PRIVACY PRESERVED SPATIO-TEMPORAL TRAJECTORY DATA PUBLICATION THROUGH TEMPORAL PERTURBATION AND GENERALIZATION WITH PLACES OF INTERESTS

Dr. Rajesh N

Associate Professor, S. A. S. S. N. D. P. Yogam College
Konni, Pathanamthitta, Kerala, India
rajeshshyni2000@gmail.com

Dr. Sajimon Abraham

Professor, SMBS, Mahatma Gandhi University
Kottayam, Kerala, India
sajimabraham@rediffmail.com

Dr. Nishad A

Faculty, Dept. of Higher Secondary Education
Government of Kerala, India
an.nishad@gmail.com

Dr. Lumy Joseph

Associate Professor, Marian College (Autonomous)
Kuttikkanam Idukki, Kerala, India
lumy.biju@mariancollege.org

Dr. Benymol Jose

Associate Professor, Marian College (Autonomous)
Kuttikkanam Idukki, Kerala, India
benymol.jose@mariancollege.org

Abstract

The prevalence of mobile devices led the authorities to collect enormous volume of spatio-temporal trajectories. This practice may lead to the exposition of valuable sensitive personal details to adversaries and thereby privacy may be endangered. But the publication of datasets is essential for developmental activities. Hence anonymization of trajectory before publishing is imperative. In this work, instead of anonymizing the whole trajectory, it focuses on some stay locations on the trajectory as most sensitive to the user and anonymized them to provide privacy. This work suggests a perfect blend of existing generalization techniques with Places of Interests along with a new temporal perturbation technique for the anonymization of sensitive stay locations by adding temporal noise values. The experiments and evaluations with real-world datasets prove that this approach reduces unnecessary anonymization of trajectories and provide high data utility, less information loss and greater privacy for the users during the trajectory publication.

***Keywords*: Anonymization; Privacy preserved; Spatio-temporal; Trajectory publication; Temporal perturbation.**

1. Introduction

The popularity of IoT enabled services and its deployment in the field of applications such as smart cities, smart homes and especially in transportation systems creates huge amount of big data. Also, the spatio-temporal data [Qu et al. (2020)] generated from various activities of the moving objects in Location Based Services (LBS) are enriched with rich semantic information and the publishing of which as such, creates a lot of privacy issues to the user. The concept of privacy may differ from one domain to another, but the ultimate aim is to protect the personal, identifiable information from the adversaries and provide the quality data for the research purposes by maintaining the maximum data secrecy. The disclosure of non-anonymized trajectory data may pose a serious privacy threat [Zhao et al. (2020); Eom et al. (2020)] to the user such as the leakage of health conditions, religious or sexual preferences, living habits etc. and it gives a chance for the adversaries to create an opportunity for planning and executing various attacks. So, we should design an appropriate method to tackle this serious issue. Thus, minimizing the information loss and maximizing the privacy and data utility are the main challenges in the publishing scenario.

The extracted results from trajectories are extremely useful in the sectors like supply chain management, intelligent transportation implementation, urban planning etc. to name a few areas. The trajectory owners need to publish the data for these developmental as well as research activities, but the published data may also reach the hands of the untrustworthy malevolent. These third party adversaries may link it with other known trajectory databases to steal the personal information and the privacy can be under threat [Bonchi et al. (2011)]. It means that, the disclosure of location data in trajectories may pose serious privacy threat to the individuals [Niu et al. (2014); Chow and Mokbel (2011)]. So it is essential to apply privacy preserved approaches on trajectory before it is being published [Soria-Comas and Domingo-Ferrer (2012); Domingo-Ferrer et al. (2010)].

For the protection of sensitive data, mere removal of external identifiers is not adequate enough for the trajectory protection. A work carried out earlier dealt with significant stay point protection only [Huo et al. (2012)]. Another work considered the location frequency of the vehicles parked and found that anonymizing higher frequency parking locations was enough for the trajectory protection [Gambs et al. (2014); Zang and Bolot (2011)], but it ignored the concept of inverse user frequency, i.e.; it did not mean that the higher frequency parking location had the chance of higher re-identification and the lower frequency, low re-identification probability. Another work in this field suggests that the above methods were not sufficient against moving attacks and a *k*-correlation model along with the mix of location frequency function and inverse user frequency function was better for the protection of location trajectory data [Sui et al. (2017)]. But all this has resulted in unnecessary protection in certain parts of trajectory data and does not consider the temporal dimension of the dataset.

For providing better trade-off between privacy and utility, we need to identify the areas which are going to be anonymized. For that purpose, we have to extract the sensitive stay locations from the trajectory. We consider that anonymizing most sensitive stay locations in a trajectory is sufficient to protect privacy. Most of the studies consider that the start and end points in a trajectory are the most sensitive locations of a user. This is not false, because in most of the trajectories the start and the end points are home and workplace of a user and are highly probable to spot an individual. But, in this work we assume that the intermediate stay locations are also very important for a user because it indicates the purpose of the trip and has to be anonymized to protect the trajectory privacy. For a person like a VVIP, this type of trajectory protection is particularly important. Since the residence and office may be protected by the securities, the attacker can plan a physical attack based on the background trajectory information obtained from the trip details. For example, a person starts their journey from the home and visits and spends some time at a hospital and then proceeds to a laboratory, may be an indication for the malevolent that the victim has some health-related issue. Also, we assume that, for user trajectories, only a few stay locations are sensitive and some, less sensitive in nature. Consequently, for the anonymization, we need to extract these stay locations.

In this work, we propose a new model which considers the location frequency within the stay location and also the frequency of the days in which the user has stayed. A sensitive stay location function considers these factors and suggests only one or few stay locations which are much sensitive to the user by considering the temporal factors also. This avoids the over protection problems that are mostly found in the earlier works. In this work, we also introduced a new approach called TemPert along with the generalized anonymization of sensitive stay location with POI (Places Of Interest).

Fig. 1 gives a hierarchical diagram about our approach. Our approaches TemPert and Anony_Gen is derived mainly from the anonymization techniques such as perturbation and generalization, which are derived under the privacy preserved data mining area called trajectory privacy. The TemPert method has been achieved by adding temporal random noises to the specific spatio-temporal points. The Anony_Gen works with the generalization of sensitive stay locations with the adjacent POIs.

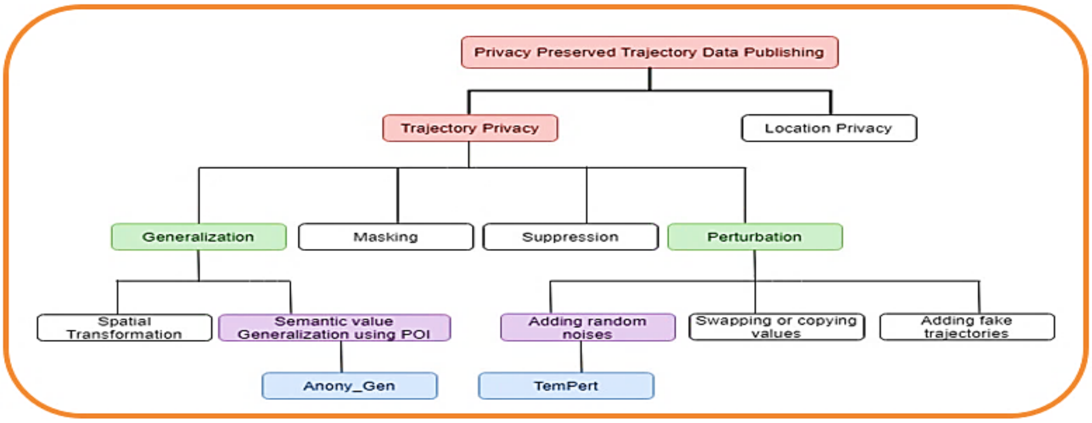


Fig. 1. The hierarchical diagram of the proposed approach

The main contributions to this work are as follows.

- We consider not only the highest stay duration points but also the wandering points around these points with low velocity in a Haversine distance measured area.
- Then we design a model which extracts sensitive and less-sensitive stay locations from each user by considering the stay duration, frequency of visits to a particular location, number of days visited to this location etc. This will enable us to find the most sensitive stay locations of the user.
- Categorize each trajectory of the user into four, based on sensitive and less-sensitive nature. The first category of trajectories has only less-sensitive stay locations. Second category, consisting of the trajectories, does not have any sensitive or less-sensitive ones. Third category consisting of trajectories has both sensitive and less-sensitive stay locations. And finally, the trajectories having only sensitive stay locations will fall into the fourth category.
- The first and second categories can be directly put into a publishable database without any anonymization. For the third category, we present a new approach called temporal perturbation, which will extract trajectory segments and add noise values to the temporal part of the sensitive and less sensitive stay locations. Then all temporal parts of the location points are re-adjusted before it is sent for publication.
- The final category works with the anonymization of sensitive stay points with “*k*” number of POIs in Minimum Bounding Rectangular (MBR) areas which are located very close to the sensitive location and is found from the Google Maps API of the particular area.
- We evaluate the performance of the model with a real-world dataset which shows a better privacy-utility trade-off between the existing similar models.

The remainder of this paper is categorized as follows: Section 2 depicts an overview of the research domain and discusses the related work. Section 3 gives the fundamental notations and the problem statements about this work. Section 4 presents the proposed methods and anonymity process in detail. The experimental results and analysis are outlined in Section 5. Finally, section 6 presents the concluding statements.

2. Related works

The privacy protection for publishing trajectories is mainly classified into three categories just like the operations of the relational database viz. add, delete, and update. The individual trajectory consists of sensitive and non-sensitive or less-sensitive stay locations and without applying proper anonymization on trajectories and publishing them as such will surely result in privacy breach. The trajectory parts or even quasi-identifiers may give a chance for adversaries to re-identify an individual.

The concept of anonymization was first proposed in the work [Samarati and Sweeney (1998)], k-anonymity, which uses clustering techniques and states that each individual cannot be identified among other k-1 individuals. It focuses mainly on publishing a new trajectory dataset by anonymizing the existing individual’s dataset [Sweeney (2002)]. A famous work [Abul et al. (2008)] which extends the k-anonymity concept by paying heed to the location imprecision (δ) to (k, δ)-anonymity. Another work [Gurung et al. (2014)] proposes a method called ICBA, which eliminates the infrequent sub-trajectories within the same time span by taking account of the frequency of spatial patterns. A method called SeqAnon [Poulis et al. (2014)], generalizes locations which are neighborhood points in trajectory and replaces it with a set containing both the points. A distinguishing work [Gramaglia et al. (2017)] proposes $k^{\tau,\varepsilon}$ - anonymity, which partitions the trajectories based on the time and uses generalization and suppression to attain k-anonymity groups within the same time intervals. A recent work [B. Liu et al. (2018)] states that night time POIs are also sensitive like homes as they stay long most of the time. Another work [Dong and Pi (2018)] also suggested that sub-trajectories, like “Home to Work” are sensitive and maintain the balance of privacy and usability. We have to remove those sub-trajectories within the same time interval and also less than *k* individuals.

