Privacy Preservation in Publishing Electronic Health Records Based on Perturbation

Lin Yao[1], Xinyu Wang[2], Zhenyu Chen[2], and Guowei Wu[2(✉)]

[1] DUT-RU International School of Information Science and Engineering, Dalian University of Technology, Dalian, China
[2] School of Software, Dalian University of Technology, Dalian, China
wgwdut@dlut.edu.cn

Abstract. The patients' health information is often kept as electronic health records (EHRs). To improve the quality and efficiency of the care, EHRs can be shared among different organizations. However, the inappropriate sharing or usage of these healthcare data could threaten people's privacy. It becomes increasingly important to preserve the privacy of the published EHRs. An attacker is apt to identify an individual from the published EHRs by partial measurement information as background knowledge, with attacks through the record linkage and attribute linkage. To resist the above types of attacks, we propose a privacy preservation with perturbation in the published healthcare data (PPHR). To protect the privacy of sensitive information, we first determine the critical sequences based on which some specific records are easy to be identified. Then, we adopt perturbation on these sequences by adding or deleting some points while ensuring the published data to satisfy l-diversity. A comprehensive set of real-life healthcare data sets are applied to evaluate the performance of our anonymization approach. Simulations show our scheme possesses better privacy while ensuring higher utility.

Keywords: Privacy Preservation · Perturbation · Electronic health records

1 Introduction

The traditional paper-based health records may cause much inconvenience in collecting and storing various types of patient data. With the development of information and communications technologies, there is a great interest in moving from paper-based health records to electronic health records (EHRs). In 2003 and 2004, EHRs were used in 18% of the estimated 1.8 billion physicians in the U.S. In 2016, over 70% of physicians have used EHRs [20]. By storing a patient's medical history in electronic form, errors due to bad handwriting can be eliminated and it is easier for doctors to follow a patient's health condition.

J. Li et al. (Eds.): SPNCE 2019, LNICST 284, pp. 125–140, 2019.
https://doi.org/10.1007/978-3-030-21373-2_12

On the one hand, disseminating these data can provide better quality of care and thereby improve the public health [17]. For example, doctors in the San Diego Beacon Community (SDBC) can provide a cheaper, faster and more efficient diagnosis by obtaining the patient's EHR from his/her healthcare provider. On the other hand, researchers can benefit from the shared EHRs. In 2012, a group of UCLA researchers set out to mine thousands of EHRs for a more accurate and less expensive way to identify people who have undiagnosed Type 2 diabetes.

While the publication of EHRs is greatly beneficial, it can still entail a privacy threat for the users if some sensitive information is released with each EHR consisting of the patient's name, measurement history of physiological indicators, medical history, and other health data information. The measurement history or medical history is in chronological order which called healthcare trajectory or patient trajectory. A recent study has summarized that approximately 87% of the population of the United States can be identified by a given data set [26]. Therefore, it is critical to conserve the privacy of published health data, especially the sensitive information. The HIPAA Privacy Rule also proposed that the privacy of individually identifiable health information should be protected [15].

The original data tables or EHRs such as in Table 1 typically consist of four types of attributes, direct identifier, quasi-identifier, sensitive attribute, and non-sensitive attribute [8,25]. Direct identifier such as name and social security number can identify an individual uniquely, which is usually removed from the published tables. Each specific Quasi-Identifier (QI) such as healthcare trajectory in Table 1 cannot uniquely identify an individual, but the combination of some points can cause identity disclosure. Sensitive Attribute (SA) such as disease in Table 1 contains the private or specific information of each individual. Non-sensitive attribute can be known for the public without any concern. Based on the above attributes, it is obvious that privacy threats are related to those attributes except the last one.

