is at least l. Other definitions are based on entropy and frequency of values [10]. However, if the distribution of sensitive values in a group is known (e.g., is highly skewed) or the sensitive values are semantically similar, privacy can still be leaked [10]. T-closeness [11] overcomes these limitations of l-diversity in the protection of attribute linkage threats by ensuring that the distance between the distribution of sensitive attributes within a group and the global distribution is smaller than t.

KLT: A trajectory is intrinsically a sequence of spatiotemporal points which can have various semantic information such as POI or road network. KLT [22] is the only approach implementing both l-diversity and t-closeness in trajectory protection. It follows the framework of GLOVE [76] to ensure k-anonymity. Further, it involves the semantic data by partitioning the whole space into several regions, each of which is denoted as an irregular polygon covering various types of POIs. Each location in a trajectory located in a specific region is associated with the heterogeneous semantic labels. When merging trajectories, it combines neighboring regions to make the resulting region satisfying l-diversity (i.e., the number of distinct POI categories in that region should exceed l). Similar operations are applied to achieve t-closeness. That is, more neighboring regions are merged until the divergence between its POI distribution and that of the global city is no larger than t. Compared with GLOVE, the total computational cost of KLT increases to $O(|D|^2 n^2 N)$, where N is the number of regions in the space. The extra cost is caused by retrieving the list of regions when computing the cost matrix and merging trajectories for achieving two additional criteria.

Other Implementations: Except KLT considering both l-diversity and t-closeness formulations, some other models also implement l-diversity. For instance, $(K, C)_L$-privacy [93] guarantees that any sub-sequence τ of any known L locations is shared by at least K trajectories and that the confidence to infer any sensitive value from τ is at most C. Any sub-sequence $q, 0 < |q| \leq L$ is a violating sequence if it does not satisfy KCL conditions, and it will be suppressed from the trajectories. Similarly, PPTD [94] suppresses a critical sub-trajectory τ if the possibility to

link an individual in the private dataset given the sub-trajectory τ is higher than a given threshold. (α,K,L)-privacy [95] guarantees that any sub-trajectory τ is contained in a group of at least k elements, the probability of inferring a sequence of L sensitive locations from τ is lower than α, and the probability to infer a sensitive value v is lower than α. (l,α,β)-privacy [96] ensures distinct l-diversity, α-sensitivity (i.e., the probability to infer sensitive value is below α), and β-similarity (i.e., the probability to infer a value within a sensitive group is below β). Authors identify critical sequences of maximum length m (upper bound to the adversary knowledge) and modify/drop them to enforce l-diversity, α-sensitivity and β-similarity. c-safety [97] protects semantic trajectories based on the generalization of visited places within a POI taxonomy. This is similar to l-diversity, but the number of sensitive places is not fixed.

3.1.3 Differential Privacy

Differential privacy [12] ensures that the presence of a record in a dataset leaks a controlled amount of information ϵ. An algorithm f satisfies ϵ-differential privacy if for any two datasets D_1 and D_2 that differ on at most one record, and all sets S of values in the image of the algorithm (i.e., $S \subseteq Range(f)$), it has

$$Pr(f(D_1) \in S) \geq e^{\epsilon} \cdot Pr(f(D_2) \in S)$$

where Pr is the probability to observe a specific output. Differential privacy is usually guaranteed by generating synthetic data from the original one with controlled amount of random noise. In the field of traditional relational database, several randomized mechanisms have been already utilized to achieve ϵ-differential privacy. For example, the *Laplace mechanism* adds noise drawn from the Laplacian distribution $Lap(\frac{\Delta f}{\epsilon})$ [12] to the original database w.r.t. the function f. Another well-known technique is the *exponential mechanism* [98] that handles complex cases where the function f maps the data to strings, trees or other non-numerical data, which makes the Laplace mechanism no longer suitable.

Existing differential privacy models for trajectories share a common procedure: 1) modeling raw trajectories to capture the statistical distribution of original data, and 2) sampling *synthetic* trajectories (i.e., data not preserving truthfulness at record level) from the constructed mobility model. On top of this basic framework, the approaches vary from many aspects such as the ways of modeling trajectories, the sampling methods or the mechanism for noise injection.

DPT: DPT [77] is one of the most famous models achieving differential privacy on trajectory data which adapts the Laplacian mechanism to publish synthetic trajectories. In DPT, the entire space is discretized at different resolutions to build the hierarchical reference systems modeling the trajectories at various speeds, each of which corresponds to a prefix tree to store the counts of trajectories moving through these grid cells based on the l-order Markov process. Furthermore, an adaptive model selection step is proposed to learn the optimal height of tree as well as dropping some useless trees with high noise and low utility in the differential private manner. The privacy budget ϵ is divided into two parts, one of which is responsible for the bias caused by the removal of trees and the other is for the Laplace-based noise added to the counts in the tree nodes. Minimizing the error defined by these two parts is the goal of model selection. Finally, after the hierarchical reference system is stable, a direction weighted sampling strategy is adopted by remembering the recent trend of directionality during sampling. Avoiding sudden unrealistic changes of direction can improve the data utility. In principle, the runtime complexity of DPT is $O(|D|n|\Sigma||O|)$, where $|\Sigma|$ denotes all the possible anchor points in the spatial domain and $|O|$ indicates the number of required synthetic trajectories.

SPLT: SPLT [78] can be regarded as a variant of differential privacy which provides some sort of indistinguishability. Generally speaking, it ensures that an adversary cannot distinguish whether a synthetic trajectory is generated by a certain individual compared with other k-1 individuals in the original dataset. SPLT synthesizes trajectories with high *semantic similarity* (sim_S) and low *geographic similarity* (sim_G) compared to the original ones. Intuitively, given two individuals' mobility data, when sim_G is very low (i.e., the two do not frequently visit the places that are spatially close), sim_S can still be high (i.e., the frequent places are semantically similar, e.g., "home" and "work"). To this end, for each seed trajectory, authors compute a 1st-order Markov model representing the probability to visit and transit between locations. An aggregated mobility model is derived by averaging all the individual models. Next, a location-semantic graph is built by regarding each location as a vertex and weighting the edges based on the semantic similarity between locations. Vertices of this graph are clustered into classes, so that locations within the same class have similar semantics and could be visited in a same way regardless of their geographic distance. Then, each seed trajectory is transformed into a sequence of semantic classes. A valid trace similar to a seed is generated by sequentially picking a location from the semantic class and meanwhile enforcing its geographical consistency with the aggregated mobility model. Finally, authors run a privacy test to decide whether to release each synthetic trace under the required *statistical dissimilarity* (based on EDR distance) and *plausible deniability* (i.e., the synthetic trace could be generated by at least $k-1$ alternative trajectories).

AdaTrace: Recently, AdaTrace is proposed in [50], [51] to mitigate the shortcomings of DPT-based approaches which offer strong differential privacy guarantee but fail to resist some targeted syntactic attacks due to their probabilistic nature. To this end, AdaTrace combines differential privacy with attack resilience along with a utility-aware generator. In brief, it first extracts various features and encodes them in the private synopsis. The noise injection is enforced to satisfy both the standard differential privacy principle and the attack resilience constraints, including Bayesian inference threat (when an adversary has prior knowledge about a privacy sensitive zone as well as the visitors), partial sniffing threat (users can be tracked in sniff regions with the help of technical tools so as to expose a sub-trajectory of her full trajectory) and outlier leakage threat (trajectories with unique characteristics are regarded as outliers and can be easily hunted). Besides, the synthesizer particularly cares about the data utility such as the distributions of trip and route length, resulting in the utility-aware and attack-resilient synthetic trajectories which are highly useful in practice compared to DPT [77] and SPLT[78].

Other Implementations: As DPT does not consider temporal information in trajectory data, SafePath is proposed in [99] to synthesize spatiotemporal trajectories by adding the timestamp location to the prefix tree. Another drawback of DPT is the poor utility reserved in the output dataset. To this end, DP-STAR [51] synthesizes trajectories by injecting noise to various utility features including density grid, mobility model, trip distribution, route length. Raw trajectories are rewritten by their representative points derived from the minimum description length metric. A density-aware grid structure is built to preserve the spatial densities in the original dataset despite the Laplacian noise added to the counts. The mobility model in DP-STAR is actually a collection of transition probabilities by aggregating and averaging each individual model and the noise is injected to Markov chain. Besides, by taking care of users' trip lengths using a median length estimation method, it preserves more utility of data. In addition to the Laplace mechanism, many researchers also work on implementing differential privacy in trajectory data through the exponential mechanism [98]. For instance, [100] formalizes anonymity group as the one with the highest utility (i.e., intra-group similarity) among the groups in all the possible partitions. Since the number of partitions is exponential, authors provide a sub-optimal solution which leverages a single partitioning instance. Similarly, in [101], authors assign utility to k-means clustering in terms of intra-cluster distance and sample a clustering partition from an exponential distribution. In [102], authors generate synthetic trajectories by incrementally sampling the next trajectory location distance and direction from exponential distributions. Finally, differential privacy can be achieved via *randomized response*, i.e. deciding by chance whether to return the *actual* outcome or a randomized one. In [103], authors sample trajectory locations and interpolate the missing ones. Since locations adjacent to sensitive ones may leak sensitive information, Lclean [104] determines the correlation between sensitive and adjacent locations. For each sensitive region, Lclean finds sequences close in space/time that either do not contain sensitive information or show strong correlations. Given the sequences, Lclean substitutes trajectory subsequences via randomized response, making it impossible for an adversary to predict sensitive regions.

3.2 Ad-hoc Models

Some ad-hoc models have been proposed to address privacy preserving publication specific to trajectory data. Here, we discuss two popular models: *mix-zone* (i.e., geographical areas where individuals must swap identifiers) and *dummy* (i.e., synthetic trajectories resembling the original ones).

3.2.1 Mix-zone

Basically, a mix-zone refers to a geographical region on the map where passing objects are enforced to change their pseudonyms to avoid being tracked by the adversaries. The attackers need to observe pseudonyms of all ingress/egress events in order to reconstruct mappings between pseudonyms (i.e., record linkage). To apply mix-zones for trajectory privacy protection, existing approaches are mainly composed of two separate parts, i.e., the placement of mix-zones and the anonymization of trajectories. As the latter process is straightforward, researchers usually focus on the former one. In practice, to balance the level of privacy protection provided by Mixzone-based models and the reserved utility of generated trajectories, the placement of mix-zones is usually regarded as an optimization problem with many constraints to be satisfied, such as *location accuracy* (i.e., the bigger the area, the lower the accuracy), *sampling accuracy* (i.e., the higher the sampling rate, the more accurate the linking is), and *computational cost* (i.e., the more mix-zones, the higher the computational cost).

