Privacy-protected statistics publication over social media user trajectory streams

Shuo Wang a,*, Richard Sinnott a, Surya Nepal b

a Computing and Information Systems, The University of Melbourne, Australia
b Data61, CSIRO, Australia

HIGHLIGHTS

• A private statistics publication approach is proposed for online trajectory stream.
• An anchor discovering and segmenting model is used to relieve heterogeneity issue.
• An adaptive privacy budget distribution mechanism is used in w-step sliding window.
• A private KNN selection model is used in improved multi-timestamp prediction.

ARTICLE INFO

Article history:
Received 13 January 2017
Received in revised form 20 July 2017
Accepted 1 August 2017
Available online xxxx

Keywords:
Location privacy
Differential privacy
Social media
Multi-timestamp prediction
Stream aggregate publication

ABSTRACT

An increasing amount of user location information is being generated due to the widespread use of social network applications and the ubiquitous adoption of mobile and wearable technologies. This data can be analysed to identify precise trajectories of individuals — where they went and when they were there. This is an obvious privacy issue, yet publication of real-time aggregate over such location streams can provide valuable resources for researchers and government agencies, e.g., in case of pandemics it would be very useful to identify who might have come into contact with an infected individual at a given time. Differential privacy techniques have become popular and widely adopted to address privacy concerns. However, there are three key issues that limit the application of existing differential privacy approaches to user trajectory data: (a) the heterogeneous nature of the trajectories, (b) uniform sliding window mechanisms do not meet individual privacy requirements and (c) limited privacy budgets and impact on data utility when applied to infinite data streams. To tackle these problems, this paper proposes a private real-time trajectory stream statistics publication mechanism utilizing differential privacy (DP-PSP). To relieve the heterogeneity issues, anchor point discovery (e.g., fixed locations like museums, parks, etc.) and road segmenting mechanisms are proposed. We provide an adaptive w-step sliding window approach that allows users to specify their own dynamic privacy budget distribution to optimize their own privacy budget. To preserve the data utility, we present multi-timestamp prediction models and private k-nearest neighbour selection and perturbation algorithms to reduce the amount of perturbation distortion induced through the differential privacy mechanism. Comprehensive experiments over real-life location-based social network user trajectories show that DP-PSP provides private data aggregate over infinite trajectory streams and boosts the utility and quality of the perturbed aggregation without compromising individual privacy.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

With the improvement of sensing technologies and widespread popularity of mobile devices with location-aware capabilities, it is now possible to harvest, store, analyse and publish user locations and their movements (trajectories) in real-time. The publication of harvested (raw or processed) data offers an unprecedented opportunity to gain insights into people’s movements that can be used in many applications such as social meetings (e.g., dating), urban planning, traffic management, managing emergency situations (e.g., earthquake, fires, etc.), and targeted marketing. In some applications like urban planning, the data is collected and analysed offline; whereas some applications like traffic management require data to be collected, analysed and published in real-time.
Such real-time applications require processing of trajectory data in the form of a continuous stream. For example, real-time data from location-based social networks like Foursquare and Tinder can be used for social meetings; user trajectory streams produced by the spatial–temporal data from applications such as Twitter can be used to estimate the historic and current road and traffic conditions and possible traffic jams and accidents that are occurring; mining trajectory stream statistics helps urban planners to efficiently use existing transport networks and support optimal route computations. However, there are inherent challenges in using and publishing user trajectory stream data, caused by the heterogeneity of trajectories and most importantly, user privacy concerns.

The publication of user trajectory streams provides opportunities for new innovation. However, trajectory data is often personal and sensitive and can reveal the successive user (spatial) locations combining with timestamps. Publication of such data may compromise individuals’ privacy (e.g., home location, political views or categories of disease based on their visited locations). Hence, there is a requirement for a privacy framework that can deal with individual’s privacy needs without compromising the data utility. It is essential that personal and sensitive information is not leaked from released statistics results, while maintaining the statistical significance of the perturbed data. In recent times, a robust privacy preserving paradigm, differential privacy [1] has been implemented to protect the privacy of sensitive trajectory data aggregate releases. The sensitive individuals’ information can be perturbed in aggregate before publishing the statistics through differential privacy framework. Using \( \varepsilon \)-differential privacy, the change of the released outcomes is guaranteed to be negligible (according to the privacy budget \( \varepsilon \)) by removing or changing any single individual attendance in the database.

However, the application of differential privacy to protect privacy of trajectories stream data in real-time is not straightforward and many open challenges remain. First, there is an inherent heterogeneity in trajectories as shown in [2] that has a negative impact on the effectiveness of trajectory similarity measures. Second, the uniform sliding window mechanism cannot meet personal user privacy requirements and a uniform privacy budget distribution is ineffective. The privacy budget and utility is also limited in existing \( w \)-event models for infinite stream publication, which requires effective and adaptive privacy budget distribution.

**Contributions.**

To tackle these problems, a novel private statistics publication framework for real-time trajectory streams under differential privacy (DP-PSP) is proposed. It is based on a variable length sliding window mechanism, called \( w \)-steps privacy sliding window. We propose algorithms to realize the proposed framework and prove that our approach not only satisfies the differential privacy needs but also provides increased utility. This framework is composed of three sub-algorithms that solve challenges faced by existing infinite stream statistics release schemes as outlined above. These include:

(a) Novel anchor point clustering and road segment mechanisms: A novel sensitive anchor point clustering method is proposed to discovery the feature-based sensitive anchor points considering both density and features instead of all locations in the trajectory database. We also use a road segment mechanism based on sensitive anchor points for road network segments to handle the heterogeneity issues of the trajectory data.