Table 1. An example of healthcare trajectory dataset

ID.	Name	Healthcare trajectory	Disease	$\cdots$
1	Alice	$a1 \rightarrow d2 \rightarrow b3 \rightarrow e4 \rightarrow f6 \rightarrow e8$	HIV	$\cdots$
2	Ben	$d2 \rightarrow c5 \rightarrow c7 \rightarrow e9$	Flu	$\cdots$
3	Cary	$b3 \rightarrow f6 \rightarrow c7 \rightarrow e8$	Hepatitis	$\cdots$
4	David	$b3 \rightarrow e4 \rightarrow f6 \rightarrow e8$	Fever	$\cdots$
5	Eric	$a1 \rightarrow d2 \rightarrow c5 \rightarrow f6 \rightarrow c7$	Flu	$\cdots$
6	Frank	$c5 \rightarrow f6 \rightarrow e9$	Hepatitis	$\cdots$
7	Gina	$f6 \rightarrow c7 \rightarrow e8$	Fever	$\cdots$
8	Henry	$a1 \rightarrow c2 \rightarrow b3 \rightarrow c7 \rightarrow e9$	Hepatitis	$\cdots$
9	Kevin	$e4 \rightarrow f6 \rightarrow e8$	Fever	$\cdots$

1.1 Motivation

It is a challenge to prevent the disclosure of a person's specific healthcare data from the published EHRs so as to preserve his/her privacy. To achieve anonymity, the original records should be modified before being published. The existing privacy preserving approaches of publishing health data are classified into generalization and suppression, anatomization and permutation, and perturbation techniques [6,8]. Generalization or suppression technique aims to hide some details of QIs. Generalization replaces some QI values with a broader category such as a parent value in the taxonomy of an attribute. Suppression eliminates a certain number of points in the trajectory for privacy. Both techniques often result in considerable information loss by modifying the trajectory or sensitive attributes [21]. Perturbation distorts the original dataset by adding noise, aggregating values, swapping values, or generating synthetic data while preserving the statistical information of the attributes. Consequently, the transformed data after perturbation can provide higher utility [19].

By adopting the idea of perturbation, we consider the problem of publishing EHRs for more accurate analysis while limiting the disclosure of sensitive health information. Specifically, we want to ensure that an adversary cannot reliably infer the presence of an individual by linking some QIs. In this paper, we focus on the privacy breach caused by the healthcare trajectory in EHRs. For example, the sequence $e4 \rightarrow f6$ as background knowledge cannot infer a specific record in Table 1. Thereby, the privacy of each record is preserved. In this paper, we introduce a novel data perturbation approach to protect the privacy of sensitive health data and resist the following two kinds of attacks, record linkage and attribute linkage [14]. Record disclosure happens when a target user can be identified from a specific sequences in healthcare trajectory. Attribute disclosure occurs when some revealed attributes can link to a specific individual or infer a victims sensitive information. In our approach, we first identify the critical sequences in the healthcare trajectory that are prone to privacy breaches. For each sequence, we use addition or subtraction to implement l-diversity so as to ensure that each sequence matches at least l types of SA values in the published data.

1.2 Contributions and Organization

In this paper, we propose a novel scheme to preserve the health or medical privacy of EHRs with a single SA, named Privacy Preservation in Publishing Electronic Health Records Based on Perturbation (PPHR). Given all the above considerations, this paper has the following contributions:

- We propose our l-diversity privacy model to protect the sensitive information such as *Disease* in EHRs. l-diversity ensures that at least l records are matched by the attacker based on the healthcare trajectory sequence as background knowledge, which can be set according to the owner' s requirement.

- Our PPHR includes two steps, determining critical sequences and anonymizing data using perturbations. To the best of our knowledge, we are the first to perform perturbation to protect the sensitive attribute in EHRs.
- We evaluate the performance through extensive simulations based on a real-world data set. Compared with **PPTD** [14], and **KCL-Local** [4], our mechanism is superior in data utility ratio with better privacy.

The remainder of this paper is organized as follows. In Sect. 2, we discuss the related work. Preliminaries are given in Sect. 3. In Sect. 4, we present the details of our approach. Simulations on data utility are presented in Sect. 5. Finally, we conclude our work in Sect. 6.

2 Related Work

In this section, we mainly introduce existing approaches to prevent the privacy leakage of the published data from the following three categories, generalization, suppression and perturbation [8,28].

Generalization-Based. Generalization is one of the most common anonymity operations to implement k-anonymity for privacy protection. Generalization replaces some QI values with a broader category such as a parent value in the taxonomy of an attribute. In [12,31], a taxonomy tree was built first and then a node in the tree was generalized to its parent node, which aimed to reach k-anonymity. In [16], a node' s attribute was replaced by its sibling' s attribute. Generalization was first proposed in [13] to process trajectories and sensitive attributes based on different privacy requirements of moving objects. Gao [7] proposed to use trajectory angle to evaluate trajectory similarity and direction, and construct an anonymity region on the basis of trajectory distance so as to achieve k-anonymity. In [9], generalization technique is applied to anonymize the trajectory data and a heuristic approach is proposed to achieve LKC-privacy model. A look-up table brute-force (LT-BF) algorithm is proposed to preserve privacy and maintain the data quality based on LKC-privacy model by applying the generalization technique in [10].