The work [Abul et al. (2008)] proposed a cluster based algorithm with uncertainty threshold and another work [Nergiz et al. (2008)] proposed generalization based algorithm, evolved from the work on trajectory anonymization and based on the concept of *k*-anonymity [Sweeney (2002); K. Wang et al. (2019)]. But all these approaches tried to anonymize the whole trajectories resulting in huge information loss, but provided higher privacy.

Another method called perturbation applies on the trajectory dataset by adding suitable noise to the original data. The authors proposed an exponential mechanism which can create noise level through sampling distance and the direction of points in the raw trajectory dataset and then adding these global noises to the spatio-temporal points on the trajectory to get the better data utility. But this differential privacy mechanism is applied to perturb ship trajectories but it is not so suitable for the road network [Jiang et al. (2013)]. Another work proposes that the outcome of the location obfuscation method is the same for the points that are spatially close to each other and follows the geo-indistinguishability concept [Miguel et al. (2013)]. This mechanism is possible by adding two-dimensional Laplace noise to the real location point of the user. After that, for protecting the trajectories they used composition theorems with the help of generalization. Some other work suggests that geo-indistinguishability needs greater amount of noise for the trajectory data protection [Bordenabe et al. (2014)]. But there are some approaches which directly add noises to the location points to achieve differential privacy [Li et al. (2017); Chen et al. (2012)]. For the work specified in this chapter, we used this strategy of adding noise values directly to the location points.

For retaining almost, the same structure, granularity of the data to its higher level, and less information loss, most researchers will choose the generalization method. For example, if a person visited a specific shop in a mall several times, to mask the behavior of this person, during the publication, the place of visit is renamed as the mall in order to maintain privacy. So, a certain level of semantics is maintained by not disclosing the vital information [Monreale et al. (2011)]. The process of generalization can be implemented in two ways: (i) Spatial transformation - to create grid cells like area blocks, adding imprecision to the existing spatial coordinates [Gramaglia et al. (2017); Huo et al. (2012); Abul et al. (2010)], (ii) based on the semantic value generalization (POI names).

There are only a few methods derived for semantic trajectories by giving special attention to the spatio-temporal features. A work [Huo et al. (2012)] which used the *k*-anonymity method in their approach and states that instead of anonymizing the whole trajectory, the anonymization of significant stops are sufficient to achieve privacy. Another recent method [Dai et al. (2018)] suggests performing trajectory reconstruction method, which takes into consideration the semantic property of the POI name during the process of perturbation by replacing sensitive stops with other points to achieve personal privacy.

Here are some works which address the problem of trajectory anonymity for publishing recently. A different work [Sui et al. (2017)] takes into account the moving preference of individuals and tries to protect the trajectories from re-identification attacks. For that, they propose a trajectory anonymity model, which protects the important parking locations in the *k*-correlation region with the help of the concept location frequency-inverse user frequency (LF-IUF). Another work [Wang et al. (2019)], suggests for protecting generated trajectory points which satisfy *k*-anonymity by interchanging the positions of the trajectory points on the *k*-core subnet of the relation network. In a work [X. Liu et al., (2018)], which proposed a method called SLAT, a sub-trajectory linkage attack tolerance framework to protect privacy based on sensitive value generalization, trajectory splitting and location suppression in order to preserve the data utility. Another work [Ghasemi Komishani et al. (2016)] proposed a model to balance data utility and privacy of the data by employing local suppression and generalization of sensitive attributes. They were given separate privacy protection mechanisms for different objects while publishing the trajectory data. The main drawback with this approach is that it does not take a relationship between time and trajectory, and also does not pay much attention to trajectory data publishing. Another approach [Hu et al. (2018)] proposes the method to satisfy different privacy requirements for different locations and time points with the help of constructing a privacy requirement matrix and satisfy the (l, δ) constraint to build an undirected trajectory graph. In the work ([Naghizade et al. (2020); Naghizade et al. (2014)], the authors proposed a method called flip-flop exchange, which substitutes sensitive stops with moves or a less-sensitive stop available in the said trajectory.

Some studies have concentrated on the semantic features in the trajectory, especially the stop points in the trajectory and try to protect these points to achieve privacy. A work proposes an obfuscation technique for the online LBSs by making provision for the spatial coordinates of a stop and derives a region based on the user profile and POI distribution [Damiani et al. (2009)]. Another work [Huo et al. (2012)] proposes to anonymize stop points in *l*-diversified zones. The authors in their work [Dai et al., (2018)] proposed to replace sensitive stop points in a trajectory with some POI coordinates using the stop taxonomy tree concept, but this technique does not account for the temporal property of the trajectory. In our work we were mindful of spatial as well as temporal factors for the effective anonymization.

An existing work [Domingo-Ferrer and Trujillo-Rasua (2012)] suggests a micro-aggregation approach for the privacy protection in trajectory data publishing with the use of micro-data anonymization. They clustered the trajectories based on the similarities with a measure of at least *k*-size and replaced these trajectories with synthetic data that accumulated in the actual trajectories or in the visited locations.

In the work [Gambs et al. (2014)], the authors proposed to prevent de-anonymization attacks by modeling an individual mobility based function using a mobility Markov chain. An attack based semi supervised learning approach was presented in the work [Hua et al. (2017)] by the authors to infer the riding trajectory of the user. A scalable sanitization method (SafePath) for publishing differentially-private trajectories by considering trajectories as a noisy prefix tree was proposed in the work [Al-Hussaeni et al. (2018)].

Another privacy preserving method was suppression-based method and it suppresses some vital points in the trajectory and releases the rest. A work by the authors [Chen et al., (2013)] proposed a tailored privacy data anonymization model which allows to adopt various data utility metrices for different data mining task through local suppression. A work [Hu et al. (2018)] proposed by the authors specifies the trajectory privacy protection problem by achieving the various privacy requirement using the method of dividing the time intervals. It forms an undirected trajectory graph by considering the (*l, δ*) factors. But this method uses Manhattan distance for the work, which is not enough to face the trajectory-oriented computations. Another work [Ghasemi Komishani et al. (2016)] uses the local trajectory suppression and sensitive attribute generalization, which suggests different privacy approaches for different moving objects and it does not adopt the time constraints in the trajectory. But the above-mentioned approaches do not consider the temporal or semantic features of the trajectory.

Some other works [Gruteser and Xuan Liu (2004); Gidófalvi et al. (2008)] propose a tailored release of trajectory by avoiding the sensitive ones and publish only the non-sensitive trajectory samples. The method they choose in these methods is that, whenever the user enters the sensitive places, the location samples are to be deleted/suppressed and the balance is sent for trajectory publication. This method seems to be straight forward and more effective but sudden responses in trajectory leads to high distortion in the data utility.

All the trajectory protection methods leads to the urgency of a vast storage and a greater computational expenses, which leads to the research of safeguarding sensitive locations in a trajectory [Huo et al. (2012); Naghizade et al. (2014); Han and Tsai (2015)]. Since the nature and purpose of a user lies with these sensitive location points. A research work [Naghizade et al. (2014)] was proposed a re-construction mechanism in trajectories to protect stop points with less-sensitive POIs. In this method, we have to measure and prefix the POI sensitivity and sampling points and fix values prior to the reconstruction, which is considered to be less suitable.

Even though new methods and approaches are suggested by various researchers across the world to cater to the problem of balancing data utility against the privacy of the data while publishing is still a hot research area because of its importance, dimensionality and complex nature.

3. Problem definitions

The traces left by the moving objects under LBS are trajectories, a sequence of spatio-temporal points. This section contains some important basic definitions that are needed to illustrate the trajectory privacy concepts used in our work.

Definition 3.1 (Trajectory). *A sequence of spatio-temporal points constitute a trajectory and which is represented as* $\mathscr{T}^{u}{}_{j}$*= {*Tp_0*,* Tp_1*, …………..,* $Tp_{n(j)}$*}, where* Tp_i *= {*$lati_i$*,* $longi_i$*,* t_i*}, u ϵ U, j= 1……*$J_{u,}$*. Here, (*$lati_i$*,* $longi_i$*) represent latitude and longitude and the timestamp at this point is represented by* t_i*.*

The notation afore mentioned represents the trajectory $\mathscr{T}$ for the user ‘u’ which is from the set of whole users U. n(j) shows that the trajectory length could vary and J_u shows the number of trajectories can be varied by users.

Definition 3.2 (Stay point). *A stay point* STp_i *is represented as {*STp_{id}*,* Tp_i*, £st}, where* STp_{id} *is the trajectory stay point identifier,* Tp_i *can be commonplace trajectory parameters such as latitude and longitude of the stay point. £st is the duration of the stay which can be calculated by taking the difference of arrival and departure time of the user at this point and it exceeds* Δt_{th}.

When the user has stayed at a location for over a mean or a specific time threshold (Δt_{th}), the trajectory stay points are formed. Its stay duration was fixed by the data publisher. If the user wishes to wander about a location at a low velocity, which may be sightseeing or calling at different rooms in a building etc. This comes under the classification of wandering points. Hence it is needed to take account of these trajectory points within the category of stay points. To this end it is needed to define one more term namely stay location.

Definition 3.3 (Stay location). *A stay location, STL_i, an MBR centroid coordinate of the entire trajectory stay points, satisfying the user defined distance threshold, $£d_{th}$. ie; STL_i = {STL_{id}, ST_i, $£d_{th}$}, where STL_{id} is the stay location identifier and ST_i is the total time span of all stay points that can be located within $£d_{th}$.*

Hospitals, hotels, shopping malls, outdoor or indoor parks, religious or worship places etc. are stay locations like the actual places.

