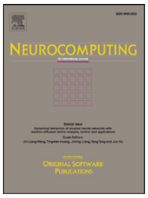
Contents lists available at ScienceDirect

Neurocomputing

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



A privacy-preserving approach for multimodal transaction data integrated analysis

Peipei Sui[a,b], Xianxian Li[c,*]

[a] School of Management Science and Engineering, Shandong Normal University, Jinan, Shandong, 250014
[b] School of Computer Science and Engineering, Beihang University, Beijing 100191, China
[c] Guangxi Key Lab of Multisource Information Mining and Security, Guangxi Normal University, Guilin, Guangxi, 541004, China

ARTICLE INFO

Article history:
Received 14 May 2016
Revised 12 September 2016
Accepted 14 September 2016
Available online 8 March 2017

Keywords:
Multimodal data
Trajectory data
Transaction data
Privacy protection

ABSTRACT

Multimodal transaction data mining has received a great deal of attention recently. Protection of private information is an essential requirement of data analysis. Existing work on privacy protection for transaction data usually focus on a single mode dataset. The existing privacy-preserving methods cannot be used directly to address privacy issues for multimodal data integration, since information leakage may be caused by data correlations among multiple heterogeneous datasets. In this work, we address privacy protection on the integration of transaction data and trajectory data. We first demonstrate a privacy leakage model caused by integration of multimodal datasets, where integrated data are modeled as a tree. To address the identity disclosure of trajectories, we partition location sequences to meet privacy demands, and copy locations to offset information loss caused by partition; then, to deal with the sensitive item disclosure of transactions, we use suppression technique to eliminate sensitive association rules. Consequently, we propose a k^m-anonymity-ρ-uncertainty privacy model to protect the privacy information in integrating transaction data with trajectory data in a tree-structured data model. Finally, we perform experiments on two synthetic integration datasets, and analyze privacy and information loss under varying parameters.

1. Introduction

Integrated analysis of multimodal transaction data is useful to provide high-quality service to customers; therefore, multiple service providers need to collaborate and share data. However, these datasets often contain a large amount of personal information, that can be misused by adversaries to reveal sensitive information concerning the individual. The challenge is to ensure privacy for combined data publishing.

Consider the Octopus Company in Hong Kong, that offers integrated payment services for transportation and day-to-day purchases. The Octopus card is used by 95% of the population of Hong Kong aged 16–65, generating over 12 million daily transactions worth a total over HK $130 million [1]. The Octopus Company accumulates extensive trajectory data daily, while retail shops which support Octopus card payments keep detailed transaction data in their own database. Therefore, a large quantity of transaction logs and movement data are accumulated, that could be shared and published to analyze movement and behavioral patterns of Hong Kong residents. While data sharing and publishing can improve the service quality, it can also lead to serious individual privacy disclosure. On the one hand, trajectory data, as a typical type of spatiotemporal data, contains lots of individual movement patterns. On the other hand, transaction data, as a typical type of set-valued data, contains lots of individual behavior patterns. Spatiotemporal and set-valued data, have high dimensions, are sparse, and unstructured. These special features create new challenges in privacy protection. Preserving privacy becomes more difficult when trajectory data integrates with transaction data. In this paper, we study privacy issues caused by the integrated data publishing of trajectory and transaction.

The work in [2] considered a scenario where location samples were drawn from a shop address set, and assumed that adversaries knew some trajectory fragments to identify users corresponding to those segments. For example, Octopus collects transactions for a user, and a partner company such as a retail chain knows all items purchased in the retail company by the user. Since the retail com-

[*] Corresponding author.
E-mail addresses: suipp@act.buaa.edu.cn (P. Sui), lixx@gxnu.edu.cn (X. Li).

http://dx.doi.org/10.1016/j.neucom.2016.09.130

id	trajectory
$traj_1$	$a_1 \to b_1 \to a_2$
$traj_2$	$a_1 \to b_1 \to a_2 \to b_3$
$traj_3$	$a_1 \to b_2 \to a_2$
$traj_4$	$a_1 \to a_2 \to b_2$
$traj_5$	$a_1 \to a_3 \to b_1$
$traj_6$	$a_3 \to b_1$
$traj_7$	$a_3 \to b_2$
$traj_8$	$a_3 \to b_2 \to b_3$

(a) Original data T

id	trajectory
$traj_1{}^A$	$a_1 \to a_2$
$traj_2{}^A$	$a_1 \to a_2$
$traj_3{}^A$	$a_1 \to a_2$
$traj_4{}^A$	$a_1 \to a_2$
$traj_5{}^A$	$a_1 \to a_3$
$traj_6{}^A$	a_3
$traj_7{}^A$	a_3
$traj_8{}^A$	a_3

(b) A's knowledge T^A

id	trajectory
$traj_1{}'$	$a_1 \to b_1 \to a_2$
$traj_2{}'$	$a_1 \to b_1 \to a_2$
$traj_3{}'$	$a_1 \to b_2 \to a_2$
$traj_4{}'$	$a_1 \to a_2 \to b_2$
$traj_5{}'$	$a_3 \to b_1$
$traj_6{}'$	$a_3 \to b_1$
$traj_7{}'$	$a_3 \to b_2$
$traj_8{}'$	$a_3 \to b_2$

(c) Transformed dataT'

Fig. 1. An example of [2].

pany knows partial trajectories followed by this user, they may infer other locations the user visited. In this paper, we consider not only the privacy issue of trajectory data but also of the detailed transaction data. For example, when a user purchases products at a shop paying with an Octopus card, the Octopus Company records the transaction trajectory and the shop keeps the detailed transaction record. In fact, the trajectory and detailed transaction data provide an electronic image of a user's life, and we can obtain valuable knowledge through the integration of these two types of data. In this scenario, associating data from different companies need to not only address the privacy risk introduced by trajectories but also privacy leakage from transactions.