UTMP: UTMP [79] formalizes the deployment of mix-zones as an optimization problem by minimizing the number of pairwise-associated vertices in a road network. Two vertices are pairwise associated if a moving object can travel from one to the other without going through any mix-zone. As the optimal placement of mix-zones is a NP-hard problem, a heuristic solution is proposed in [79] to reduce computational cost. The road network is partitioned into disconnected components by looking for the articulation points (or called cut vertices) through a depth-first search. For each component, it finds a maximal independent set by iteratively adding non-adjacent vertices such that all the vertices that are not in the independent set are selected. To maintain the budget constraint K, it iteratively removes the vertex introducing the least number of pairwise associations from the candidate set until the total number of mix-zones is less than a given value K. As can be seen, determining the mix-zones is irrelevant to the original trajectory dataset but only depends on the structure of road network. Hence, the total computational cost of mix-zone placement in UTMP is $O(|V|(|V| + |E|))$, where $|V|$ is the number of anchor points and $|E|$ represents the number of edges connecting those points in the road network. Furthermore, the cost of anonymizing the trajectory dataset D with an average length of n is $O(|D|nK)$ in total, since it only needs to replace trajectory points with mix-zones.

Other Implementations: Some follow-up Mix-zone algorithms have been developed to further improve privacy protection. For instance, MobiMix [105] models a mix-zone as a k-anonymous region, where k individuals enter in some order, swap pseudonyms, and none leaves it before another k individuals have entered. The placement, geometry, and time spent inside mix-zones affect the privacy level. It is naturally easy to perform a first-in first-out attack if staying time is constant. Randomness ensures reordering, however, individuals are unable to often spend random time inside a road network, and do not follow uniform transition probability when entering/exiting the mix-zone (e.g., in case of trafficked routes [106]). MobiMix introduces the *time window bounded non-rectangular* mix-zone model: for each road junction, a mix-zone region starts from the center of the junction and expands to the outgoing road segment. The length of a zone is proportional to the average road-segment speed, providing the best protection against timing attacks. By contrast, [32] attempts to figure out the vulnerabilities of Mix-zone methods by conducting an attack under the assumption that moving objects follow the shortest path between origins and destinations. It claims that an attacker can compare the minimum path between known OD pairs using the Dijkstra algorithm in a road network and the minimum DTW distance between anonymized trajectories.

3.2.2 Dummy

Basically, the objective of dummy anonymization is similar to those synthesizing trajectories. However, unlike DPT [77] or SPLT [78], no mathematical formulation is adopted in dummy models. Instead, the generation of dummy candidates for each input trajectory is defined and executed in various ad-hoc ways. The effectiveness of dummy-privacy models highly relates to the potential capability to rule out unqualified trajectories.

DTPP: DTPP [31] generates dummy trajectories based on the assumption of some exposed locations. When producing dummy trajectories for a real trajectory, those exposed locations are remained in dummies while all the others are replaced by their neighboring points picked from the located grid cell. Meanwhile, all the generated dummy traces are verified by whether to be connective in the road network and be feasible in terms of the maximum speed derived from the true trajectory. Basically, DTPP generates k-1 dummy trajectories to form an anonymous trajectory set including the real one (whereas in k-anonymity no synthetic trajectory is generated). Each unexposed location in a trajectory should have at least $l-1$ alternatives in its dummies to ensure the diversity. Note that any unqualified trajectory or too sensitive location according to the anonymity requirements will be suppressed directly. Theoretically, DTPP is a very time-consuming model as the generation of k-1 dummies for each single trajectory takes $O(n^3m^2)$ time complexity, where n is the average length of trajectories and m denotes the average number of anchor points within a grid cell. Hence, processing a dataset D with DTPP costs $O(|D|n^3m^2)$ in total.

Other Implementations: Instead of considering the road network, [107] generates dummy trajectories resembling individuals moving in free space given three privacy parameters: *short-term disclosure* (i.e., the probability of successfully identifying a true individual location), *long-term disclosure* (i.e., the probability to identify a trajectory depending on its intersection with others), and *distance deviation* (i.e., the distance between dummy and real trajectories for a given individual). Authors introduce the *random pattern* strategy, which selects dummy start/end points and intermediate movements as random moves towards the end point. On the other hand, several implementations aim to reduce the number of generated dummies by applying different strategies. In [108], authors introduce the *K-intersected* strategy, where, given K intersection points as input, a dummy trajectory is generated by composing two sub-dummy trajectory sets: one between two intersection points (a sub-dummy is obtained by performing random moves from the start to the end point), and one of sub-dummies that do not contain intersection points. In [109], authors introduce the *adaptive generation* strategy for dummy trajectories. For each given rotation angle and location in a trajectory, authors synthesize a new candidate dummy trajectory satisfying the distance distortion, and then perturb trajectory locations to achieve more uniformly distributed trajectories by moving these locations in sparse areas. In [110], authors attempt to generate dummies resembling known individual movements between known stop locations.

4 Evaluation Metrics

Naturally, a good privacy protection model should be able to balance two metrics: *privacy* (how much private information is leaked) and *utility* (how much information is retained/lost). On one hand, returning completely random data guarantees privacy but results in null utility. On the other hand, retaining raw data maximizes utility but ensures no additional privacy. Therefore, privacy-preserving publication of trajectories aims to anonymize spatiotemporal dataset to release an altered version that prevents the disclosure of sensitive information while preserving its usefulness for certain analytic tasks. In this section, we provide a systematic summarization of privacy and utility metrics that have been used in the literature to evaluate the performance of existing privacy models designed for trajectories, some of which are also considered in our experiments.

4.1 Privacy Metrics

We first introduce some typical examples in different classes of privacy metrics along with the privacy models to which they can be applied (Table 3). The metrics provide a privacy evaluation additional to the privacy guarantees achieved in the formal privacy models, namely k tunes the size of the anonymity group in k-anonymity; l tunes the "well-represented" sensitive values in l-diversity; t tunes the distance of sensitive-attribute distributions between original and anonymized data in t-closeness; and ϵ tunes the amount of leaked information in differential privacy.

Group-based Metrics: For a group-based privacy model (e.g., k-anonymity, l-diversity and t-closeness), all the individuals within an anonymity group are indistinguishable from one another. The *anonymity group size* bounds the probability of identifying an individual within a group (namely, 1 divided by the group size) [90].

Sensitive Attribute Disclosure: When protecting sensitive attributes attached to the released trajectory records (e.g., l-diversity and t-closeness), the *disclosure risk* of an anonymized trajectory should be considered. In [96], authors identify the risk in terms of both identity disclosure and attribute disclosure given a sub-trajectory τ:

$$Pr_{dis}(\tau) = \max(\frac{1}{|S(\tau)|}, \frac{\alpha}{|D(\tau)|})$$

where $D(\tau)$ is the set of trajectories including τ, and $S(\tau)$ returns the set of sensitive values belonging to $D(\tau)$. α is a smoothing parameter.

Attack Success: Success metrics quantify how effective/accurate an attack model is. For instance, *identification accuracy* measures how many individuals can be accurately identified (i.e., linked back to the original records) after anonymization [111], [112], [32], [79]. In practice, this kind of metrics mostly depends on the adopted attack model. However, every privacy model usually aims at a specific attack, which leads to the lack of formal quantification and makes the comparison of privacy models difficult. Based on our summarization in Table 2, all the aforementioned privacy protection models can be applied to counter record linkage attack (i.e., re-identification attack), making it the best option to evaluate attack success ratio for comparing different models.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2022.3174204, IEEE Transactions on Knowledge and Data Engineering

TABLE 3
Privacy metrics with their scope of application.

Type	k-anonymity	l-diversity and t-closeness	Differential Privacy	Mix-zone	Dummy
Group-based	✓	✓	-	-	-
Sensitive information	-	✓	-	-	-
Attack success	✓	✓	✓	✓	✓
Mutual information	✓	✓	✓	✓	✓

Mutual Information: [101] uses mutual information to understand how much information an anonymized dataset leaks about the original one. In general, given two random variables X and Y, mutual information measures their mutual dependence, i.e. to what extent knowing one variable reduces uncertainty about the other. Hence, given two trajectories denoted as time series $x(t)$ and $y(t)$ with $t = \{1, ..., N\}$, the mutual information is defined as:

$$MI(x,y) = \sum_{t} \sum_{x(t)} \sum_{y(t)} Pr(x(t), y(t)) log \frac{Pr(x(t), y(t))}{Pr(x(t))Pr(y(t))}$$

$Pr(x(t)), Pr(y(t))$ are generic in [101] but can be specified according to what is measured. In particular, $Pr(x(t))$ of trajectories can represent the probability/frequency that individuals in the dataset occur in location $x(t)$ at time t, and $Pr(x(t), y(t))$ measures the joint probability.

4.2 Utility Metrics

It is crucial for any privacy model to preserve sufficient data utility, which is usually measured from two perspectives in the literature: the quality of trajectory data and the quality of data mining results for a specific trajectory operation. Note that these utility metrics are model-agnostic, i.e., they can be applied to evaluate any type of anonymization models.

4.2.1 The Quality of Data

We classify the data-based utility metrics into two categories: *statistical* metrics and *spatial* metrics.

Statistical Metrics: Basically, the quality of data before and after anonymization can be compared based on some statistical features. Anonymization is inevitably accompanied by *information loss*, which should be minimized to preserve enough data utility. Defining information loss varies according to the purpose and the way of achieving anonymization. For example, [113] as a suppression technique regards the information loss as the sum of distance between each suppressed trajectory and the original one. In [16], the average information loss is defined as the shrink of the probability that an object can be determined in a certain position. [114] evaluates point-level information loss based on the *translation ratio* which is the percentage of modified points in each trajectory after anonymization. In particular, it is computed as follows:

$$INF = 1 - \frac{1}{|D|} \sum_{\tau \in D} \frac{|\tau \cap \tau^*|}{|\tau|}$$

where τ^* is the anonymized trajectory belonging to the same user of τ, $|D|$ is the dataset size, $|\tau|$ is the trajectory length, and $\tau \cap \tau^*$ represents the set of common points between τ and τ^* (i.e., how many original points are preserved in the anonymized trajectory).

