IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received December 24, 2020, accepted December 28, 2020, date of publication December 31, 2020,
date of current version January 13, 2021.

*Digital Object Identifier 10.1109/ACCESS.2020.3048394*

# Risk-Aware Individual Trajectory Data Publishing With Differential Privacy

**JIANZHE ZHAO [1,4], JIE MEI [2], STAN MATWIN [3,5], YUKAI SU [1], AND YUANCHENG YANG [1]**

[1] Software College, Northeastern University, Shenyang 110169, China
[2] Microsoft Corporation, Redmond, WA 98052, USA
[3] Department of Computer Science, Dalhousie University, Halifax, NS B3H 4R2, Canada
[4] Neusoft Corporation, Shenyang 110179, China
[5] Institute of Computer Science, Polish Academy of Sciences, 00-901 Warsaw, Poland

Corresponding author: Jianzhe Zhao (zhaojz@swc.neu.edu.cn)

This work was supported in part by the Project supported by the Natural Science Foundation of Liaoning Province, China, under Grant 2020-MS-091, and in part by the Fundamental Research Funds for the Central Universities, China, under Grant N2017015.

**ABSTRACT** Large-scale spatiotemporal data mining has created valuable insights into managing key areas of society and the economy. It has encouraged data owners to release/publish trajectory datasets. However, the ill-informed publication of such valuable datasets may lead to serious privacy implications for individuals. Moreover, as a major goal of data protection, balancing privacy and utility remains a challenging problem due to the diversity of spatiotemporal data. However, the user dimension was not considered for traditional frameworks, which limits the application at the global level as opposed to the user level. Many researchers overcome this issue by assuming that a user in the dataset generates only one trajectory. Actually, a user always generates multiple and repetitive trajectories during observation. Only considering one trajectory for one user may cause insufficient privacy protection at the trajectory level alone, as a user's privacy can be manifested in many trajectories collectively. In addition, it demonstrates strong user correlation when using multiple and repetitive trajectories. If not considered, additional information will be lost, and the utility will be decreased. In this article, we propose a novel privacy-preserved trajectory data publishing method, i.e., IDF-OPT, which can reduce global least-information loss and guarantee strong individual privacy. Comprehensive experiments based on an actual trajectory publishing benchmark demonstrate that the proposed method maintains high practicability in trajectory data mining.

**INDEX TERMS** Differential privacy, trajectory data publishing, data correlation, utility optimization.

## I. INTRODUCTION

With the development of information technology and its penetration into daily life, sensor devices connected to the Internet, such as smartphones and wearable devices, are widely used, which results in a vast amounts of personal data with geographic location and time stamps being collected and stored [1].

Large-scale spatiotemporal datasets with abundant temporal and spatial information provide the basis for the research of trajectory data mining [2], [3]. The knowledge of regularity and aggregation of individuals or groups contained in the statistical information of trajectory data offers valuable insights into key areas of society and economy, such as transportation and urban planning [4], [5], health and welfare [6], [7], epidemiology, and natural disaster management [8]–[10]. Thus, the tendency of sharing large spatiotemporal datasets among multiple entities is becoming increasingly obvious. However, malicious attackers can also mine and spy on the sensitive information hidden in the data, such as home or work addresses, preferences, social relationships, income, and medical conditions, etc. Data mining and privacy protection have become two sides of the same coin. Any inappropriate data release may cause serious infringement of users' privacy. Meanwhile, application-oriented data publishing must be the priority of privacy protection. At present, with the development of privacy protection technology, balancing privacy and utility in the design of privacy protection methods has become a major goal. At the same time, in the context of big data, owing to the drive toward shared data, there are many additional privacy leaks caused by data association. The issue of related privacy leaks has become a new research hotspot for data privacy protection [11], [12].

Trajectory data refers to a sequence of geographic location coordinates of a moving object in a specified time slice arranged in time stamp order. The dataset formed by

The associate editor coordinating the review of this manuscript and approving it for publication was Chintan Amrit.

several trajectories is represented as a trajectory database. Determined by its nature, the trajectory database has the characteristics of high dimensionality, sparseness, and sequentiality [13]. In recent years, privacy preserved data publication has become a research hotspot in the fields of network security and data security [14], [15]. Many researchers have accumulated a diversity of research results and manifested rich achievements in privacy protection level, data utility, publishing data characteristics and so on [16], [17].

The present study of the privacy protection data release mechanism is mainly based on anonymity and differential privacy. Data publishing based on anonymity has led to some research work based on partitioned privacy models [18]–[20], such as k-anonymity [18], l-diversity [19], and so on. However, the design of the trajectory anonymity method is highly dependent on the background knowledge, which shows the vulnerability to the background knowledge of the attacker. Differential privacy (DP) [21], [22] is proposed for the privacy disclosure of statistical databases. Its advantage is that it provides background knowledge which is independent of attack and a strict and quantifiable privacy protection method. Trajectory data protection based on differential privacy technology enables the statistics of the published trajectory dataset and original trajectory dataset to meet the upper limit of the indistinguishable threshold by adding noise to the target database to ensure that the modification of a trajectory record in the dataset will not have a significant impact on the statistical results.

The existing trajectory data publishing methods based on differential privacy technology can be divided into two types. One of the types is the trajectory data publishing technology of position protection [23]–[26], which treats a trajectory as a database and each position in the trajectory as a separate record. The research on location privacy protection provides reasonable privacy protection solutions for location-based services (LBS). However, some studies have demonstrated the difference between location privacy and trajectory release [26]. The goal of another type is to publish a set of trajectories and treat each trajectory as a separate record, releasing a privacy protected synthetic trajectory dataset [27], [28] or trajectory data statistical dataset [29]–[34] based on differential privacy technology. In this type, trajectory statistics publishing technologies based on differential privacy that have been widely applied and recognized include n-gram [30] and DPT [31]. This kind of algorithm focuses on the high-dimensional characteristics of trajectory data and uses tree data or hierarchical structure to add noise to trajectory statistics to protect privacy and retain trajectory information to the maximum extent.

However, based on the review of existing research results, it is found that the current approach focuses on the privacy protection at the trajectory level but ignores the user dimension. Among them, the trajectory data publishing technology of location protection focuses on the location obfuscation mechanism on a trajectory to satisfy location privacy; the operation of differential privacy protection of the tracking dataset focuses on modifying a tracking record and adding noise to make the released tracking dataset indistinguishable from the original tracking dataset to satisfy the tracking privacy. This causes the continued existence of the problem setting of trajectory data release of traditional privacy protection: removing the user dimension and separating the connection between user and trajectory for privacy protection [18]–[20], [23]–[26]; supposing that one user can only generate one trajectory [27]–[34], the default privacy protection operation on a trajectory can protect the privacy of individual users. Actually, by observing the actual cases of the trajectory dataset, such as the D4D-Senegal dataset [35], from the perspective of user dimension, each individual would produce multiple trajectories during the observation time, many of which were repeated as shown in Fig.1. We observed repeated trajectories for the same user and the same trajectory for different users.

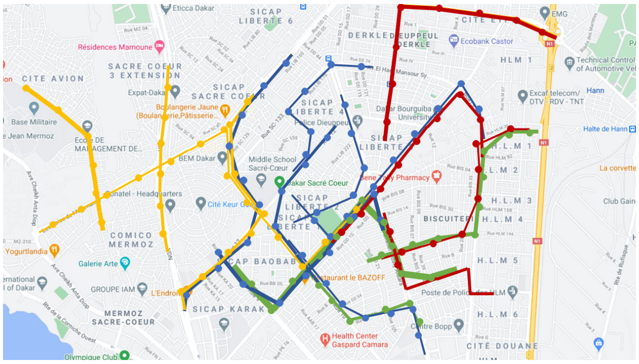


**FIGURE 1.** Observations of multiple and repeated trajectories from users.

We believe that adding the user dimension is very important for mining knowledge of the regularity and aggregation of individuals. However, in the traditional trajectory data publishing scenario, that is, data publishing without user dimensions, only all/a group of user statistics can be obtained, so the trajectory data application can only be carried out for all/a group of users. The release of such data cannot obtain personal statistical information, which limits its application.

To better release large-scale spatiotemporal data for application, we intend to define such a trajectory dataset as the individual trajectory dataset (ITD) and carry out research on privacy protection data release. Compared with the traditional trajectory dataset, the ITD added the dimension of users, as shown in Table 1. Through in-depth analysis and research, we found that due to the addition of the user dimension in addition to the traditional features of the personal trajectory database, the trajectories produced by individual users in the ITD have the characteristics of *multiplicity* and *repeatability*; at the same time, there is a certain *correlation* between trajectory data of different individual users.

Based on the literature review, we found that because of the addition of user dimension and new features of the ITD, the traditional differential privacy technology faces the following three challenges in trajectory data publishing under ITD scenarios:

**TABLE 1.** **ITD vs. traditional trajectory dataset.**

(a) ITD

| U-ID | T-ID | TRAJECTORY |
|---|---|---|
| $u_1$ | $T_{11}$ | $1 \to 3 \to 5 \to 7$ |
| $u_1$ | $T_{12}$ | $1 \to 3 \to 5$ |
| $u_1$ | $T_{13}$ | $5 \to 7$ |
| $u_1$ | $T_{14}$ | $3 \to 5 \to 7$ |
| $u_1$ | $T_{15}$ | $1 \to 3 \to 5 \to 7$ |
| $u_1$ | $T_{16}$ | $2 \to 4 \to 6$ |
| $u_2$ | $T_{21}$ | $1 \to 3 \to 5 \to 7$ |
| $u_2$ | $T_{22}$ | $1 \to 3 \to 5 \to 7$ |
| $u_2$ | $T_{23}$ | $3 \to 5 \to 7$ |
| $u_2$ | $T_{24}$ | $3 \to 5 \to 7$ |
| $u_2$ | $T_{25}$ | $5 \to 7$ |
| $u_2$ | $T_{26}$ | $1 \to 3$ |
| $u_3$ | $T_{31}$ | $2 \to 4 \to 6$ |
| $u_3$ | $T_{32}$ | $1 \to 3$ |
| $u_3$ | $T_{33}$ | $5 \to 7$ |
| $u_3$ | $T_{34}$ | $5 \to 7$ |
| $u_3$ | $T_{35}$ | $1 \to 3$ |
| $u_4$ | $T_{41}$ | $2 \to 4 \to 6$ |
| $u_4$ | $T_{42}$ | $1 \to 3 \to 5$ |
| $u_4$ | $T_{43}$ | $1 \to 3 \to 5 \to 7$ |

(b) traditional

| T-ID | TRAJECTORY |
|---|---|
| $T_1$ | $1 \to 3 \to 5 \to 7$ |
| $T_2$ | $1 \to 3 \to 5$ |
| $T_3$ | $5 \to 7$ |
| $T_4$ | $3 \to 5 \to 7$ |
| $T_5$ | $1 \to 3 \to 5 \to 7$ |
| $T_6$ | $2 \to 4 \to 6$ |
| $T_7$ | $1 \to 3 \to 5 \to 7$ |
| $T_8$ | $1 \to 3 \to 5 \to 7$ |
| $T_9$ | $3 \to 5 \to 7$ |
| $T_{10}$ | $3 \to 5 \to 7$ |
| $T_{11}$ | $5 \to 7$ |
| $T_{12}$ | $1 \to 3$ |
| $T_{13}$ | $2 \to 4 \to 6$ |
| $T_{14}$ | $1 \to 3$ |
| $T_{15}$ | $5 \to 7$ |
| $T_{16}$ | $5 \to 7$ |
| $T_{17}$ | $1 \to 3$ |
| $T_{18}$ | $2 \to 4 \to 6$ |
| $T_{19}$ | $1 \to 3 \to 5$ |
| $T_{20}$ | $1 \to 3 \to 5 \to 7$ |

*Note: in Table 1, user id is denoted as* **U-ID,** *the trajectory id of one user is denoted as* **T-ID,** *and the instances of individual trajectories are denoted as* **TRAJECTORY.**

- How to ensure adequate privacy protection for individuals in the dataset. In the scenario where one user generates one trajectory and one record belongs to one user, differential privacy can provide a strict and quantifiable privacy protection method for sensitive personal information in the dataset. However, the individual trajectory data release scenario described by the ITD, in which individual users generate multiple repeated trajectories, presents new requirements for differential privacy operations. On the one hand, traditional methods ignore the multiplicity of individual trajectory production. Current trajectory publishing technology defines privacy in a single trajectory and quantifies it. Faced with the problem of multiple trajectories produced by a user and repeated trajectories, existing research approaches are insufficient in quantifying the level of personal privacy protection. On the other hand, due to the neglect of the individual dimension, privacy protection remains at the trajectory level. However, the privacy protection target for individual users should be multiple trajectories rather than a single trajectory, and privacy protection at the trajectory level alone cannot provide sufficient privacy protection for each individual. At present, research on the privacy protection methods for multiple trajectories for individual users is insufficient.
- How to measure the correlation of differential privacy protection levels among individual users. In the ITD scenario, considering the perspective of user dimension, different individual trajectory data have high repeatability, leading to correlation between the data of different individuals, and the relevant data will generate additional privacy leakage of different degrees. Therefore, the level of privacy protection of individuals is not only affected by their own privacy parameters but is also affected by their relevant individual datasets. At present, research on the correlation level of privacy protection of individual users is insufficient.
- How to improve the utility of the overall published data. In the ITD scenario, to ensure that the publishing algorithm meets the requirement of differential privacy protection, repeatability of individual trajectories will lead to a huge loss of data utility. On the one hand, simply extending the traditional trajectory data publishing method to the ITD scenario is highly sensitive. On the other hand, adding noise to a trajectory in the trajectory dimension based on differential privacy protection technology may lead to more statistical information loss for individuals who repeat the trajectory. As a result, ITD privacy protection based on the existing differential privacy method will cause a large loss of data utility.