Definition 3.4 (Stay location frequency). *The stay location frequency, StLF is the result obtained when the total frequency of visits made by the user each day at a specific stay location, STL_i is divided by the total number of visits made by the user at various locations.*

The frequency of visits made by the user at various STL_j in matrix form is represented as FM.

$$\mathrm{FM}\ (U_i,\ STL_j) \quad = \quad \begin{array}{c} \\ d_1 \\ d_2 \\ \\ d_m \end{array} \begin{array}{c} STL_1 \quad STL_2 \ \ldots\ldots\ldots \ STL_n \\ \left[\begin{array}{l} vf_{11} \quad vf_{12} \ \ldots\ldots\ldots\ldots vf_{1n} \\ vr_{21} \quad vf_{22} \ldots\ldots\ldots\ldots vf_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ vf_{m1} \quad vf_{m2} \ \ldots\ldots\ldots.. \ vf_{mn} \end{array}\right] \end{array}$$

The Eq. (1) is used to calculate the stay location frequency from the user trajectories.

$$StLF\ (U_i,\ STL_j) = \frac{\sum_{k=1}^{m} Vf_{kj}}{\sum_{j=1}^{n}\sum_{k=1}^{m} Vf_{kj}}. \qquad (1)$$

The user identifier is denoted with U_i, in Eq. (1) where i = 1 to u, and u is the total count of users; vf, stands for the visit frequency at STL_j; m, the number of visits at STL_j and n, is the sum of visits made by the specific user at a different STL_j. Thus, StLF value lies between 0 and 1 denoting the significance of this stay location to the user. But, this ratio is insufficient to compute if a location is sensitive or not. This leads us to go after yet another term known as day frequency.

Definition 3.5 (Day frequency). *A day frequency, DvF is calculated by the division of the total of presents of visits by the particular user in sampling days, d_k at the particular stay location, STL_j and the sum of sampling days, D.*

The presence of visits by the user at STL_j on d_k days can be represented as a matrix, DM.

$$\mathrm{DM}\ (U_i,\ STL_j) \quad = \quad \begin{array}{c} \\ d_1 \\ d_2 \\ \\ d_m \end{array} \begin{array}{c} STL_1 \quad STL_2 \ \ldots\ldots\ldots \quad STL_n \\ \left[\begin{array}{l} dv_{11} \quad dv_{12} \ \ldots\ldots\ldots\ldots \ dv_{1n} \\ dv_{21} \quad dv_{22} \ldots\ldots\ldots\ldots dv_{2n} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ dv_{m1} \quad dv_{m2} \ldots\ldots\ldots\ldots \ dv_{mn} \end{array}\right] \end{array}$$

The entries in the matrix DM (U_i, STL_j) stands for the presence of visits that the user U_i has made at the particular location STL_j on each day, d_k. The presence and absence of visits are represented by 1 and 0 respectively.

$$DvF\ (U_i,\ StL_j) = \frac{\sum_{k=1}^{m} dv_{kj}}{D}. \qquad (2)$$

In Eq. (2), U_i is the user identifier, where i = 1 to u; u is the total number of users and j= 1 to n; n is the count of all stay locations and m is the total sum of day visits. And dv_{kj} represents the presence of visits at STL_j and D, the count of whole travelling days by the specific user. The DvF value also would lie between 0 and 1 representing the significance of the particular stay location to the user. Though this value too is not enough to identify if a location is sensitive or less-sensitive. Thus, we need another notion termed sensitive stay location.

Definition 3.6 (Sensitive stay location). *A stay location STL_i is said to be Sensitive stay location, $SStL_i$, if the user U_i has stayed over for a long time as stay points and/or as wandering points while considering other stay locations; the user may have performed multiple visits during a period in this stay location and the visits to this stay location should be more than a single day.*

The sensitive and less-sensitive stay locations are distinguished with Eq. (3) using sensitivity threshold (δ), the mean of SStLF.

$$\text{SStLF}\ (U_i, \text{STL}_j) = \text{StLF}(U_i, \text{STL}_j) \times \text{DvF}(U_i, \text{STL}_j). \qquad (3)$$

While extracting the sensitive stay locations, there are many less-sensitive locations as well. The less-sensitive locations, NSStL_i are stay locations besides sensitive stay locations. Usually, the number of less-sensitive stay locations is more in number than the sensitive ones. The intention of ours is to anonymize only the sensitive stay locations before it is published. This is imperative as the sensitive ones that have more delicate information regarding an object and the leakage will cause a lot of privacy breach. It can be exploited by the adversaries.

Definition 3.7 (Temporal noise). *The temporal noise, $¥_i$ is a temporal value which we derived to make necessary adjustments (adding or subtracting) on the spatio-temporal parameters of the trajectory segment (sub-trajectory) containing sensitive and less-sensitive stay locations. The $¥_i$ is calculated by the mean of temporal deviations of the stay points from sensitive to less-sensitive stay locations within the trajectory segment.*

The adjustments with the noise value led to the process of temporal perturbation and which is a new idea that we are presenting.

Definition 3.8 (Temporal perturbation). *The temporal perturbation means making adjustments with the noise value on the temporal parameters of the trajectory's stay locations.*

The temporal perturbation method will surely eliminate the unnecessary anonymization of spatio-temporal points, which occurred mostly in the generalization, obfuscation and dummy trajectory anonymization methods. So, the spatial parameters are published without any alteration and spatial parameters can be used for research purposes in an un-altered manner. This is a great relief to researchers who need exact spatial coordinates.

Definition 3.9 (Sensitive stay zone area). *The anonymization through generalization method is achieved through the creation of the stay zone. A sensitive stay zone area is created in the specified area when the trajectory contains only the sensitive stay locations. In this situation, we created the stay zone area over the sensitive stay location with "k" number of the nearest POI locations as the stay location, which have been identified by using the Google Maps API of specified location.*

In this method, we attach the nearest (found through Haversine distance measure) "*k*" POI locations and build an MBR over these location points, including the sensitive points.

4. Proposed solutions

The proposed solution, Anony_Pub consisting of three major phases, namely sensitive and less-sensitive stay location extraction - Sen_Stln_Extrn, Temporal perturbation-TemPert and the anonymization through generalization of sensitive stay locations with POIs - Anony_Gen. These phases are explained as follows.

4.1. *Overview of the solution*

The proposed approach mainly aims at preprocessing each user's trajectories from the IpTrDb and extract sensitive and less-sensitive stay locations from the user trajectories. We assume that the malevolent are capable enough to leak the sensitive information by linking known databases. So, our main intention is to protect only the sensitive stay location information from the adversaries by anonymizing these stay locations. Also, it is to be noted that we are performing this anonymization by analyzing the trajectory nature and applying temporal perturbation or anonymous generalization through the creation of the MBR zone depending on the nature. From the experiments we also found that the anonymization process was only needed for some trajectories which contained sensitive stay locations and thereby we avoided unnecessary anonymization of all the stay points. This will certainly result in less information loss and maximum privacy gain than the existing methods.

This approach portrayed in Fig. 2 actually works in two major stages. One is the identification of stay locations and extraction of sensitive and less-sensitive stay locations from the input trajectory database, IpTrDb. For that, we use the Algorithm: 1, Sens_Stln_Extrn.

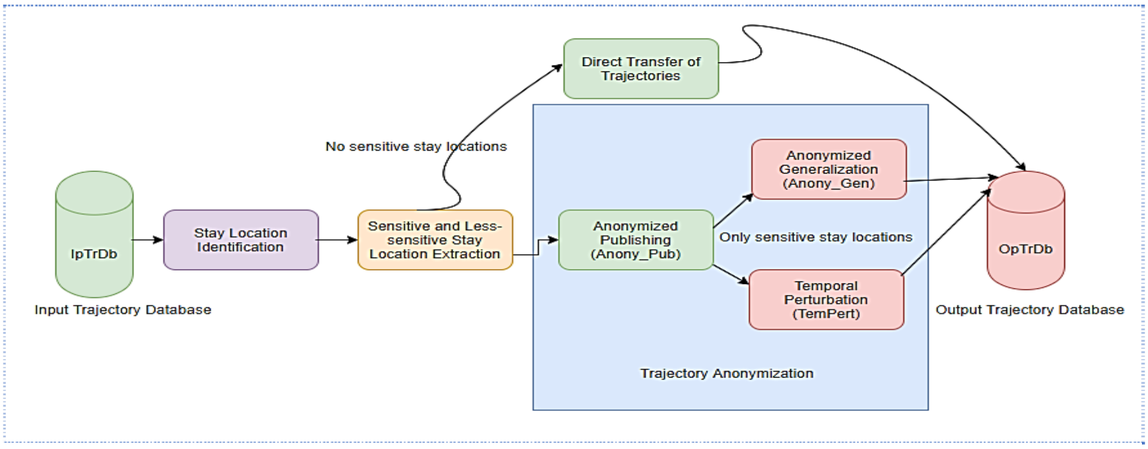

Direct Transfer of Trajectories
No sensitive stay locations
Anonymized Generalization (Anony_Gen)
IpTrDb
Stay Location Identification
Sensitive and Less-sensitive Stay Location Extraction
Anonymized Publishing (Anony_Pub)
Only sensitive stay locations
OpTrDb
Input Trajectory Database
Temporal Perturbation (TemPert)
Output Trajectory Database
Trajectory Anonymization

Fig. 2. An overview of the proposed approach

In the second stage, we anonymized the sensitive stay locations with the help of less-sensitive stay locations and POIs. The second stage considers the category of four types of trajectories. In category-1, there are only sensitive stay locations, category-2 contains only less-sensitive stay locations, category-3 contains both sensitive and less-sensitive stay locations and finally, a few trajectories do not have sensitive or less-sensitive stay locations which fall in category-4. Categories 2 and 4, did not need any type of anonymization before publishing, which can directly transfer to the output trajectory database for publishing. The category 3 and 1 need some type of anonymization before it is being published. But in the second stage, the category-3 uses the new approach called temporal perturbation, which creates temporal noise values and it gets added to the temporal duration of each stay location within the sub trajectory and finally the actual temporal parameters of the stay locations are readjusted to make it temporally perturbed and replace this trajectory segment in the original trajectory and pass it for the publication. The new method called temporal perturbation is applied to the sub-trajectory segments having sensitive and less-sensitive stay locations. The temporal values of each stay locations with sensitive and less-sensitive stay location's stay duration are readjusted with new values and all of them have equal stay duration, but the entry and exit time of the trajectory segment is unaltered. So, this will definitely make utter confusion among the adversaries to distinguish which stay locations are sensitive. The main advantage of this method is that we are implementing perturbation only on the temporal part of the trajectory segment, not with any of the spatial part. So, definitely information loss will be very less compared to any other method and we will get high data utility because of unaltered spatial information which are more beneficial in spatial related research and applications.

In the final category, the trajectories of the form category-1 are having only sensitive stay locations. In this situation, we use Google Maps API or OpenStreetMap API[a] to find important "*k*" POIs from the surroundings. We used Google Maps API for getting the POIs throughout this thesis. With the help of POIs, the sensitive stay locations are anonymized in the MBR area and the MBR's coordinate values will replace these sensitive stay locations and thus the anonymized trajectory is sent for publication.