Spatial Metrics: From another perspective, trajectory data intrinsically has some spatial properties, which are expected to be sufficiently consistent after anonymization. Hence, several spatial utility metrics have been proposed and utilized in existing works. [115] aims at capturing the distance-based distortion of spatial shapes between original and anonymized trajectories. Any location removal in the anonymized version will be applied to a constant penalty. Plus, authors stress two desirable utility features: *location preservation* expects fewer fake locations replacing any original location to facilitate applications accurately; and *reachability* requires any anonymized trajectory to guarantee the geographical distance from its i-th location to the next is controlled. In [51], two spatial indicators are proposed in a similar way: 1) *trip error* is to quantify the preservation of start/end regions for each trip, which is defined as the grid-based Jensen-Shannon divergence between trip distributions of original and anonymized datasets; 2) *diameter error* is also measured by the Jensen-Shannon divergence between the diameter distributions, where the diameter of a trajectory is computed as the farthest pairwise distance.

4.2.2 The Quality of Data Mining Results

Apart from the above metrics evaluating the utility of data itself, another category of utility metrics pay attention to the performance of some trajectory operations such as querying, clustering, and pattern mining.

Query-based Metrics: Naturally, the accuracy of answering some generic queries can demonstrate whether the anonymized dataset is still useful. [116] proposes two categories of operators for querying trajectories. The first contains two point-based queries: $Where(\tau, t)$ returns the exact location of trajectory τ at time t; and $When(\tau, l)$ returns the time at which the object stays at location l in τ. The second type is a set of spatiotemporal range query operators to qualitatively describe an object's relative position with respect to a region from different aspects. The *average relative error* in [117] quantifies the accuracy of query answers as the average number of trajectories incorrectly retrieved by a certain *COUNT* query q in a workload Q:

$$error(q) = \frac{|q(D^*) \cap q(D)|}{|q(D)|},$$
$$error(Q) = \frac{\sum_{q \in Q} error(q)}{|Q|}$$

where $q(D)$ and $q(D^*)$ represent the result sets when using the query q to retrieve the original dataset D and the anonymized dataset D^*, respectively.

Clustering-based Metrics: The utility of data can be measured by the quality of clustering results obtained from the original and anonymized dataset, respectively. [90] focuses on two indicators: 1) the *precision* to measure how the

singularity of a cluster is mapped into an anonymized cluster; and 2) the *recall* to measure how the cohesion of a cluster is preserved. Similarly, [118] considers a utility metric, *global fitness*, measuring the quality of clustering. It generates some representative regions (RR) using a density-based clustering method on the end points of all trajectory segments and then generalizes the RRs to satisfy the k-anonymity. The fitness of a generalized cluster is based on the consistency of internal and external degrees, which indicates the number of sub-trajectories that arrive or depart from this region. In other words, it does not require exactly the same clusters after anonymization. Instead, the distribution of in-degree and out-degree should not change too much.

Mining-based Metrics: Frequent pattern mining is a popular task applied in trajectory analysis. [119] utilizes the $precision = N_m/N_r$ and $recall = N_m/N_a$ to measure the performance of privacy-preserving pattern mining. Here, N_r and N_a denote the total number of patterns in the raw mining results and the anonymized ones; N_m is the number of matched patterns occurring in both sets. Recently, [51] defines *frequent pattern support* as the average relative error with respect to the divergence of top-k patterns' support:

$$E = \frac{1}{k} \sum_{P \in FP(k,D)} \frac{s(D,P) - s(D^*,P)}{s(D,P)}$$

where the supports of a certain pattern P in the original dataset D and the anonymized dataset D^*, denoted as $s(D,P)$ and $s(D^*,P)$, are computed by the number of P's occurrences in D and D^* respectively; and the set $FP(k,D)$ consists of the top-k frequent patterns discovered from D.

5 Experiments

In this survey, we have conducted extensive empirical evaluation to show the pros and cons of each privacy protection model. Here, we will detail the dataset, evaluation metrics and compared methods used in our experiments, and report our experimental results and analysis comprehensively.

5.1 Experiment Setting

Datasets: Various types of trajectory datasets have been used to evaluate the performance of existing privacy models, such as taxi trips, user check-ins, phone call records, Bluetooth readings, etc. In this work, we adopt two publicly-available trajectory datasets, *T-Drive* and *Geolife*, to systematically compare the trajectory protection models discussed in Section 3. T-Drive [120] was generated by 10,357 taxis during the period of 2-8 February 2008 within Beijing, China. There are 94,177 raw trajectories consisting of 15 million GPS points. On average, the sampling rate is 3.1 minutes per point and the Euclidean distance between two continuous points is about 600 meters. We also generate some synthetic datasets from T-Drive with different characteristics (i.e., dataset size and sampling rate), to evaluate the sensitivity and scalability of the privacy protection models. Different from vehicle trajectories offered by T-Drive, Geolife is a check-in dataset generated by 182 users during five years. These trajectories were recorded by different GPS-enabled devices, and most of them were logged in a second-based dense representation. We regard each daily record as a trajectory of a user, resulting in 18,670 trajectories in total.

Evaluation Metrics: We compare the privacy models from various performance criteria including privacy metrics, utility metrics, and computational cost. Based on the existing evaluation metrics summarized in Section 4, we choose some representative measures as the privacy and utility metrics to compare all the models:

- Privacy metrics: As the attack success ratio can apply to all types of privacy models (formal and ad-hoc) and the record linkage attack (i.e., re-identification attack) is the most mainstream threat, we use the state-of-the-art re-identification algorithm [37] to evaluate the linking attack accuracy (*LA*);
- Utility metrics: 1) Point-based information loss (*INF*) [114] measures the percentage of modified points in the anonymized trajectory; 2) Diameter distribution error (*DE*) and trip distribution error (*TE*) [51] at the spatial level, where the diameter of a trajectory is defined as the maximum distance between two composing points and the trip of a trajectory is the pair of its start/end points; 3) F-measure of frequent patterns (*FFP*) [119] mines the top-ranked frequent itemsets of points in a trajectory.

Compared Methods: We report in this section the most *relevant* privacy models used for trajectory protection. Relevance is defined in terms of: (i) representativeness (i.e., for each type of privacy models, we select the implementations at the core of more recent contributions) and (ii) number of citations (i.e., how popular the privacy model is). The algorithms chosen for empirical comparison are:

- *W4M* [75] and *GLOVE* [76] as they represent the well-known major contributions to trajectory protection under k-anonymity principle;
- *KLT* [22] as it is the only attempt that adapts both l-diversity and t-closeness to trajectories against the semantic attack;
- *DPT* [77] as its noisy prefix-tree is at the core of many contributions on differential privacy for trajectories;
- AdaTrace [50], [51] as it is the latest differential privacy model further combined with attack resilience;
- *Mixzone* [79] as it provides a well-studied multiple mix-zone placement;
- *Dummy* [31] as it is the most well-known approach for the generation of dummy trajectories against the attack of exposed locations.

All the algorithms are implemented in Java[1], and evaluated on a server with two Intel(R) Xeon(R) CPU E5-2630, 10 cores/20 threads at 2.2GHz each, 378GB memory, and Ubuntu 16.04 operating system.

Parameter Setting: All the privacy models need to determine some hyper-parameters that play very different roles in the anonymization process. Parameter selection is not an easy task for a fair comparison among these models. Hence, we first refer to the original papers and conduct a series of preliminary experiments to understand the functionality of parameters within each model, respectively. Considering the

1. Github link of the open-source library will be added later.

trade-off between performance (i.e., the privacy protection level, the data utility reserved and the running efficiency), we finally fix these parameters as follows:

- $k = 5$ (*W4M*, *GLOVE*, *KLT* and *Dummy*);
- $l = 3$ (*KLT* and *Dummy*);
- $t = 0.1$ (*KLT*);
- $\delta = r = 500$ m (δ in *W4M* and radius r in *Mixzone*);
- $\epsilon = 5.0$ (*DPT* and *AdaTrace*);
- $m = 1000$ (total number of mix-zones in *Mixzone*).

5.2 Results and Analysis

In order to comprehensively compare the anonymization models, we evaluate their privacy protection level, utility loss, and time cost when varying the trajectory *dataset size* and *sampling rate*, respectively.

5.2.1 Sensitivity to Dataset Size

We examine the scalability of the privacy models as well as their sensitivity to the dataset size (i.e., number of objects). In particular, we generate six datasets with varying sizes by randomly sampling 100, 200, 500, 1000, 1500, and 2000 taxis respectively from T-Drive, along with all their original trajectories, and then apply each of the anonymization models. The results are depicted in Fig. 1.

Privacy Protection: Recall that we employ the current state-of-the-art re-identification algorithm [37] to simulate the linking attack. Each taxi is represented by a single trajectory (reflecting its whole moving history) in the dataset. After the anonymization, we search for the most similar trajectory in the original dataset D for each anonymized one in D^*. If two matched trajectories belong to the same object in the original and anonymized datasets, it will be regarded as a successful linkage. The linking accuracy is calculated by $LA = \frac{|D_s^*|}{|D^*|}$, where D_s^* denotes the set of anonymized trajectories that are successfully linked. Apparently, the higher the LA, the less protection the anonymization model offers. It is worth noting that, for generative privacy models (i.e., DPT and AdaTrace) producing the synthetic trajectories, we conduct a threshold-based linking attack with a predefined similarity threshold 0.5, which means two trajectories with similarity more than 0.5 will be regarded as correctly linked pairs when calculating LA.