(b) Adaptive privacy budget distribution for adaptive \( w \)-steps sliding windows: We propose an adaptive \( w \)-steps sliding window approach to allow users to specify their own length. In addition, an adaptive privacy budget distribution mechanism is adopted for flexible and dynamic budget allocation. We propose a novel private stream statistics publication algorithm that skips the perturbation and releases stage for timestamps that can be accurately predicted by private \( k \)-nearest neighbour models. In our approach, there is no perturbation on skipped timestamps, which can save privacy budgets for future perturbation and release.

(c) Novel multi-timestamp prediction algorithm: We propose a multi-timestamp prediction model along with private \( k \)-NN selection and perturbation algorithm to approximate the perturbed statistics results. The basic idea is to increase the number of perturbation skips for timestamps whose perturbed statistics can be predicted well, based on given privacy guarantees. Specifically, a private neighbour selection and perturbation mechanism is proposed to privately select neighbours for use in differential privacy. This mechanism can enhance the prediction accuracy through adopting truncating mechanisms and adjusting post-processing, which can identify high quality neighbours with differential privacy guarantees.

Finally, experiments over real social media user trajectories are conducted, comprising spatial-tagged Twitter data harvested on major national Cloud facilities in Australia to show the efficiency and improved utility of DP-PSP. The rest of this paper is organized as follows. Section 2 describes the related work, focusing especially on differential privacy of trajectory and location data and the adoption of differential privacy compared to other approaches used for stream data. Section 3 introduces the preliminary concepts and background approaches used in the work. Section 4 presents the differentially private statistics publication for real-time trajectory streams. Section 5 presents the evaluation metrics that have been used and the experimental results of the DP-PSP using real-life stream data. Finally, Section 6 draws conclusions and outlines areas of future work.

### 2. Related work

To protect the individual privacy of location data, several differential privacy based solutions have been proposed in [3–6]. However, most current data release approaches under differential privacy are based on one-time static data publication, e.g., [2,7,8]. Differential privacy has been used for stream data, which can be divided into user-level (preserve the privacy of all of the individual visited location) and event-level (protect a single visited location). Existing works mainly focus on infinite streams event-level privacy [9–11], and finite streams user-level privacy [12]. Few recent works [9,13,14] adopted \( w \)-event privacy or similar models to release infinite stream data under differential privacy. [9] used a \( w \)-event mechanism for infinite stream release, but the uniform \( w \)-event model cannot meet the personal privacy requirements. Fan et al. [15] proposed the FAST framework for publishing time-series data at a user-level. FAST uses sampling and filtering components to reduce the noise; given a specified number of samples, the filtering component predicts the future data and corrects its prior data by noisy samples. The authors report that their adaptively sampling scheme preserves high utility at the same privacy level. However, this scheme takes the total amount of timestamps \( T \) as input, which is unsuitable for infinite stream scenarios.

[13] proposed a flexible privacy model of \( l \)-trajectory privacy to ensure every desired length of trajectory for stream agglomeration publication, using dynamic budget allocation and approximate publishing to reduce the privacy cost. However, the approximate mechanism adopted in current works is single timestamp prediction, which is inefficient both in accuracy and computing speed. Most of these works directly use an exponential mechanism; this means the prediction accuracy will be limited as the random selection may choose some inaccurate candidates with lower scores with higher probability. Furthermore, existing works...
do not consider the high inherent heterogeneity of trajectories, as shown in [2]. The heterogeneity has a negative impact on the effectiveness of trajectory similarity measures. Applying coarse and finer-grained sampling approaches leads to significantly different results.

To the best of our knowledge, this paper is the first work to support a real-time infinite trajectory stream statistics publication by using multi-timestamp prediction mechanisms, variable $w$-steps privacy and dynamic privacy distribution to boost the data utility and accuracy, with the heterogeneous nature of user trajectories taken into account.

3. Background

This section introduces the preliminary concepts and background approaches used in following parts. The first three parts describe the definition of differential privacy, $k$-nearest neighbour model and its implementation for prediction of traffic conditions as well as for multi-timestep predictions. The last section presents the motivation and problem to be solved.

3.1. Differential privacy

Differential privacy was proposed by Dwork et al. in [1]. It is based on the idea that valuable knowledge can be gained from datasets (including individual identifying information), while this sensitive data can be protected. It offers rigorous privacy assurances that one individual cannot be recognized whenever this individual is in or is deleted from the dataset, namely the results will not change much.

The formalized definition of differential privacy is that if an individual is deleted from a database, there is no output that will not change much. Specifically, a private function $F$ with $\varepsilon$-differential privacy for databases $D_1$ and $D_2$, differing at most one element from each other, satisfies differential privacy if for all outcomes of the database $S (S \subset Range(F))$ there is: $Pr[F(D_1) \in S] \leq e^{\varepsilon} Pr[F(D_2) \in S]$. Differential privacy has two important properties: (1) sequential composition. Differential privacy provided by a set of mechanisms $M_i$ on an input set $D$ is $\sum_{i=1}^{|D|}$ parallel Composition. If every mechanism $M_i$ acts on a disjoint subset $D_i \subset D$, the privacy provided will be $\max(\varepsilon_i)$-differential privacy for all $M_i$. (2)

3.2. $k$-nearest neighbour model and prediction of traffic conditions

The $k$-NN model is an efficient statistical mechanism that can be used for short-term aggregate value prediction. The historical states can be used as a sample database, which is described by an appropriate vector space. K-sample records can be discovered between the current record and all samples based on the Euclidean distance, and the future states can be predicted by averaging $k$-NNs. The main advantage of $k$-NN is that it is based on the history of published data and parameter $K$ without any mathematical models. However the results of the native $k$-NN models are generally hysteric for time series streams and the forecasting accuracy is limited due to their oversimplified nature [16,17]. Usually, a correlation coefficient weighting mechanism [18] can be adopted to figure out the distance similarity:

$$d_{a,b} = ||a, b||_2,$$

where the $d_{a,b}$ is the distance between states at $a$ and $b$.