In this article, we propose a new privacy-preserved trajectory data publishing framework, i.e., risk-aware individual differential privacy optimization (IDF-OPT). It provides a solution to publish the sanitized ITD in the way of differential privacy methodology, which suppresses the riskiest trajectory under a specific threshold and simultaneously makes the protection secrets indistinguishable. The methodology proposed will cost the least-information loss in the global case and provides strong individual privacy guarantee.

The major contributions of the paper are as follows:

- The paper proposes a new privacy preserved trajectory data publishing method via differential privacy, i.e., IDF-OPT. It suppresses the high risk trajectory of individuals and adds noises to statistical dataset ensuring the indistinguishability to provide a strong privacy guarantee for each individual.
- The paper proposes a correlated differential privacy leakage model which provides a more delicate describing and dynamic standard for measuring the complex correlated data. It is an appropriate tool used to measure the correlated differential privacy leakage of individual drawing from the ITD scenario.
- The paper designs Pareto multiobjective optimization model and proposes an individual DF-optimization algorithm which uses to obtain a group of Pareto efficient parameters of correlated individuals maximizing the preserved utility of data.

## II. RELATED WORK

### A. PRIVACY-PRESERVED TRAJECTORY DATA PUBLISHING

Pufferfish privacy [36] gives a series of strict definitions related to privacy for data publishing, abstracting privacy protection as the indistinguishable pairs of potential secrets reaching a certain threshold through certain methods to achieve the privacy protection of the protection target. The existing research on trajectory data publishing for privacy protection uses anonymous technology or differential privacy technology to make the target trajectory in the trajectory dataset reach the indistinguishability threshold to achieve privacy protection.

The anonymity mechanism is mainly utilized to generate an anonymous set by means of generalization, bucketization, suppression, etc., so that the indistinguishable pairs in potential secrets reach the anonymity threshold. Researchers have proposed a variety of trajectory anonymity methods to ensure that the published trajectories meet specific threshold requirements [37]–[39]. Different algorithms improve the indistinguishability of protection targets by selecting different aggregation criteria. Among them, anonymous trajectory sets are generated according to specific generalization patterns, including partition-based generalization [40], hierarchy-based generalization [41] and spatial generalization [42]. To achieve a better generalization effect, a quasi-identifier (QIDs) mechanism [43] using the timestamp as a quasi-identifier and a local extension mechanism [13] were proposed. Nergiz *et al.* [38] proposed a data compression technique based on the attacker's background knowledge to model a set of trajectory ions in the trajectory dataset. However, existing opinions indicate that the design of trajectory anonymity methods is highly dependent on background knowledge, which shows vulnerability to attackers' background knowledge. Pellungrini *et al.* [44] measured the risk of GPS data privacy leakage based on the background knowledge of the attack and proposed a privacy quantification method based on background knowledge.

Differential privacy (DP) [21], [22] is a concept of privacy proposed to solve the problem of privacy disclosure in statistical databases. It provides a strict and quantifiable privacy protection method which is independent of attack background knowledge. In recent years, a large number of trajectory data publishing methods based on differential privacy technology have been proposed, which can be classified into two different types of trajectory data publishing. One of the types publishes trajectory data recorded by position [23]–[26]. This type of research makes use of the geographical indistinguishability and the expected inference of the next position to design the location confusion mechanism to satisfy the location privacy, which provides a reasonable privacy protection solution for LBS. However, location privacy differs from trajectory publishing in the following aspects. First, some location privacy does not consider the sequential nature of trajectory data. Studies have shown that only interfering with a single instantaneous location of a mobile user is still vulnerable to tracking and inference attacks [26]. Second, many location privacy works convert the original location into a location set or hidden area, and so this information is not easily used by trajectory mining applications that take the original location tracking as input. Finally, location privacy conceals the fact that users participate in the trajectory database, and privacy protected data publishing algorithms should provide traces of users' participation in the database.

Another type publishes datasets with different trajectories as records. The core of the differential privacy algorithm is to publish synthetic trajectory datasets based on the Laplace mechanism or exponential mechanism [27], [28] or to publish trajectory statistics [29]–[34]. Among them, some researchers aim to release the approximated trajectory of a real trajectory, that is, the synthetic trajectory, and release approximate trajectory data satisfying differential privacy to protect privacy, but the loss of the approximate trajectory in geographical location is irreversible, and the synthetic trajectory cannot retain the location information of the real trajectory, so the applicability of data analysis based on location statistical information is low. Different from publishing synthetic trajectories, publishing statistical trajectory data preserves the geographic location information of the trajectory to the greatest extent. This technology protects privacy by adding noise to the statistical trajectory information. At present, the most widely used method to achieve differential privacy is the Laplacian mechanism [27]–[29], [31]–[34], which implements the privacy protection of trajectory data by adding random noise sampled in the Laplacian distribution to the trajectory count.

Chen *et al.* [29] grouped sequences with the same prefix into the same branch and proposed a trajectory counting and noise algorithm based on a prefix tree structure. This is the first work that uses differential privacy technology to publish a large number of position sequences. Although their disinfection algorithm only retains counting queries and frequent item pattern mining, the data receiver can perform other data mining tasks on the disinfected output dataset. Chen *et al.* extended this work using the n-gram model so that the sequences stored in the tree can be of different lengths, and constructed a synthetic dataset based on Markov assumptions [30]. He *et al.* took advantage of the novelty of the hierarchical reference system and developed a trajectory publishing system DPT for privacy protection using the position discretization of the hierarchical organizational grid [31]. Shao *et al.* published a trajectory with a weak differential privacy protection concept by injecting noise into the trajectory position [32]. Hua *et al.* [33] reconstructed the trajectory by defining the utility function to achieve the minimum geometric distance, and then released the trajectory count with noise. n-gram and DPT are considered to be more advanced trajectory release technologies based on differential privacy. On this basis, Al-Hussaini *et al.* studied the privacy protection of passenger trajectory information disclosure and proposed the Safepath [34] algorithm to publish a differential privacy conversion trajectory. The algorithm modeled the trajectory as a noisy prefix tree to minimize the impact of data utility. This kind of research [29]–[31], [34] provides a reasonable hierarchical structure to reconstruct trajectories by adding noise to frequent prefixes or n-grams. This method effectively reduces the output domain and provides high practical value for frequency pattern mining in the global domain. Considering that the main application scenario of trajectory data is based on frequent pattern mining, the released noise count is more suitable for the application of trajectory data mining.

However, these methods present certain limitations. First, the research implicitly assumes that the original trajectory to be published contains a common prefix or n-gram. Second, the study assumes that the location comes from a small, discrete domain, such as hundreds of subway stations and bus stops [29], [30], [34]. In today's GPS driven mobile system, positions are collected in the form of (longitude, latitude) pairs at any location. Third, it does not consider the application value of repetitive personal trajectories within the same time segment, nor does it consider the issue of individual privacy protection due to diversity, repetition, and relevance.

The above research shows that the design of anonymity mechanisms generally assumes the background knowledge that the attacker may have, and selects the aggregation standard under this assumption, so that the technology exhibits the vulnerability to background knowledge. The advantage of differential privacy protection technology is that it is independent of the background knowledge of the attack, and it provides a strict and quantifiable privacy protection method. However, in the current differential privacy mechanism, the potential secret is declared as a trajectory or trajectory count, and the true count and the noise count declared as the trajectory or the real and reconstructed trajectory are indistinguishable. The level of privacy protection remains at the trajectory level instead of the individual level, and the protection object of interference with global statistical data is not an individual, but a trajectory of the individual. Owing to the ignorance of the user dimension, the data mining and application of individual trajectories is limited, and the research on the relevance and utility optimization of personal trajectory data release is insufficient.

### B. DIFFERENTIAL PRIVACY OF CORRELATED DATA PUBLISHING

Differential privacy provides a rigorous mathematical method of defining indiscernibility to protect privacy, ensuring that adding or removing any single record does not affect the results of the analysis. However, the recent research [45]–[47] shows that differential privacy is vulnerable when multiple datasets are correlated, though it provides a strong privacy guarantee with respect to the independent datasets.

Kifer and Machanavajjhala [45] first raised the important issue that the strong correlations make the sensitive data more readily distinguished from output. To remedy this defect, Kifer and Machanavajjhala [36] utilized differential privacy and defined a new privacy framework named Pufferfish, which considers the correlated data. To maximize the utility under privacy constraints, He *et al.* [47] proposed a new definition of Blowfish Privacy to tune privacy-utility tradeoffs. Chen *et al.* [48] demonstrated that differential privacy still provides a privacy guarantee for the correlated data and requires some adjustment.

Correlation is easy to define and measure when two different datasets contain an identical record about some user. However, it is more complex to measure the indirect correlation, which is defined as two different records about some user or his correlated users. For instance, information streams of some user's activity, e.g., GPS records and social network records, are correlated with each other. Kifer and Machanavajjhala [45] have shown that the privacy of correlated individuals may be compromised when their records are correlated.

To measure the privacy of the records with indirect correlation, substantial work has been conducted. Zhu *et al.* [49] utilized a correlated degree matrix to present the relationships between correlated datasets. In this case, the sensitivity of a query is changed into correlated sensitivity, which is the maximum among record sensitivities. Yang *et al.* [50] proposed Bayesian differential privacy leakage (BDPL) for correlated datasets. The idea is to utilize a Bayesian approach to analyze an uncertain query, accompanied with some given and unknown tuples.

Many state-of-the-art algorithms quantifying DP under temporal correlation have been proposed [11], [51]–[53]. Song *et al.* [11] present a detailed study about how to apply Pufferfish to achieve privacy and build up the robustness properties of Pufferfish against adversarial beliefs. They propose a mechanism called Markov Quilt to protect privacy for correlated data. Cao *et al.* [51], [52] analyze the privacy leakage of a DP mechanism under temporal correlation that can be modeled using Markov Chain and call the unexpected privacy loss temporal privacy leakage (TPL). They design data releasing mechanisms that convert any existing DP mechanism into effective one against TPL. Bozkir *et al.* [53] propose a novel transform-coding based differential privacy mechanism to further adapt it to the statistics of eye movement feature data by comparing various low-complexity methods, which provides the best utility-privacy trade-off in the eye tracking literature.

The proposed Markov models and Bayesian Network (BN) provide a probabilistic way for correlation measurement. Although probabilistic method is possible to model both user-user correlation (when the nodes in the model are individuals) and temporal correlation (when the nodes in the model are individual data at different time points), but the individual data is limited to a single data sequence, therefore it is not suitable for the setting of the context in our work, in which each user generates many trajectories. At the same time, aiming at mining the regularity and aggregation of personal knowledge, we believe that statistics of individual' trajectory has great significance. So before private operation, without considering the certain location in time we remove the timestamp of the position point within the trajectory. And another weakness is that the measurement of probabilistic correlation model between datasets is static, not dynamic. Even if some dataset adjusts its privacy level, the static matrix does not change the privacy relationship. Following the previous works, Wu *et al.* [54] provide their own definition of correlated differential privacy and provide a dynamic way to measure the correlation of datasets. However, it is not suitable for our case, in which the sub dataset of individual datasets pertains to a specific user, and the correlation defined between users by sharing one or more same records.

As shown in the existing research, the privacy level of some datasets is influenced not only by their own privacy parameters but also by their neighboring datasets when they are correlated with each other. We need a more delicate describing and dynamic standard for measuring the complex correlated data.

### C. UTILITY OPTIMIZATION

One of the fundamental challenges of privacy-preserved data publishing is utility optimization, since privacy-preserved mechanisms inevitably cause the utility loss of data [55]–[57].