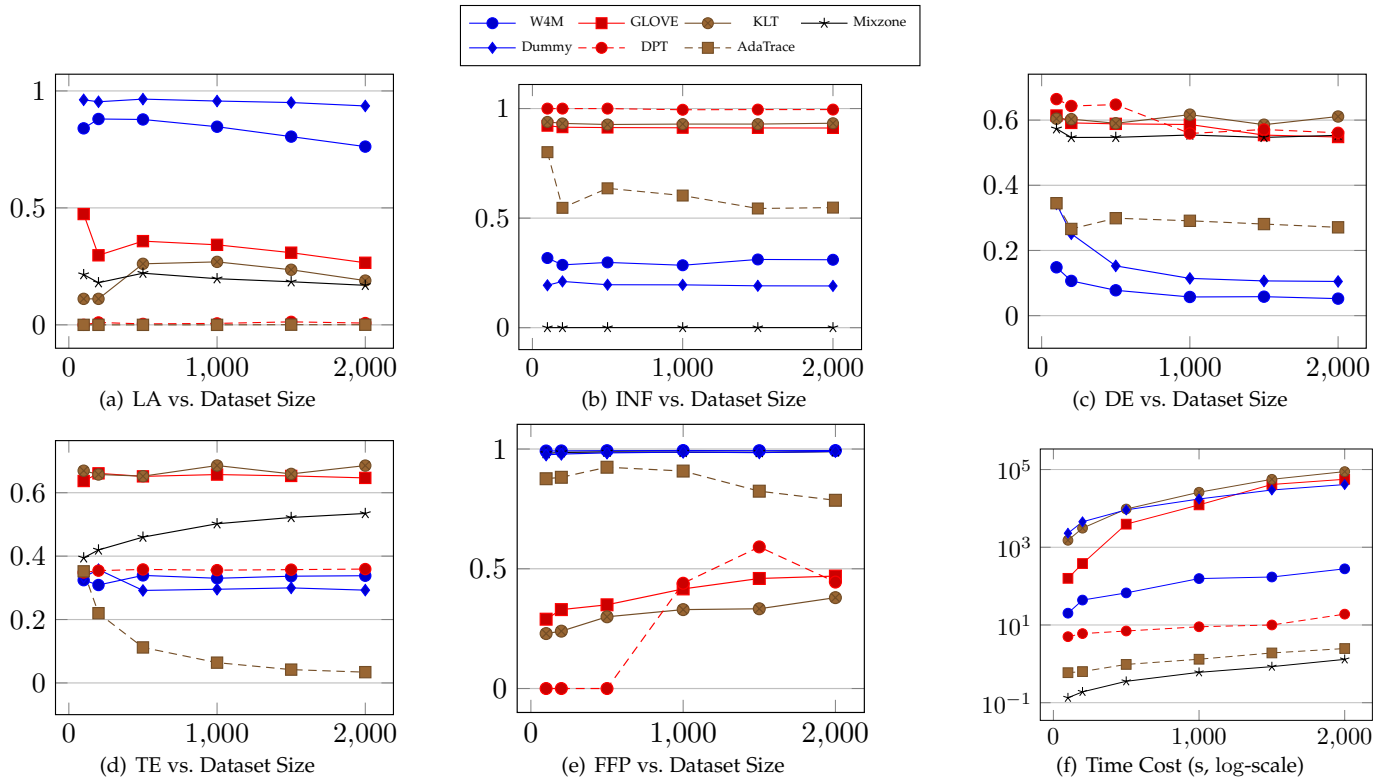
Overall, the linking accuracy drops slightly with the increase of dataset size. This is consistent with our expectation as the attack model needs to choose from more candidates to determine the matched individuals. Among all the anonymization models, *Dummy* and *W4M* provide much worse privacy protection than the others, with a linking accuracy of more than 80%. *W4M* modifies each trajectory to make it more similar with its pivot in a cluster and two spatially matched points would be reserved in the resulting trajectory for the purpose of utility preservation. As a consequence, many original points are actually unchanged, making it highly possible to run a successful linkage. In *Dummy*, points selected for composing dummy trajectories are usually close to the unexposed true locations in space, and hence dummies are mostly located together within a small area and easy to be linked. *KLT* performs much better than *GLOVE* in terms of privacy protection, thanks to the newly-incorporated l-diversity and t-closeness mechanisms. *Mixzone* delivers a similar performance with *KLT*. Apparently, two differential privacy models, i.e., *DPT* and *AdaTrace*, provide the most perfect protection against the re-identification attack, since the generation procedures completely reconstruct synthetic trajectories following the differentially private statistics without preserving any personal information of the original ones.

Utility Loss: Most approaches are relatively stable in terms of utility loss regardless of the varying number of objects to be protected. Information loss (INF), as the utility metric at the point level, demonstrates good performance on the ad-hoc models. *Dummy* generates dummy trajectories to make the real trips hidden within a k-size group and the participant points are also spatially close to the original ones, resulting in a large percentage (around 80%) of spatial point preservation. *Mixzone* has no point-level alteration but only splits a trajectory into several sub-trips along with pseudonym identifiers, and hence it reserves almost all raw points. Among the formal models, only *W4M* can retain around 70% of raw points in the anonymized trajectories. *GLOVE* and *KLT* brutally generalize the spatial points to regions for the purpose of privacy protection at the cost of losing more than 90% point-level information. As for the two generative differential privacy models, *AdaTrace* clearly outperforms *DPT* with a better balance between privacy guarantee and utility preservation. In fact, it even defeats almost all the other models except *Mixzone*, *Dummy* and *W4M* in terms of INF, due to its intrinsic design where the utility is particularly optimized. Regarding the divergence of diameter and trip (DE and TE, respectively) and the F-measure of top frequent patterns (FFP), the performance of *Dummy* and *W4M* are similarly desirable as well, since *W4M* anonymizes trajectories within its cylinder, leading to few changes in shape, diameter as well as the start/end positions; *Dummy* well controls the generation of dummy trajectories sufficiently close to the real ones at each timeslot. This shows a clear trade-off between the power of privacy protection and utility preservation for all these models. It is worth noting that when only a small number of objects are anonymized, *DPT* cannot discover any frequent sequential patterns occurring in the original dataset, which is caused by the incomprehensive mobility model captured by *DPT* from the extremely small original dataset. Hence, both pros and cons of *DPT* are quite obvious (i.e., strong privacy guarantee while large utility loss, and a higher requirement for data volume). Another notable observation is that the l-diversity and t-closeness mechanisms bring extra privacy protection gain but have a negative impact on the utility preserving, as demonstrated by the slightly worse results of *KLT* than those of *GLOVE* for all the utility metrics.

Time Cost: *Mixzone*, *AdaTrace* and *DPT* can efficiently process the trajectories, as *Mixzone* only needs to linearly scan the trajectories and split them if passing a pre-defined mix-zone area, while *AdaTrace* and *DPT* can generate synthetic traces as many as required after the features are extracted and the mobility models are built. The efficiency of *W4M* is also acceptable in practice, since it only takes around 2 minutes for anonymizing 2000 objects' mobility data. However, *GLOVE*, *KLT*, and *Dummy* are too time-consuming to serve for the anonymization of a real-life

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2022.3174204, IEEE Transactions on Knowledge and Data Engineering

Fig. 1. Impact of dataset size.



trajectory dataset (In fact, these three models were evaluated on some very small-scale datasets, e.g., with at most hundreds of objects, in their original papers). In comparison, the efficiency of *GLOVE* is better than that of *KLT* which takes some extra time to guarantee the l-diversity and t-closeness criteria during anonymization. *Dummy* is relatively less sensitive to the growth of dataset size $|D|$ than *GLOVE* and *KLT*, and its efficiency surpasses both *GLOVE* and *KLT* after $|D|$ increases to over 1200.

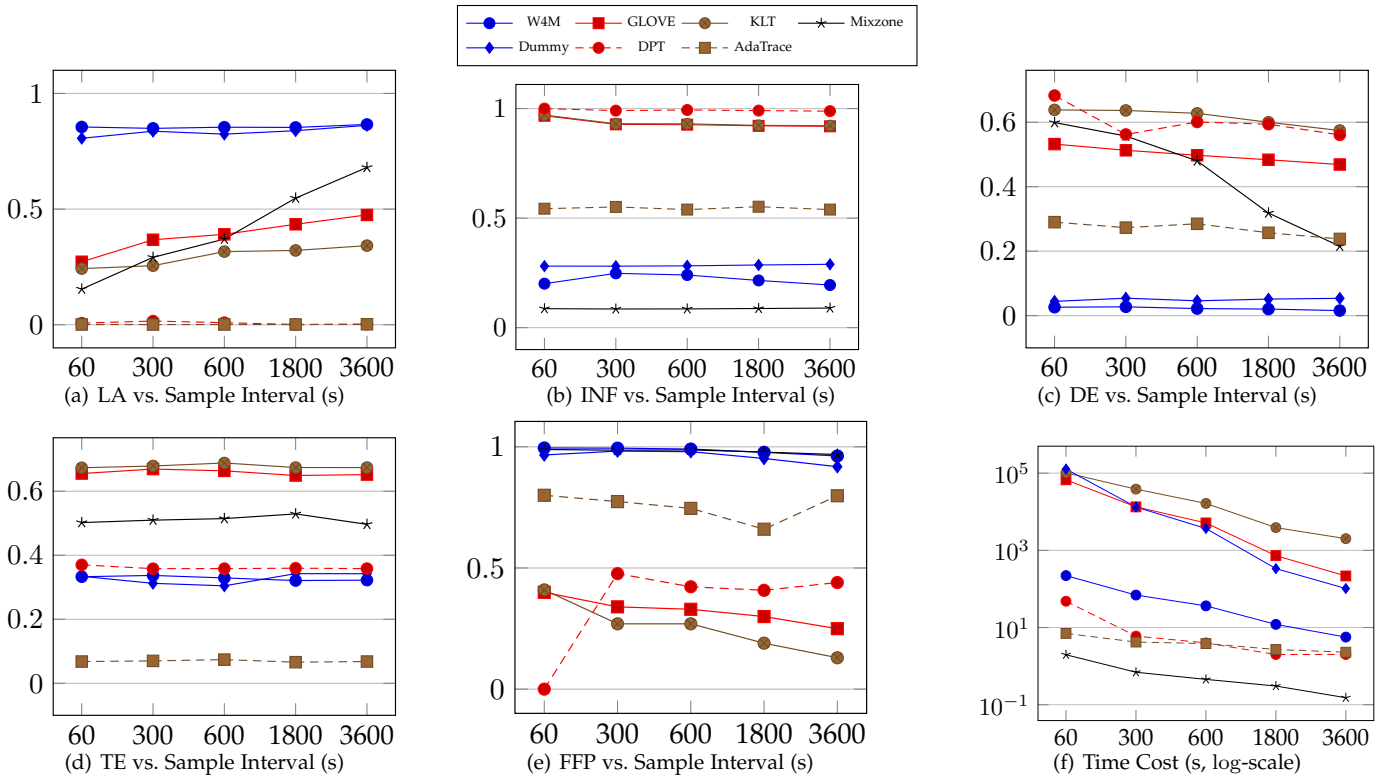
5.2.2 Sensitivity to Trajectory Sampling Rate

We expect that the sampling rate of trajectories, namely the average time interval between two consecutive points in the trajectories, may have some influence on the uncertainty of trajectory data. In another word, low-sampling-rate trajectories might lose most details of their movement, while on the contrary, more detailed trajectories with higher sampling rate always provide richer information that can be exploited as a weapon against personal privacy. Additionally, higher sampling rate leads to denser dataset and longer trajectories, which also poses great challenges to the efficiency of the anonymization models. Hence, in this part, we explore the capability of each privacy model in tackling trajectories with different sampling rates. Given the original T-Drive dataset which is around 3 minutes per point, we generate another five datasets with sampling intervals of 60, 300, 600, 1800, 3600 seconds, respectively. In particular, when preprocessing the T-Drive dataset, we insert extra samples into the raw trajectories based on the road network structure [121] to reach the denser sampling rate of 60s, while a straightforward down-sampling method is adopted to construct all the other sparser datasets. The empirical results are illustrated in Fig. 2.