A prediction algorithm is used to demonstrate how the $k$-NN model predicts the future states.

$$S_m(t + 1) = \sum_{g=1}^{k} \frac{d_{g,m}^{-1}}{\sum_{g=1}^{k} d_{g,m}^{-1}} S_g,$$

where $S_m(t + 1)$ is the statistics value at timestamp $t + 1$ on choosing spatial location $m$ and $S_g$ is the predicted statistics value of nearest neighbours $g$ searched in the history database.

Many existing works have successfully applied the $k$-NN model in short-term traffic-flow predictions. [19] combined the condition-monitoring methods with the $k$-NN model to improve the accuracy of atypical traffic conditions. [20] used $k$-NN non-parametric regression to build a dynamic multiple-interval traffic-flow prediction model that could recognize the rich information of historical data. [21] confirmed that the $k$-NN non-parametric regression model was superior to Kalman filtering and seasonal auto-regressive integrated moving average (ARIMA) in terms of both prediction accuracy. These studies demonstrated that the $k$-NN model performs well in short-term traffic-flow forecasting.

3.3. Multi-timestamps $k$-NN prediction model

Existing spatial and temporal relations in road networks helps to predict the current road section based on upstream and downstream road segments. Existing works mainly focus on short-term traffic state prediction based on temporal correlations, instead of combining temporal correlations and spatial correlations [18,22–24]. However, most of these works focus on a single timestamp prediction model and the distance similarity evaluation metrics do not include any weight assignment strategies. In this paper, we build a segmenting mechanism using anchor points on road networks and propose an improved multi-timestamp prediction model considering all connected segments that could potentially impact the current segment states. Our approach provides an extension of the single-timestamp prediction based on the structure of the road network.

Fig. 1 shows a common multi-timestamp $k$-NN prediction model to predict successive $w$ timestamps states for each segment. First, the traffic states of the current segment $m$ and all connected segments to it are harvested for the prediction. Secondly, a $k$-NN prediction model is built to calculate the $k$ nearest neighbours for $m$ and all $k$-NNs for the entire connected segment candidates for $m$. Finally, the prediction $\hat{S}_m(t + 1), \hat{S}_m(t + 2), \ldots, \hat{S}_m(t + n)$ is derived for the current segment $m$.

This prediction model takes the temporal correlation and all connected segments that could potentially affect the current segment states (e.g., upstream and downstream flows). It can be divided into a temporal prediction, a connected segment prediction and the entire connected segment prediction. In the temporal prediction, the $w$-steps successive statistics prediction $\hat{S}_m(t + 1), \hat{S}_m(t + 2), \ldots, \hat{S}_m(t + w)$ can be obtained based on $w$-steps successive statistics after the $k$-NNs of the current states vector $S_m(t - n), S_m(t - n + 1), \ldots, S_m(t)$ is computed. For the connected segment prediction, given two connected segments of the current segment, e.g., the upstream and downstream segments, the $w$-steps successive statistics prediction $\hat{S}_m(t + 1), \hat{S}_m(t + 2), \ldots, \hat{S}_m(t + w)$ for the current segment $m$ can be obtained based on $w$-steps
successive statistics after the $k$-NNs of current states matrix is computed.

$$
\begin{bmatrix}
S_{m-1}(t-n) & S_{m-1}(t-n+1) & \cdots & S_{m-1}(t) \\
S_{m}(t-n) & S_{m}(t-n+1) & \cdots & S_{m}(t) \\
S_{m+1}(t-n) & S_{m+1}(t-n+1) & \cdots & S_{m+1}(t)
\end{bmatrix}.
$$

(3)

In the entire connected segment prediction, the $w$-steps successive statistics prediction $\hat{S}_w(t + 1), \hat{S}_w(t + 2), \ldots, \hat{S}_w(t + w)$ for the current segment $m$ can be obtained based on $w$-steps successive statistics after the $k$-NNs of the current states matrix is computed. This includes all previous states of segments directly or indirectly connected to $m$.

The traffic statistics states for multi-segments can be demonstrated by a spatio-temporal matrix $ST(m, n)$, where $m$ is the number of timestamps ($m - 1$ is the previous timestamp) and $n$ is the number of segments ($n - 1$ is the nearby segment). The records $v_{ij}$ in $ST(m, n)$ represent the statistics value (i.e., the number of users) at the $i$th timestamp on segment $j$.

$$
ST(m, n) = \begin{bmatrix}
v_{11} & v_{12} & \cdots & v_{1n} \\
v_{21} & v_{22} & \cdots & v_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
v_{m1} & v_{m2} & \cdots & v_{mn}
\end{bmatrix}
$$

(4)

As described above, $k$-NN model can be used for single-time-step prediction that predicts traffic conditions on the target road link of the next one-time interval as well as multi-time-step prediction that aims to predict the traffic condition of more than one time interval, i.e., the extension of the single-time-step prediction. For short-term traffic condition prediction, the later can generate prediction results of longer time. An improved multi-time-step $k$-NN prediction takes the possible connected segments into account to figure out multi-step prediction for multi-segments at the same time. The details are given in Section 4.2.

3.4. Problem statement

The problem to be solved by this work is based on the following scenario where an honest-but-curious server collects data $D_t$ of individuals’ spatio-temporal information continuously at each timestamp $t \in [1, T]$, as shown in Fig. 2(A). There is a requirement to publish a statistics of time series streams that can be generated from an individuals’ spatio-temporal data stream as illustrated in Fig. 2(B). To address the concerns of individual’s privacy, the server needs to perturb the raw stream statics time series data before release, i.e., to generate noise-enhanced statistics under differential privacy, as demonstrated in Fig. 2(C).