For trajectory data publishing, the trajectory anonymity methods adopt various data utility metrics for different trajectory data mining tasks, aiming at preserving both instances of location-time doublets and frequent sequences in a trajectory database. Chen *et al.* [13] proposed a local suppression method which eliminates the exact instances that cause privacy breaches without penalizing others. Thus, local suppression much more effectively preserves data utility when compared to global suppression.

The existing differential privacy methods of solving the trade-off problem between utility and privacy are generally divided into two categories. The first category is using the exponential mechanism [58], [59], in which the utility function is defined to assign higher scores with the exponentially greater probability of being selected to an output so that the final output is close to the optimum utility. Li *et al.* [60] proposed a differential privacy trajectory publishing methodology using the utility function to merge locations generating closer trajectory partitions that effectively reduces the trajectory information loss after generalization. The second category is choosing the optimal privacy parameter to maximize the utility. The differential privacy trajectory data publication method proposed by Chen *et al.* [30] used the inherent constraints of a prefix tree to conduct constrained inferences to select privacy parameters, which leads to better utility. However, they only consider the privacy-utility trade-off of a single dataset without the privacy influence of the other datasets.

## III. PROBLEM DEFINITION

In this section, we first focus on a set of assumptions which indicate how the data were generated, how the data are correlated, and what potential attacks are possible in actual scenarios. We describe evolution scenarios by giving the definition of the individual trajectory dataset, the definition of individual privacy risk derived from actual scenarios, and the definition of risk-aware correlated individuals. Aiming at protecting each individual, we then describe the privacy requirement for making discriminative pairs indistinguishable, along with the utility requirement about losing less individual information. We summarize the privacy issue of individual trajectory data published in the problem statement.

### A. INDIVIDUAL TRAJECTORY DATASET

The individual trajectory dataset contains multiple users' trajectories during observation. Different from the traditional trajectory dataset, the ITD includes one user dimension.

*Definition 1 (Trajectory): A trajectory $T_{ij} = (l_1, t_1) \rightarrow (l_2, t_2) \rightarrow \cdots \rightarrow (l_{|T|}, t_{|T|})$ is a sequence of location-time pairs of length $|T|$. $U$ is the universe of users in the trajectory dataset, $\forall u_i \in U$ is the $i^{th}$ user of $U$, where $i = 1, \ldots, n$ and $j$ is the $j^{th}$ trajectory of $u_i$.*

*In the trajectory $T_{ij}$, $\forall (l_k, t_k)$ states in which $u_i$ appears at location $l_k$ at time $t_k$, where $1 \leq k \leq |T|$, $L$ is the universe location space $\forall l_k \in L$, a spatial point denoted by latitude and longitude coordinates.*

Length $|T|$ is a meaningful short time period during observation, such as one day in a year. For $\forall k \in |T|$, the observation interval is the time span between adjacent observation points, denoted as $|t_k - t_{(k+1)}|$. There exists the situation that in a time period which is $n$ times the observation interval, a user remains at a location without any movement. In comparison with the length $|T|$, the observation time period is quite a long time period. In this sense, taking no account of the difference of time intervals, we only focus on the mobility of an individual, and the trajectories in different meaningful short periods might be totally identical.

*Definition 2 (Individual Trajectory Dataset ): A trajectory dataset $D = \{D_1, D_2, \cdots, D_n\}$ consists of a group of datasets, each of which contains all trajectories generated by a single user in the observation time period, where $D_i$ is a subtrajectory dataset about $u_i$. If a trajectory dataset demonstrates the following characteristics, it is defined as ITD:*

**Multiplicity** *In the individual trajectory dataset, each individual generates k trajectories (k≥0) in the observation time period.*

**Repeatability** *In the individual trajectory dataset, each individual generates k trajectories in the observation time period, many of which are totally identical with each other. The repeatability of an individual's trajectory represents behavior regularity and a mobility pattern.*

**Correlation** *By sharing m identical trajectories ($k \geq m \geq 0$), individuals are correlated. The level of the correlation is measured by the numbers of the same trajectories. The more similar the patterns of the group individuals, the stronger correlations they have.*

### B. INDIVIDUAL PRIVACY RISK

Empirically, for any ITD, if a trajectory $T_{ij}$ frequently appears in $D_i$, it will reveal more behavioral habits of $u_i$. In contrast, relatively less is revealed if the same pattern frequently appears in the whole ITD. If the appearance of $T_{ij}$ is unique in $D$, this indicates that the adversaries have a significant probability of identifying the individual $u_i$ from one attack trajectory $T_{ij}$. Therefore, the following assumption is made in this work.

*Assumption 1: The privacy threat of individuals in an ITD is derived from the number of trajectories generated by individuals.*

Under the assumption, the difference of the time interval effect on the trajectory data can be ignored, and a trajectory denoted by location sequences in different time intervals may be completely the same; therefore, the numbers of the same trajectory can be counted.

*Definition 3 (Count of Trajectory): The number of appearances of trajectory $T_{ik}$ in $D_i$ is denoted as $TC_{ik}$, where $T_i$ is the set of different trajectories generated by $u_i$, $u_i \in U$, $i = 1, \cdots, n$, and the $k^{th}$ different trajectory of $u_i$ denoted as $T_{ik}$, $\forall T_{ik} \in T_i, k \leq j$. Here, $j$ is the total number of multiple trajectories of $u_i$, many of which might be repetitive.*

Considering the actual scenarios, we propose a flexible framework to measure the individual privacy risk, which is given as follows.

*Definition 4 (Individual Privacy Risk of ITD): The privacy risk (or risk of reidentification) of an individual $u_i$ denoted as $v_i$ is the reidentification risk from all of the different trajectories of $u_i$. It is the sum of the reidentification probability of $T_{ik}$, where $k = 1, \cdots, l$ and $l$ is the number of different trajectories of $u_i$ in the observation time period:*

$$v_i = \sum_{k=1}^{l} v_{ik} = \sum_{k=1}^{l} \Pr_D(u_i \big| T_{ik}). \tag{1}$$

In formula 1, $\Pr_D(u_i \mid T_{ik})$, i.e., $v_{ik}$ is the reidentification probability of $T_{ik}$.

*Definition 5 (Individual Trajectory Privacy Risk): The individual trajectory privacy risk of $T_{ik}$ is the reidentification probability of $T_{ik}$, denoted as $\Pr_D(u_i \big| T_{ik})$, i.e., $v_{ik}$. It is the probability of the specific trajectory $T_{ik}$ being re-identified in an ITD, and the value is the ratio of the counts of $T_{ik}$ appearing in $D_i$ and $D$, which can be calculated as follows:*

$$\begin{aligned} v_{ik} &= \Pr_D(u_i \big| T_{ik}) \\ &= \frac{TC_{ik}}{\sum_{j=-i} \mathrm{Match}(D_j \big| T_{ik}) + TC_{ik}}; \end{aligned} \tag{2}$$

$$\mathrm{Match}(D_j \big| T_{ik}) = \begin{cases} true & T_{ik} = T_{jl} \text{ for } \forall T_{jl} \in T_j \\ false & otherwise \end{cases}. \tag{3}$$

The number of appearances of $T_{ik}$ in $D_i$ is the count of the trajectory, i.e., $TC_{ik}$, which is described in Definition 3; the number of appearances of $T_{ik}$ in $D$ can be divided into two parts: the counts of the trajectory in $D_i$ and $D_{-i}$, where $D_{-i} = D \setminus \{i\}$.

The count of the specific trajectory in $D_i$ can be obtained by searching the same trajectory for all users in $D_{-i}$ according to the following matching function shown in formula 3, and then summing the numbers of the matching results. According to the risk measurement of Definition 4 and Definition 5, it is intuitive to reach the same conclusion as Assumption 1: that the individual trajectory privacy risk is monotonically increasing with the number of appearances of $T_{ik}$ in $D_i$, i.e., $TC_{ik}$; it is degrading with the number of appearances of the same trajectory in $D_{-i}$.

*Definition 6 (Individual Riskiest Trajectory): The individual riskiest trajectory is a trajectory which takes the maximum value of individual trajectory privacy risk in $T_i$, denoted as $T_{risk-i}$.*

*Definition 7 (Risk-aware Correlated Individuals): For $\forall u_i, u_j \in U$, if $\exists T_{ij} \in T_j$ makes $\mathrm{Match}(D_j \big| T_{ik}) = true$, then $u_i$ and $u_j$ are risk-aware correlated individuals; otherwise, they are risk-aware independent.*

### C. PRIVACY REQUIREMENT

To demonstrate a well-defined privacy requirement, first, we discuss a privacy-preserved data publication framework in this section, which is a series of operations of the differential privacy mechanism; we then illustrate the specific privacy requirements of ITD publishing in detail and provide the related definitions of the privacy protection for individuals.

A privacy-preserved data publication framework is a privacy protection framework which includes a series of operations of differential privacy mechanism design for satisfying a specific privacy requirement:

- A setting of response or publication from the data curator between the users and the database. Generally, the setting might be interactive or noninteractive.
- An operation on the raw dataset to create synthetic data for the purpose of generating two discriminative pairs to protect potential secrets, such as suppression: that is, global sensitivity.
- A methodology of making the discriminative pairs indistinguishable. For the differential privacy methodologies, the privacy parameter gives its privacy criterion to measure the indistinguishability.
- The mechanism for private perturbation of data. If the data are correlated, the correlations of data influencing privacy are considered. A general perturbation algorithm is given for utility optimization.

From the discussion in Section I mentioned above, the ultimate goal is providing a privacy guarantee for each individual in the ITD, and the potential secrets should be the set of all possible counts of the trajectories for each individual. Therefore, the privacy requirements need to consider the following operations.

In the ITD publishing in our work, we create a noninteractive setting by assuming that the data mining tasks would be fixed before publishing to the specific user. The interactive setting is more flexible than the noninteractive setting; however, the number of queries of the interactive setting is limited because excessive queries lead to a large amount of noise. Due to the special characteristics of ITD, we suppress some trajectories to reduce the disclosure probability of individuals: that is, the sensitivity of the DP mechanism. We should obtain a group of differential privacy parameters making the discriminative pairs indistinguishable and provide a rigorous undistinguished upper bound under the least-cost suppression, providing a strong guarantee for each individual. However, it is known that the same trajectories followed by different users cause correlation with each other. We should solve the problem that one user obtains a stronger guarantee, which causes the global utility to decline. Therefore, we should design the method to generate a group of optimal DP parameters.

To illustrate the specific privacy requirements of ITD publishing, we provide the related definitions as follows.

*Definition 8 (Differential Privacy): A random mechanism $M$ satisfies $\epsilon$-differential privacy if*

$$\mathrm{DP}(M) = \sup_{D_1,D_2,S} \log \frac{\Pr\left[r \in S \middle| D_1\right]}{\Pr\left[r \in S \middle| D_2\right]} \leq \epsilon \tag{4}$$

*for any datasets $D_1$ and $D_2$ differing in at most one record, and for any possible sanitized dataset $r \in Range(M)$.*

Global Sensitivity: For any function $f : D_i \to \mathbb{R}^d$, for all $D_1$ and $D_2$ differing in at most one record, the sensitivity of $f$ is

$$\Delta f = \max_{D_1,D_2} \left\|f(D_1) - f(D_2)\right\|_1. \tag{5}$$

*Definition 9 (Mechanism): Let D be a database, where a randomized function $M(D)$ is a (randomized) perturbation mechanism on D, if the output $r = M(D)$ follows a conditional distribution* $\Pr(r \in S \middle| D)$.

*Definition 10 (Laplace Mechanism): By adding Laplace noise to the output of a function to achieve differential privacy, the Laplace mechanism which takes a database D, a function $f$, and the privacy budget $\epsilon$ as inputs is designed for those functions whose outputs are real. For any function $f$ : $D_i \to \mathbb{R}^d$, the following mechanism provides $\epsilon$-differential privacy, in which $\epsilon = 1/\lambda$.*

$$M(D) = f(D) + Laplace(\frac{\Delta f}{\varepsilon}). \tag{6}$$

*Specifically, the Laplace noise is sampled from the Laplace distribution, denoted as Laplace($\lambda$) with the probability density function:*

$$\Pr(x \middle| \lambda) = e^{\frac{-|x|}{\lambda}}, \tag{7}$$

*which has mean zero and standard deviation $\sqrt{2}\lambda$, and $\lambda$ is the scale parameter determined by sensitivity $\Delta f$ and privacy budget $\varepsilon$, i.e., $\lambda = \Delta f/\varepsilon$.*

Scale parameter $\lambda$ determines the curve of the distribution, where a large $\lambda$ flattens the curve and leads to significant noise. When the global sensitivity is fixed ($\Delta f = 1$, with one trajectory specified for one individual), the indistinguishability is determined by the privacy parameter, such as $\varepsilon$ in the Laplace mechanism. Privacy parameter $\varepsilon$ is the upper bound of the indistinguishability, which measures the privacy leakage; also, as a privacy budget, it determines the curve of the distribution. Decreasing $\varepsilon$ will flatten the Laplace distribution curve and cause substantial noise. Given the fixed privacy budget $\varepsilon$, a large $\Delta f$ will flatten the curve and again lead to substantial noise. Therefore, the individual protection level and utility for analytics are decided by the sensitivity $\Delta f$ and privacy parameter $\varepsilon$.

