Transportation Research Part C 39 (2014) 63–79

Contents lists available at ScienceDirect

Transportation Research Part C

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

Anonymizing trajectory data for passenger flow analysis

Moein Ghasemzadeh [a], Benjamin C.M. Fung [b,*], Rui Chen [c], Anjali Awasthi [a]

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

ARTICLE INFO

Article history:
Received 24 August 2013
Received in revised form 6 December 2013
Accepted 6 December 2013

Keywords:
Data privacy
Anonymity
Trajectory
Passenger flow

ABSTRACT

The increasing use of location-aware devices provides many opportunities for analyzing and mining human mobility. The trajectory of a person can be represented as a sequence of visited locations with different timestamps. Storing, sharing, and analyzing personal trajectories may pose new privacy threats. Previous studies have shown that employing traditional privacy models and anonymization methods often leads to low information quality in the resulting data. In this paper we propose a method for achieving anonymity in a trajectory database while preserving the information to support effective passenger flow analysis. Specifically, we first extract the passenger flowgraph, which is a commonly employed representation for modeling uncertain moving objects, from the raw trajectory data. We then anonymize the data with the goal of minimizing the impact on the flowgraph. Extensive experimental results on both synthetic and real-life data sets suggest that the framework is effective to overcome the special challenges in trajectory data anonymization, namely, high dimensionality, sparseness, and sequentiality.

1. Introduction

Over the last few years, transit companies have started using contactless smart cards or RFID cards, such as the EasyCard in Taiwan, the Public Transportation Card in Shanghai, and the OPUS card in Montréal. In 2008, *Société de transport de Montréal* (*STM*), the public transit agency in Montréal, deployed the *Smart Card Automated Fare Collection* (*SCAFC*) system (Pelletier et al., 2011) in its transportation network. Senior and junior passengers have to register their personal information when they first purchase their cards so that an appropriate fare is charged based on their statuses. In the SCAFC system, each passenger leaves a trace of reading in the form of (ID, loc, t), which identifies the passenger's identity, location, and time when she scans her smart card. The trajectory of a passenger is then stored as a sequence of visited locations, sorted by time, in a central database.

Constructions occur and new trends emerge as a city develops. Thus, passenger flows in a city are not static and are subject to change depending on all these uncertainties and developments. Transit companies have to periodically share their passengers' trajectories among their own internal departments and external transportation companies in order to perform a comprehensive analysis of passenger flows in an area, with the goal of supporting trajectory data mining (Giannotti et al., 2007; Lee et al., 2008, 2007; Tang et al., 2012; Zheng et al., 2013) and traffic management (Burger et al., 2013; Li et al., 2007a). For instance, by using a probabilistic flowgraph, as shown in Fig. 1, an analyst can identify the major trends in passenger flows and hot paths in a traffic network. However, sharing passenger-specific trajectory data raises new privacy

* Corresponding author. Tel.: +1 5143983360.
E-mail addresses: mo_gh@ciise.concordia.ca (M. Ghasemzadeh), ben.fung@mcgill.ca (B.C.M. Fung), ruichen@comp.hkbu.edu.hk (R. Chen), awasthi@ciise.concordia.ca (A. Awasthi).