The relevant definitions used to solve this problem are given as follows.

Definition 1 (Statistics of Time Series Stream). Given $D_t$ comprises individual spatial–temporal data consisting of locations at contiguous timestamps $t \in [1, T]$ harvested by trusted servers as shown in Fig. 2(A). Here each row is a unique users’ profile. Statistics of time series streams $Q^{dnl} = \{Q(D_1), \ldots, Q(D_T)\}$ can be generated based on different unique user profiles, where $|L_t|$ is the number of sensitive locations or areas needing to be monitored, and $Q(D_t)$ are counts (the number of non-zero values in each column) of $loc \in L_t$ in $D_t$ at timestamp $t$, as shown in Fig. 2(B). Fig. 2(C) shows the perturbed published statistics results obtained by adding noise.

For the infinite data stream $S = \{D_1, D_2, \ldots, D_t\}$, the trusted server harvests real-time updates at discrete timestamps. At each timestamp $t$, the update is $D_t$ with $|L_t|$ columns (each column is a location, i.e., an event) and numerous rows for different users represented by $(userid, location, time)$. If the user appeared in location $loc \in L_t$ (corresponding column is $j$) at $t$, then the event $j$ happens with value 1, and 0 otherwise.

Definition 2 (Variable $w$-steps Sequence). The $w$-steps sequence is the successive locations sequence visited by one user $u$ with length $w$, denoted by $u^w = \{(u, loc_1, t), \ldots, (u, loc_w, v)\}$ where $loc_t$ and $loc_{ct}$ are $L_t$ at timestamp at which the $w$-steps sequence ends. If the length of the $w$-steps sequence is flexible for different user, then it can be called variable $w$-steps sequence.

Definition 3 (Neighbouring Databases at Timestamp $t$). Given a database $D_t$ harvested at timestamp $t$, and $D^*_t$ is a copy of $D_t$ but removing or adding a row from $D_t$. Then $D_t$ and $D^*_t$ are neighbouring databases. Note that we assume one single individual only has at most one location at every timestamp, hence the sensitivity of $Q(D_t)$ is 2, as modifying single user in Neighbouring Databases will cause the maximum change of $Q(D_t)$ by 2. For $w$ timestamps in $w$-sliding windows, the total sensitivity is $2 + w$.

Definition 4 ($w$-steps Sequence $\epsilon$-differential Privacy). Given an algorithm $O$ whose input is $S_t = \{D_1, D_2, \ldots, D_t\}$ with possible perturbed output $P_t = \{p_1, p_2, \ldots, p_T\}$. If there is $Pr(O(S_t) = P_t) \leq Pr(O(S^{*}_t) = P_t)$ for any $w$-steps Sequence Neighbouring Stream Prefixes $S_t$ and $S^{*}_t$, then $O$ satisfies $w$-steps Sequence $\epsilon$-differential Privacy ($w$-steps privacy in short).

4. Differentially private statistics publication for trajectory streams

Fig. 3 shows our proposed differential privacy-oriented approach (DP-DSP) framework for protection of individual privacy of real-time user trajectories aggregate data publication. The DP-DSP can publish real-time aggregate statistics at the current timestamp with differential privacy guarantees. We describe the key steps in our framework. Steps 1 and 2 correspond to anchor point discovery and road segmenting mechanisms. Based on the archived location data, a Density Joinable Cluster (DjCluster) [25] is used by the trusted server to cluster the locations and extract the anchor points (step 1). Using the discovered sensitive anchor points, the road network is divided into a number of segments. Each segment is represented by its start and end points; each spatial location in trajectory stream is calibrated to the nearest anchor point (step 2). A trusted server harvests and then aggregates real-time statistics of users’ trajectories continuously, consisting of spatial–temporal locations (step 3). In Step 4, adaptive privacy budget distribution is used to assign a privacy budget for the $w$-step sliding window publication mechanism. Step 5–8 correspond to multi-timestamp prediction models and private $k$-NN selection and perturbation algorithms. The following step is used to generate the connected segment to find all directly or indirectly connected segments for the current segment that potentially affect the state of the current timestamp (step 5). In addition, an improved multi-timestamp prediction model is proposed based on a Gaussian weighted Euclidean distance for similarity evaluation (step 6). Next, a private $k$-nearest neighbour selection and perturbation algorithm (PKSP) under differential privacy is built. This is achieved in two steps. The first step is to select $k$-NNs states for the current timestamp with differential privacy guarantees using exponential mechanisms (step 7). Following this, the private aggregate perturbation (step 8) and publication (step 9) are applied.

4.1. Discovering anchor point and segmenting

The private road segments construction can be achieved through three steps, e.g., private anchor points discovery, connected segments equivalent distance calculation and segments weight assignment.
4.1.1. Sensitive anchor points discovery

The first step is sensitive anchor points discovery by optimized DJ-Clustering mechanism. DJ-Cluster (density-joinable) \[25\] uses connectivity notion of connected components instead of the clique graph used in DBSCAN. The density-based neighbourhood $NB$ of a location $p$, namely $NB(p)$, is defined by $NB(p) = \{x \in P \mid \text{dist}(x, p) \leq \text{Eps}\}$ where $P$ is the entire set of locations and $\text{Eps}$ is the distance threshold. If there is a point $x$ such that both $NB(p)$ and $NB(q)$ contain $x$ then $NB(p)$ is density-joinable to $NB(q)$, namely $J(N(p), N(q))$. The DJ-Cluster, namely density and join-based cluster DC, is defined by $\forall p \in P$ and $\forall q \in P$. In order to improve the utility of anchor points, the density-connected relation is modified to coherence-connected relations, namely $NB(p) = \{x \in P \mid \text{dist}(x, p) \leq \text{Eps}\}$ is modified to

\[
NB(p) = \{x \in P \mid \text{dist}(x, p) \leq \text{Eps} \&\& \text{coherence}(x, p) \leq \mu\} \tag{5}
\]

where $\mu$ is the coherence threshold and $\text{coherence}(x, p)$ is given as:

\[
\text{coherence}(p_1, p_2) = \exp\left(-\frac{\text{ED}(p_1, p_2)}{\delta}\right)^\alpha |\sin\theta|^\beta. \tag{6}
\]