Through the discussion in Section III, it is easy to infer that the individuals are risk-aware correlated in an ITD. There is an urgent need for accurate measurement of the correlation in terms of privacy to compute the real privacy level of an ITD. To measure the privacy of the record with the complex correlation of two datasets, many kinds of literature have studied [50], [51] utilizing the Bayesian and dynamic methods to analyze the correlations of the datasets which share one or more records of some user or his correlated users. However, they are not appropriate for our case, in which the two datasets correspond to different users sharing one or more of the same records.

Here, we define correlated individual differential privacy leakage to measure correlated privacy preservation of an ITD as follows:

*Definition 11 (Correlated Individual Differential Privacy Leakage): Suppose that $D = D_1, D_2, \ldots, D_n$ is an ITD, in any subdataset of which all of the records are about the same user. Different subdatasets may include one or more of the same items. Datasets $D_i^1$ and $D_i^2$ $(1 \leq i \leq n)$ generated from $D_i$ differ by at most one record. A correlated privacy mechanism $M$ is a randomized function on D, and the range is S. The correlated individual differential privacy leakage of $M_{i \in n}$ is*

$$\mathrm{CIDP}_{\mathcal{A}_i}(M) = \sup_{D_i^1,D_i^2,S,D_{-i}} \log \frac{\Pr\left[M(D_i^1) \in S|D_{-i}\right]}{\Pr\left[M(D_i^2) \in S|D_{-i}\right]}, \tag{8}$$

*in which $D_{-i} = D \setminus \{i\}$.*

For $\forall D_i$, $\mathrm{CIDP}_{\mathcal{A}_i}(M)$ measures the privacy leakage of the correlated individual datasets $D_{-i}$, since removing or adding at most one item in an objective individual substantially leads to the privacy leakage of correlated individuals.

As a result, a privacy mechanism $M$ yields $\epsilon$-correlated differential privacy if and only if

$$|\mathrm{CIDP}_{\mathcal{A}_i}(M)| \leq \epsilon. \tag{9}$$

### D. ANALYTIC REQUIREMENT

Since we aim at privacy-preserved ITD publishing, which allows the adoption of various forms of trajectory data mining, especially the analytics for individuals, we perform a general trajectory mining tasks, i.e., frequent sequential pattern mining, and measure the preserved framework utility. The set of *Top-K* frequent sequential patterns is a general measure of both trajectory anonymity and differential privacy methodology [13], [29].

*Definition 12 (Top-K Frequent Sequential Patterns): Given a positive number $K$, the set of* Top-K *frequent sequential patterns on the raw dataset $D$ and sanitized dataset $\tilde{D}$ are denoted as $F_K(D)$ and $F_K(\tilde{D})$; also, individual Top-K frequent sequential patterns are denoted as $F_K(D_i)$ and $F_K(\tilde{D_i})$ for $\forall D_i \in D$.*

We measure preserved utility in terms of true positive and false positive by the ratio of true positive and $K$, since $|F_K(D)| = |F_K(\tilde{D})| = K$. True positive is the number of frequent sequential patterns in $F_K(D)$ that are correctly identified in $F_K(\tilde{D})$, i.e., $|F_K(D) \cap F_K(\tilde{D})|$. False positive is the number of infrequent sequential patterns in D that are mistakenly included in $F_K(\tilde{D})$.

### E. PROBLEM STATEMENT

For satisfying the privacy requirement and analytic requirement, we propose IDF-OPT. We consider the least-cost sensitivity of each individual; we set noninteractive curator publication; we design a risk function and utility function to measure privacy preserved level and utility preserved level, proposing the Individual DF-optimization algorithm to obtain the Pareto optimization solution; we propose the correlated individuals differential privacy leakage model, measuring the correlated individuals differential privacy leakage; and design sanitization algorithm adding noises for statistical ITD publishing.

## IV. METHODOLOGY

In this section, we demonstrate the detail of proposed IDF-OPT. Firstly, we introduce the sketch of our solution. We then describe the details of our methodology, including definitions, algorithms, and the example.

### A. SKETCH OF IDF-OPT

Before delineating the details IDF-OPT, a sketch is presented. A system architecture of ITD-OPT is obviously represented in Fig.2; as a noninteractive curator publication, it is a series of operations based on a group of theorems and algorithms. It decides a group of Pareto privacy parameters $\epsilon_i^*$ and publishes the sanitization data with high utility and low privacy risk for data analytics.

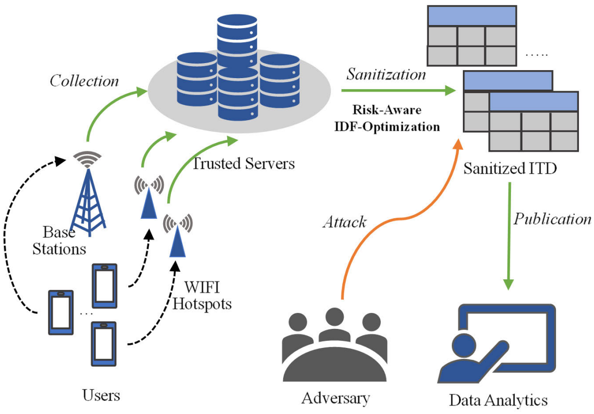


**FIGURE 2.** A system architecture of risk-aware ITD-OPT.

As shown in Fig.3, it is the sketch of IDF-OPT, consists of the following important stages:

1) **Calculation**
Calculate the individual privacy risk for each $u_i \in U$, $i = 1, \ldots, n$, and for each $T_{ik} \in T_i$.
2) **Suppression**
If the trajectory privacy risk is greater than the threshold $1/k$, add it into the set of risk trajectories denoted as $R_i$, and label the individual riskiest trajectory $T_{risk-i}$ in each round. For each individual in each round, suppress $T_{risk-i}$ until the reidentification probability of the trajectory is under the threshold, then return $\Delta f_i$.

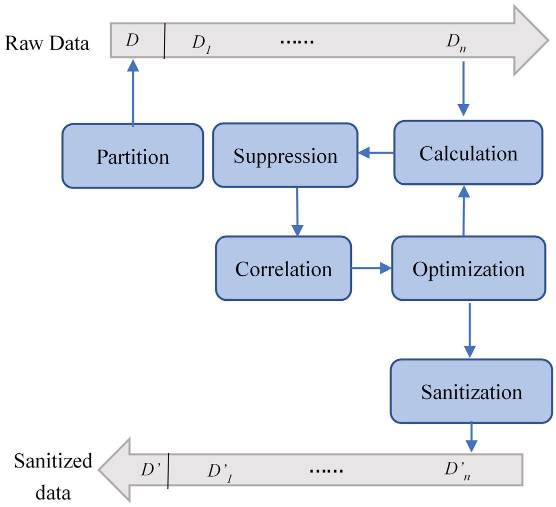


**FIGURE 3.** Sketch of IDF-OPT.

3) **Individual Correlation Computation**
For given $D = \{D_1, D_2, \ldots, D_n\}$, whose subdatasets are labeled with $T_{risk-i}$, search the risk-aware correlated individuals and compute the correlated individual differential privacy leakage CIDP$_{\mathcal{A}_{i\in n}}(M)$.
4) **Individual DF-optimization**
By giving a group an initial $\epsilon_i$, substitute each $\epsilon_i$ and CIDP$_{\mathcal{A}_i}(M)$ into the objective functions, i.e., risk function and utility function. Generate a set of Pareto efficient parameters $\epsilon_i^*$ via the individual DF-optimization algorithm. Return to step **II**, then label the next riskiest trajectory in $R_i$ as $T_{risk-i}$ until all of the trajectories in $R_i$ are under the threshold
5) **Sanitization**
Sanitize $D_i$ for each $u_i \in U$ by adding Laplace noises which are drawn from the Laplace distribution with Pareto efficient parameters $\epsilon_i^*$.

### B. INDIVIDUAL RISK FUNCTION AND UTILITY FUNCTION

For the purpose of meeting the privacy requirement and the utility requirement, we define the individual risk function and utility function in this section to measure the privacy preserved level and utility level.

For $\forall D_i \in D$, the designed risk function $r : D_i \times \epsilon_i \rightarrow \mathbb{R}$ is a measurement of the individual privacy risk of a sanitized ITD, which needs to return a risk value under a specific differential privacy mechanism for each $\epsilon_i$. The value of risk coefficient $\epsilon_i$ identifies the risk variation of each individual under the specific privacy preservation mechanism. Interpretively, the risk coefficient $\epsilon_i$ is proportional to privacy budget $\varepsilon_i$ and inversely proportional to sensitivity $\Delta f_i$. Higher $\epsilon_i$ leads to an elevated upper bound, leading to more information leakage, which is accompanied by higher risk to the individual. The sensitivity $\Delta f_i$ is the suppression value to $TC_{risk_i}$: normally, it is the lower bound to achieve the reidentification threshold. Therefore, we designed risk function as shown.

*Definition 13 (Individual Risk Function): The individual risk function indicates the individual privacy-preserved level of the mechanism $M_i$, which is given as follows:*

$$R(D_i, M_i) = v_i \cdot \epsilon_i, \tag{10}$$

*where $v_i$ is the privacy value of $u_i$ and $v_i > 0$: it is a constant derived from $D_i$.*

For every individual, the risk value of privacy before the mechanism of differential privacy is $v_i$, i.e., $R(D_i) = v_i$. Under $\epsilon_i$-differential privacy, the risk value should be less then $v_i$, which gives a privacy level constraint for each individual:

$$0 < \epsilon_i \leq 1 \tag{11}$$

As discussed in Section III, the individuals are risk-aware correlated, and we also define the correlated individual differential privacy leakage to measure the correlated privacy preservation of multiple individuals. Considering the correlated privacy leakage under mechanism $M_i$ and formula 10, the individual risk function is given as shown:

$$\mathrm{R}(u_i) = v_i \cdot CIDP_{\mathcal{A}_i}(M_i) \tag{12}$$

Therefore, the risk variation of each individual under the specific privacy preserved mechanism is indicated by the differential privacy protection level of a group of correlated individuals. Furthermore, the constraint of correlated individual differential privacy leakage is:

$$\epsilon_i \leq CIDP_{\mathcal{A}_i}(M_i) \leq 1 \tag{13}$$

The trade-off problem between privacy and utility is an inherent problem of the privacy preserved publication framework. To preserve the utility achieving the analytic requirement, existing research designs the utility function to measure utility. Under the Laplace mechanism, the utility is measured by noise, and we therefore design the noise function to minimize the magnitude of the global noise.

*Definition 14 (Noise Function): The noise function, indicating the magnitude of the global noise of an ITD, measures the utility of the data for analytics, which is given as follows:*

$$N(M_i) = \sum_{i=1}^{n} noise_i; \tag{14}$$

$$noise_i = \lambda_i = \frac{\Delta f_i}{\varepsilon_i}. \tag{15}$$

In formula 14, the scale parameter $\lambda_i$ determines the distribution curve of each respective individual in the mechanism $M$: a large $\lambda_i$ flattens the curve and leads to noise of large magnitude.

## C. INDIVIDUAL CORRELATION LEAKAGE MODEL

In this section, we propose an individual correlation leakage model which describes the correlated differential privacy leakage between any subset $D_i$ of an ITD and all the neighbors of $D_i$, i.e., $D_{-i}$. $D_i$ and $D_{-i}$ are indirectly correlated with each other, since they have the same records of different users. This is similar, but different from other indirect correlations. Therefore, we define the privacy leakage of correlated individuals for the ITD scenario in Definition 11. Then, we design the individual correlation leakage model to describe in detail the complex and dynamic influences of correlated individual privacy leakage in an ITD; this is a weighted sum of the privacy leakage of $D_i$. It is improved by the general models describing the dataset correlation, such as Bayesian Differential Privacy (BDPL).

