COMPUTERS & SECURITY 55 (2015) 175–192

Available online at www.sciencedirect.com

ScienceDirect

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

Computers & Security

Differentially private maximal frequent sequence mining

*Xiang Cheng, Sen Su *, Shengzhi Xu, Peng Tang, Zhengyi Li*

State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

ARTICLE INFO

Article history:
Received 12 January 2015
Received in revised form 6 August 2015
Accepted 25 August 2015
Available online 30 September 2015

Keywords:
Differential privacy
Maximal frequent sequence mining
Frequent sequence mining
Length reduction
Threshold relaxation

ABSTRACT

In this paper, we study the problem of designing a differentially private algorithm for mining maximal frequent sequences, which can not only achieve high data utility and a high degree of privacy, but also provide high time efficiency. To solve this problem, we present a new differentially private algorithm, which is referred to as DP-MFSM. DP-MFSM consists of three phases: pre-processing phase, expected frequent sequence mining (ESM) phase, and candidate extraction and verification (CEV) phase. Specifically, in the pre-processing phase, we first extract some statistical information from the input database, and use the extracted information to determine the values of some variables which will be used in the ESM phase. Then, in the ESM phase, we randomly partition the input database into several sub-databases, and use a partition-based ESM technique to find expected frequent sequences, which are a subset of candidate frequent sequences and more likely to be frequent. At last, in the CEV phase, we extract candidate maximal frequent sequences from the discovered expected frequent sequences, and use a splitting-based technique to verify which candidates are actually frequent in the input database. Through privacy analysis, we show that our DP-MFSM algorithm is ε-differentially private. Extensive experiments on real-world datasets illustrate that our DP-MFSM algorithm can substantially outperform alternative approaches.

1. Introduction

With the increasing ability to collect personal data, sequential data with sensitive personal information, such as trajectories and DNA sequences, become more prevalent in recently years. Mining frequent sequences from such data can be applied in several applications, such as user behavior analysis and biological sequence analysis. Compared to mining frequent sequences, mining maximal frequent sequences can not only provide valuable insights, but the number of maximal frequent sequences can be significantly smaller than the number of frequent sequences as well (Han et al., 2007). However, directly releasing the maximal frequent sequences will pose concerns on the privacy of the users participating in the data (Bhaskar et al., 2010). For example, the released maximal frequent sequences can be linked with a large amount of data available creating opportunities for adversaries to break the individual privacy of the users and disclose sensitive information (Bonomi and Xiong, 2013a). To this end, in this paper, we focus on the problem of privacy-preserving maximal frequent sequence mining.

Differential privacy (Dwork, 2006; Dwork et al., 2006) has become the *de facto* standard for research in data privacy, as it provides strong and provable guarantees of privacy. Our goal is to design a differentially private algorithm for mining

* *Corresponding author.* Tel.: +8613910932716.
E-mail addresses: chengxiang@bupt.edu.cn (X. Cheng), susen@bupt.edu.cn (S. Su), dear_shengzhi@bupt.edu.cn (S. Xu), tangpeng@bupt.edu.cn (P. Tang), lizhengyi@bupt.edu.cn (Z. Li).
http://dx.doi.org/10.1016/j.cose.2015.08.005
0167-4048/

maximal frequent sequences, which can not only achieve high data utility and a high degree of privacy, but also provide high time efficiency. To achieve this goal, intuitively, we could design a two-phase differentially private algorithm based on the widely used *a priori*-based GSP algorithm (Srikant and Agrawal, 1996). In the first phase, we find all frequent sequences under differential privacy. In particular, during the mining process, we perturb the support of candidate sequences by adding random noise, and use the noisy support to determine whether a candidate sequence is frequent. In the second phase, we extract maximal frequent sequences from the discovered frequent sequences. However, this algorithm might suffer from poor data utility in real-world datasets due to the existence of very long sequences (Zeng et al., 2012). In addition, since this algorithm is designed based on the GSP algorithm, which requires multiple passes over the entire dataset, it is very time consuming when the dataset is large. Besides, we also notice that some existing approaches (Chen et al., 2012) can be used indirectly to mine maximal frequent sequences under differential privacy. However, we are not aware that any of them can achieve all the requirements in our problem.

