Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

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-steps sliding window.
- A private KNN selection model is used in improved multi-timestamps 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.

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.

[*] Corresponding author.
E-mail addresses: shuow4@student.unimelb.edu.au (S. Wang), rsinnott@unimelb.edu.au (R. Sinnott), Surya.Nepal@csiro.au (S. Nepal).

http://dx.doi.org/10.1016/j.future.2017.08.002
0167-739X/© 2017 Elsevier B.V. All rights reserved.

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 ϵ-differential privacy, the change of the released outcomes is guaranteed to be negligible (according to the privacy budget ϵ) 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-NNs* 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 aggregate 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