**TABLE 2.** Example of suppression on ITD and risk-aware correlations.

| $u_i$ | $T_{ik}$ | Trajectory | $TC_{ik}$ | Label | Correlation |
|---|---|---|---|---|---|
| $u_1$ | $T_{11}$ | 1 → 3 → 5 → 7 | 2 | | |
| $u_1$ | $T_{12}$ | 1 → 3 → 5 | 1-1 | $T_{risk-1}$ | strong with 4 |
| $u_1$ | $T_{13}$ | 5 → 7 | 1 | | |
| $u_1$ | $T_{14}$ | 3 → 5 → 7 | 1 | | |
| $u_1$ | $T_{15}$ | 2 → 4 → 6 | 1 | | |
| $u_2$ | $T_{21}$ | 1 → 3 → 5 → 7 | 2 | | |
| $u_2$ | $T_{22}$ | 3 → 5 → 7 | 2-1 | $T_{risk-2}$ | weak with 1 |
| $u_2$ | $T_{23}$ | 5 → 7 | 1 | | |
| $u_2$ | $T_{24}$ | 1 → 3 | 1 | | |
| $u_3$ | $T_{31}$ | 2 → 4 → 6 | 1 | | |
| $u_3$ | $T_{32}$ | 1 → 3 | 2-1 | $T_{risk-3}$ | weak with 2 |
| $u_3$ | $T_{33}$ | 5 → 7 | 2 | | |
| $u_4$ | $T_{41}$ | 2 → 4 → 6 | 1 | | |
| $u_4$ | $T_{42}$ | 1 → 3 → 5 | 1-1 | $T_{risk-4}$ | strong with 1 |
| $u_4$ | $T_{43}$ | 1 → 3 → 5 → 7 | 1 | | |

*Note: in Table 2, $u_i$ represents U-id in Table 1, k is the $k^{th}$ different trajectory of each individual, $T_{ik}$ represents the instance of the $k^{th}$ trajectory and $TC_{ik}$ is the count of $T_{ik}$.*

Bayesian differential privacy provides a general model to describe the correlation of datasets. Let $G(D, L)$ be a Gaussian correlation model, in which $L$ is a Laplace matrix of $G(D, L)$, i.e.,

$$L = \begin{bmatrix} w_1 & -w_{12} & \dots & -w_{1n} \\ -w_{12} & w_2 & \dots & -w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -w_{1n} & -w_{2n} & \dots & w_n \end{bmatrix}. \tag{16}$$

We present in Table 2 results obtained from analyzing four users' strong and weak correlations in the ITD shown in Table 1, where $TC_{ik}$ is trajectory count and the Label represents the marked individual riskiest trajectory, i.e., $T_{risk-i}$, where trajectory suppression has been marked in red. According to the individual riskiest trajectory marked by each user, its strong and weak risk-aware correlated individuals can be found and recorded as Correlation. By analyzing the differential privacy definition, individuals who suppress the same trajectory demonstrate strong correlation with each other, such as $u_1$ marked as *strong with 4*; an individual's suppressed trajectory is another user's trajectory, demonstrating that the user has a weak correlation with another user, such as $u_2$ marked as *weak with 1*.

We represent in Fig.4(a) the above individual user correlation description. With a bilateral influence, the strong correlation is marked as $S$, and the correlation coefficient is 1; the weak correlation is marked as $W$, its correlation has a unilateral influence, and the correlation coefficient is

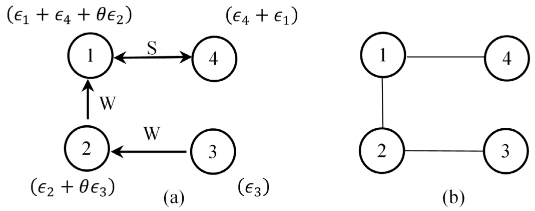

$(\epsilon_1+\epsilon_4+\theta\epsilon_2)$
$(\epsilon_4+\epsilon_1)$
1
S
4
W
W
2
3
$(\epsilon_2+\theta\epsilon_3)$
(a)
$(\epsilon_3)$
1
4
2
3
(b)

**FIGURE 4. Description of individual correlations.**

$\theta$ $(0 < \theta < 1/(n-1))$; the correlation depicted by BDLP is shown in Fig.4(b).

Although many general models such as $\mathrm{BDPL}_{\mathcal{A}_i}(M)$ provide a general standard to measure the differential privacy leakage for the correlated datasets, unfortunately, a more delicate describing and dynamic standard is required to measure the complex correlated data for the following reasons:

- First, $\mathrm{BDPL}_{\mathcal{A}_i}(M)$ is the upper bound of the privacy leakage, because not all members in $D_0$ are the correlated neighbors of $D_i$. The real value of correlation leakage should be less than the upper bound due to eliminating the nonneighbor's effect. It is easy to reach this conclusion and prove Proposition 1.

  *Proposition 1: For any differential privacy mechanism M, the measurement of correlated individual differential leakage is less than Bayesian differential privacy leakage.*

$$\mathrm{CIDP}_{\mathcal{A}_i}(M) \leq \mathrm{BDPL}_{\mathcal{A}_i}(M) \tag{17}$$

- Second, the correlation measurements should be dynamic values rather than static values. The different neighbors' privacy protection levels cause the variation of the values. The value of $\mathrm{CIDP}_{\mathcal{A}_i}(M)$ is decided by a group of variables, i.e., the correlated neighbors' and its own privacy parameters.
- Third, in the Gaussian correlation model, the value of weight $\forall w_{ij} \in L$ which represents the correlation between the tuples *i* and *j* is either 0 or 1. Therefore, they can only describe the adjacent correlation, but not the weighted correlation. It is necessary to design a correlation model in which the large correlation $w_{ij}$ means that $D_j$ has a large effect on the mean of $D_0$, where $D_j \in D_0$, $D_j$ and $D_i$ are correlated.
- Finally, the Gaussian correlation model treats the correlations as a nondirected graph. However, in some special cases, the effect on the risk of two individuals on an edge is not always two-way.

For the reasons above, we proposed an individual correlation leakage model based on the Bayesian differential privacy leakage.

*Definition 15 (Individual Correlation Leakage Model): The individual correlation leakage model is a general model used to describe the individual correlation in an ITD*

$$\mathrm{CIDP}_{\mathcal{A}_{i\in n}}(M) = L \times R, \tag{18}$$

*in which R is an individual weighted correlation matrix and L is a Laplace matrix, which is dynamically decided by adjacent matrix $L = (w_{ij})$ and the individual weighted correlation matrix R, which provides a scale on the L.*

*Definition 16 (Individual Weighted Correlation Matrix): Let individual weighted correlation matrix R be an $n \times n$ matrix which describes the scale of influence on L. Each row represents the protection level of a specific individual dataset, and each column represents the directed influence of neighbors to each individual.*

$$R = \begin{bmatrix} \epsilon_{11} & -\epsilon_{12} & \cdots & -\epsilon_{1n} \\ -\epsilon_{12} & \epsilon_{22} & \cdots & -\epsilon_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -\epsilon_{1n} & -\epsilon_{2n} & \cdots & \epsilon_{nn} \end{bmatrix} \tag{19}$$

*Note: the first index of element $\epsilon_{ij}$ is the differential privacy level of $D_i$, and the second index is the weighted directed influence on $D_j$.*

We describe a group setting for some specific situations.

- For the one-way edge in $L$ which gives the directed influence, the other side will be set to 0.
- According to the vertex degree, the scale will be repetitively computed $w_i$ times. The scale of $diag(R)$ should be divided by $diag(L)$.
- A set of weight $\theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ on an edge represents the coefficients on the scale. For $\forall\theta_j \in \theta$, it represents the same level of influence on the scale.

Obviously, our method provides a more delicate description for the complex correlated data and is appropriate for describing the risk-aware correlations drawn from the application scenario. According to the Definition 15, 16, for the correlation described in Figure 4(a), the CIDP is described as follows:

$$L = \begin{bmatrix} 2 & -1 & 0 & -1 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}$$

$$R = \begin{bmatrix} \frac{\epsilon_1}{w_1} & 0 & -\epsilon_1 & -\epsilon_1 \\ -\theta\epsilon_2 & \frac{\epsilon_2}{w_2} & 0 & -\epsilon_2 \\ -\epsilon_3 & -\theta\epsilon_3 & \frac{\epsilon_3}{w_3} & -\epsilon_3 \\ -\epsilon_4 & -\epsilon_4 & -\epsilon_4 & \frac{\epsilon_4}{w_4} \end{bmatrix} \tag{20}$$

We can calculate correlated individual differential privacy leakage by

$$CIDP_{\mathcal{A}_{i\in n}}(M) = \sum_{j=1}^{n} w_{ij} \times \epsilon_{ji}. \tag{21}$$

Algorithm 1 presents the details of searching the risk-aware correlated individuals for a specific individual and computing the correlated individual differential privacy leakage $\mathrm{CIDP}_{\mathcal{A}_{i\in n}}(M)$. For each individual in an ITD, its candidate set $U_{-i}$ is generated. For every member in $U_{-i}$, if they share the

**ALGORITHM 1** Individual Correlation Leakage Model

```
Input: Individual trajectory dataset $D = \{D_1, D_2, \ldots, D_n\}$
       with labeled $T_{risk-i}$, weak correlation coefficient $\theta$,
       $i \in U$ set a group of initial privacy parameters $\epsilon_i$
Output: CIDP$_{\mathcal{A}_{i \in n}}(M)$
for $u_i \in U$ do
    Generate a candidate set $U_{-i}$,
    $w_i = 0; w_j = 0$
    for $\forall j = -i$ do
        if $T_{risk-i} = T_{risk-j}$;     // specifies that $u_i$
         and $u_j$ are strongly correlated with
         each other
        then
            $w_{ij} = w_{ji} = -1; w_i = w_i + |w_{ij}|$
            $\epsilon_{ij} = -\epsilon_i; \epsilon_{ji} = -\epsilon_j$;
        else
            if $T_{risk-i} = T_{jk}, \forall T_{jk} \in T_j$ ;     // specifies
             that $u_i$ is weakly correlated with
             $u_j$
            then
                $w_{ij} = w_{ji} = -1$
                $w_i = w_i + |w_{ij}|; w_j = w_j + |w_{ji}|$
                $\epsilon_{ij} = -\theta\epsilon_i; \epsilon_{ji} = 0$;
            else
                $w_{ij} = w_{ji} = 0$
                $\epsilon_{ij} = \epsilon_i; \epsilon_{ji} = \epsilon_j$
            end
        end
    end
end
$\epsilon_{ii} = \frac{\epsilon_i}{w_i}$
CIDP$_{\mathcal{A}_{i \in n}}(M) = \sum_{j=1}^{n} w_{ij} \times \epsilon_{ji}$
Return CIDP$_{\mathcal{A}_{i \in n}}(M)$
```

same riskiest trajectory, then they are strongly related with each other; else, if the riskiest trajectory of $u_i$ is one of the trajectories $u_j$, then $u_i$ is weakly related with $u_j$, which means that there is an edge from $u_i$ to $u_j$; else, they are risk-aware independent.

### D. INDIVIDUAL DF-OPTIMIZATION ALGORITHM

For publishing the sanitized ITD, one of the most important stages is deciding appropriate optimal privacy parameters to strike a wonderful balance between privacy and utility. However, different from the traditional case, in the ITD the individual's subdataset is risk-aware correlated, as shown in the above sections. When we focus on the selection of the one individual privacy parameter, we need to consider the privacy parameter of the correlated neighbors, since someone's dataset privacy guarantee depends not only on his own privacy parameter; we also need to consider the effect of privacy parameters on the global utility. To ensure that the global individuals achieve privacy and utility optimization, we evolve the problem into a multiobjective optimization problem and design the individual DF-optimization algorithm to obtain a group of optimal parameters of correlated individuals.

The Pareto optimal is a well-suited multiobjective optimization solution to discuss this case. It is a state of global utility allocation in which it is impossible to reallocate the privacy parameters to make any one individual privacy guarantee better off without making global utility worse.

In Pareto optimization, the central concept is called the nondominated solution. This solution must satisfy the following two conditions: (i) there is no other solution that is superior, at least in one objective function; (ii) it is equal or superior with respect to other objective function values. Usually, the solution is not unique and consists of a set of acceptable optimal solutions (Pareto effective). From the point of view of measurement risks and utility, these criteria are incompatible and can be grouped into two different categories: objective functions and constraints (restrictions).

As a Pareto optimal multiobjective problem, we propose the following form:

**Objective Function**

$$\min_i \left[R(u_i), N(M)\right] \tag{22}$$

**Constraints**

$$0 < \epsilon_i \leq 1 \tag{23}$$

$$\epsilon_i < CIDP_{\mathcal{A}_i}(M) \leq 1 \tag{24}$$

We design an individual DF-optimization algorithm which sets initial privacy parameters $\epsilon_i$ as the input and computes CIDP$_{\mathcal{A}_{i \in n}}(M)$ many times by calling Algorithm 1, and returns a group of Pareto efficient parameters $\epsilon_i^*$; the details of the individual DF-optimization algorithm are shown in Algorithm 2.