Privacy Protection: Interestingly, the privacy protection that *GLOVE*, *KLT* and *Mixzone* provide drops with the increasing sampling intervals, while others are hardly influenced by that. Taxi trajectories used in the experiment are mainly based on passengers' demands, making them more random and less personalized. Taxis run on the road network, and inserting/removing several points uniformly from the original trajectories will not affect much on the overall spatiotemporal distribution of the data, leading to a relatively stable linking accuracy for most of the privacy models. On the contrary, *Mixzone* incurs an increase in linking accuracy from 15% to 68% when the sampling rate drops from 60s to 1h. Objects using the *Mixzone* mechanism would change their pseudonyms whenever passing mix-zones, meaning that the whole trajectory would be cut into subsequences belonging to different fake identifiers. Intuitively, the extent to which trajectories are divided by the predefined mix-zones partially depends on the density of trajectory data. That is, sparser trajectories are less likely to enter a certain mix-zone and be partitioned, as there are much fewer points in total. As a result, more original points would remain in the anonymized trajectories, which causes higher possibility to re-identify the objects. In *GLOVE* and *KLT*, k-anonymous trajectories are merged together based on the pre-computed stretch costs. According to the definition of cost, merging two denser trajectories inevitably loses much more spatiotemporal accuracy, which indicates that the resulting trajectory will be more dissimilar to the two original trajectories. Thus, it makes sense that *GLOVE* and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2022.3174204, IEEE Transactions on Knowledge and Data Engineering

Fig. 2. Impact of trajectory sampling rate.



KLT offer their best privacy protection when the sampling rate is 60s per point and the LA smoothly increases with the growth of sampling intervals. Finally, *DPT* and *AdaTrace* still greatly outperform all the other anonymization models when resisting linking attack in spite of the sampling rates. This is achieved by the differential privacy guarantee as well as the completely reconstructed trajectory data.

Utility Loss: The overall ability of these privacy models to preserve data utility is not greatly affected by sampling intervals. *Mixzone* still reserves the most percentage of raw points after anonymization as it does not conduct any point-level perturbation but only trajectory segmentation. *W4M* and *Dummy* as the second tier perform well in information loss (INF), followed by the *GLOVE* model. In comparison, *KLT* and *AdaTrace* have to modify almost all original points to satisfy their respective privacy principles. It is worth noting that *DPT* is defeated by *AdaTrace* on almost every utility metric with varying sampling rates, demonstrating that the utility-aware synthesizer in *AdaTrace* contributes a lot to the utility preservation, especially on the distribution of trip and trajectory diameter. Recall that the sampling rate affects the density of trajectories but barely the trip distribution, where a trip is defined as a grid-based origin/destination pair in the trajectory. Hence, the divergence of trip distribution (TE) between the original and the anonymized datasets does not change much for almost all the privacy models. The divergence of diameter distribution (DE), on the contrary, shows some notable decrease in *Mixzone*, *GLOVE*, and *KLT* when the sampling interval increases from minute-level to hour-level. As explained, the sparser trajectories pass mix-zone regions less possibly so as to retain more geographic diameter features of original ones. Therefore, the trajectories anonymized by *Mixzone* shows an obvious drop in DE with the increase of sampling interval. Regarding *GLOVE* and *KLT*, with the trajectories becoming sparser, the modification of points due to the anonymization would cause less fluctuation in their spatial coverage as well as diameter distribution and thus leading to a smaller DE. As for the frequent pattern mining (FFP), *GLOVE* and *KLT* notably show a decreasing trend when the sampling interval grows, whilst others keep stable or fluctuate slightly. In particular, the densest dataset generated by *DPT* can hardly retain any frequent patterns when the sampling rate is very high, caused by the excessive trajectory noise introduced into the mining algorithms. The gap between *KLT* and *GLOVE* is enlarged with the increase of sampling interval, especially on DE and FFP metrics, mainly because *KLT* has to compromise more information than *GLOVE* in order to further satisfy l-diversity and t-closeness.

Time Cost: As expected, denser dataset takes much more time to finish the anonymization no matter which model is adopted. Admittedly, the average trajectory length is proportional to the density of trajectories. Hence, this is also consistent with our theoretical complexity analysis for these models as discussed in Section 3. A notable thing is that the efficiency performance of *Dummy*, *GLOVE* and *KLT* are quite similar on denser datasets (i.e., with the sampling interval of less than 600s), while *Dummy* quickly surpasses the other two, especially the *KLT*, after the sampling interval grows to over 300s. This implies that *Dummy* might be more suitable for handling sparser trajectory data. Theoretically, the *Dummy* algorithm runs in cubic time of n, where n is the

average length of trajectories, while the time cost of either *GLOVE* or *KLT* is only quadratic to n.

5.2.3 Performance on Different Types of Trajectory Data

Naturally, the characteristics vary a lot among different types of trajectory data. Vehicle traces are automatically collected by some GPS-enabled loggers on a regular sampling basis, while location check-ins on social networks are labeled by users themselves with few temporal regularity. Even the trajectories of taxis and that of private cars have many differences, especially at the semantic level. Therefore, we choose two types of trajectory data and explore whether the privacy models perform differently on check-in data (i.e., Geolife) compared to the results on taxi data (i.e., T-Drive) with $|D| = 1000$ for a more comprehensive evaluation of the privacy models.

From Table 4, we observe that most approaches show an increase in privacy protection when processing Geolife data, coming with the increase of point-based information loss simultaneously. It indeed makes sense that the linking accuracy drops when more raw points are lost during anonymization. A trajectory generated by a taxi, on the other hand, is hard to be properly k-anonymous due to its intrinsic randomness and wide spatial range, while individual check-in history would be full of semantics and much easier to be hidden within an anonymous group. This can explain why group-based approaches (i.e., *W4M*, *GLOVE*, *KLT* and *Dummy*) show a clear drop in linking accuracy (LA) in Table 4. Another interesting observation is that, *DPT* and *AdaTrace* as two generative differential privacy models lose more statistical information (i.e., INF, DE and TE) but strengthen the ability to preserve frequent patterns (i.e., FFP) in the Geolife check-in data, as evidenced by the increase in all the four utility metrics in Table 4. It further verifies the claim made in [74] that differential privacy is more suitable for privacy-preserving data mining (PPDM) than privacy-preserving data publishing (PPDP). The noise injection mechanism for providing differential privacy guarantee inevitably brings too many noises into the anonymized data to preserve some basic statistics, while certain intrinsic hidden features like frequent patterns may survive since both approaches consider the Markov chain mobility model in their designs. Meanwhile, such hidden features are more obvious in the Geolife data as they reflect moving semantics.

TABLE 4
Performance comparison over different types of trajectories.

Dataset	Model	LA	INF	DE	TE	FFP
T-Drive (taxi)	W4M	0.847	0.285	0.057	0.330	0.994
	GLOVE	0.342	0.912	0.587	0.657	0.416
	KLT	0.269	0.929	0.617	0.686	0.330
	Mixzone	0.197	0.005	0.554	0.502	0.987
	Dummy	0.957	0.196	0.114	0.296	0.987
	DPT	0.006	0.995	0.559	0.356	0.591
	AdaTrace	0.000	0.603	0.291	0.064	0.908
Geolife (check-in)	W4M	0.253	0.492	0.018	0.273	0.770
	GLOVE	0.299	0.954	0.493	0.561	0.383
	KLT	0.214	0.971	0.508	0.588	0.279
	Mixzone	0.104	0.085	0.371	0.378	0.960
	Dummy	0.717	0.311	0.021	0.381	0.830
	DPT	0.068	0.998	0.598	0.371	0.800
	AdaTrace	0.001	0.760	0.368	0.235	0.958

5.3 Discussion

As a brief summary of the experiments detailed above, we compare the overall performance of representative trajectory protection models and examine how they are affected by the variation of dataset size (i.e., total number of objects) and sampling rate (i.e., average time interval between two consecutive points), respectively. The linkage attack model [37], [66] we choose in the experiments is quite generalized and can be countered by all the anonymization algorithms. We also evaluate their capability of utility preservation from four different perspectives: information loss (INF) at the point-based statistical level, diameter error (DE) and trip error (TE) measuring the spatial coverage and trip distribution respectively, and f-measure of frequent patterns (FFP) examining the usability for trajectory mining tasks.

Basically, these models show very different characteristics in practice. Some models (i.e., *W4M* and *Dummy*) are able to preserve desirable data utility but cannot resist the re-identification attack well. On the other hand, *DPT* provides strong guarantee of privacy protection without considering much on data utility. The k-anonymity models (i.e., *GLOVE* and KLT) can well-balance privacy and utility. In particular, *KLT* outperforms *GLOVE* when countering the linking attack by further incorporating l-diversity and t-closeness into the k-anonymity mechanism, at the increase of utility loss. This verifies the necessity of considering location semantics when protecting trajectory privacy. However, the price of the superior performance in *GLOVE* and *KLT* is the increase of computational complexity, as illustrated in both theory and practice. This is also the first time that the efficiency of trajectory anonymization models is highlighted and systematically evaluated. Overall, *AdaTrace* and *Mixzone* achieve the best trade-off between privacy protection, utility preservation and model efficiency. In particular, as two representative instances of generation-based differential privacy models, *AdaTrace* defeats *DPT* in almost every aspect, which is mainly contributed by considering the attack resilience constraints and designing the utility-aware trace synthesizer in *AdaTrace*.

6 Conclusion, Insights and Future Work

In this paper, we provide a comprehensive summarization and a systematic empirical study of the existing privacy protection models for trajectory publication. Specifically, we identify three types of sensitive information that can be discovered from trajectories (i.e., identity, personal profile and social relationship) as well as the typical attack models widely-used to expose such information (i.e., record linkage, attribute linkage, table linkage, group linkage, and probabilistic attack). We then discuss in detail how the well-known formal privacy models (i.e., k-anonymity, l-diversity, t-closeness, and differential privacy) and ad-hoc models (i.e., mix-zone and dummy) are adapted to trajectory protection. In our experiments on two real-life trajectory datasets, various privacy and utility metrics are utilized to compare the performance of these models and showcase their pros and cons for privacy-preserving data publishing.