0968-090X/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.trc.2013.12.003

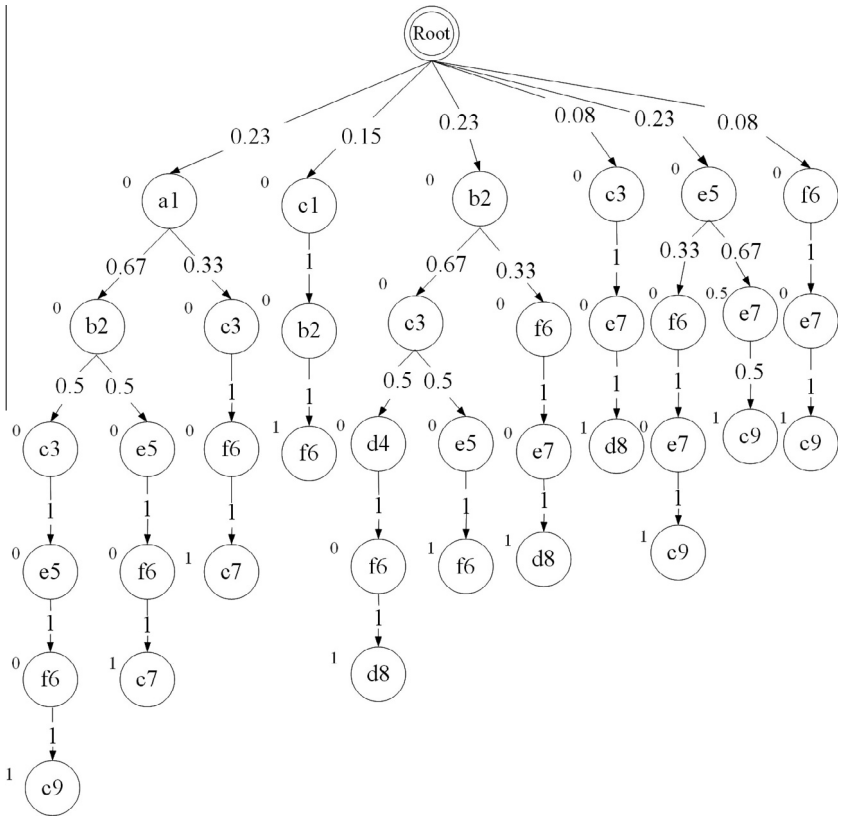

Fig. 1. Probabilistic flowgraph of Table 1.

concerns that cannot be appropriately addressed by traditional privacy protection techniques. Example 1.1 illustrates a potential privacy threat in the context of trajectory data.

Example 1.1 (*Identity linkage attack*). Table 1 shows an example of thirteen passengers' trajectories, in which each trajectory consists of a sequence of spatio-temporal doublets (or simply doublets). Each doublet has the form $(loc_i t_i)$, representing the visited location loc_i with timestamp t_i. For example, $ID\#4$ indicates that the passenger has visited locations c, e, and d at timestamps 3, 7, and 8, respectively. With adequate background knowledge, an adversary can perform a type of privacy attack, called *identity linkage attack*, on the trajectory database and may be able to uniquely identify a victim's record as well as his/her visited locations. Preventing identity linkage attacks is very important in trajectory data sharing because it is easy to be performed by an attacker and upon success, it allows the attacker to learn all other locations and timestamps of the victim. Hence, it is the main goal of this paper. Suppose an adversary knows that the data record of a target victim, Alice, is in

Table 1
Raw trajectory database T.

Series ID #	Series trajectory
1	$a1 \rightarrow b2 \rightarrow c3 \rightarrow e5 \rightarrow f6 \rightarrow c9$
2	$e5 \rightarrow f6 \rightarrow e7 \rightarrow c9$
3	$e5 \rightarrow e7$
4	$c3 \rightarrow e7 \rightarrow d8$
5	$b2 \rightarrow c3 \rightarrow d4 \rightarrow f6 \rightarrow d8$
6	$c1 \rightarrow b2 \rightarrow f6$
7	$a1 \rightarrow b2 \rightarrow e5 \rightarrow f6 \rightarrow e7$
8	$f6 \rightarrow e7 \rightarrow c9$
9	$e5 \rightarrow e7 \rightarrow c9$
10	$b2 \rightarrow f6 \rightarrow e7 \rightarrow d8$
11	$a1 \rightarrow c3 \rightarrow f6 \rightarrow e7$
12	$c1 \rightarrow b2 \rightarrow f6$
13	$b2 \rightarrow c3 \rightarrow e5 \rightarrow f6$