In the individual DF-optimization algorithm, the initial privacy parameter is a very small value which results in substantial global noise. According to Proposition 2, it searches a group of Pareto efficient parameters $\epsilon_i^*$ effectively by giving the step parameter $\beta$. The step parameter $\beta$ is the varying granularity, since $\epsilon_i^*$ is a continuous variable in the interval (0, 1]. By adding step parameter $\beta$ and computing CIDP$_{\mathcal{A}_{i \in n}}(M)$, it reduces the global noise until the calculation CIDP$_{\mathcal{A}_{i \in n}}(M_i)$ exceeds the upper bound.

For our program design, we make the following compromises which require the participation of the decision maker: (i) The ultimate goal of a multiobjective optimization algorithm is to maximize the global utility under the privacy-protected mechanism, which suppresses the riskiest trajectory for each individual and ensures the risk to be under the threshold. Therefore, the mechanism provides the strong privacy guarantee by suppression, and the differential privacy ensures the indistinguishability. (ii) Based on the premise of maximizing the global utility, the algorithm makes a best effort to achieve the minimum indistinguishability: that is, the lowest indistinguishable upper bound of the raw dataset and noisy dataset.

The individual DF-optimization algorithm should improve the minimum global noise; at the same noise level, the algorithm should improve to the minimum risk of each individual until CIDP$_{\mathcal{A}_{i \in n}}(M_i)$ exceeds the upper bound.

By following the above compromises, we can ignore many solutions to the multiobjective optimization problem and

**ALGORITHM 2** Individual DF-Optimization Algorithm

```
Input: ITD with $T_{risk-i}$, $u_i \in U$ set initial privacy parameters
       $\epsilon_i$, Set step parameter $\beta$
Output: Pareto efficient parameters $\epsilon_i^*$
for $u_i \in U$ do
    Computing CIDP$_{\mathcal{A}_{i\in n}}(M)$ with initial $\epsilon_i$
    $\epsilon_i^* = \epsilon_i$
end
for $u_i \in U$ do
    if CIDP$_{\mathcal{A}_i}(M_i) == 0$ or CIDP$_{\mathcal{A}_i}(M_i) == \epsilon_i$ then
        Continue
    else
        if CIDP$_{\mathcal{A}_i}(M_i) > \epsilon_i^*$ and CIDP$_{\mathcal{A}_i}(M_i) \leq 1$ then
            do
                $\epsilon_i^* = \epsilon_i^* + \beta$
                $\epsilon_j^* = \epsilon_j^* + \beta$ ;    // $u_j$ is correlated
                 with $u_i, \forall j \in -i$
                Computing CIDP$_{\mathcal{A}_{i\in n}}(M_i)$
            while CIDP$_{\mathcal{A}_i}(M_i) \leq 1$;
        else
            if CIDP$_{\mathcal{A}_i}(M_i) > 1$ then
                do
                    $\epsilon_i^* = \epsilon_i^* - \beta$
                    $\epsilon_j^* = \epsilon_j^* - \beta$ ;           // $u_j$ is
                     correlated with $u_i, \forall j \in -i$
                    Computing CIDP$_{\mathcal{A}_{i\in n}}(M_i)$
                while CIDP$_{\mathcal{A}_i}(M_i) > 1$;
            end
        end
    end
end
Return $\epsilon_{i\in n}^*$
```

involve the exact method in the decision process. Analyzing the noise function $N(M) = \sum_{i=1}^{n} noise_i = \sum_{i=1}^{n} 1/\epsilon_i$, we can obtain Proposition 2.

*Proposition 2: To achieve the maximum global utility, that is the minimum sum of the noise of individuals, the best solution is to divide equally given a certain amount of privacy budget.*

*Letting:* $\sum_{i=1}^{n} \epsilon_i = \alpha$, $\alpha > 0$, *we have* $\epsilon^* = min \sum_{i=1}^{n} 1/\epsilon_i$, $\epsilon_1^* = \epsilon_2^* = \cdots = \epsilon_n^* = \alpha/n$.

### E. SANITIZATION ALGORITHM

For the purpose of publishing sanitized data, we design the sanitization algorithm combining the four steps of IDF-OPT. The overview of the sanitization algorithm is given in Algorithm 3. For a given raw individual trajectory dataset ITD with $TC_{risk-i}$, a group of Pareto efficient parameters $\epsilon_i^*$, and risk threshold 1 / $k$, it returns a sanitized dataset ITD.

The process of sanitization is improved by several rounds, each of which suppresses the riskiest trajectory for an individual. The number of rounds is decided by the number of elements in the risk trajectory set. In the first round, the process deals with the first riskiest trajectory for individuals; for the second round, it deals with the second riskiest trajectory; this continues all the way to the end. In the sanitization process, we add noise to the counts of the trajectories for each individual in each round using the training Pareto optimal parameter, since the discriminative pairs are counts of trajectories in raw and sanitized ITD and make them indistinguishable. We can thus control the overall noise effectively and obtain the strong guarantee for each individual.

**ALGORITHM 3** Sanitization Algorithm

```
Input: Pareto efficient parameter $\epsilon_i^*$, risk threshold $p$, ITD
       with $TC_{risk-i}$
Output: Sanitized ITD
for $u_i \in U$ do
    Computing $v_{ik} = \Pr_D(u_i | T_{ik})$ and $TC_{ik}$
    if $v_{ik} \geq p$ then
        Put $T_{ik}$ into $R_i$
    end
end
do
    for $u_i \in U$ do
        if $R_i \neq \varnothing$ then
            Labeling riskiest $T_{ik}$ as $T_{risk-i}$
        end
    end
    Computing $\epsilon_i^*$ using Algorithm 2
    $c = TC_{risk-i}$
    do
        $c = c - 1$
    while risk of $T_{risk-i}$ with count $c \leq p$;
    $\Delta f_i = TC_{risk-i} - c$
    for $\forall T_{risk-i} \in R_i$ do
        $NC_{ik} = TC_{ik} +$ Laplace$(\frac{\Delta f_i}{\epsilon_i^*})$
        Removing $T_{risk-i}$ from $R_i$
    end
while $\exists R_i \neq \varnothing$; $R_i \in \{R\}_n$;
Return Sanitized ITD
```

The continued example is shown in Tables 3-4: we set the risk threshold $p = 1/k = 0.5$ for $k = 2$. In the first round, as shown in Table 3, we suppress 1 $\rightarrow$ 3 $\rightarrow$ 5 for $u_1$; 3 $\rightarrow$ 5 $\rightarrow$ 7 for $u_2$; 1 $\rightarrow$ 3 for $u_3$ and 1 $\rightarrow$ 3 $\rightarrow$ 5 for $u_4$. Each one suppresses the true counts of the riskiest trajectory to keep the risk below the threshold; then, for $u_1$, $u_2$ and $u_4$, all of the trajectories are under the threshold. After suppression, we add Laplace noise drawn from $Lap(1/\epsilon_i^*)$ to the counts of the trajectories, and the parameters denoted as $\epsilon_i^*$ are obtained from Algorithm 2, i.e., the individual DF-optimization algorithm. However, for $u_3$, the risk value of the trajectory 5 $\rightarrow$ 7 is still higher than the threshold. In the second round, as shown in Table 4, we suppress the true count of riskiest trajectory 5 $\rightarrow$ 7 for user 3, then add $Lap(1/\epsilon_3^{**})$ to the true counts of $u_3$' trajectories.

## V. EXPERIMENT EVALUATION

In this section, we examine the performance of IDF-OPT in terms of individual information loss and global utility measure. To meet the analytic requirement, we perform frequent sequential pattern mining, i.e., *Top-K*, for both individuals and the global users to measure the individual information loss and global utility decline due to the suppression and noisy operation. Comprehensive experiments based on actual trajectory publishing benchmarks are comparable with

**TABLE 3. Continuous example of suppression and sanitization ITD (first round).**

| $u_i$ | $T_{ik}$ | Trajectory | $TC_{ik}$ | $v_{ik}$ | Label | Correlation | $NC_{ik}$ |
|---|---|---|---|---|---|---|---|
| $u_1$ | $T_{11}$ | $1 \to 3 \to 5 \to 7$ | 2 | 2/5 | | | $2 + Lap(1/\epsilon_1^*)$ |
| $u_1$ | $T_{12}$ | $1 \to 3 \to 5$ | 1-1 | 1/2 | $T_{risk-1}$ | strong with 4 | $0 + Lap(1/\epsilon_1^*)$ |
| $u_1$ | $T_{13}$ | $5 \to 7$ | 1 | 1/4 | | | $1 + Lap(1/\epsilon_1^*)$ |
| $u_1$ | $T_{14}$ | $3 \to 5 \to 7$ | 1 | 1/3 | | | $1 + Lap(1/\epsilon_1^*)$ |
| $u_1$ | $T_{15}$ | $2 \to 4 \to 6$ | 1 | 1/3 | | | $1 + Lap(1/\epsilon_1^*)$ |
| $u_2$ | $T_{21}$ | $1 \to 3 \to 5 \to 7$ | 2 | 2/5 | | | $2 + Lap(1/\epsilon_2^*)$ |
| $u_2$ | $T_{22}$ | $3 \to 5 \to 7$ | 2-1 | 2/3 | $T_{risk-2}$ | weak with 1 | $1 + Lap(1/\epsilon_2^*)$ |
| $u_2$ | $T_{23}$ | $5 \to 7$ | 1 | 1/4 | | | $1 + Lap(1/\epsilon_2^*)$ |
| $u_2$ | $T_{24}$ | $1 \to 3$ | 1 | 1/3 | | | $1 + Lap(1/\epsilon_2^*)$ |
| $u_3$ | $T_{31}$ | $2 \to 4 \to 6$ | 1 | 1/3 | | | $1 + Lap(1/\epsilon_3^*)$ |
| $u_3$ | $T_{32}$ | $1 \to 3$ | 2-1 | 2/3 | $T_{risk-3}$ | weak with 2 | $1 + Lap(1/\epsilon_3^*)$ |
| $u_3$ | $T_{33}$ | $5 \to 7$ | 2 | 2/4 | | | $1 + Lap(1/\epsilon_3^*)$ |
| $u_4$ | $T_{41}$ | $2 \to 4 \to 6$ | 1 | 1/3 | | | $1 + Lap(1/\epsilon_4^*)$ |
| $u_4$ | $T_{42}$ | $1 \to 3 \to 5$ | 1-1 | 1/2 | $T_{risk-4}$ | strong with 1 | $0 + Lap(1/\epsilon_4^*)$ |
| $u_4$ | $T_{43}$ | $1 \to 3 \to 5 \to 7$ | 1 | 1/5 | | | $1 + Lap(1/\epsilon_4^*)$ |

**TABLE 4. Continuous example of suppression and sanitization ITD (Second Round).**

| $u_i$ | $T_{ik}$ | Trajectory | $TC_{ik}$ | $v_{ik}$ | Label | Correlation | $NC_{ik}$ |
|---|---|---|---|---|---|---|---|
| $u_1$ | $T_{11}$ | $1 \to 3 \to 5 \to 7$ | 2 | 2/5 | | | |
| $u_1$ | $T_{12}$ | $1 \to 3 \to 5$ | 1 | 1/2 | | | |
| $u_1$ | $T_{13}$ | $5 \to 7$ | 1 | 1/4 | | | |
| $u_1$ | $T_{14}$ | $3 \to 5 \to 7$ | 1 | 1/3 | | | |
| $u_1$ | $T_{15}$ | $2 \to 4 \to 6$ | 1 | 1/3 | | | |
| $u_2$ | $T_{21}$ | $1 \to 3 \to 5 \to 7$ | 2 | 2/5 | | | |
| $u_2$ | $T_{22}$ | $3 \to 5 \to 7$ | 1 | 2/3 | | | |
| $u_2$ | $T_{23}$ | $5 \to 7$ | 1 | 1/4 | | | |
| $u_2$ | $T_{24}$ | $1 \to 3$ | 1 | 1/3 | | | |
| $u_3$ | $T_{31}$ | $2 \to 4 \to 6$ | 1 | 1/3 | | | $1 + Lap(1/\epsilon_3^*) + Lap(1/\epsilon_3^{**})$ |
| $u_3$ | $T_{32}$ | $1 \to 3$ | 2 | 2/3 | | | $1 + Lap(1/\epsilon_3^*) + Lap(1/\epsilon_3^{**})$ |
| $u_3$ | $T_{33}$ | $5 \to 7$ | 2-1 | 2/4 | $T_{risk-3}$ | weak with 1,2 | $1 + Lap(1/\epsilon_3^*) + Lap(1/\epsilon_3^{**})$ |
| $u_4$ | $T_{41}$ | $2 \to 4 \to 6$ | 1 | 1/3 | | | |
| $u_4$ | $T_{42}$ | $1 \to 3 \to 5$ | 1 | 1/2 | | | |
| $u_4$ | $T_{43}$ | $1 \to 3 \to 5 \to 7$ | 1 | 1/5 | | | |

*Note: in Tables 3 and 4, $v_{ik}$ represents risk value of the trajectory, and the noisy count of a trajectory is denoted as $NC_{ik}$.*

previous works with respect to both method of anonymization and DP.