Here, $\alpha$ and $\beta$ are tuning parameters, $\text{ED}$ is the Euclidean Distance, and $\delta$ is a scaling factor [26]. In addition, if the $\text{coherence}(p_1 \text{ and } p_2)$ is greater than the threshold $\mu$, we can say they are directly coherence-connected, denoted by $p_1 \text{ } \square \text{ } p_2$. If the angle $\theta$ between $p_1$ and $p_2$ is in $(0, \pi)$, there is a potential intersection (it is a T-junction or cross-roads). If $\theta$ tends to $\pi/2$, there is a higher probability of finding an intersection.

Based on sensitive anchor points discovery, the road network can be divided into various segments. Each segment can be represented by its start and end points. Every spatial location in a trajectory stream will be calibrated to the nearest anchor point.

4.1.2. Generating connected segments

The traffic states of the current segments can be affected by the connected segments, e.g., upstream and downstream segments. The spatio-temporal connection correlation is related to the road network structure and the characteristics among the time series streams [24]. The road network can be segmented by sensitive anchor points for traffic state statistics. Based on these segments, the spatio-temporal connection correlation is built, as shown in the following figure. The segment can be represented by its midpoint. The connection correlation can be divided into direct connections and indirect connections using the different grades in Fig. 4. Here $g = 1$ (star) represents the current segment, $g = 2$ (circle) contains
all direct connected segments, \( g = 3 \) (triangles) corresponds to all direct connected segments for \( \text{grade} = 2 \), and the process continues for grades 4, 5, \ldots, \( N \). The correlation coefficient is adopted and the equivalent distance is used to recognize the connection correlation between the current segment and other related segments instead of only the artificial grade shown in Fig. 4.

Given two segments \( a \) and \( b \), \( A = \{a_1, a_2, \ldots, a_n\} \) and \( B = \{b_1, b_2, \ldots, b_n\} \) are two historical time series at \( n \) timestamps. The correlation coefficient between the historical time series of two segments is defined as follows:

\[
 r = \frac{\text{Cov}(A, B)}{\sqrt{\text{Var}(A) \times \text{Var}(B)}} = \frac{\sum_{i=1}^{n} (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n} (a_i - \bar{a})^2 \times \sum_{i=1}^{n} (b_i - \bar{b})^2}}. \tag{7}
\]

Here the \( \bar{a} \) and \( \bar{b} \) is the average value of \( A \) and \( B \) respectively. The equivalent distance \( eDis \) is given as follows:

\[
eDis = \sqrt{\frac{1}{g^2} \cdot (\text{ED}))}, \tag{8}
\]

where \( ED \) is the Euclidean distance between the midpoints of the compared segments, \( g \) the connective grades of the segments and \( r \) the correlation coefficient. The connected segments generation is demonstrated in the following algorithm:

![Algorithm 1: Optimized DJ-Cluster.](image)

**Algorithm 1:** Optimized DJ-Cluster.

4.2. Improved multi-timestamp prediction

Common single-timestamp \( k\text{-NN} \) prediction models cannot be used for this scenario, as they can only predict multi-segments individually, which is computationally expensive and time consuming. Hence, we propose a novel multi-timestamp \( k\text{-NN} \) prediction model with improved accuracy, as well predicting the states for the next multi-timestamp at a given time.

Generally, the Euclidean distance is used as the states similarity metric between the current state and other states to find \( K \) nearest neighbours. However, it cannot evaluate changing trends of these states. Therefore, a Gaussian weighted Euclidean distance [27] is adopted to measure the similarity between two states. This metric assigns different weights for spatio-temporal states based on a Gaussian function using a space-weighted matrix \( M_s \).

\[
 M_s = \begin{bmatrix}
 w_{s_1} & w_{s_2} & \cdots & w_{s_n} \\
 w_{i_2} & \cdots & \cdots & \cdots \\
 \vdots & \ddots & \ddots & \cdots \\
 w_{i_n} & \cdots & \cdots & \cdots
\end{bmatrix}, \tag{9}
\]

where \( w_{si} = \frac{1}{4\pi \sigma_i^2} \exp \left( -\frac{(\text{eDis}(c, s_i))^2}{4\sigma_i^2} \right), \ i \in \{1, n\} \).

Here, the \( \text{eDis}(c, s_i) \) is the equivalent distance between the current segment and the \( p\text{th} \) segment, and \( \sigma_i \) is a space weighted parameter. Given the current spatio-temporal matrix \( ST(m, n) \) and \( p\text{th} \) spatio-temporal matrix \( ST_p(m, n) \), the spatio-temporal state similarity \( SS \) among these states can be obtained by the Gaussian weighted Euclidean distance by:

\[
 SS_{c,p} = \| (ST(m, n) - ST_p(m, n)) \times M_s \|_2. \tag{10}
\]
Based on the Gaussian weighted Euclidean distances calculated by the above mechanism, the $K$ nearest neighbours $SK(m)$ can be selected for multi-segments that are recognized as the connected segments by the previous step from the published location data, according to step 7 in our framework.