4.2. *Sensitive and less-sensitive stay location extraction*

For the extraction of sensitive and less-sensitive stay locations, we had to go through the processes like trajectory preprocessing, stay point extraction, finding of wandering points, formulation of the stay location and the categorize these stay locations into sensitive and less-sensitive stay locations. All these processes are described in the Algorithm: 1, Sen_Stln_Extrn.

Algorithm: 1 - Sen_Stln_Extrn

Input: *Input trajectories from IpTrDb*

Output: *Updated $\mathscr{T}^u_j$ with sensitive stay locations, $SStL_i$ and less-sensitive stay locations, $NSStL_i$*

```
for i=1, j=1 to n do
    preprocess and read all trajectories $\mathscr{T}^u_j \in U$ from ITrD
    find the stay duration $£st_i$, and spatial distance $£dt_i$ of each spatio-temporal point $p_i$ from the previous
    point $p_{i-1}$ and update $Stp_i(p_i, £st_i, £dt_i)$
end for
initialize user defined or mean($£st_i$) time threshold, $\Delta t_{th}$ according to the user choice, and the user defined
distance threshold, $\Delta d_{th}$
```

[a] https://openstreetmap.org/

```
for i=1, j=1 to n, $\mathcal{T}^{u}_{j} \epsilon U$ do
    if ($£dt_i <= \Delta d_{th}$) then consider those points as wandering points and find sum of all $£st_i$'s -> $ST_i$
    if ($ST_i >= \Delta t_{th}$) then create an MBR by collecting $Stp_i$'s within $\Delta d_{th}$ and find its centroid as $STL_i$, stay location
    replace each $p_i$ c $STL_i \epsilon \mathcal{T}^{u}_{j}$ by stay location $STL_i = \{STL_{id}, ST_i, £d_{th}\}$
    calculate stay location frequency, StLF, day frequency, DvF and sensitivity frequency, SStLF
    calculate sensitivity threshold δs as Mean (SStLF)
    if (SStLF >= δs) then update $STL_i \rightarrow SStL_i$ ($SStL_{id}$, $STL_i$), sensitive stay location else update
        $STL_i \rightarrow NSStL_i$ ($NSStL_{id}$, $STL_i$), less-sensitive stay location
    return updated $\mathcal{T}^{u}_{j}$ with $SStL_i$ and $NSStL_i$
end for
```

In the above algorithm, the stay points, Stp_i on each of the trajectory is found by taking the difference of each spatial point's temporal part, ie; | $dept_i$ - $arrt_i$ | and if this duration, £st is greater than or equal to the user defined time threshold, Δt_{th} then consider it as the stay location. Sometimes there are some wandering points whose velocity is almost negligible and distance is within the user defined distance threshold, Δd_{th}. The stay point consists of the distance measure, *£dt* for considering these wandering points as stay locations. Here we took Δt_{th} as user-defined.

After the identification of these stay locations, we identified the sensitive and less-sensitive locations. Here we calculated the stay location frequency, StLF, day frequency, DvF and sensitivity frequency, SStLF using equations (1), (2) and (3). We set the sensitivity threshold, δs by considering the mean value of SStLF. For finding the sensitive and less-sensitive stay locations we checked the (SStLF >= δs) then considered STL_i as a sensitive stay location, $SStL_i$ and otherwise considered STL_i as less-sensitive stay location, $NSStL_i$. Update this information on the trajectories.

The following matrix FM represents the frequency of visit to a specific location STL_j over a specific time duration for a user, U_1 as an example which was mentioned in the definition part.

$$FM(U_1, STL_j) = \begin{array}{c|cccccc|} & STL_1 & STL_2 & STL_3 & STL_4 & STL_5 & STL_6 \\ d_1 & 2 & 0 & 0 & 1 & 0 & 1 \\ d_2 & 0 & 1 & 0 & 0 & 0 & 1 \\ d_3 & 1 & 0 & 2 & 0 & 0 & 0 \\ d_4 & 0 & 0 & 1 & 1 & 1 & 0 \\ d_5 & 2 & 1 & 0 & 0 & 1 & 2 \\ d_6 & 0 & 1 & 0 & 0 & 0 & 1 \\ d_7 & 1 & 0 & 1 & 1 & 0 & 0 \end{array}$$

Also, the matrix *DM* represents the presence of visits made by the user, U_1 at STL_j on each day in a matrix form for the calculation of DF as an example referred to in the definition part.

$$DM(U_1, STL_j) = \begin{array}{c|cccccc|} & STL_1 & STL_2 & STL_3 & STL_4 & STL_5 & STL_6 \\ d_1 & 1 & 0 & 0 & 1 & 0 & 1 \\ d_2 & 0 & 1 & 0 & 0 & 0 & 1 \\ d_3 & 1 & 0 & 1 & 0 & 0 & 0 \\ d_4 & 0 & 0 & 1 & 1 & 1 & 0 \\ d_5 & 1 & 1 & 0 & 0 & 1 & 1 \\ d_6 & 0 & 1 & 0 & 0 & 0 & 1 \\ d_7 & 1 & 0 & 1 & 1 & 0 & 0 \end{array}$$

Table 1. shows an illustration of the sensitive stay location extraction algorithm (Sens_Stln_Extrn) for identifying the sensitive and less-sensitive stay locations based on the matrices FM and DM. If (SStLF >= δs), then those stay locations are considered as sensitive and others are less-sensitive one's. In this example, STL_1 and STL_6 are sensitive stay locations.

		Frequency of visits						
		STL_1	*STL_2*	*STL_3*	*STL_4*	*STL_5*	*STL_6*	
Day	1	2	0	0	1	0	1	
	2	0	1	0	0	0	1	
	3	1	0	2	0	0	0	
	4	0	0	1	1	1	0	
	5	2	1	0	0	1	2	
	6	0	1	0	0	0	1	
	7	1	0	1	1	0	0	
	Total visits at the Stay location	6	3	4	3	2	5	23
	No. of days visited	4	3	3	3	2	4	
	StLF	0.261	0.130	0.174	0.130	0.087	0.217	
	DvF	0.571	0.429	0.429	0.429	0.286	0.571	
	SStLF	**0.149**	0.056	0.075	0.056	0.025	**0.124**	
	δs = mean(SStLF)	**0.081**						

Table 1. An illustration of sensitive stay location extraction

4.3. *Anonymization process in the publication of trajectories*

In order to publish the trajectory information in a privacy preserved manner, we have to perform anonymization with the spatio-temporal points in the trajectory. Here, we present a new method by combining the anonymous generalization cum temporal perturbation method. This can be achieved through implementing the following Algorithm: 2 for the trajectories.

```
Algorithm: 2 - Anony_Publn
Input: $\mathcal{T}^u_j$ with $SStL_i$ and $NSStL_i$
Output: OpTrDb with Anonymized $\mathcal{T}^u_j \in U$ for publication
Call Sen_Stln_Extrn
for each $\mathcal{T}^u_j \in U$ do
    Check $\mathcal{T}^{\pi}_j$ contains any $SStL_i$, then count number of sensitive stay locations, Cs and contains any
    $NSStL_i$, then count number of less-sensitive stay locations, Cns
    if (Cs==0) then put directly that $\mathcal{T}^u_j \rightarrow$OpTrDb
    otherwise, if (Cns==0) then perform Anony_Gen else perform Tempert
end for
Return $\mathcal{T}^u_j$ to OpTrDb
```

4.4. *Temporal perturbation process*

This perturbation process starts with identifying the total number of sensitive stay locations, *$SStL_i$* and less-sensitive stay locations *$NSStL_i$,* in each trajectory. According to the *k* value, it selects "n" number of less-sensitive stay locations. The collection of sensitive and less-sensitive stay locations in a trajectory forms a sub-trajectory. Then find the median value, Md of the stay duration, *£st_i* in each sensitive and less-sensitive stay location. Calculate the deviations from the stay duration to the mean which is considered as the temporal noise, ¥$_i$ for the temporal perturbation method. Finally, readjust the temporal part of each stay point with regard to ¥$_i$. Update the values in the sub-trajectory and the trajectory $\mathcal{T}^{\pi}_j$. The temporal perturbation (TemPert) process is diagrammatically represented in the Fig. 3.

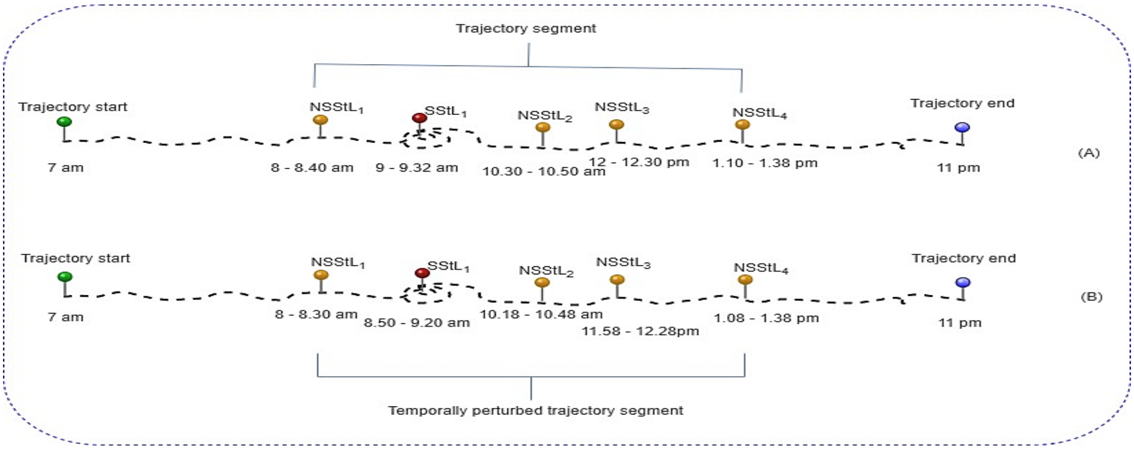


Fig. 3. Temporal perturbation process (A)Raw trajectory (B)Publishable trajectory

The proposed Algorithm: 3 describes the trajectory anonymization approach based on the temporal perturbation method.