Three groups of experiments are performed to verify the validity of IDF-OPT, two of which study the comparison of changes of individual information loss extent between classical methods and our methods when different parameters change, while the other studies the comparison of global utility between different methods. We verify the loss of personal information and the change of utility caused by the change of two parameters, risk threshold $p$ value and $K$ most frequent pattern, and compare the results with the classical algorithms.

First, since the sensitivity of the differential privacy mechanism we designed depends on the suppression level of an individual's riskiest trajectory, noise is added to trajectories to satisfy differential privacy. Therefore, by setting the same anonymization threshold, the classical algorithms will retain the same privacy preserved level as our method, and the experimental results show that our method achieves reduced information loss for individuals. In addition to the classical algorithm $k$-anonymity, $(K, C)_L$ is also known as local suppression. It achieves a tailored privacy model for trajectory data anonymization; in comparison with the previous works in the literature, the proposed local suppression method can significantly improve the data utility in the case of anonymous trajectory data. Therefore, we compare individual information loss changes with risk threshold $p$ values between the classical anonymity mechanism algorithm $k$-anonymity, $(K, C)_L$ [13], and our approach at the same suppression level to study the impact of suppression level on individual information loss. Since the change of suppression level is not a significant parameter of influence of the classical DP algorithms, we do not compare it with the classical DP algorithms when the risk threshold $p$ changes. We also did not examine the effect of the change in the risk threshold $p$ in the global case.

Second, we study how individual and global utility change with $K$ frequent items, and we compare our method with the classical methods $(K, C)_L$, n-gram [30] and DPT [31]. n-gram and DPT are considered to be the relatively most advanced trajectory publishing technologies based on differential privacy. They reconstruct trajectories by defining a reasonable hierarchical structure and adding noise to frequent prefixes or n-grams, which effectively reduces the output domain, providing high practical value for frequent pattern mining in the global domain. We compare our method with $(K, C)_L$, DPT and n-gram with respect to the individual and global utility decline by setting the general privacy criterion measures for the DP mechanism and anonymity and study the effects of varying $K$ upon the methods.

### A. D4D DATASET

Since none of the previous works can preserve the user dimension for data analysis, such as individuals' frequent patterns, we cannot directly compare our method with them using the same trajectory dataset. Therefore, we carry out the experiments using the spatiotemporal data, including users' dimension. The D4D-Senegal challenge includes open innovation data from Orange's mobile phone users in Senegal. D4D contains 300,000 randomly selected users' trajectories at the site level for one year on a rolling 2-week basis. We generate 3,000 users and capture the daily trajectories for each individual during two weeks in January 2013. Because the area of users' activities is too large in D4D, such that one group of users has no intersection of activities with other groups, it is meaningless for hiding users and reducing risks. Therefore, the rule of capturing is that the users must be active in the same area. We randomly select 300 users from the capturing dataset. We also preserve the users' dimension for further analytics: therefore, this is referred to as ITD in our work.

### B. INDIVIDUAL INFORMATION LOSS

To verify the efficiency of the proposed method with respect to personal information loss, we first evaluate in terms of individual information loss by varying the risk threshold $p$ value for the anonymization threshold $1/k$ for both $k$-anonymity and $(K, C)_L$. According to the scale of the dataset, the selection of parameters for the contrast method $(K, C)_L$ is $L = 3$, $C = 60\%$. The following parameters are used for all users, respectively: $p$ from 0 to 1 for $K = 5$ and $K = 10$ performed on classical $k$-anonymity, $(K, C)_L$ and IDF-OPT. During the 2-week observation, each individual generates no more than 20 trajectories per day, and therefore $1/p$ is set from 1 to 5.

We randomly select 4 users, and the experimental results are shown in Fig.5 and Fig.6. As shown in the results, IDF-OPT performs significantly better than the existing suppression method at the same level of anonymity for every selected user. In particular, by calculation it achieves 51.8% and 59.3% improvement on average for 300 users and all $1/p$ values with $K = 5$ and $K = 10$ most frequent patterns, respectively.

To further verify the effectiveness of the proposed method, we compare the proposed method with $(K, C)_L$, DPT and n-gram for further study of the effectiveness of our method in preserving the $K$ most frequent patterns. According to the scale of the dataset, the selection of parameters for the contrast method is noisy prefix tree height $h = 5$, privacy budget $\epsilon = 1.0$ for n-gram and $n = 5$ for DPT.

We also randomly select 4 users, and the experimental results are shown in Fig.7. By varying $K$ and setting $p = 0.3$, we show the comparison between the classical $(K, C)_L$, DPT, n-gram and IDF-OPT for individuals: our method was able to retain more information for each individual.

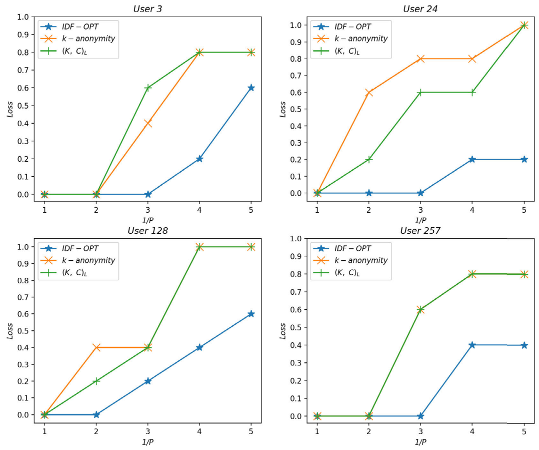


**FIGURE 5.** Individual information loss vs. $k$-anonymity and $(K, C)_L$ ($K = 5$).

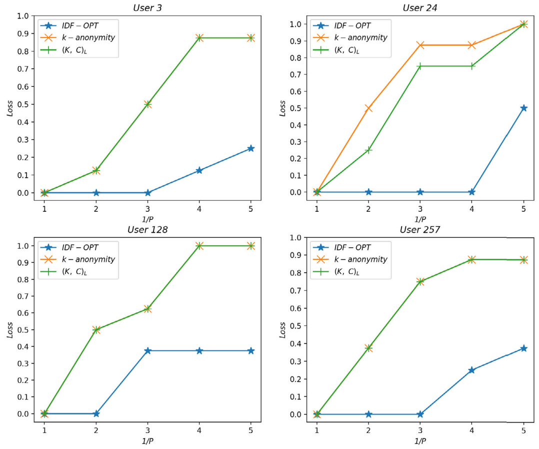


**FIGURE 6.** Individual information loss vs. $k$-anonymity and $(K, C)_L$ ($K = 10$).

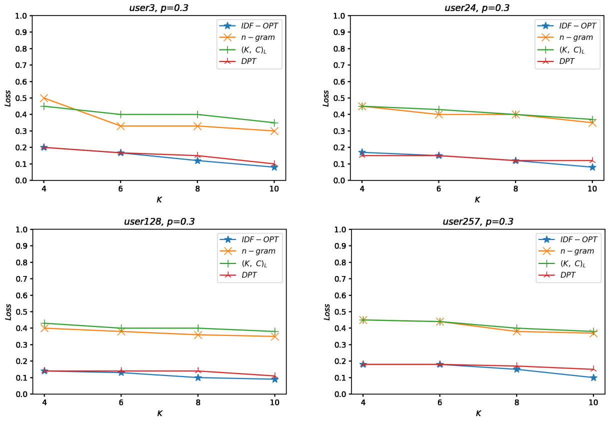


**FIGURE 7.** Individual information loss vs. $(K, C)_L$, DPT and n-gram ($p = 0.3$).

Through the analysis of the results, we believe that the local suppression mechanism will cause a large amount of information loss and reduce the data utility of individuals when it performs suppression operations on individual repeated trajectories. Compared with the existing differential privacy protection technology, the advantage of our method lies in considering the correlation problem of different users repeating the same trajectory and relaxing sensitivity to conserve privacy budget. In our work, the correlation between different users is dynamically measured through the individual correlation model. The multiobjective optimization model is designed to solve for the optimal privacy protection parameters of different users, adding less noise to improve data utility.

### C. UTILITY MEASURE

In addition to individual information loss for specific individuals, the data utility is measured for global users by performing *Top-K* frequent pattern mining for raw ITD and sanitized ITD. We examine the impact of utility loss of global users on $K$ and compare the ratio of the true positive with classical methods. By varying $K$ and setting $p = 0.5$ and $p = 0.25$, the experiment shows the comparison between $(K, C)_L$, DPT, n-gram and IDF-OPT for global users. For the global users, there are more than 3,000 trajectories generated by 300 users during observation, and therefore $K$ is set to a larger value than those of individuals: the value of $K$ is varied from 5 to 35.

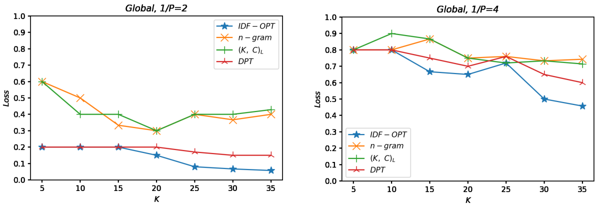


**FIGURE 8. Utility loss vs. $(K, C)_L$, DPT and n-gram ($p = 0.5, p = 0.25$).**

As shown in Fig.8, for both $p = 0.5$ and $p = 0.25$, IDF-OPT performs better than classical methods. The experiments also show that the utility loss is sensitive to the values of $K$ and $p$. When either $K$ or $p$ become larger, the utility loss is minimized.

Experimental results show that the proposed method not only maintains good performance in terms of individual information loss but also in terms of global data utility. This is because the ultimate goal of a multiobjective optimization algorithm is to maximize the global utility under the privacy-protected mechanism, which suppresses the riskiest trajectory for each individual to decrease the risk to below the threshold. Therefore, the method provides an effective way to improve the minimum global noise.

## VI. CONCLUSION

In this article, we summarize the challenges of privacy preserved individual trajectory data publishing to solve the problems of the existing trajectory data publishing in both application and technology. We design the risk-aware IDF-optimization method to reduce the risk of personal privacy disclosure while retaining the statistical characteristics of data for data analysis and its application. In the framework, we define individual risk to quantify privacy and the measurement of correlated individual privacy risk. By extending the scope of research from the trajectory protection level to the level of individual privacy protection, we provide sufficient privacy protection for individuals. At last, we include the risk function and noise function in the Pareto optimization problem for achieving data utility optimization. In our work, we perform a large number of experiments based on the actual trajectory publishing case of D4D, demonstrating that this method maintains high practicability in the task of trajectory data mining and that its performance is better than those of the existing privacy protection methods.

## REFERENCES

[1] M. Lv, L. Chen, T. Chen, D. Zeng, and B. Cao, "Discovering individual movement patterns from cell-id trajectory data by exploiting handoff features," *Inf. Sci.*, vol. 474, pp. 18–32, Feb. 2019.

[2] S. Claude and I. W. Geoffrey, *Encyclopedia of Machine Learning and Data Mining*. Berlin, Germany: Springer, 2017, p. 1283.

[3] Y. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, pp. 1–29, 2015.

[4] K. Dong, B. Zhang, Y. Shen, Y. Zhu, and J. Yu, "GAT: A unified GPU-accelerated framework for processing batch trajectory queries," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 1, pp. 92–107, Jan. 2020.

[5] M. Samir, S. Sharafeddine, C. M. Assi, T. M. Nguyen, and A. Ghrayeb, "UAV trajectory planning for data collection from time-constrained IoT devices," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 34–46, Jan. 2020.

[6] B. Chaix, J. Méline, S. Duncan, C. Merrien, N. Karusisi, C. Perchoux, A. Lewin, K. Labadi, and Y. Kestens, "GPS tracking in neighborhood and health studies: A step forward for environmental exposure assessment, a step backward for causal inference?" *Health Place*, vol. 21, pp. 46–51, May 2013.

