Data & Knowledge Engineering 94 (2014) 89–109

Contents lists available at ScienceDirect

Data & Knowledge Engineering

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

Editorial

Privacy-preserving trajectory stream publishing

Khalil Al-Hussaeni [a], Benjamin C.M. Fung [b,*], William K. Cheung [c]

[a] CIISE, Concordia University, Montreal H3G 1M8, Canada
[b] School of Information Studies, McGill University, Montreal H3A 1X1, Canada
[c] Department of Computer Science, Hong Kong Baptist University, Hong Kong

ARTICLE INFO

Article history:
Received 26 November 2013
Received in revised form 14 September 2014
Accepted 19 September 2014
Available online 28 September 2014

Keywords:
Data sharing
Data mining
Privacy protection
Spatio-temporal databases
Data stream

ABSTRACT

Recent advancement in mobile computing and sensory technology has facilitated the possibility of continuously updating, monitoring, and detecting the latest location and status of moving individuals. Spatio-temporal data generated and collected on the fly are described as *trajectory streams*. This work is motivated by the concern that publishing individuals' trajectories on the fly may jeopardize their privacy. In this paper, we illustrate and formalize two types of privacy attacks against moving individuals. We devise a novel algorithm, called *Incremental Trajectory Stream Anonymizer* (*ITSA*), for incrementally anonymizing a sequence of sliding windows on trajectory stream. The sliding windows are dynamically updated with joining and leaving individuals. The sliding windows are updated by using an efficient data structure to accommodate massive volume of data. We conducted extensive experiments on simulated and real-life data sets to evaluate the performance of our method. Empirical results demonstrate that our method significantly lowers runtime compared to existing methods, and efficiently scales when handling massive data sets. To the best of our knowledge, this is the first work to anonymize high-dimensional trajectory stream.

1. Introduction

The improvement of information technology in the past years has facilitated sharing data among organizations, firms, and to the public. Location-aware devices, such as GPS and mobile phones, *constantly* report spatio-temporal data of a moving object or the individual carrying this object. In many cases, it is important to publish the automatically-collected data on the fly for various purposes, such as traffic analysis, live monitoring of moving objects, and mining recent events in a data stream. This process becomes of vital importance especially when it is essential to take immediate actions or follow certain detection or prevention measures. Nevertheless, releasing the automatically-collected raw data by a data holder for analysis and service improvement may compromise individuals' privacy. We assume that recipients of a published data stream are untrustworthy, and they may attempt to identify target victims or infer their sensitive information. In this paper, we study the challenges in anonymizing a stream of trajectories, and propose an efficient algorithm to anonymize a trajectory stream with the goal of minimizing data distortion.

Fig. 1 shows an overview of the trajectory stream environment. A trajectory stream S is a continuous sequence of *triples*, in which each *triple* has the form $\langle id, loc, t\rangle$, indicating a person p with id is at location loc at timestamp t. A combination of loc and t is called a doublet. We assume that the trajectory stream S is published for stream mining [16] or the trajectory paths are simply displayed on screen. We propose a trajectory stream anonymization method based on a *sliding window* [18,5]. The literature has

[*] Corresponding author at: 3661 Peel St., Montreal, QC, Canada H3A 1X1.
E-mail addresses: k_alhus@ciise.concordia.ca (K. Al-Hussaeni), ben.fung@mcgill.ca (B.C.M. Fung), william@comp.hkbu.edu.hk (W.K. Cheung).

http://dx.doi.org/10.1016/j.datak.2014.09.004
0169-023X/© 2014 Elsevier B.V. All rights reserved.