After constructing the $SK(m)$, instead of using the mean value of $k$-NNs as the predicted value, we adopt a Gaussian function weight assign mechanism for $k$-NNs. The distribution of weights to $q$th $k$ – $NN$s done as follows:

$$\gamma_q = \frac{1}{4 \pi a^2} \exp \left( -\frac{||S_m||^2}{4a^2} \right),$$

(11)

where $a_2$ is the adjustment weighted parameter. The predicted value for the following $l$th timestamps state $ST_{m,l}$ is defined as:

$$ST_{m,l} = \left( \sum_{q=1}^{k} SK(q, l)\gamma_q \right) / \left( \sum_{q=1}^{k} \gamma_q \right), \quad q \in [1, k],$$

(12)

where $SK(q, l)$ is the $l$th successive statistics for the $q$th $k$ – $NN$ of segment $m$.

For instance, in the $w$-steps sliding window scenario from the $k$th timestamp, there is a segment $l$, at timestamp $t_l$ required to predict the next $w-1$ timestamps (e.g., $t_{l+1}$, $t_{l+2}$, ..., $t_{l+w}$) states. Here, the connected segments candidates for $l_0$ are discovered from the connected segments step generation, e.g., $l_1, l_2, ..., l_{w-1}$. The spatio-temporal state matrix $ST_{m,n}$ based on the states of $n$ previous successive timestamps is derived as follows:

$$ST(m, n) = \begin{bmatrix}
        t_{l_{m-1}} & l_1 & l_2 & \cdots & l_n \\
        t_{l_{m-2}} & v_{m1} & v_{m2} & \cdots & v_{mn} \\
        \vdots & \vdots & \vdots & \ddots & \vdots \\
        t_{l_i} & \vdots & \vdots & \cdots & \vdots \\
    \end{bmatrix},$$

(13)

Following this we can use the above $k$-NNs prediction mechanism to select the $k$ nearest states of $m$, i.e., $SK(l_0)$, based on the Gaussian weighted Euclidean distances mechanism, and the predicted values $ST_{m,i} = S_k(t + i)$, $i \in [0, w - 1]$. This can be obtained using Eq. (10) where the $SK(q, i)$ is the $i$th successive value of the last column in $q$th $k$-NN matrix. In a similar manner, predictions of all segments can be computed.

4.3. Adaptive privacy budget distribution

Given a privacy budget $\epsilon$ assigned for the $w$-steps sliding window mechanism, this can be divided into $w$ parts at each timestamp. At timestamp $t$, the privacy budget can consist of fixed budget $\epsilon_{i,1} = \epsilon/2$ for private $k$-NNs selection and prediction to decide whether the statistic at $t$ can be republished by the prediction, and an adaptive budget for private publication with $\epsilon_{i,2} = \epsilon/2$. For the private $k$-NNs selection, there are two steps required to access the original real values, i.e., private selection and private error comparison. Therefore, the $\epsilon_{i,1}$ should be divided into $\epsilon_{i,1} = \epsilon_{i,1}/2/k$ for each round of private selection and $\epsilon_{i,1} = \epsilon_{i,1}/(2/w)$ for the private error comparison for each timestamp. Note that the $w$-steps prediction can be achieved by executing only one time prediction on the first timestamp of each $w$-sliding window.

For the private publishing, an adaptive budget saving mechanism is adopted in an exponential decay fashion. Firstly, the total spent budget used in the active sliding window is calculated by $\epsilon_{spent}^{t} = \sum_{w=1}^{t} \epsilon_{w,2}$, where $w \geq 2$. Then the remaining budget is given by $\epsilon_{rest}^{t} = \epsilon/2 - \epsilon_{spent}^{t}$. If one timestamp can be republished by $k$-NN prediction, the perturbation will be skipped and its budget saved for a future perturbation and then set to 0. Otherwise, half of the remaining budget will be assigned to the current perturbation, namely $\epsilon_{i,2} = \epsilon_{rest}^{t}/2$ at timestamp $t$. The worst case here is all timestamps will be perturbed, then the total budget used in the private publication stage of active sliding window is $\epsilon/2^2 + \epsilon/2^3 + \cdots + \epsilon/2^{w+1} \leq \epsilon/2$; hence, this privacy budget distribution satisfies $\epsilon$-differential privacy.

4.4. Private k-NN selection and publication

To handle the privacy concerns in $k$-NNs selection, a private $k$-nearest neighbour selection and perturbation algorithm (PKSP) under differential privacy is proposed. The PKSP can achieve $\epsilon$-differential privacy while retaining considerable utility for prediction purposes. It can be achieved by following two steps: private neighbour selection and prediction perturbation.

4.4.1. Private k-NN selection

Private neighbour selection uses the exponential mechanism to privately select nearest neighbours to build the $k$-NNs dataset from candidates list for the state at timestamp $t$ under differential privacy (step 7 in our framework). Unlike selecting $k$ nearest neighbours from the sorted candidates list according to the state similarity, an exponential mechanism first assigns a probability for each candidate based on a score function and then randomly performs the selection according to the arranged probability distributions. Given the current states matrix $ST$ and one published states matrix $ST_{p}$, where $p \in [0, t - 1]$ is a past timestamp, we use the Gaussian weighted Euclidean distance $SS_{k,p}$ to be the score function: $q(t, p) = -SS_{k,p}, p \in [t-1]$. The probability of the chosen states matrix $ST_{p}$ will be assigned according to the score function to preserve differential privacy.

