Information Sciences 467 (2018) 238–255

Contents lists available at ScienceDirect

Information Sciences

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

K-VARP: *K*-anonymity for varied data streams via partitioning

Ankhbayar Otgonbayar*, Zeeshan Pervez, Keshav Dahal, Steve Eager

University of the West of Scotland Paisley, Scotland, UK

ARTICLE INFO

Article history:
Received 6 October 2017
Revised 19 April 2018
Accepted 25 July 2018
Available online 3 August 2018

Keywords:
Internet of things
Data privacy
Data streams
Anonymization
Missing values

ABSTRACT

The Internet-of-Things (IoT) produces and transmits enormous amounts of data. Extracting valuable information from this enormous volume of data has become an important consideration for businesses and research. However, extracting information from this data without providing privacy protection puts individuals at risk. Data has to be sanitized before use, and anonymization provides solution to this problem. Since, IoT is a collection of numerous different devices, data streams from these devices tend to vary over time thus creating varied data streams. However, implementing traditional data stream anonymization approaches only provide privacy protection for data streams that have predefined and fixed attributes. Therefore, conventional methods cannot directly work on varied data streams. In this work, we propose *K*-VARP (*K*-anonymity for VARied data stream via Partitioning) to publish varied data streams. *K*-VARP reads the tuple and assigns them to partitions based on description, and all tuples must be anonymized before expiring. It tries to anonymize expiring tuple within a partition if its partition is eligible to produce a *K*-anonymous cluster. Otherwise, partition merging is applied. In *K*-VARP we propose a new merging criterion called *R*-likeness to measure similarity distance between tuple and partitions. Moreover, flexible re-using and imputation free-publication is implied in *K*-VARP to achieve better anonymization quality and performance. Our experiments on a real datasets show that *K*-VARP is efficient and effective compared to existing algorithms. *K*-VARP demonstrated approximately three to nine and ten to twenty percent less information loss on two real datasets, while forming a similar number of clusters within a comparable computation time.

1. Introduction

The technological revolution of the Internet-of-Things (IoT hereafter) has become an inseparable part of the modern world. We are living in an era in which enormous volumes of data are generated and transmitted in the form of streams [1]. Everything that we do in our lives leaves a trace, forming a digital data stream, such as, the browser history of internet users, bank transactions and energy consumption logs of houses. Extracting this valuable knowledge from the streaming data can provide a realistic and approximate insight into individuals' activities [2] and the behaviour of a society [3]. Many organizations publish and exchange data for business and research purposes; however, processing individuals' information without compromising privacy is a primary concern for IoT [4,5].

* Corresponding author.
E-mail addresses: ankhbayar.otgonbayar@uws.ac.uk (A. Otgonbayar), zeeshan.pervez@uws.ac.uk (Z. Pervez), keshav.dahal@uws.ac.uk (K. Dahal), steve.eager@uws.ac.uk (S. Eager).

https://doi.org/10.1016/j.ins.2018.07.057
0020-0255/

The most popular technique to provide privacy protection for publishing data is anonymization [6–10]. Anonymization removes or replaces the information, which can be exploited by an attacker, to compromise the privacy of a user. Therefore, individuals remain hidden from potential threats when their data is published for analytical or business purposes. Confidential or identifier information of individuals, which must not be published to the public domain, is called sensitive information. Non-sensitive information which can be exploited by an attack is called quasi-identifiers (*QID* hereafter). Anonymization approaches are classified into two major classes; static data anonymization and data stream anonymization [11–13]. Static data anonymization works with pre-recorded datasets having pre-defined *QID*s. The quality of the static data anonymization is measured by information loss which indicates the usability of anonymized data. Data stream anonymization processes the data on the fly (i.e., publishes the data as it arrives) [13–16]. The quality of the data stream anonymization is defined by a tradeoff between data freshness and data usability. Some publishers may want fast anonymization - although it gives more disrupted data; whereas, some publishers may prioritize data usability rather than data freshness to get data which is more precise. For example, the data stream of a mission critical system requires a minimum delay to publish data that can be used to take immediate action against potential threats. On the other hand, sales transaction data can be processed with a longer delay when data usability is prioritized. Sliding window is the most widely used technique for data stream anonymization, it keeps an anonymization algorithm consistent and tolerant when dealing with fast and high dimensional data streams [17–19]. This technique is an accumulation based mechanism for anonymizing data streams, which prevents the overflow of memory and helps to publish data continuously.