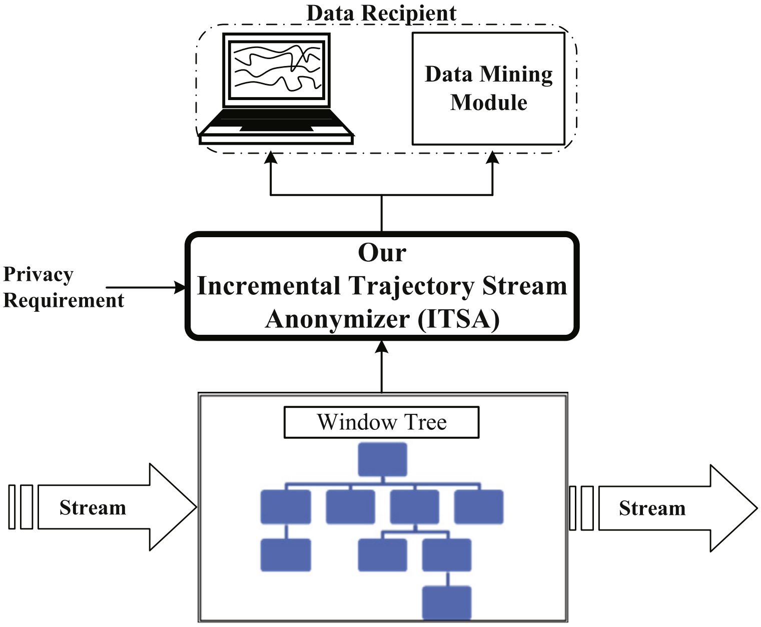

Fig. 1. Mining trajectory stream over a sliding window.

defined two types of sliding windows: *count-based* and *time-based* [6,18]. The former type defines a window that includes the N most recent data elements while the latter type defines a window that includes all elements belonging to the most recent N timestamps. We adopt a time-based sliding window because it is more general than count-based. However, using a count-based window has no impact on our approach. Hence, our approach models data stream as a sequence of sliding windows in which the most recent window includes the triples having the most recent N timestamps.

The Copenhagen International Airport is testing a mechanism for monitoring travelers' movements in real-time by following their Wi-Fi trails with the goals of improving airport design and security, directing the flow of travelers, and providing customized services to travelers [30]. Yet, disclosing the raw trajectory stream to some third-party service provider, such as an airline company or an outsourced security firm, may compromise the travelers' privacy. The following example illustrates two types of privacy attacks that an adversary can carry out by having access to the data stream.

Example 1.1. Table 1 shows the trajectories of eight travelers sorted by their *ids*. For simplicity, this example considers timestamps 1–4; however, in reality, timestamps continue indefinitely. Let us assume that sensitive information is being collected from travelers along with trajectories. The sensitive information is displayed in the sensitive attribute *sen _ att*. A potential sensitive attribute could be *Disability* where travelers with *Epilepsy*, for instance, may require special attention to facilitate their journey. The data holder (the airport) can specify a set of sensitive values from the sensitive attributes. Upon publishing the anonymized data, sensitive values should not be associated with their respected travelers. Suppose s_1 is the only sensitive value in this example.

Let the size of the sliding window be $N = 3$. The first window $W_{1 \rightarrow 3}$ includes doublets with timestamps 1–3, as indicated by the dashed box in Table 1. As the window slides with step size = 1, the second window $W_{2 \rightarrow 4}$ now includes doublets with timestamps 2–4 with no traces of doublets having timestamp 1. We note that the absence of doublets within a given window (the empty spots in Tables 1 and 2) indicates no change in a traveler's location.

Suppose an adversary has access to the trajectory stream in the form a sliding window, as in Table 1. It is possible to identify a target victim's trajectory and/or sensitive value by performing the following privacy attacks.

Identity linkage takes place when the collected trajectories contain a sequence of doublets with a rare appearance. This allows an adversary to uniquely identify a target victim. For example, suppose that the current window is $W_{2 \rightarrow 4}$, and that an adversary knows that a target victim has visited location e at timestamp 4. $W_{2 \rightarrow 4}$ contains only one trajectory ($id = 8$) with doublet $e4$. Hence, the adversary is able to learn the victim's other visited locations and sensitive value.

Attribute linkage takes place if there is a group of records, sharing the same sequence of doublets, that contains infrequent sensitive values. These values can be associated with their pertinent individuals with high confidence. This type of privacy attacks is also known as homogeneity attack [24,25]. Suppose that an adversary knows that a target victim has visited locations b and d at timestamps 2 and 4, respectively. $W_{2 \rightarrow 4}$ shows that one of two records that contain $\langle b2 \rightarrow d4 \rangle$ has the sensitive value s_1. Hence, the adversary is able to infer that the target victim has s_1 with 50% confidence. □