Fig. 1shows the example introduced by Terrovitis and Mamoulis [2]. Octopus provides a dataset T shown in Fig. 1a to data miners for research purposes. Each sequence element is a shop address, visited by the corresponding user. Shops denoted by a_i belong to the same company A (e.g., 7-Eleven). They assumed A is an adversary, and has background knowledge T^A as described in Fig. 1b. For example, if T is published, A will know $traj_5^A$ is matched to $traj_5$, as there is only one trajectory that travels through a_1 to a_3 in the original dataset. So A can infer that the user who matched $traj_5^A$ has also visited b_1, and this leads to the user's privacy leakage. To resist partial trajectory attacks, they suppressed locations in T wherever privacy leaks occurred and finally transformed it to a published dataset T' shown in Fig. 1c. After suppressing b_3 from $traj_2$, a_1 from $traj_5$, and b_3 from $traj_8$, T' ensures adversary A cannot reconstruct any location with a certainty higher than 50%.

However, in this paper we consider the integration publication of trajectories and transactions. Fig. 2a shows the integrated data, where a_i denotes a shop address. We use an item set to denote items purchased by the corresponding user in a shop, where I_i shows items which can be published directly, also called public item, and some sensitive items denoted by S_i. We also assumed that A is an adversary, and has background knowledge U^A described in Fig. 2b. If we do not consider detailed transaction data, then privacy risk is the same as [2]. But, when adding detailed transaction information to trajectories, a new privacy leakage will be occur. For example, if U' shown in Fig. 2c is published, the projection $a_1 \to a_2$ in U^A is mapped to four trajectories $u'_1 - u'_4$ in U', in which b_1 and b_2 appear with probability 50%, but the projection after adding items $u_2^A : a_1\{I_2, S_1\} \to a_2\{I_1, I_2, S_1\}$ is mapped to the only record u'_2 in U'. Therefore, A is 100% sure that the user who followed u_2^A visited b_1 and purchased sensitive item S_5. In this case, if we convert u_2' =($a_1\{I_2, S_1\} \to b_1\{I_1, I_2, S_5\} \to a_2\{I_1, I_2, S_1\}$) to u_2*= ($a_1\{I_2, S_1\} \to b_1\{I_1, I_2, S_5\} \to a_2\{I_2, S_1\}$) by suppressing I_1 from $a_2\{I_1, I_2, S_1\}$, then the demands of privacy are satisfied, the final anonymization data are shown in Fig. 2d.

To publish trajectory and transaction data safely, we need to eliminate two types of privacy risks, trajectory identity disclosure and sensitive item disclosure. Trajectory identity disclosure occurs when a user is linked to his or her transaction trajectory. Sensitive item disclosure occurs when a user is linked to a set of items that is considered to be sensitive. In this work, we consider privacy protection for publishing integration data containing both transaction and trajectory data.

To summarize, we made the following contributions:

- We illustrate challenges in privacy-preserving data when publishing both of transaction and trajectory data, and propose a tree-structure to model the integration of transactions and trajectories.
- We propose a partition and copy method to defend trajectory identity disclosure when assuming adversaries know some sub-trajectories. Our method ensures that the number of trajectories with the same sub-trajectory is not less than k when the length of the sub-trajectory is not larger than m.
- We design an algorithm applying suppression for sensitive item protection when assuming adversaries know some public items. Our method ensures that probability of inferring sensitive items based on any association rule is not larger than a fixed threshold.

The rest of the paper is structured as follows. Section 2, briefly reviews related works. Section 3 provides preliminary notions and defines the problem formulation. Section 4 presents our proposed approach. Section 5 presents experimental results. Finally, Section 6 provides conclusions.

2. Related work

In this section, we first introduce privacy protection methods for data publishing. Then, we focus on the privacy problem of trajectory and transaction data publishing. Finally, we discuss existing works on secure data integration of trajectory and transaction.

Data privacy-preserving techniques have been developed to make relational data anonymous. To protect individuals from privacy leaks caused by data publishing, there are primarily two types of privacy models [3]. One is the syntactic-based model, which is represented by k-anonymity [4]; and the other is semantic-based model, which is represented by differential privacy model [5]. K-anonymity is the first proposed privacy model for data publishing. It requires that all records in a published table cannot be distinguished from at least k-1 records. However, k-anonymity does nothing limits with sensitive attributes, so attackers may learn the relationship between sensitive data and individuals through consistency attack and background knowledge attack. Consistency attack and background knowledge attack are the terms of privacy