Toward this end, we present a new differentially private algorithm for mining maximal frequent sequences, which is referred to as DP-MFSM. DP-MFSM consists of three phases: pre-processing phase, expected frequent sequence mining (ESM) phase, and candidate extraction and verification (CEV) phase. Specifically, in the pre-processing phase, we first extract some statistical information from the input database, and use the extracted information to determine the values of some variables which will be used in the ESM phase. Then, in the ESM phase, we randomly partition the input database into several sub-databases, and use a partition-based ESM technique to find expected frequent sequences, which are a subset of candidate frequent sequences and more likely to be frequent. In particular, during the process of mining expected frequent sequences, a length reduction method and a threshold relaxation method are used to improve the utility-privacy tradeoff. At last, in the CEV phase, we extract candidate maximal frequent sequences from the discovered expected frequent sequences, and use a splitting-based technique to verify which candidates are actually frequent in the input database. DP-MFSM takes a sequence database, a user-specified threshold, and a privacy parameter ε as input, and outputs the discovered maximal frequent sequences by making three passes over the input database in total. Through privacy analysis, we show that our DP-MFSM algorithm is ε-differentially private. The results of extensive experiments on real-world datasets show that the proposed DP-MFSM algorithm achieves all the design goals and substantially outperforms alternative approaches.

To summarize, our key contributions are:

- We present a new differentially private algorithm for maximal frequent sequence mining, which is referred to as DP-MFSM. In DP-MFSM, two novel techniques, namely, partition-based expected frequent sequence mining and splitting-based candidate verification, are proposed to improve the tradeoff between utility and privacy. To our best knowledge, this is the first attempt to solve the maximal frequent sequence mining problem under differential privacy.
- Through privacy analysis, we prove that DP-MFSM satisfies ε-differential privacy. Experimental results on real datasets illustrate that DP-MFSM can not only achieve high data utility and a high degree of privacy, but also provide high time efficiency.

The rest of paper is organized as follows. We review related work in Section 2. Section 3 presents necessary background on differential privacy and briefly reviews the problem of maximal frequent sequence mining. In Section 4, we propose a straightforward approach for mining maximal frequent sequences under differential privacy. Section 5 presents the details of our DP-MFSM algorithm. In Section 6, we give the privacy analysis of the proposed DP-MFSM algorithm. Experimental results are reported in Section 7. Finally, we conclude the paper in Section 8.

2. Related work

As the *de facto* standard notion of privacy in privacy-preserving data analysis, differential privacy has received considerable attention recently. There are two settings: interactive and non-interactive (Dwork, 2008). In the interactive setting where the database is held by a trusted server, users pose queries about the data, and the answers to the queries are modified to protect the privacy of the database participants. In the non-interactive setting, the data custodian either computes and publishes some statistics on the data, or releases an anonymized version of the raw data. Several differentially private frequent sequence mining algorithms have been proposed under the interactive setting in the literature. Bonomi and Xiong (2013b) propose a two-phase differentially private algorithm for mining frequent consecutive sequences. They first utilize a prefix tree to find candidate sequences, and then leverage a database transformation technique to refine the support of candidate sequences. In our previous work (Shengzhi Xu et al., 2015), we propose a sampling-based candidate pruning technique for mining frequent non-consecutive sequences. Different from the above two works, in this work, we focus on designing a new algorithm which can directly mine maximal frequent sequences. Kellaris et al. (2014) put forward the novel notion of w-event privacy over infinite streams, which protects any event sequence occurring in w successive time instants. Unlike this work, which solves the problem of continuous publication of statistics over infinite streams, we focus on the static environment in which the datasets are relatively stable.

There are also some studies on applying differential privacy to non-interactive frequent sequence mining (Chen et al., 2012). In particular, Chen et al. (2012) propose a differentially private sequence database publishing algorithm based on a prefix tree. In Chen et al. (2012), the authors employ a variable-length n-gram model to extract the necessary information of the sequence databases, and utilize an exploration tree to reduce the amount of added noise. However, while these two studies focus on the publication of sequence databases, our work aims at the release of maximal frequent sequences.

There is another series of studies on finding frequent itemsets and subgraphs under differential privacy. To meet the