Algorithm: 3 - TemPert

Input: $\mathcal{T}^{u}_{j}$ *with* $SStL_i$ *and* $NSStL_i$
Output: *Temporally perturbed trajectory* $\mathcal{T}^{u}_{j}$

```
for $\forall$ $SStL_i$, $NSStL_i \in \mathcal{T}^{u}_{j}$ do
    extract sub trajectory segment $ST^{u}_{j}$ containing at least "n" number of $SStL_i$ and "k" number of $NSStL_i$,
    where (n, k) >= 1
    calculate $Md = Median$ (£$st_i$ ) $\in SStL_i$ , $NSStL_i$
    find temporal noise, ¥$_i$ = (£$st_i$ - Md) $\in SStL_i$, $NSStL_i$
   perform temporal perturbation on each $SStL_i$(£$st_i$) and $NSStL_i$ (£$st_i$) $\in ST^{u}_{j}$ with ¥$_i$.
   re-adjust $t_i \in SStL_i$ , $NSStL_i \in ST^{u}_{j}$
   update all $t_i \in ST^{u}_{j}$ and replace it in $\mathcal{T}^{u}_{j}$
end for
return $\mathcal{T}^{u}_{j}$
```

4.5. *Anonymization through generalization*

This method is needed when the trajectory contains only sensitive stay locations and no less-sensitive stay locations. For the generalized anonymization, read each trajectory from IpTrDb, which contains $SStL_i$ and create MBR stay zones (StZ) over the $SStL_i$ by including "*k*" POIs. These POIs are selected based on the nearest distance from the $SStL_i$. The POIs are identified by using the Google Maps API used in the program for accessing the nearest POIs. Then update the trajectory segment which is included in the stay zone with the coordinates of stay zone contains the $SStL_i$. The proposed algorithm is given as Algorithm: 4.

Algorithm: 4 - Anony_Gen

Input: $\mathcal{T}^{u}_{j}$ *with* $SStL_i$, *POIs from Google Maps API*
Output: *Anonymized trajectory* $\mathcal{T}^{u}_{j}$

```
for i=1, j=1 to n do
    read each trajectories of $\mathcal{T}^{u}_{j} \in U_i$ from IpTrDb
    $\forall SStL_i \in \mathcal{T}^{u}_{j}$, create a stay zone $StZ_i$ over the $SStL_i$ with "k" number of adjacent POI's
    Replace all $SStL_i$, $Stp_i \in StZ_i$ with the coordinates of $ul_c$ and $br_c \in StZ_i$
end for
Return updated $\mathcal{T}^{u}_{j}$ to OpTrDb
```

4.6. *Utility analysis*

The measure of information loss and range query distortion / data error rate are the two best ways to measure the data utility. Data utility is an important criterion in the anonymization process because it measures the actual deviation from the original data to anonymized data.

The trajectory average information loss [X. Liu et al. (2018)] is measured as the ratio of number of spatio-temporal points in the anonymized dataset to the original dataset.

$$\text{Information Loss(InfL)} = \frac{|Loc_{Ip}| - |Loc_{Op}|}{|Loc_{Ip}|} \qquad (4)$$

In this Eq. (4), Loc_{Op} is the number of locations in the anonymized output dataset and Loc_{Ip} is the actual locations in the input trajectory dataset. The numerator part will give the count of the locations that were deleted or altered or suppressed due to anonymization.

The purpose of publishing data is usually for efficient querying or for actual analysis. So, the query error rate measure is to evaluate the data utility or anonymization quality. This is done by comparing the query results from the original dataset to the anonymized dataset. The query distortion/error rate is calculated using the spatial range query, ie; a user defined COUNT (*) query. The query is performed by taking the parameters as the spatial cloaking region of the size R_g and the time interval is set as t_s and t_e. Usually there are two kinds of queries for computing range query distortion. They are Possibly_Sometimes_Inside and Definitely_Always_Inside. The Query error rate [Huo et al. (2012); Tan et al. (2019)] is calculated using the following Eq. (5).

$$\mathrm{QEr}_{Qr} = \frac{Min\ ((Q(D),\ \ Q(D*))}{Max\ ((Q(D),\ \ Q(D*))} \tag{5}$$

Here *Q(D)* to *Q(D*)* is the deviation of values performed both in the input trajectory data, *D* and with the output data (anonymized version), *D** after the query processing.

5. Experiments and result analysis

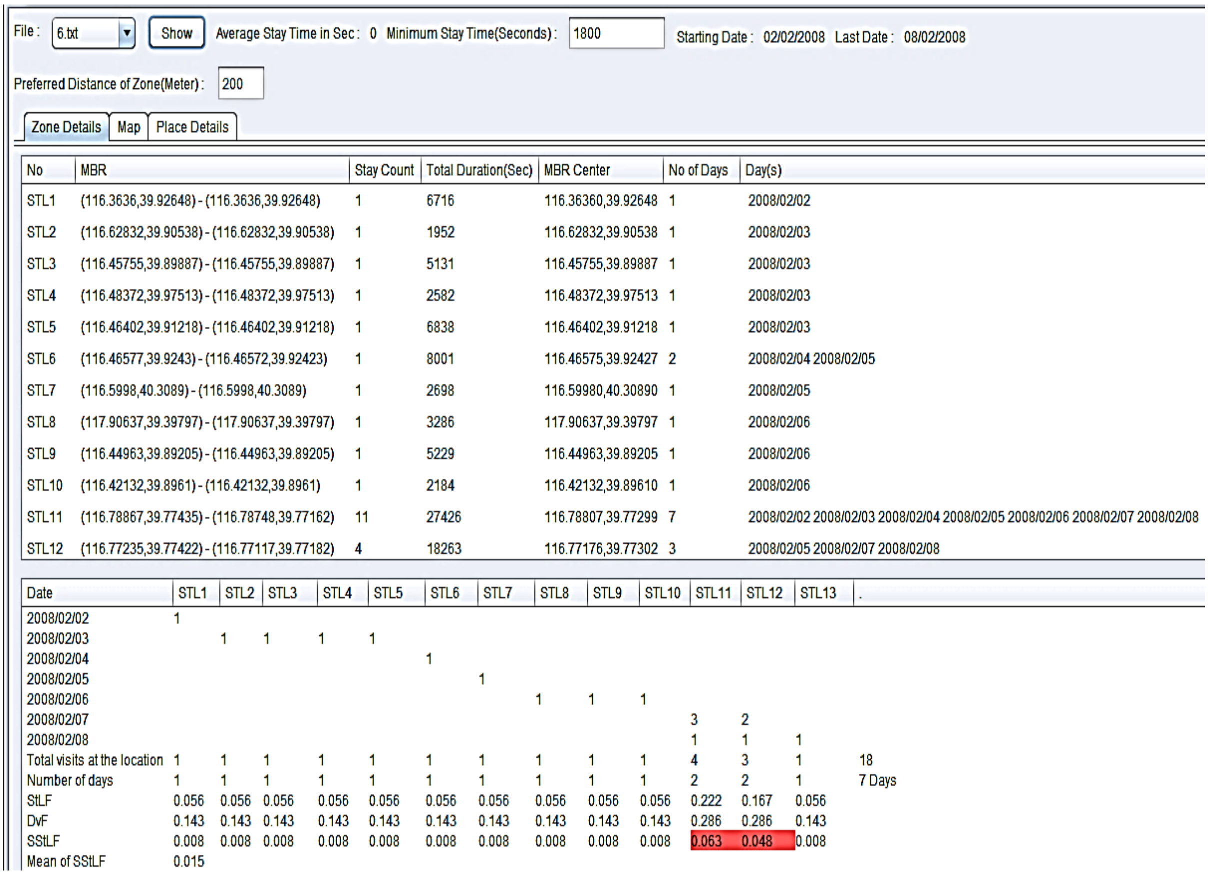
This section discusses the experiments conducted with real dataset and the outcome we got.

5.1. *Experimental setup and outputs*

For the experiment, we used T-Drive trajectory dataset, a real-world dataset from Microsoft (Yuan et al., 2011;Yuan et al., 2010). Each observation was taken on an average of 177 seconds time interval. The experiments were run on an Intel’s 8th generation core-i5 processor with up to 4 GHz speed and the machine is equipped with Windows 10 operating system and having 8 GB of RAM.

From the large trajectory database, we randomly took a maximum of 1000 users out of 10,357 users. The selected trajectories are the representation of the whole dataset and the rest of the trajectories were considered as history trajectories.

The temporal threshold for stay point, Δt_{th} is taken as the mean value of the stay duration of each stay point. So, it may vary in accordance with the number of users and their nature of travel. The distance threshold, Δd_{th} is taken as 200m. This is because, when we consider a wandering point as a stay point, usually the object may roam in a place or a building which is set up below an aerial distance of 200m. Also, for calculating all distance measures in this work we used Haversine measure. While extracting sensitive and less-sensitive stay locations, we considered the sensitivity threshold, δs as the mean of SStLF value.



File : 6.txt Average Stay Time in Sec : 0 Minimum Stay Time(Seconds) : 1800 Starting Date : 02/02/2008 Last Date : 08/02/2008

Preferred Distance of Zone(Meter) : 200

No	MBR	Stay Count	Total Duration(Sec)	MBR Center	No of Days	Day(s)
STL1	(116.3636,39.92648) - (116.3636,39.92648)	1	6716	116.36360,39.92648	1	2008/02/02
STL2	(116.62832,39.90538) - (116.62832,39.90538)	1	1952	116.62832,39.90538	1	2008/02/03
STL3	(116.45755,39.89887) - (116.45755,39.89887)	1	5131	116.45755,39.89887	1	2008/02/03
STL4	(116.48372,39.97513) - (116.48372,39.97513)	1	2582	116.48372,39.97513	1	2008/02/03
STL5	(116.46402,39.91218) - (116.46402,39.91218)	1	6838	116.46402,39.91218	1	2008/02/03
STL6	(116.46577,39.9243) - (116.46572,39.92423)	1	8001	116.46575,39.92427	2	2008/02/04 2008/02/05
STL7	(116.5998,40.3089) - (116.5998,40.3089)	1	2698	116.59980,40.30890	1	2008/02/05
STL8	(117.90637,39.39797) - (117.90637,39.39797)	1	3286	117.90637,39.39797	1	2008/02/06
STL9	(116.44963,39.89205) - (116.44963,39.89205)	1	5229	116.44963,39.89205	1	2008/02/06
STL10	(116.42132,39.8961) - (116.42132,39.8961)	1	2184	116.42132,39.89610	1	2008/02/06
STL11	(116.78867,39.77435) - (116.78748,39.77162)	11	27426	116.78807,39.77299	7	2008/02/02 2008/02/03 2008/02/04 2008/02/05 2008/02/06 2008/02/07 2008/02/08
STL12	(116.77235,39.77422) - (116.77117,39.77182)	4	18263	116.77176,39.77302	3	2008/02/05 2008/02/07 2008/02/08

Date	STL1	STL2	STL3	STL4	STL5	STL6	STL7	STL8	STL9	STL10	STL11	STL12	STL13	.
2008/02/02	1													
2008/02/03		1	1	1	1									
2008/02/04						1								
2008/02/05							1							
2008/02/06								1	1	1				
2008/02/07											3	2		
2008/02/08											1	1	1	
Total visits at the location	1	1	1	1	1	1	1	1	1	1	4	3	1	18
Number of days	1	1	1	1	1	1	1	1	1	1	2	2	1	7 Days
StLF	0.056	0.056	0.056	0.056	0.056	0.056	0.056	0.056	0.056	0.056	0.222	0.167	0.056	
DvF	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.286	0.286	0.143	
SStLF	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.008	0.063	0.048	0.008	
Mean of SStLF	0.015													

Fig. 4. Sensitive stay location identification process for the User no. 6

As an example, the Fig. 4 shows the identification process of sensitive and less-sensitive stay locations from the stay points for the user no. 6. Here you can see that there are 13 stay locations in which the stay locations STL_{11} and STL_{12} are identified as sensitive stay locations by our algorithm Sens_Stln_Extrn and the rest of the stay locations from STL_{1} to STL_{10} and STL_{13} are less-sensitive stay locations.