6.1 Observations and Insights

We provide some insights on the superiority, limitations and proper application scenario of each type of trajectory privacy protection model, based on our observations and analysis in the experiments.

k-anonymity shows promising performance against the linkage attack (i.e., re-identification attack) in trajectory data, and meanwhile achieves a good trade-off between privacy protection and utility preservation. It is a quite simplistic principle of data privacy, which relies on making k elements indistinguishable via some common techniques such as generalization and suppression. However, its limitations are also obvious, especially when applying to trajectory data. First, it makes no assumption on the apriori adversary knowledge and cannot resist attribute linkage. Second, it is difficult to formally define the quasi-identifier and equivalent class in trajectory data since individuals' movements are highly unique and personalized. Finally, how to merge trajectories with the least utility loss still needs further study. Practical tasks which emphasize the truthfulness at record-level (i.e., no synthetic records generated) or can afford some privacy leakage for stable utility preservation would prefer to choose k-anonymity based approaches.

l-diversity and *t-closeness* are proposed to fix the vulnerabilities of k-anonymity, in particular the attribute linkage attack. Anonymization models that apply both mechanisms to trajectory data are rare, as the identification of sensitive attributes in trajectories is still a challenging task. In addition, they still have no quantification of the information leaked by accessing/querying an anonymized dataset, which is crucial to trajectory data. For example, an experienced attacker with background knowledge is able to potentially infer private information (e.g., an individual's presence/absence in a trajectory dataset) by repeatedly querying the data. Nevertheless, applying these two principles upon k-anonymity indeed gains more privacy protection due to the additional complex anonymization rules, which compromises the data usefulness and efficiency to some extent.

Differential privacy is one of the most powerful models which has no assumption on the type/amount of the adversary knowledge. It usually generates synthetic dataset from the original one through introducing random noises by Laplace or exponential mechanisms. Although it shows obvious superiority in tackling the linkage attack when applied to trajectory protection, it still suffers from a huge utility loss due to the tremendous modifications of the original points. Furthermore, as stated in [73], even under the guarantee of differential privacy, some attributes can still be exposed and become risky if the attacker aims to mine the properties of a population rather than targeting a person. Purely relying on differential privacy is not the ultimately safest choice. Thus, some attack-resilient models have been proposed to specify these threats and blood into the model design for the purpose of enhanced privacy protection.

Dummy, as an ad-hoc model specifically applied to trajectory data, aims at generating duplicate candidates to hide the original ones. However, dummy trajectories are still spatially and temporally close to the real one, leading to low privacy protection. *Mixzone* as another typical model can efficiently anonymize trajectories with acceptable data utility after the mix-zone regions are defined. Nevertheless, a reliable third-party is always needed to record the mappings between all the true identities and extensive pseudonyms so as to reconstruct the trajectories for analysis. Besides, *Mixzone* splits a trajectory into segments with unique pseudonyms. This not only causes an adversary to lose the tracking target but also damages data utility. Overall, some ad-hoc models expect to capture special properties of trajectory data for better performance but hardly provide theoretical guarantee or dramatically defeat formal models. It still has a long way to go and cooperating with well-defined privacy principles would be a better choice.

6.2 Open Challenges and Future Directions

We summarize some open challenges observed in this work, and introduce some future directions for follow-up studies in the field of privacy-preserving trajectory data publishing:

Model Adaption: Existing formal privacy models (i.e., k-anonymity, l-diversity, t-closeness, and differential privacy) have demonstrated their superiority in relational database, while it is still a challenging task to effectively adapt them to the trajectory data. The main issue of applying k-anonymity, l-diversity and t-closeness mechanisms lies in the inconsistency between relational modeling and trajectories. Unlike tabular records with well-defined attributes, a trajectory is intrinsically a sequence of spatiotemporal points, making it difficult to formally define both *quasi-identifiers* and *sensitive attributes*. Naturally, quasi-identifiers should be relatively unique and representative of an individual. [66] presents a pioneering work that extracts "signatures" from trajectories and utilizes them as quasi-identifiers to prevent the re-identification attack. Combining signatures with k-anonymity models and merging quasi-identifiers is a promising research direction yet to be explored. Similarly, POIs have been used in existing trajectory anonymization models to simulate sensitive attributes. Indeed, POIs reflect location semantics which can potentially expose some sensitive information such as an individual's religious or political orientation, health status, etc. However, aggregating all the POIs (as in existing work) may introduce extensive noises into model formalization. Instead, a selection mechanism to identify the real sensitive attributes should be studied. As for differential privacy, despite its proved superiority in relational data, how to accurately model people's collective spatiotemporal behavior and location semantics in trajectories and how to effectively introduce random noise for privacy guarantee are still challenging.

Model Efficiency: Based on our empirical results on real-life trajectory data, k-anonymity models (i.e., *GLOVE* and *KLT*) and ad-hoc model *Mixzone* achieve satisfactory trade-off between privacy protection and utility preservation, when countering the linkage attack. However, the cost is a huge increase of computational complexity. Hence, improving the efficiency of these models is definitely a promising direction for follow-up research, as real-world trajectory datasets are inevitably large-scale and the volume continues to grow with more data being collected over time. Although the running time of *Mixzone* is linearly proportional to the dataset size once the set of mix-zones is determined, finding the best mix-zones (i.e., optimal mix-zone placement) is still a challenging and time-consuming process, which calls

for effective approximation algorithms to be developed for addressing this NP-hard problem. As for *GLOVE* and *KLT*, the most costly operation in both models is the calculation of pairwise trajectory merge cost for identifying k-anonymous *equivalent classes* (i.e., clustering) so as to minimize utility loss. However, real trajectories are usually localized, and merging trajectories that are far away from each other in either space or time would naturally result in huge utility loss. In other words, merge costs only need to be calculated between nearby trajectories. Hence, it is also a promising direction to utilize such trajectory "locality" in *GLOVE* and *KLT*, and design effective pruning/indexing techniques to reduce the computation of trajectory merge costs.

Model Evaluation: It is necessary to evaluate and compare with state-of-the-art anonymization models in terms of both privacy protection and utility preservation. Data utility has been extensively considered in existing work, and this survey provides a comprehensive summary of utility metrics as well as a detailed classification that targets at different aspects of trajectory utility. Whereas, most privacy metrics are model-specific, except the attack success ratio discussed in Section 4.2. The absence of a standard privacy definition makes it difficult to measure privacy, compare between the anonymization algorithms, or make an informed choice for model selection. Therefore, a set of more generalized privacy metrics (e.g., mutual information, information entropy) need to be devised for a fair model comparison.

ACKNOWLEDGMENT

This work was partially supported by the Australian Research Council under grants DP200103650 and LP180100018.

REFERENCES

[1] A. F. Westin, "Privacy and freedom," *Washington and Lee Law Review*, vol. 25, no. 1, p. 166, 1968.
[2] Z. Yang, S. Zhong, and R. N. Wright, "Anonymity-preserving data collection," in *Proc. KDD*. ACM, 2005, pp. 334–343.
[3] R. Mendes and J. P. Vilela, "Privacy-preserving data mining: Methods, metrics, and applications," *IEEE Access*, vol. 5, pp. 10 562–10 582, 2017.
[4] R. Ostrovsky and W. E. S. III, "A survey of single-database private information retrieval: Techniques and applications," in *Proc. PKC*, ser. Lecture Notes in Computer Science, vol. 4450. Springer, 2007.
[5] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, pp. 14:1–14:53, 2010.
[6] P. Samarati and L. Sweeney, "Generalizing data to provide anonymity when disclosing information (abstract)," in *Proc. PODS*. ACM Press, 1998, p. 188.
[7] P. Samarati, "Protecting respondents identities in microdata release," *IEEE Trans. Knowl. and Data Eng.*, vol. 13, no. 6, 2001.
[8] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, vol. 10, no. 5, 2002.
[9] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," in *Proc. ICDE*. IEEE Computer Society, 2006, p. 24.
[10] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, p. 3, 2007.
[11] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. ICDE*. IEEE Computer Society, 2007, pp. 106–115.
[12] C. Dwork, "Differential privacy," in *Proc. ICALP*, ser. Lecture Notes in Comput. Sci., vol. 4052. Springer, 2006, pp. 1–12.
[13] A. E. Cicek, M. E. Nergiz, and Y. Saygin, "Ensuring location diversity in privacy-preserving spatio-temporal data publishing," *VLDB J.*, vol. 23, no. 4, pp. 609–625, 2014.
[14] Z. Montazeri, A. Houmansadr, and H. Pishro-Nik, "Achieving perfect location privacy in wireless devices using anonymization," *IEEE Trans. Inf. Forensics and Secur.*, vol. 12, 2017.
[15] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, p. 1376, 2013.
[16] R. Yarovoy, F. Bonchi, L. V. S. Lakshmanan, and W. H. Wang, "Anonymizing moving objects: how to hide a MOB in a crowd?" in *Proc. EDBT*, ser. ACM International Conference Proceeding Series, vol. 360. ACM, 2009, pp. 72–83.
[17] H. Kido, Y. Yanagisawa, and T. Satoh, "Protection of location privacy using dummies for location-based services," in *Proc. ICDE Workshops*. IEEE Computer Society, 2005, p. 1248.
[18] Y. Ouyang, Y. Xu, Z. Le, G. Chen, and F. Makedon, "Providing location privacy in assisted living environments," in *Proc. PETRA*, ser. ACM Int. Conf. Proc. Series, vol. 282. ACM, 2008, p. 39.
[19] V. Bindschaedler, R. Shokri, and C. A. Gunter, "Plausible deniability for privacy-preserving data synthesis," *Proc. VLDB*, vol. 10, no. 5, pp. 481–492, 2017.
[20] M. E. Nergiz, M. Atzori, and Y. Saygin, "Towards trajectory anonymization: a generalization-based approach," in *Proc. SPRINGL*. ACM, 2008, pp. 52–61.
[21] R. Chen, B. C. M. Fung, B. C. Desai, and N. M. Sossou, "Differentially private transit data publication: a case study on the montreal transportation system," in *Proc. KDD*. ACM, 2012.
[22] Z. Tu, K. Zhao, F. Xu, Y. Li, L. Su, and D. Jin, "Protecting trajectory from semantic attack considering k-anonymity, l-diversity, and t-closeness," *IEEE Trans. Netw. and Service Management*, vol. 16, no. 1, pp. 264–278, 2019.
[23] J. Jiang, G. Han, H. Wang, and M. Guizani, "A survey on location privacy protection in wireless sensor networks," *J. Netw. Comput. Appl.*, vol. 125, pp. 93–114, 2019.
[24] S. Zakhary and A. Benslimane, "On location-privacy in opportunistic mobile networks, a survey," *J. Netw. Comput. Appl.*, vol. 103, pp. 157–170, 2018.
[25] V. H. Le, J. den Hartog, and N. Zannone, "Security and privacy for innovative automotive applications: A survey," *Comput. Communications*, vol. 132, pp. 17–41, 2018.
[26] C. Bettini, "Privacy protection in location-based services: A survey," in *Handbook of Mob. Data Privacy*. Springer, 2018, pp. 73–96.
[27] V. Primault, A. Boutet, S. B. Mokhtar, and L. Brunie, "The long road to computational location privacy: A survey," *IEEE Commun. Surv. Tutorials*, vol. 21, no. 3, pp. 2772–2793, 2019.
[28] M. Fiore, P. Katsikouli, E. Zavou, M. Cunche, F. Fessant, D. L. Hello, U. M. Aivodji, B. Olivier, T. Quertier, and R. Stanica, "Privacy of trajectory micro-data: a survey," *arXiv preprint arXiv:1903.12211*, 2019.
[29] M. Guo, X. Jin, N. Pissinou, S. Zanlongo, B. Carbunar, and S. S. Iyengar, "In-network trajectory privacy preservation," *ACM Comput. Surv.*, vol. 48, no. 2, pp. 23:1–23:29, 2015.
[30] G. Poulis, S. Skiadopoulos, G. Loukides, and A. Gkoulalas-Divanis, "Apriori-based algorithms for k^m-anonymizing trajectory data," *Trans. Data Privacy*, vol. 7, no. 2, pp. 165–194, 2014.
[31] X. Liu, J. Chen, X. Xia, C. Zong, R. Zhu, and J. Li, "Dummy-based trajectory privacy protection against exposure location attacks," in *Proc. WISA*, ser. Lecture Notes in Computer Science, vol. 11817. Springer, 2019, pp. 368–381.
[32] E. P. de Mattos, A. C. S. A. Domingues, and A. A. F. Loureiro, "Give me two points and i'll tell you who you are," in *Proc. IV*. IEEE, 2019, pp. 1081–1087.
[33] M. Douriez, H. Doraiswamy, J. Freire, and C. T. Silva, "Anonymizing NYC taxi data: Does it matter?" in *Proc. DSAA*. IEEE, 2016, pp. 140–148.
[34] C. J. Riederer, Y. Kim, A. Chaintreau, N. Korula, and S. Lattanzi, "Linking users across domains with location data: Theory and validation," in *Proc. WWW*. ACM, 2016, pp. 707–719.
[35] M. Maouche, S. B. Mokhtar, and S. Bouchenak, "Ap-attack: A novel user re-identification attack on mobility datasets," in *Proc. MobiQuitous*, Melbourne, Australia, 2017, pp. 48–57.
[36] F. Xu, Z. Tu, Y. Li, P. Zhang, X. Fu, and D. Jin, "Trajectory recovery from ash: User privacy is NOT preserved in aggregated mobility data," in *Proc. WWW*. ACM, 2017, pp. 1241–1250.
[37] F. Jin, W. Hua, J. Xu, and X. Zhou, "Moving object linking based on historical trace," in *Proc. ICDE*. IEEE, 2019, pp. 1058–1069.
[38] M. Francia, E. Gallinucci, M. Golfarelli, and N. Santolini, "Dart: De-anonymization of personal gazetteers through social trajectories," *J. of Inf. Secur. and Appl.*, vol. 55, p. 102634, 2020.