IoT consists of multiple internet enabled sensing and actuating devices used by individuals for different purposes. For instance, smart car, smart heating, fire alarms and security cameras for smart homes and offices, wearable devices to measure the physical performance of a person, and data generated from smart cities to provision personalized services to the inhabitants. IoT data streams generate data streams with missing values due to its unstable and uncontrollable properties. There are three main factors that cause missingness on IoT data streams:

- *Individuals' preference:* each individual have varying types of devices depending on their preferences;
- *Different usage pattern:* each individual can choose to use different devices at any given time;
- *Uncertain environmental condition:* environmental conditions can cause devices to malfunction or lose connectivity.

Therefore, we call it varied data stream due to the varying sets of *QID*s in each tuple of missing data stream. Anonymizing data with missing values is always an interesting topic for researchers [20]. The main challenge for anonymizing incomplete data is handling the missingness in data streams originating from multiple streams i.e., IoT devices used by a user. Researchers identified three main methods to handle missingness of static data:

a) *Imputation:* values are calculated to fill the missingness [21];
b) *Marginalization:* ignore missingness while anonymizing [22];
c) *Partitioning:* splits data into disjoint partitions based on tuple's description [21].

However, there has been no substantial work published on handing missingness for data stream anonymization. Incomplete dataset anonymization techniques can be extended to work on varied data streams; however, this will cause more information loss, weak privacy protection and a high computational time.

To the best of our knowledge, there is no known algorithm proposed specifically to anonymize varied data stream. To address this, we are proposing *K*-VARP (*K*-anonymity for VARied data stream via Partitioning) for anonymizing varied data streams. Our target is to anonymize and publish varied data streams with minimum delay and less information loss. The *K*-VARP algorithm uses both partitioning and marginalization methods to anonymize varied data streams under a time based sliding window. As previously discussed, a time based sliding window is the most convenient technique for anonymizing data streams, allowing us to publish data streams with minimum delay and less information loss. Our proposed algorithm *K*-VARP provides privacy preserving capabilities to real world applications that utilizes varied data streams. For example, social network analysis [23,24], patient monitoring [25,26] and smart city [27,28].

An overview of the *K*-VARP algorithm is illustrated in Fig. 1. *K*-VARP has two main phases, partitioning and anonymizing. In partitioning, *K*-VARP assigns receiving tuples to partitions using their *QID* set with their received timestamp attached. This phase plays the role of a buffer, and helps to store received tuples in an organized form to perform fast and efficient anonymization. In time-based sliding window, the maximum time for each tuple to stay in the buffer is defined by a time constraint, denoted as δ (see Fig. (1)). Each expiring tuple has its own anonymization round. The anonymization round is invoked when a tuple is about to expire according to time window criteria δ.

There are three modules to anonymize an expiring tuple t' regarding the size of its partition P', and each of these modules has an option to anonymize an expiring tuple by re-using recently published *K*-anonymous clusters.

- *In-Partition clustering:* This module is designed to publish cluster with no missing value. However, it is invoked only if partition has enough tuples to form *K*-anonymous cluster for expiring tuple.
- *Merge clustering:* Partition merging is inevitable when dealing with varied data stream, and this module is designed to merge the most suitable partitions to anonymize expiring tuples with less information loss.
- *Single anonymization:* This module is designed to publish expiring tuple when partition merging is not possible for expiring tuple.

For more details, please refer to Section 4.