5.2. *Measure of efficiency*

The efficiency of the algorithm is tested with Semantic Trajectory Anonymizing based on *k*-anonymity Model (STAKM) method (Tan et al., 2019) and Flip-Flop Exchange (FFE) method [Naghizade et al. (2014);Naghizade et al. (2020)] for the run time in similar test conditions, which are shown in Fig. 5 and Fig. 6. The evaluations are performed for the trajectories of 100 and 1000 users. The result shows that our method shows less run time than the STAKM method but more time than FFE method. This is because our method has a greater number of extractions, MBR zone creations and anonymizations. This does not imply that our algorithm is not performing well, but it uses four types of assessments within the trajectory and choosing appropriate methods for the right trajectories makes it very useful and dependable.

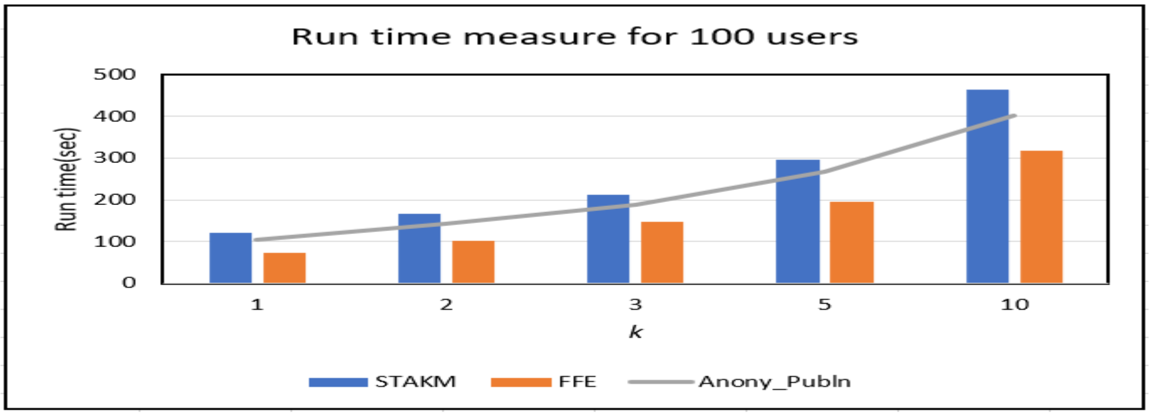


Fig. 5. Run time evaluation with 100 user's dataset

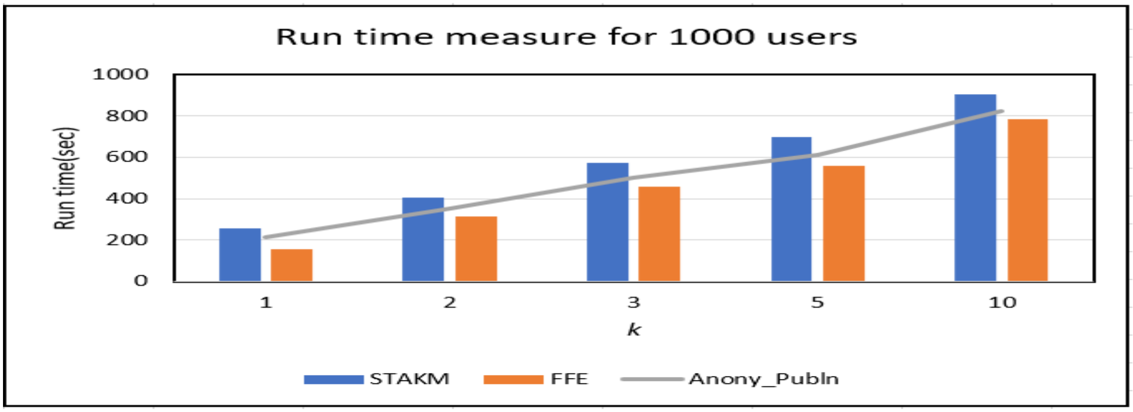


Fig. 6. Run time evaluation with 1000 user's dataset

5.3. *Measure of data utility*

The measure of data utility comprising two types of measures namely measure of information loss and measure of data error rate. For the purpose of evaluating information loss for this work (Anony_Publn), we have to measure the information loss for TemPert analysis and Anony_Gen analysis first. The Table 2. shows the extraction summary of various parameters used in this work.

No. of users	Total Location points	Total trajectories	Sensitive stay locations with SStLF >= δs	No. of less sensitive stay locations	No. trajectories possible with TemPert	No. trajectories possible with Anony_Gen	Total trajectories unaltered
1	1431	9	1	6	1	2	6
100	162018	717	108	360	45	117	555
1000	1302435	6996	1044	4212	366	825	5805

Table 2. Extracted parameters for the anonymization

The Table 3. shows the computation pattern of the information loss analysis of the TemPert algorithm. Since this approach alters only the temporal part, the actual information loss is the half of total information loss.

No. of trajectories with 1 sensitive and		Total points in the trajectory	Total points altered	Information loss	Actual information loss
1 less sensitive stay location	243	30618	2250	0.073486	0.036743
2 less sensitive stay locations	99	13662	1020	0.074660	0.037330
3 less sensitive stay locations	24	3384	255	0.075355	0.037678
Total	366	47664	3525		
Average information loss using TemPert				0.037250	

Table 3. Information loss summary for TemPert

The Fig. 7 gives an idea about the information loss occurred with the process of TemPert. The categorization of strategy of our approach results in fewer trajectories for TemPert analysis and some trajectories for the anonymization with Anony_Gen. More trajectories are left non-anonymized and these can be put into a publishable trajectory database. All the experiments are conducted here with a randomly chosen dataset size of 1000 users from T-Drive.

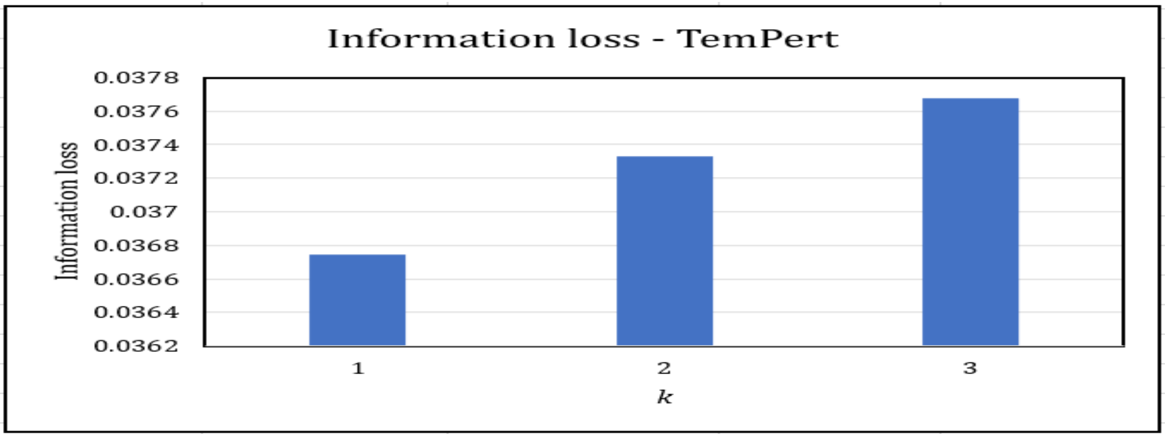


Fig. 7. Information loss measure for TemPert

The next analysis shown in Table 4. summarises the parameters needed for the evaluation of information loss for our next anonymization approach for the category consisting of the trajectories having sensitive stay locations only.

Privacy parameter, *k*	Anony_Gen– Average stay zone size(km^2)	Anony_Gen altered locations	Anony_Gen Info. loss
1	0.21	10163	0.080522
2	0.42	11220	0.088897
3	1.02	12432	0.098499
5	1.63	14718	0.116609
10	2.04	18306	0.145035

Table 4. Information loss summary for Anony_Gen (126225 location points)

Finally, information loss evaluation is done as concatenated analysis consisting of all categories of trajectories. For this purpose, we used the Eq. (4) and the summary of results is shown in Table 5.

Privacy parameter, *k*	STAKM - altered locations	FFE - altered locations	Anony_Publn - altered locations	STAKM - Info. loss	FFE - Info. loss	Anony_Publn - Info. loss
1	22672	27247	11288	0.017407	0.020920	0.008667
2	24214	28068	11730	0.018591	0.021550	0.009006
3	25347	29362	12560	0.019461	0.022544	0.009643
5	27069	31652	14718	0.020783	0.024302	0.011300
10	30178	34860	18306	0.023170	0.026765	0.014055

Table 5. Information loss summary for Anony_Publn

For the information loss calculation of Anony_Publn, we took half of the altered location tally from TemPert Information loss computation and added it along with the altered locations from Anony_Gen computation. The Fig. 8 shows average information loss measure for the analyses.