[39] H. Wang, C. Gao, Y. Li, G. Wang, D. Jin, and J. Sun, "De-anonymization of mobility trajectories: Dissecting the gaps between theory and practice," in *NDSS'18*. The Internet Society.

[40] S. Gambs, M. Killijian, and M. N. del Prado Cortez, "Show me how you move and I will tell you who you are," *Trans. Data Privacy*, vol. 4, no. 2, pp. 103–126, 2011.

[41] A. Y. Xue, R. Zhang, Y. Zheng, X. Xie, J. Huang, and Z. Xu, "Destination prediction by sub-trajectory synthesis and privacy protection against such prediction," in *Proc. ICDE*. IEEE Computer Society, 2013, pp. 254–265.

[42] L. Franceschi-Bicchierai, "Redditor cracks anonymous data trove to pinpoint muslim cab drivers," *Online at: http://mashable.com/2015/01/28/redditor-muslim-cab-drivers*, 2015.

[43] H. Zang and J. Bolot, "Anonymization of location data does not work: a large-scale measurement study," in *Proc. MobiCom*. ACM, 2011, pp. 145–156.

[44] J. Krumm and D. Rouhana, "Placer: semantic place labels from diary data," in *Proc. UbiComp*. ACM, 2013, pp. 163–172.

[45] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. Xie, "You are where you go: Inferring demographic attributes from location check-ins," in *Proc. WSDM*, Shanghai, China, 2015, pp. 295–304.

[46] Y. Dai, J. Shao, C. Wei, D. Zhang, and H. T. Shen, "Personalized semantic trajectory privacy preservation through trajectory reconstruction," *World Wide Web*, vol. 21, no. 4, pp. 875–914, 2018.

[47] M. Francia, M. Golfarelli, and S. Rizzi, "Summarization and visualization of multi-level and multi-dimensional itemsets," *Inf. Sci.*, vol. 520, pp. 63–85, 2020.

[48] M. Gramaglia, M. Fiore, A. Tarable, and A. Banchs, "Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories," in *Proc. INFOCOM*. IEEE, 2017, pp. 1–9.

[49] M. Terrovitis, G. Poulis, N. Mamoulis, and S. Skiadopoulos, "Local suppression and splitting techniques for privacy preserving publication of trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 7, pp. 1466–1479, 2017.

[50] M. E. Gursoy, L. Liu, S. Truex, L. Yu, and W. Wei, "Utility-aware synthesis of differentially private and attack-resilient location traces," in *Proc. SIGSAC*. ACM, 2018, pp. 196–211.

[51] M. E. Gursoy, L. Liu, S. Truex, and L. Yu, "Differentially private and utility preserving publication of trajectory data," *IEEE Trans. Mob. Comput.*, vol. 18, no. 10, pp. 2315–2329, 2019.

[52] A. Pyrgelis, C. Troncoso, and E. D. Cristofaro, "Knock knock, who's there? membership inference on aggregate location data," in *Proc. NDSS*. The Internet Society, 2018.

[53] I. Bilogrevic, K. Huguenin, M. Jadliwala, F. Lopez, J. Hubaux, P. Ginzboorg, and V. Niemi, "Inferring social ties in academic networks using short-range wireless communications," in *Proc. WPES*. ACM, 2013, pp. 179–188.

[54] "Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation)," https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng, Accessed 09/12/2019.

[55] Y. Yu, L. Cao, E. A. Rundensteiner, and Q. Wang, "Detecting moving object outliers in massive-scale trajectory streams," in *Proc. KDD*. ACM, 2014, pp. 422–431.

[56] Y. Zheng, L. Zhang, X. Xie, and W. Ma, "Mining interesting locations and travel sequences from GPS trajectories," in *Proc. WWW*. Madrid, Spain: ACM, 2009, pp. 791–800.

[57] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Ma, "Recommending friends and locations based on individual location history," *ACM Trans. Web*, vol. 5, no. 1, pp. 5:1–5:44, 2011.

[58] K. Chang, L. Wei, M. Yeh, and W. Peng, "Discovering personalized routes from trajectories," in *Proc. LBSN*. ACM, 2011.

[59] J. Dai, B. Yang, C. Guo, and Z. Ding, "Personalized route recommendation using big trajectory data," in *Proc. ICDE*. Seoul, South Korea: IEEE Computer Society, 2015, pp. 543–554.

[60] S. Moosavi, R. Ramnath, and A. Nandi, "Discovery of driving patterns by trajectory segmentation," in *Proc. SIGSPATIAL PhD Symposium*. ACM, 2016, pp. 4:1–4:4.

[61] S. Moosavi, B. Omidvar-Tehrani, R. B. Craig, and R. Ramnath, "Annotation of car trajectories based on driving patterns," *CoRR*, vol. abs/1705.05219, 2017.

[62] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *Proc. KDD*. ACM, 2011, pp. 1082–1090.

[63] J. He, W. W. Chu, and Z. Liu, "Inferring privacy information from social networks," in *Proc. ISI*, ser. Lecture Notes in Computer Science, vol. 3975. Springer, 2006, pp. 154–165.

[64] W. Chang, J. Wu, and C. C. Tan, "Friendship-based location privacy in mobile social networks," *Int. J. Secur. Networks*, vol. 6, no. 4, pp. 226–236, 2011.

[65] H. W. Kuhn, "The hungarian method for the assignment problem," in *50 Years of Integer Programming*. Springer, 2010.

[66] F. Jin, W. Hua, T. Zhou, J. Xu, M. Francia, M. Orowska, and X. Zhou, "Trajectory-based spatiotemporal entity linking," *IEEE Trans. Knowl. and Data Eng.*, 2020.

[67] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, 2010.

[68] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[69] A. Sadilek and J. Krumm, "Far out: Predicting long-term human mobility," in *Proc. AAAI*, Toronto, Ontario, Canada, 2012.

[70] A. Hern, "Fitness tracking app strava gives away location of secret us army bases," *The Guardian*, vol. 28, 2018.

[71] S. Fortunato, "Community detection in graphs," *CoRR*, vol. abs/0906.0612, 2009.

[72] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, pp. 14:1–14:53, 2010.

[73] G. Cormode, "Personal privacy vs population privacy: learning to attack anonymization," in *Proc. KDD*. ACM, 2011.

[74] C. Clifton and T. Tassa, "On syntactic anonymity and differential privacy," in *Proc. ICDEM*. IEEE, 2013, pp. 88–93.

[75] O. Abul, F. Bonchi, and M. Nanni, "Anonymization of moving objects databases by clustering and perturbation," *Inf. Syst.*, vol. 35, no. 8, pp. 884–910, 2010.

[76] M. Gramaglia and M. Fiore, "Hiding mobile traffic fingerprints with GLOVE," in *Proc. CoNEXT*. ACM, 2015, pp. 26:1–26:13.

[77] X. He, G. Cormode, A. Machanavajjhala, C. M. Procopiuc, and D. Srivastava, "DPT: differentially private trajectory synthesis using hierarchical reference systems," *Proc. VLDB*, vol. 8, no. 11, pp. 1154–1165, 2015.

[78] V. Bindschaedler and R. Shokri, "Synthesizing plausible privacy-preserving location traces," in *Proc. IEEE Symposium on Security and Privacy*. IEEE Computer Society, 2016, pp. 546–563.