[7] Q. Lin, D. Zhang, K. Connelly, H. Ni, Z. Yu, and X. Zhou, "Disorientation detection by mining GPS trajectories for cognitively-impaired elders," *Pervas. Mobile Comput.*, vol. 19, pp. 71–85, May 2015.

[8] Y. Liu, Y. Zhao, L. Chen, J. Pei, and J. Han, "Mining frequent trajectory patterns for activity monitoring using radio frequency tag arrays," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 11, pp. 2138–2149, Nov. 2012.

[9] X. Song, Q. Zhang, Y. Sekimoto, and R. Shibasaki, "Prediction of human emergency behavior and their mobility following large-scale disaster," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2014, pp. 5–14.

[10] M. N. Burns, M. Begale, J. Duffecy, D. Gergle, C. J. Karr, E. Giangrande, and D. C. Mohr, "Harnessing context sensing to develop a mobile intervention for depression," *Med. Internet Res*, vol. 3, no. 3, pp. 55–61, 2019.

[11] S. Song, Y. Wang, and K. Chaudhuri, "Pufferfish privacy mechanisms for correlated data," in *Proc. ACM Int. Conf. Manage. Data*, May 2017, pp. 1291–1306.

[12] Y. Xiao and L. Xiong, "Protecting locations with differential privacy under temporal correlations," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1298–1309.

[13] 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, May 2013.

[14] C. Chen, Y. Luo, Q. Yu, and G. Hu, "Privacy-preserving trajectory data publication based on 3D-Grid partition," *Intell. Data Anal.*, vol. 23, no. 3, pp. 503–533, 2019.

[15] S. Li, H. Shen, and Y. Sang, "A survey of privacy-preserving techniques on trajectory data," in *Proc. Parallel Archit., Algorithms Program., PAAP*, 2019, pp. 461–476.

[16] Z. Ma, T. Zhang, X. Liu, X. Li, and K. Ren, "Real-time privacy-preserving data release over vehicle trajectory," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8091–8102, Aug. 2019.

[17] K. Gu, L. Yang, Y. Liu, and B. Yin, "Efficient trajectory data privacy protection scheme based on laplace's differential privacy," *Informatica*, vol. 42, no. 3, PP. 1–13, 2018.

[18] 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. Service Manage.*, vol. 16, no. 1, pp. 264–278, Mar. 2019.

[19] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "$l$-diversity: Privacy beyond $k$-anonymity," *ACM Trans. Knowl. Discovery from Data(TKDD)*, vol. 1, no. 1, pp. 1–13, 2007.

[20] X. Wang, Z. Zhang, Y. Luo, and Q. Yu, "Hierarchical interpolation point anonymity for trajectory privacy protection," *Intell. Data Anal.*, vol. 23, no. 6, pp. 1397–1419, Nov. 2019.

[21] C. Dwork, "Differential privacy," in *Proc. Int. Colloq. Automata, Lang. Program. (ICALP)*, 2006, pp. 1–12.

[22] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Conf. Theory Cryptogr. (TCC)*, 2006, pp. 265–284.

[23] E. Naghizade, L. Kulik, E. Tanin, and J. Bailey, "Privacy- and context-aware release of trajectory data," *ACM Trans. Spatial Algorithms Syst.*, vol. 6, no. 1, pp. 1–25, Feb. 2020.

[24] S. Ghane, L. Kulik, and K. Ramamohanarao, "TGM: A generative mechanism for publishing trajectories with differential privacy," *IEEE Internet Things J.*, vol. 7, no.4 , pp. 2611–2621, Apr. 2020.

[25] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," in *Proc. Conf. Comput. Commun. Secur., ACM SIGSAC*, vol. 2, no. 1, 2013, pp. 901–914.

[26] R. Shokri, G. Theodorakopoulos, C. Troncoso, J.-P. Hubaux, and J.-Y. L. Boudec, "Protecting location privacy: Optimal strategy against localization attacks," in *Proc. ACM Conf. Comput. Commun. Secur. CCS*, 2012, pp. 617–627.

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

[28] V. Bindschaedler and R. Shokri, "Synthesizing plausible privacy-preserving location traces," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2016, pp. 546–563.

[29] R. Chen, B. C. M. Fung, and B. C. Desai, "Differentially private trajectory data publication," 2011, *arXiv:1112.2020*. [Online]. Available: https://arxiv.org/abs/1112.2020

[30] R. Chen, G. Acs, and C. Castelluccia, "Differentially private sequential data publication via variable-length n-grams," in *Proc. ACM Conf. Comput. Commun. Secur. CCS*, 2012, pp. 638–649.

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

[32] D. X. Shao, K. F. Jiang, T. Kister, S. Bressan, and K.-L. Tan, "Publishing trajectory with differential privacy: A priori vs. A posteriori sampling mechanism," in *Proc. Int. Conf. Dataset Expert Syst. Appl. (DESA)*, 2013, pp. 357–365.

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

[34] 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.*, vol. 143, pp. 126–139, Oct. 2018.

[35] Y.-A. de Montjoye, Z. Smoreda, R. Trinquart, C. Ziemlicki, and V. D. Blondel, "D4D-senegal: The second mobile phone data for development challenge," 2014, *arXiv:1407.4885*. [Online]. Available: http://arxiv.org/abs/1407.4885

[36] D. Kifer and A. Machanavajjhala, "Pufferfish: A framework for mathematical privacy definitions," *ACM Trans. Database Syst.*, vol. 39, no. 1, pp. 1–36, Jan. 2014.

[37] J. W. Byun, A. Kamra, E. Bertino, and N. Li, *Efficient $k$-Anonymization Using Clustering Techniques*. Berlin, Germany: Springer, 2007, pp. 188–200.

[38] M. E. Nergiz, M. Atzori, and Y. Saygin, "Towards trajectory anonymization: A generalization-based approach," *Trans. Data Privacy*, vol. 2, no. 1, pp. 52–61, 2008.

[39] Y. Wang, M. Li, S. Luo, Y. Xin, H. Zhu, Y. Chen, G. Yang, and Y. Yang, "LRM: A location recombination mechanism for achieving trajectory $k$-anonymity privacy protection," *IEEE Access*, vol. 7, pp. 182886–182905, 2019.

[40] M. Gramaglia, M. Fiore, A. Tarable, and A. Banchs, "$k^{\tau}, \epsilon$-anonymity: Towards privacy-preserving publishing of spatiotemporal trajectory data," 2017, *arXiv:1701.02243*. [Online]. Available: https://arxiv.org/abs/1701.02243

[41] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: Improving geographical prediction with social and spatial proximity," in *Proc. 19th Int. Conf. World Wide Web - WWW*, 2010, pp. 61–70.

[42] H. Hu, J. Xu, S. T. On, J. Du, and J. K.-Y. Ng, "Privacy-aware location data publishing," *ACM Trans. Database Syst.*, vol. 35, no. 3, pp. 1–42, Jul. 2010.

[43] Y. Sei, H. Okumura, T. Takenouchi, and A. Ohsuga, "Anonymization of sensitive quasi-identifiers for l-Diversity and t-Closeness," *IEEE Trans. Depend. Sec. Comput.*, vol. 16, no. 4, pp. 580–593, Jul. 2019.

[44] R. Pellungrini, L. Pappalardo, F. Pratesi, and A. Monreale, "A data mining approach to assess privacy risk in human mobility data," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 3, pp. 31:1–31:27, 2018.

[45] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proc. Int. Conf. Manage. Data SIGMOD*, 2011, pp. 193–204.

[46] D. Lv and S. Zhu, "Achieving correlated differential privacy of big data publication," *Comput. Secur.*, vol. 82, pp. 184–195, May 2019.

[47] X. He, A. Machanavajjhala, and B. Ding, "Blowfish privacy: Tuning privacy-utility trade-offs using policies," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2014, pp. 1447–1458.

[48] R. Chen, B. C. M. Fung, P. S. Yu, and B. C. Desai, "Correlated network data publication via differential privacy," *VLDB J.*, vol. 23, no. 4, pp. 653–676, Aug. 2014.

[49] T. Zhu, P. Xiong, G. Li, and W. Zhou, "Correlated differential privacy: Hiding information in non-IID data set," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 2, pp. 229–242, Feb. 2015.

[50] B. Yang, I. Sato, and H. Nakagawa, "Bayesian differential privacy on correlated data," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, May 2015, pp. 747–762.

[51] Y. Cao, M. Yoshikawa, Y. Xiao, and L. Xiong, "Quantifying differential privacy under temporal correlations," in *Proc. IEEE 33rd Int. Conf. Data Eng. (ICDE)*, Apr. 2017, pp. 821–832.

[52] Y. Cao, M. Yoshikawa, Y. Xiao, and L. Xiong, "Quantifying differential privacy in continuous data release under temporal correlations," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 7, pp. 1281–1295, Jul. 2019.

[53] E. Bozkir, O. Günlü, W. Fuhl, R. F. Schaefer, and E. Kasneci, "Differential privacy for eye tracking with temporal correlations," 2020, *arXiv:2002.08972*. [Online]. Available: https://arxiv.org/abs/2002.08972

[54] X. Wu, T. Wu, M. Khan, Q. Ni, and W. Dou, "Game Theory Based Correlated Privacy Preserving Analysis in Big Data," *IEEE Trans. Big Data*, vol. 7790, no. 5, pp. 1–16, 2017.

[55] T. Li and N. Li, "On the tradeoff between privacy and utility in data publishing," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining - KDD*, 2009, pp. 517–526.

[56] B. Lin and D. Kifer, "Information measures in statistical privacy and data processing applications," *ACM Trans. Knowl. Discovery Data*, vol. 9, no. 4, p. 28, 2015.

[57] B.-R. Lin and D. Kifer, "Geometry of privacy and utility," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Dec. 2013, pp. 281–284.

[58] F. McSherry and K. Talwar, "Mechanism design via differential privacy," in *Proc. 48th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2007, pp. 94–103.

[59] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally utility-maximizing privacy mechanisms," *SIAM J. Comput.*, vol. 41, no. 6, pp. 1673–1693, Jan. 2012.

[60] M. Li, L. Zhu, Z. Zhang, and R. Xu, "Achieving differential privacy of trajectory data publishing in participatory sensing," *Inf. Sci.*, vols. 400–401, pp. 1–13, Aug. 2017.

**JIANZHE ZHAO** was born in Tonghua, Jilin, China, in 1982. She received the bachelor's and master's degrees in management science and engineering from the Beijing Institute of Technology, Beijing, China, in 2005 and 2009, respectively, and the Ph.D. degree in business management from Northeastern University, Shenyang, China, in 2015. Since 2009, she has been a Lecturer with the Software College, Northeastern University. Her research interests include big data, data privacy, and data mining.

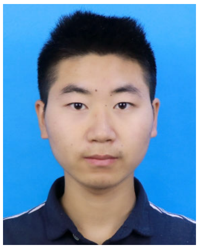
**JIE MEI** received the bachelor's and master's degrees from the Computer Science Department, Dalhousie University, in 2015 and 2017, respectively. He is currently an applied Scientist with Microsoft AI Cognitive Services. His research interests include natural language processing and data mining.

**STAN MATWIN** received the M.Sc. and Ph.D. degrees from the Department of Computer Science, Warsaw University, in 1972 and 1977, respectively.

He has been serving in a state professorship position in the Republic of Poland since 2012. He was a tenured Full Professor with the University of Ottawa, from 1992 to 2013. He has been the Canada Research Chair, a Professor with the Computer Science Faculty, Dalhousie University, and the Director of the Institute for Big Data Analytics at Dalhousie, since 2013. His research interests include machine learning, data mining, text mining, and their applications.

Dr. Matwin is a Fellow of the European Coordinating Committee for Artificial Intelligence. He was the General Co-Chair of IEEE Data Science and Advanced Analytics in 2016 and the General Chair of ACM SIGKDD 2017. He has been the Area Editor of IEEE Transactions on Knowledge and Data Engineering since 2014 and a member of the Editorial Board of the *International Journal of Social Network Mining* (IJSNM) since 2010 and the *Journal of Intelligent Information Systems* (Springer) since 2012. His awards and honors include the General Chair of Knowledge Discovery in Databases, and the Program Committee Chair and the Area Chair of a number of international conferences in AI and Machine Learning.

**YUKAI SU** was born in Guangxi, China, in 2000. He is currently pursuing the bachelor's degree in software engineering with Northeastern University, Shenyang, China. His research interests include big data and machine learning.



**YUANCHENG YANG** was born in Anhui, China, in 1999. He is currently pursuing the bachelor's degree in software engineering with the Software College, Northeastern University, China. He has strong coding experience. His research interests include machine learning and software architecture.