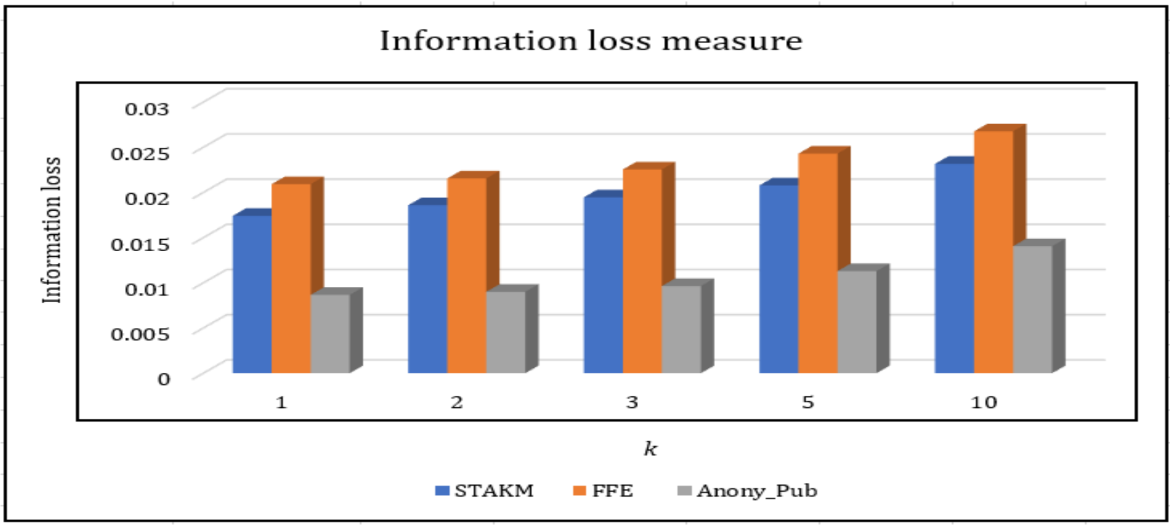


Fig. 8. Information loss measure for Anony_Publn

The Query error rate is another measure for evaluating the data utility during publishing of trajectories. For this we used the Eq. (5) and the experimental setup is prepared with spatial radius value (R_g) ranging from 300 to 4000 and the temporal range is between 1 to 4 hours (t_s and t_e). We experimented with this setup for 500 queries and each one having 500 trials. The evaluation result portrayed in Fig. 9 shows better results for our approach. An example for range query is shown as Q1 and Q2 for Possibly_Sometimes_Inside (P_S_I) and Defintely_Always_Inside (D_S_I). We ran both these queries for the input trajectory dataset IpTrDb and OpTrDb for calculating the query error rate.

Q1: select count(*) from IpTrDb where P_S_I(IpTrDb.$T^{u}{}_{j}$, R_g, t_s, t_e)
Q2: select count(*) from OpTrDb where D_A_I(OpTrDb.$T^{u}{}_{j}$, R_g, t_s, t_e)

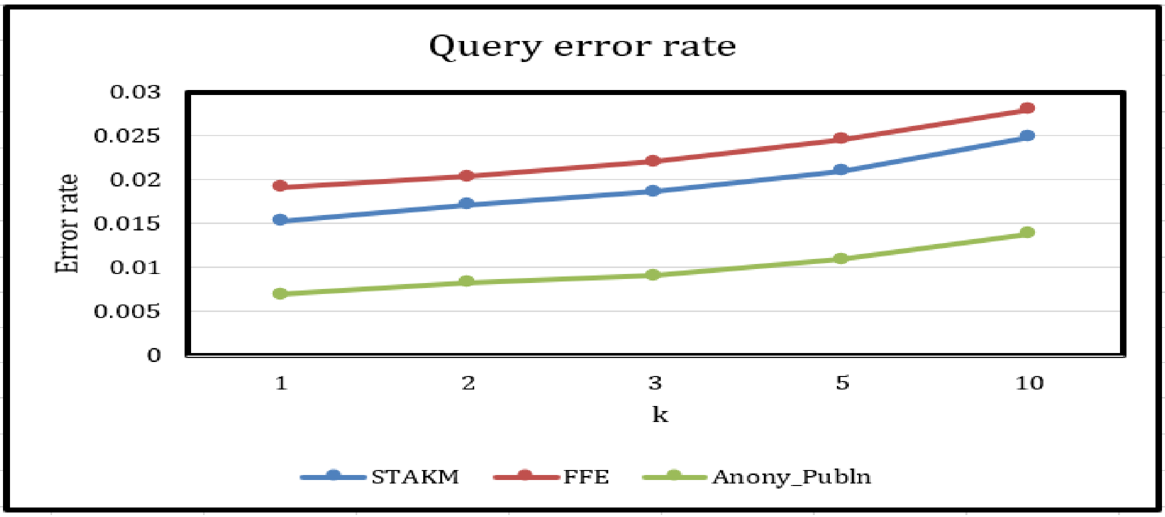


Fig. 9. Query error rate measure

The results show that our method has very less information loss and data error rate against the privacy parameter, *k* . In all the information loss analyses shows that our approaches had comparable results than the approaches we tested. This is because the categorization methodology that we applied to the user trajectories prevented the unnecessary anonymization of the stay points or stay locations within the trajectory. Also, the method *TemPert* has very little information loss in its spatial coordinates and no loss or error in the temporal part. Less anonymization results in high data utility and better privacy for the moving objects.

6. Conclusion

The advancements in the field of GPS embedded systems have resulted in a huge revolution in data collection and publication of these details are essential for the technological developments. The publishing of spatio-temporal data for the external users' needs a privacy preserved nature, because privacy of every individual is precious. The existing location privacy and trajectory privacy methods are not adequate enough to resist the privacy attacks. Some methods provide better privacy but more vulnerable to information quality and some others give less information loss but more privacy leakage. The researchers are developing new methods and approaches to balance this privacy gain and less information loss. In this scenario, our new *TemPert* method provides a better balance between privacy and information loss. The main advantage of *TemPert* is that it does not change any spatial coordinates. So, it is well suited for the applications which use spatial data more than temporal data. Moreover, the number of generalizations with POI for sensitive stay locations required in the dataset is few, so distortion for spatial as well as temporal part in the anonymized set is low. Some other trajectories also do not require any anonymization, which can directly be put into an anonymized set. The analysis also reveals that our method has clear advantage over the known approaches in terms of less information loss and query error rate. In future, we plan to tackle the privacy analysis and anonymization of real-time streaming of data and the inclusion of more semantic parameters for better extraction of sensitive stay locations.

Conflicts of Interest

The authors have no conflicts of interest to declare.

References

[1] Abul, O., Bonchi, F., Nanni, M. (2010). Anonymization of moving objects databases by clustering and perturbation. Inf. Syst. 35, 884–910. https://doi.org/10.1016/j.is.2010.05.003

[2] Abul, O., Bonchi, F., Nanni, M. (2008). Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases, in: 2008 IEEE 24th International Conference on Data Engineering. pp. 376–385. https://doi.org/10.1109/ICDE.2008.4497446

[3] Al-Hussaeni, K., Fung, B.C.M., Iqbal, F., Dagher, G.G., Park, E.G. (2018). SafePath: Differentially-private publishing of passenger trajectories in transportation systems. Comput. Netw. 143, 126–139. https://doi.org/10.1016/j.comnet.2018.07.007

[4] Bonchi, F., Lakshmanan, L.V.S., Wang, H. (Wendy) (2011). Trajectory anonymity in publishing personal mobility data. ACM SIGKDD Explorations Newsletter, Vol. 13, No. 1, pp. 30-42. https://doi.org/10.1145/2031331.2031336

[5] Bordenabe, N.E., Chatzikokolakis, K., Palamidessi, C. (2014). Optimal Geo-Indistinguishable Mechanisms for Location Privacy, in: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, CCS '14. pp. 251–262. https://doi.org/10.1145/ 2660267.2660345

[6] Chen, R., Acs, G., Castelluccia, C. (2012). Differentially private sequential data publication via variable-length n-grams, in: Proceedings of the 2012 ACM Conference on Computer and Communications Security, CCS '12. pp. 638–649. https://doi.org/10.1145/2382196.2382263.

[7] Chen, R., Fung, B.C.M., Mohammed, N., Desai, B.C., Wang, K. (2013). Privacy-preserving trajectory data publishing by local suppression. Inf. Sci., Data Mining for Information Security 231, 83–97. https://doi.org/10.1016/j.ins.2011.07.035

[8] Chow, C.-Y., Mokbel, M.F. (2011). Trajectory privacy in location-based services and data publication.

[9] Dai, Y., Shao, J., Wei, C., Zhang, D., Shen, H.T. (2018). Personalized semantic trajectory privacy preservation through trajectory reconstruction. World Wide Web 21, 875–914. https://doi.org/10.1007/s11280-017-0489-2

[10] Damiani, M.L., Bertino, E., Silvestri, C. (2009). Protecting location privacy against spatial inferences: the PROBE approach, in: Proceedings of the 2nd SIGSPATIAL ACM GIS 2009 International Workshop on Security and Privacy in GIS and LBS,. pp. 32–41. https://doi.org/10.1145/ 1667502.1667511

[11] Domingo-Ferrer, J., Sramka, M., Trujillo-Rasúa, R. (2010). Privacy-preserving publication of trajectories using micro-aggregation, in: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS, SPRINGL '10. pp. 26–33. https://doi.org/10.1145/1868470.1868478

[12] Domingo-Ferrer, J., Trujillo-Rasua, R. (2012). Micro-aggregation- and permutation-based anonymization of movement data. Inf. Sci. 208, 55–80. https://doi.org/10.1016/ j.ins.2012.04.015

[13] Dong, Y., Pi, D. (2018). Novel Privacy-preserving algorithm based on frequent path for trajectory data publishing. Knowl.-Based Syst. 148, 55–65. https://doi.org/10.1016/j.knosys.2018.01.007

[14] Eom, C.S.-H., Lee, C.C., Lee, W., Leung, C.K. (2020). Effective privacy preserving data publishing by vectorization. Inf. Sci. 527, 311–328. https://doi.org/10.1016/j.ins.2019.09.035

[15] Gambs, S., Killijian, M.-O., Núñez del Prado Cortez, M. (2014). De-anonymization attack on geolocated data. J. Comput. Syst. Sci., Special Issue on Theory and Applications in Parallel and Distributed Computing Systems 80, 1597–1614. https://doi.org/10.1016/j.jcss.2014.04.024

[16] Ghasemi Komishani, E., Abadi, M., Deldar, F. (2016). PPTD: Preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression. Knowl.-Based Syst. 94, 43–59. https://doi.org/10.1016/j.knosys.2015.11.007

[17] Gidófalvi, G., Huang, X., Pedersen, T.B. (2008). Privacy: preserving trajectory collection, in: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '08. pp. 1–4. https://doi.org/10.1145/ 1463434.1463491

[18] Gramaglia, M., Fiore, M., Tarable, A., Banchs, A. (2017). $k^{t, \varepsilon}$-anonymity: Towards Privacy-Preserving Publishing of Spatiotemporal Trajectory Data. ArXiv170102243 Cs.