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

[80] A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity," in *Proc. PODS*. ACM, 2004, pp. 223–228.

[81] O. Abul, F. Bonchi, and M. Nanni, "Never walk alone: Uncertainty for anonymity in moving objects databases," in *Proc. ICDE*. IEEE Computer Society, 2008, pp. 376–385.

[82] F. Li, F. Gao, L. Yao, and Y. Pan, "Privacy preserving in the publication of large-scale trajectory databases," in *Proc. BigCom*, ser. Lecture Notes in Comput. Sci., vol. 9784. Springer, 2016.

[83] T. Chiba, Y. Sei, Y. Tahara, and A. Ohsuga, "Trajectory anonymization: Balancing usefulness about position information and timestamp," in *Proc. NTMS*. IEEE, 2019, pp. 1–6.

[84] S. Mahdavifar, M. Abadi, M. Kahani, and H. Mahdikhani, "A clustering-based approach for personalized privacy preserving publication of moving object trajectory data," in *Proc. NSS*, ser. Lecture Notes in Computer Science, vol. 7645. Springer, 2012.

[85] D. Kopanaki, V. Theodossopoulos, N. Pelekis, I. Kopanakis, and Y. Theodoridis, "Who cares about others' privacy: Personalized anonymization of moving object trajectories," in *Proc. EDBT*. OpenProceedings.org, 2016, pp. 425–436.

[86] Z. Hu, J. Yang, and J. Zhang, "Trajectory privacy protection method based on the time interval divided," *Computers & Security*, vol. 77, pp. 488–499, 2018.

[87] Y. Dong and D. Pi, "Novel privacy-preserving algorithm based on frequent path for trajectory data publishing," *Knowl.-Based Syst.*, vol. 148, pp. 55–65, 2018.

[88] Z. Huo, Y. Huang, and X. Meng, "History trajectory privacy-preserving through graph partition," in *Proc. MLBS*. ACM, 2011.

[89] S. Gao, J. Ma, C. Sun, and X. Li, "Balancing trajectory privacy and data utility using a personalized anonymization model," *J. Netw. and Comput. Appl.*, vol. 38, pp. 125–134, 2014.

[90] A. Monreale, G. L. Andrienko, N. V. Andrienko, F. Giannotti, D. Pedreschi, S. Rinzivillo, and S. Wrobel, "Movement data anonymity through generalization," *Trans. Data Privacy*, vol. 3, no. 2, pp. 91–121, 2010.

[91] Y. Wang, W. Hua, F. Jin, J. Qiu, and X. Zhou, "An efficient approach for spatial trajectory anonymization," in *Proc. WISE*. Springer, 2021, pp. 575–590.

[92] R. Tan, Y. Tao, W. Si, and Y.-Y. Zhang, "Privacy preserving semantic trajectory data publishing for mobile location-based services," *Wireless Networks*, vol. 26, no. 8, pp. 5551–5560, 2020.

[93] R. Chen, B. C. M. Fung, N. Mohammed, B. C. Desai, and K. Wang, "Privacy-preserving trajectory data publishing by local suppression," *Inf. Sci.*, vol. 231, pp. 83–97, 2013.

[94] E. G. Komishani, M. Abadi, and F. Deldar, "PPTD: preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression," *Knowl.-Based Syst.*, vol. 94, pp. 43–59, 2016.

[95] X. Liu, L. Wang, and Y. Zhu, "SLAT: sub-trajectory linkage attack tolerance framework for privacy-preserving trajectory publishing," in *Proc. NaNA*. IEEE, 2018, pp. 298–303.

[96] L. Yao, X. Wang, X. Wang, H. Hu, and G. Wu, "Publishing sensitive trajectory data under enhanced l-diversity model," in *Proc. MDM*. IEEE, 2019, pp. 160–169.

[97] A. Monreale, R. Trasarti, D. Pedreschi, C. Renso, and V. Bogorny, "C-safety: a framework for the anonymization of semantic trajectories," *Trans. Data Privacy*, vol. 4, no. 2, pp. 73–101, 2011.

[98] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. FOCS*. IEEE Computer Society, 2007.

[99] K. Al-Hussaeni, B. C. M. Fung, F. Iqbal, G. G. Dagher, and E. G. Park, "Safepath: Differentially-private publishing of passenger trajectories in transportation systems," *Comput. Netw.s*, vol. 143, pp. 126–139, 2018.

[100] J. Hua, Y. Gao, and S. Zhong, "Differentially private publication of general time-serial trajectory data," in *Proc. INFOCOM*. IEEE, 2015, pp. 549–557.

[101] M. Li, L. Zhu, Z. Zhang, and R. Xu, "Achieving differential privacy of trajectory data publishing in participatory sensing," *Inf. Sci.*, vol. 400, pp. 1–13, 2017.

[102] K. Jiang, D. Shao, S. Bressan, T. Kister, and K. Tan, "Publishing trajectories with differential privacy guarantees," in *Proc. SSDBM*. ACM, 2013, pp. 12:1–12:12.

[103] D. Shao, K. Jiang, T. Kister, S. Bressan, and K. Tan, "Publishing trajectory with differential privacy: A priori vs. A posteriori sampling mechanisms," in *Proc. DEXA*, ser. Lecture Notes in Computer Science, vol. 8055. Springer, 2013, pp. 357–365.

[104] Q. Han, D. Lu, K. Zhang, X. Du, and M. Guizani, "Lclean: A plausible approach to individual trajectory data sanitization," *IEEE Access*, vol. 6, pp. 30 110–30 116, 2018.

[105] B. Palanisamy and L. Liu, "Mobimix: Protecting location privacy with mix-zones over road networks," in *Proc. ICDE*. IEEE Computer Society, 2011, pp. 494–505.

[106] A. R. Beresford and F. Stajano, "Mix zones: User privacy in location-aware services," in *Proc. PerCom Workshops*. IEEE Computer Society, 2004, pp. 127–131.

[107] T. You, W. Peng, and W. Lee, "Protecting moving trajectories with dummies," in *Proc. MDM*. IEEE, 2007, pp. 278–282.

[108] P. Lei, W. Peng, I. Su, and C. Chang, "Dummy-based schemes for protecting movement trajectories," *J. Inf. Sci. Eng.*, vol. 28, no. 2, pp. 335–350, 2012.

[109] X. Wu and G. Sun, "A novel dummy-based mechanism to protect privacy on trajectories," in *Proc. ICDM Workshops*. IEEE Computer Society, 2014, pp. 1120–1125.

[110] R. Kato, M. Iwata, T. Hara, A. Suzuki, X. Xie, Y. Arase, and S. Nishio, "A dummy-based anonymization method based on user trajectory with pauses," in *Proc. SIGSPATIAL*. ACM, 2012.

[111] F. M. Naini, J. Unnikrishnan, P. Thiran, and M. Vetterli, "Where you are is who you are: User identification by matching statistics," *IEEE Trans. Inf. Forensics and Secur.*, vol. 11, 2016.

[112] S. Wang and R. O. Sinnott, "Protecting personal trajectories of social media users through differential privacy," *Computers & Security*, vol. 67, pp. 142–163, 2017.

[113] M. Terrovitis and N. Mamoulis, "Privacy preservation in the publication of trajectories," in *Proc. MDM*. IEEE, 2008, pp. 65–72.

[114] P. Han and H. Tsai, "SST: privacy preserving for semantic trajectories," in *Proc. MDM*. IEEE Computer Society, 2015, pp. 80–85.

[115] J. Domingo-Ferrer and R. Trujillo-Rasua, "Microaggregation- and permutation-based anonymization of movement data," *Inf. Sci.*, vol. 208, pp. 55–80, 2012.

[116] G. Trajcevski, O. Wolfson, K. H. Hinrichs, and S. Chamberlain, "Managing uncertainty in moving objects databases," *ACM Trans. Database Syst.*, vol. 29, no. 3, pp. 463–507, 2004.

[117] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Proc. ICDE*. IEEE Computer Society, 2006, p. 25.

[118] Y. Xin, Z. Xie, and J. Yang, "The privacy preserving method for dynamic trajectory releasing based on adaptive clustering," *Inf. Sci.*, vol. 378, pp. 131–143, 2017.

[119] S. Gurung, D. Lin, W. Jiang, A. R. Hurson, and R. Zhang, "Traffic information publication with privacy preservation," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 44:1–44:26, 2014.

[120] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: driving directions based on taxi trajectories," in *Proc. SIGSPATIAL*. San Jose, CA, USA: ACM, 2010, pp. 99–108.

[121] H. Su, K. Zheng, H. Wang, J. Huang, and X. Zhou, "Calibrating trajectory data for similarity-based analysis," in *Proc. SIGMOD*, 2013, pp. 833–844.

Fengmei Jin received her Bachelor of Engineering from Sun Yat-Sen University in 2016 and Master of Engineering from Renmin University of China in 2019. Currently, she is a PhD candidate at The University of Queensland. Her research interests include spatiotemporal databases, pattern mining, and data privacy.

Wen Hua is a Senior Lecturer and ARC DECRA Fellow at The University of Queensland. She received her PhD and Bachelor degrees in Computer Science from Renmin University of China in 2015 and 2010, respectively. Her main research interests include database systems, information extraction, data integration, and spatiotemporal data management.

Matteo Francia is an adjunct professor and post-doc research fellow at The University of Bologna, Italy, where he received the MSc and BSc with honors and the PhD in Computer Science and Engineering in 2021. His research focuses on analytics of unconventional data, with particular reference to trajectory, social, and sensory data.

Pingfu Chao Pingfu Chao received the BE degree in Automation from Tianjin University in 2012, the ME degree in Software Engineering from East China Normal University in 2015, and the PhD degree in Computer Science from The University of Queensland in 2020. Currently, he is working as an Associate Professor at Soochow University, China. His research interests include spatiotemporal data management and trajectory data mining.

Maria E Orlowska is a Professor at Polish-Japanese Academy of Information Technology in Warsaw, Poland. She was a Professor of Information Systems at The University of Queensland from 1988 to 2016. She is a Fellow of the Australian Academy of Science. Her main research interests include databases and business IT systems with a focus on modeling and enforcement issues of business processes.

Xiaofang Zhou is a Chair Professor at The Hong Kong University of Science and Technology. Before joining HKUST, he was a Professor of Computer Science at The University of Queensland from 1999 to 2020. His research interests include spatial and multimedia databases, high performance query processing, data mining, data quality management, and machine learning. He is a Fellow of IEEE.