$k$-NN models aim to select $k$ neighbours to predict the current states, whose accuracy depends on the quality of chosen neighbours. However, if the exponential mechanism is implemented in a direct way, the prediction accuracy will be limited as the random selection may choose some inaccurate candidates with lower scores. Therefore, a truncating mechanism [28] is adopted to improve the quality of $k$-NNs, thereby improving the utility. Let $ST_{k,i}$ denote the $k$th nearest neighbour, the truncated parameter $\tau$ is used in the truncated mechanism to find the high quality neighbours. Through this truncating mechanism, the candidates with score lower than $ST_{k,i} - \tau$ will be truncated to $ST_{k,i} - \tau$, where the truncated score is $q(t, p) = max(q(t, p), ST_{k,i} - \tau)$. The truncating mechanism can guarantee that records with scores less than $ST_{k,i} - \tau$ will not be selected in $SK(t)$; whereas records with scores more than $ST_{k,i} + \tau$ are certainly selected. This can significantly improve the quality of the selected candidates.

The inputs of the algorithm are the privacy budget $\epsilon_s$, the parameter of $k$-NNs $k$, the candidates list $CL$, the current states matrix $ST_{t}$ at timestamp $t$ and the truncated parameter $\tau$. The first step is to get the score vector $SV_{t}$ according to the score function $q$, followed by sorting $SV_{t}$. We then divide the candidates list into two subsets: $C_1$ and $C_0$. The former comprises the states of the matrices with scores more than $ST_{k,i} - \tau$, and the latter consists of the rest. For the records in $C_1$, a probability is assigned according to the exponential distribution $exp(-\frac{\epsilon_s}{\Phi(q,t)})$. $C_0$ is regarded as a single candidate and its probability is decided by $exp(-\frac{\epsilon_s}{\Phi(q,t)})$. If $C_0$ is selected, we randomly choose one record from $C_0$ according to the exponential mechanism under differential privacy. Finally, one candidate is selected according to the assigned probabilities and then the algorithm moves to the next step. The private selection is a $\epsilon$-defferential privacy mechanism where the $\epsilon$ is the sum of privacy budget used for $m$ timestamps according to the sequential composition.
4.4.2. Private aggregate perturbation and publication

The next step is the private publication (steps 8 and 9 in our framework). First, we need to decide whether the use of prediction in publishing is beneficial based on the Mean Evaluation Indicator (MEI) as follows:

$$\text{MAE}(v_t \in ST_{i+m}, v'_t \in STP_{i+m}) = \frac{1}{|L_i|} \sum_{i=1}^{|L_i|} |r_{t,i} - n_{t,i}|$$  \hspace{1cm} (14)

$$\text{MEI}(v_t \in ST_{i+m}, v'_t \in STP_{i+m}) = \frac{\text{MAE}(v_t, v'_t) + \text{LAP}(2/(n * \epsilon_{t,i}^2))}{2|L_i| * \epsilon_{t,i}^2}$$  \hspace{1cm} (15)

Here, $|L_i|$ is the number of segments at timestamp $t$, $v_t$ is the vector of real statistic values at timestamp $t$, and $v'_t$ is the predicted value of real statistic at timestamp $t$, and $n_{t,i} \in v'_t$ is the predicted statistic value, i.e., the released noisy value used to republish, at segment $i$ at timestamp $t$. $\epsilon_{t,i}$ is the prediction error caused by the difference between $v_t$ and $v'_t$. $\text{MAE}$ is the vector of real predicted values at timestamp $t$, and $n_{t,i} \in v'_t$ is the predicted statistic value, i.e., the released noisy value used to republish, at segment $i$ at timestamp $t$. $\text{MEI}$ is the total error of republishing $(\text{MEI})$ as follows:

$$\text{MEI} = \frac{\sum_{i=1}^{|L_i|} \left| r_{t,i} - n_{t,i} \right|}{|L_i|}$$  \hspace{1cm} (16)

$\text{MAE}$ is fulfilled with an exponential mechanism. Thus the private $k$-NN selection satisfies $\epsilon/2$-differential privacy as a whole. In the privacy error comparison step, the sensitive data is accessed by the MAE query, it is necessary to add randomness to ensure differential privacy. According to the definition of differential privacy, the sensitivity of $\text{MAE}$ between any two neighbouring databases $D_1$ and $D_2$ is $2/|L_i|$. Thus the Laplacian perturbation mechanism for private error comparison, i.e., adding Laplace noise with associated scale parameter $\epsilon_{t,i}^2$-differential privacy.

In the private publishing step, implementing Laplace mechanism, i.e., adding Laplace noise with the scale parameter $\epsilon_{t,i}^2$, ensures $\epsilon_{t,i}^2$-differential privacy under the sensitivity of 2. Therefore, according to the sequential composition of differential privacy, DP-PSP satisfies $\epsilon/2 + \epsilon_{t,i}^2 \leq \epsilon$-differential privacy. \hfill $\square$

5. Experimental evaluation

5.1. Evaluation dataset

The real-time geo-tagged tweets collection and pre-processing is executed in a Cloud environment. We ran 20 VMs (medium-sized virtual machines with 8GB memory and eight virtual CPUs with 250GB volume storage and 100GB object storage), running 18 slave machines and 2 master nodes. After pre-processing, the test dataset was composed of 207 users and 120,687 locations ordered by their timestamps, in which 90% of data was used for sensitive anchor point discovery and road segments, and the rest for testing. These data were saved with the following structure: 

| Userid | Longitude | Latitude | Timestamp |

5.2. Utility evaluation metrics

To evaluate the utility of the private publication mechanism DP-PSP, we adopt the Root Mean of Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Deviation (MAD) to measure the information loss when performing the DP-PSP. The RMSE is applied to evaluate the differences between using the original and anonymized datasets at all timestamps. RMSE is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{|T|} \sum_{i=1}^{|T|} \sum_{j=1}^{|L_i|} \left| r_{t,i,j} - n_{t,i,j} \right|^2}$$  \hspace{1cm} (17)