[19] Gruteser, M., Xuan Liu (2004). Protecting privacy, in continuous location-tracking applications. IEEE Secur. Priv. 2, 28–34. https://doi.org/10.1109/MSECP.2004.1281242

[20] Gurung, S., Lin, D., Jiang, W., Hurson, A., Zhang, R. (2014). Traffic Information Publication with Privacy Preservation. ACM Trans. Intell. Syst. Technol. 5, 3, Article 44, 26 pages. https://doi.org/10.1145/2542666

[21] Han, P.-I., Tsai, H.-P. (2015). SST: Privacy Preserving for Semantic Trajectories, in: 2015 16th IEEE International Conference on Mobile Data Management. pp. 80–85. https://doi.org/10.1109/MDM.2015.18

[22] Hu, Z., Yang, J., Zhang, J. (2018). Trajectory privacy protection method based on the time interval divided. Comput. Secur. 77, 488–499. https://doi.org/10.1016/j.cose.2018.05.001

[23] Hua, J., Shen, Z., Zhong, S. (2017). We Can Track You if You Take the Metro: Tracking Metro Riders Using Accelerometers on Smartphones. IEEE Trans. Inf. Forensics Secur. 12, 286–297. https://doi.org/10.1109/TIFS.2016.2611489

[24] Huo, Z., Meng, X., Hu, H., Huang, Y. (2012). You Can Walk Alone: Trajectory Privacy-Preserving through Significant Stays Protection, in: Lee, S., Peng, Z., Zhou, X., Moon, Y.-S., Unland, R., Yoo, J. (Eds.), Database Systems for Advanced Applications, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 351–366. https://doi.org/10.1007/978-3-642-29038-1_26

[25] Jiang, K., Shao, D., Bressan, S., Kister, T., Tan, K.-L. (2013). Publishing trajectories with differential privacy guarantees, in: Proceedings of the 25th International Conference on Scientific and Statistical Database Management, SSDBM. ACM, pp. 1–12. https://doi.org/10.1145/2484838.2484846

[26] Li, M., Zhu, L., Zhang, Z., Xu, R. (2017). Achieving differential privacy of trajectory data publishing in participatory sensing. Inf. Sci. 400–401, 1–13. https://doi.org/10.1016/j.ins.2017.03.015

[27] Liu, B., Zhou, W., Zhu, T., Gao, L., Xiang, Y. (2018). Location Privacy and Its Applications: A Systematic Study. IEEE Access 6, 17606–17624. https://doi.org/10.1109/ ACCESS.2018.2822260

[28] Liu, X., Wang, L., Zhu, Y. (2018). SLAT: Sub-Trajectory Linkage Attack Tolerance Framework for Privacy-Preserving Trajectory Publishing, in: 2018 International Conference on Networking and Network Applications (NaNA). pp. 298–303. https://doi.org/10.1109/ NANA.2018.8648724

[29] Monreale, A., Trasarti, R., Pedreschi, D., Renso, C. (2011). C-safety: a framework for the anonymiza- tion of semantic trajectories, Transactions on Data Privacy, 4, pp. 73–101, http://www.tdp.cat/ issues11/tdp.a077a11

[30] Naghizade, E., Kulik, L., Tanin, E. (2014). Protection of sensitive trajectory datasets through spatial and temporal exchange, in: Proceedings of the 26th International Conference on Scientific and Statistical Database Management, SSDBM '14. pp. 1–4. https://doi.org/10.1145/ 2618243.2618278

[31] Naghizade, E., Kulik, L., Tanin, E., Bailey, J. (2020). Privacy- and Context-aware Release of Trajectory Data, ACM Trans. Spatial Algorithms Syst. 6, 1, Article 3, 25 pages, https://doi.org/10.1145/ 3363449.

[32] Nergiz, M.E., Atzori, M., Saygin, Y. (2008). Towards trajectory anonymization: a generalization-based approach, in: Proceedings of the SIGSPATIAL ACM GIS 2008 International Workshop on Security and Privacy in GIS and LBS, pp. 52–61. https://doi.org/10.1145/ 1503402.1503413

[33] Niu, B., Li, Q., Zhu, X., Cao, G., Li, H. (2014). Achieving k-anonymity in privacy-aware location-based services, in: IEEE INFOCOM 2014 - IEEE Conference on Computer Communications. pp. 754–762. https://doi.org/10.1109/INFOCOM.2014.6848002

[34] Poulis, G., Skiadopoulos, S., Loukides, G., Gkoulalas-Divanis, A. (2014). Apriori-based algorithms for k^m-anonymizing trajectory data. Trans. Data Priv. 7, 165–194.

[35] Qu, Y., Yu, S., Zhou, W., Tian, Y. (2020). GAN-Driven Personalized Spatial-Temporal Private Data Sharing in Cyber-Physical Social Systems. IEEE Trans. Netw. Sci. Eng. 7, 2576–2586. https://doi.org/10.1109/TNSE.2020.3001061

[36] Samarati, P., Sweeney, L., 1998. Generalizing data to provide anonymity when disclosing information, in: In Proc. PODS. p. 188.

[37] Soria-Comas, J., Domingo-Ferrer, J. (2012). Probabilistic k-anonymity through micro-aggregation and data swapping, in: 2012 IEEE International Conference on Fuzzy Systems. pp. 1–8. https://doi.org/10.1109/FUZZ-IEEE.2012.6251280

[38] Sui, P., Li, X., Bai, Y. (2017). A Study of Enhancing Privacy for Intelligent Transportation Systems: k -Correlation Privacy Model Against Moving Preference Attacks for Location Trajectory Data. IEEE Access 5, 24555–24567. https://doi.org/10.1109/ACCESS.2017.2767641

[39] Sweeney, L. (2002). k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10, 557–570. https://doi.org/10.1142/S0218488502001648

[40] Tan, R., Tao, Y., Si, W., Zhang, Y.-Y. (2019). Privacy preserving semantic trajectory data publishing for mobile location-based services. Wirel. Netw. https://doi.org/10.1007/s11276-019-02058-8

[41] Wang, K., Zhao, W., Cui, J., Cui, Y., Hu, J. (2019). A K-anonymous clustering algorithm based on the analytic hierarchy process. J. Vis. Commun. Image Represent. 59, 76–83. https://doi.org/ 10.1016/j.jvcir.2018.12.052

[42] Wang, S., Chen, C., Zhang, G., Xin, Y. (2019). Interchange-Based Privacy Protection for Publishing Trajectories. IEEE Access 7, 138299–138314. https://doi.org/10.1109/ ACCESS.2019.2942720

[43] Yuan, J., Zheng, Y., Xie, X., Sun, G. (2011). Driving with knowledge from the physical world, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11. pp. 316–324. https://doi.org/10.1145/2020408.2020462

[44] Yuan, J., Zheng, Y., Zhang, C., Xie, W., Xie, X., Sun, G., Huang, Y. (2010). T-drive: driving directions based on taxi trajectories, in: Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, GIS '10. pp. 99–108. https://doi.org/10.1145/1869790.1869807

[45] Zang, H., Bolot, J. (2011). Anonymization of location data does not work: a large-scale measurement study, in: Proceedings of the 17th Annual International Conference on Mobile Computing and Networking, MobiCom '11. pp. 145–156. https://doi.org/10.1145/2030613.2030630

[46] Zhao, X., Pi, D., Chen, J. (2020). Novel trajectory privacy-preserving method based on prefix tree using differential privacy. Knowl.-Based Syst. 198, 105940. https://doi.org/10.1016/ j.knosys.2020.105940

Authors Profile

Dr. Rajesh N., M. C. A., Ph. D. Currently working as an Associate Professor in the Department of Computer Applications, S. A. S. S. N. D. P. Yogam College, Konni, Pathanamthitta, Kerala, India under the affiliation of Mahatma Gandhi University, Kottayam, Kerala. He did his M. C. A. from University of Madras and Ph. D. in Computer Science from Mahatma Gandhi University, Kottayam. He has 23 years of undergraduate and 13 years of postgraduate teaching experience. He has now 7 years of Research experience especially in the field of Privacy Preserved Spatio-temporal Trajectory data mining and Publication. His research interests are Big data Management, Spatio-temporal data mining, E-learning, Data Analytics etc. He had published more than 15 papers in the various reputed International, National and State Journals and Conference proceedings and most of them were indexed by Scopus/ESCI/WoS. His Orcid-id is 0000-0003-0542-6994.

Dr. Sajimon Abraham, M. C. A., M.Sc. (Maths), M.B.A., Ph. D. (Computer Science). Currently working as a Professor in the Department of IT and Computer Science, School of Management and Business Studies, Mahatma Gandhi University, Kottayam, Kerala, India. He is also a Research Guide in the School of Computer Sciences, Mahatma Gandhi University. Under his research guidance, 6 research scholars were awarded with Ph. D. and 9 others doing their research. His main research areas are Privacy preserved data mining, Big data Analytics, Spatio-temporal data mining, E-learning, Unstructured data mining etc. He had published over more than 120 papers in the various reputed International, National and State Journals and Conference proceedings and most of them are indexed by Scopus/ESCI/WoS.

Dr. Nishad A., M. C. A., MTech. (CSE), Ph. D. (ORCiD: 0000-0002-8942-9166). Presently working as a Faculty in Computer Science under the Department of Higher Secondary Education, Government of Kerala, India. He has completed the research from Mahatma Gandhi University, Kottayam, Kerala. His area of research includes Bigdata Analysis, Moving Object Data Mining and Trajectory Clustering, Privacy preserved data mining. He had published more than 12 papers in top indexed international and national journals and conference proceedings.

Dr. Lumy Joseph, M. C. A., M.Phil. (CS), Ph. D. She is an Associate Professor in the Department of Computer Applications at Marian College, Kuttikkanam (Autonomous), in Kerala, India. She received her doctorate in “An Intelligent E-Learning Environment for Enhancing Learner Performance” from Mahatma Gandhi University, Kottayam. Learning analytics, machine learning, educational data mining, and e-learning, Privacy preserved data mining are some of her research areas. She has many articles in international journals and conference proceedings.

Dr. Benymol Jose, M.Sc. (CS)., M.Phil. (CS), Ph. D. She did her Ph D. in Computer Science from Mahatma Gandhi University, Kottayam on the topic “Unstructured Data Mining in Bigdata: A NoSQL Perspective”. She is currently working as Associate Professor in Department of Computer Applications, Marian College, Kuttikkanam, Idukki, Kerala, India. She had published many papers in journals and conference proceedings including IEEE, Elsevier, and ACM. Her main research work focuses on unstructured data mining, NoSQL databases and Bigdata Analytics, Privacy preserved data mining. She has 23 years of teaching experience and 7 years of Research Experience.