MAPE is defined as:

$$\text{MAPE} = \frac{1}{|T| * |L_i|} \sum_{i=1}^{|T|} \sum_{j=1}^{|L_i|} \left| r_{t,i,j} - n_{t,i,j} \right| / r_{t,i,j}$$  \hspace{1cm} (18)

MAD is defined as:

$$\text{MAD} = \frac{1}{|T| * |L_i|} \sum_{i=1}^{|T|} \sum_{j=1}^{|L_i|} \left| r_{t,i,j} - n_{t,i,j} \right|$$  \hspace{1cm} (19)

5.3. Utility evaluation

In this section, the utility of the DP-PSP is evaluated using a real-time tweet trajectory dataset under the metrics defined above. In addition, we compared the DP-PSP with Uniform mechanisms. The budget distribution in the Uniform mechanism was fixed for all timestamps, and kept linear in $w$-steps. Each experiment was run 100 times. We report the average results using RMSE, MAPE and MAD. In order to show the precision and utility of the released real-time aggregate over location streams, the traffic data were aggregated for 21 consecutive days with 20 min as one statistics.
period (i.e., our time domain consisted of 1512 timestamps). The total number of roads segments was 326.

**Fig. 5** shows the RMSE, MAPE and MAD values of Optimized DJ-Cluster clustering (denoted by O-DJ in **Fig. 5**) for sensitive anchor point discovery compared to the commonly-adopted density-based clustering solutions DBSCAN, K-means and DJ-Cluster with other parts unchanged. Compared to K-Means, DBSCAN and DJ-Cluster, O-DJ achieved lower RMSE, MAPE and MAD. Namely, the O-DJ clustering approach reduced the distortion by $17 - 70\%$ compared to DBSCAN, K-means and DJ-Cluster. As noted, the privacy leak level $\epsilon$ can be composed of $\epsilon = \epsilon_{1,1} + \epsilon_{1,2} + \epsilon_{2,2}$. If not explicitly specified, the default value of privacy is set at $\epsilon = 1$ according to the experimental results of varying $\epsilon$; the default values of $K$ and $w$ are set at $k = 20$ and $w = 10$ respectively since the RMSE, MAPE and MAD achieved a relative low level when varying $K$ and $w$. The average performance of DP-PSP over different parameters $\epsilon$, $k$ and $w$ are as follows.

**Fig. 6** shows the RMSE, MAPE and MAD when varying $\epsilon$ at given $k$ and $w$ respectively. As $\epsilon$ increases, the privacy preserving level weakens, the RMSE, MAPE and MAD decrease since less noise is needed. Compared with the Uniform mechanism, the RMSE, MAPE and MAD of DP-PSP are approximately 3 to 4 times less, which illustrates that the utility and accuracy are both improved. This is also verified by **Fig. 7**, where the real statistic values and perturbed statistic values using DP-PSP and the Uniform mechanism are presented. As we can see from these figures, the DP-PSP outperforms the Uniform approach by over three times with fewer errors in RMSE, MAPE and MAD.

**Fig. 8** shows the RMSE, MAPE and MAD with varying $w$ at given $\epsilon$ and $k$ respectively. If $w$ is within a given $k$ and $\epsilon$ increases, the RMSE, MAPE and MAD will increase. The reason is that the length of the sliding window increases as $w$ becomes large and less budget will be assigned for each timestamp. Therefore, more noise needs to be introduced to achieve higher privacy preserving levels as $w$ increases, which leads to larger distortion of the perturbations.

**Fig. 9** shows the RMSE, MAPE and MAD with varying $k$ at given $\epsilon$ and $w$ respectively. At a given privacy budget and sliding window length $w$, with the increase of $k$, the RMSE, MAPE and MAD show fluctuations. However the RMSE, MAPE and MAD improve with $k$-NN prediction mechanisms than without $k$-NN prediction mechanisms (i.e., $k = 1$). In addition, the RMSE, MAPE and MAD will reach an optimized utility, i.e., lower RMSE, MAPE and MAD errors, at around $k = 80$. Consequently, the performance of the Uniform mechanism falls sharply, especially for a small given privacy budget and larger $w$ steps. Thus DP-PSP is superior due to its data-dependent accurate
approximation and efficient road networks segments, as well as its support for adaptive privacy budget distribution.

6. Conclusions

In this paper, we present a privacy-protected aggregate publication mechanism over infinite streams that can achieve better utility and accuracy (DP-PSP). An anchor point clustering and road segment mechanism is used in DP-PSP to address the heterogeneity of trajectories. DP-PSP provides an adaptive $w$-steps sliding window approach to allow users to specify their own length with an improved privacy budget distribution mechanism that can use the overall privacy budget more efficiently. We demonstrated that more privacy budget is saved and utility is enhanced by utilizing multi-time stamp prediction models and private $k$-NN selection and perturbation algorithms, since the amount of perturbation distortion required through adding random noise is reduced. Empirical studies using real-life location based social networks users’ trajectories confirm that DP-PSP enables accurate and private data analytics on infinite streams, whilst decreasing the total privacy cost without compromising utility under the metrics of RMSE, MAPE and MAD. Potential future work includes the design of private spatio-temporal data publication solutions for social media users which combines the numerical information (e.g., latitude–longitude coordinates) with non-numerical attributes (e.g., categorical or semantic content) in social media data. Another direction is to apply DP-PSP to solve more complex data analytical tasks in distributed Cloud scenarios.

References


Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Shuo, Wang

Title:
Exploration and protection of location privacy in online social networks

Date:
2018

Persistent Link:
http://hdl.handle.net/11343/212415

File Description:
Appendices-author accepted manuscripts-Chapter 5

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.