Suppression-Based. Suppression approaches aim to replace some attributes with some special symbolic characters. It was first adopted to satisfy the constraint of breach probability in [29]. [4] was the first paper to adopt suppression to prevent record linkage and attribute linkage attacks. In [2], k^m-anonymity was proposed to suppress the critical location points chosen from the quasi-identifiers in order to resist the attacks based on the background knowledge of m moving points. In [24], locations suppression and trajectories splitting are used to protect privacy and ensure the accuracy of query answering and frequent subsets.

Perturbation. Data perturbation is considered as a relatively easy and effective technique in protecting sensitive electronic data from unauthorized users. There are two main types of data perturbation appropriate for EHR data protection. The first type is known as the probability distribution approach and the second

type is called the value distortion approach. Perturbation distorts the data by adding noise, swapping values, or generating synthetic data [6]. In [1,3,30], noise was added to protect the privacy of sensitive attribute by achieving ϵ-differential privacy. Sensitive attribute values are exchanged among records to achieve data swapping [5]. Random edge perturbation was used to resist structural identification attack in [27].

Summary of Related Work. To prevent a specific individual from being re-identified from the published tables, the key solution is to protect the privacy of some sensitive information. In addition, the design of an anonymization approach should consider the balance between the data utility and the privacy preservation. Generalization and suppression often result in considerable information loss by modifying quasi-identifiers or sensitive attributes, which often causes severe loss of data analysis [21]. Comparatively, perturbation can maintain the statistical properties of published data without changing any sensitive attribute. In this paper, we adopt the perturbation technique to achieve l-diversity of the sensitive attribute. Compared with k-anonymity, l-diversity is practical and can addresses the shortcomings of k-anonymity with respect to the background knowledge such as record linkage attack and attribute linkage attack [18].

3 Preliminaries

In this section, we introduce some knowledge on the database of EHRs and two kinds of attacks.

3.1 EHRs Database

Patient healthcare trajectory [22] is a recent emergent topic, focusing on the patient trajectory based on disease management and care. A healthcare trajectory is similar to a moving path, which consists of many different positions at different timestamps. By regularly collecting the corresponding trajectory of one patient, the hospital can trace the patient disease and determine the relationship between disease and patient trajectory. The definitions of healthcare trajectory and electronic health record are given as follows:

Definition 1 *(**Healthcare Trajectory**). A healthcare trajectory is published based on the time order. Each trajectory point (such as a*1*) has two essential components, a measurement result (such as a) and a time stamp (such as* 1*), which indicate where a subject is get a measurement result at a given time instant.*

$$t = (r_1, t_1) \rightarrow (r_2, t_2) \rightarrow \cdots \rightarrow (r_k, t_k). \tag{1}$$

where k is the length of trajectory, t_i is a time stamp and r_i represents a measurement result of a data owner.

Definition 2 ***(Electronic Health Record(EHR)*** [23]. *An EHR is compose of several attributes such as ID, Name, Healthcare Trajectory, Disease and other attributes in Table 1.*

$$EHR = <ID, Name, t = (r_1, t_1) \rightarrow (r_2, t_2) \rightarrow \cdots \rightarrow (r_k, t_k), SA, ...>, \quad (2)$$

where SA represents the sensitive attribute such as Disease.

3.2 Privacy Attack

In this paper, we focus on protecting sensitive attributes in publishing EHRs such as those in Table 1. The attacker uses a sequence of at least one point in the healthcare trajectory as background knowledge to launch record linkage attack and attribute linkage attack and thereby infer the sensitive attribute of the data owner such as *Disease*.

- **Record linkage attack.** The attacker matches a specific record according to the trajectory sequence in the publishing data and can directly identify the specific data owner. When some trajectory sequences in the data occur at a low frequency, the attacker can easily identify the specific record of the data owner from the data. For example, we assume that the attacker knows that a data owner has a sequence $c2 \rightarrow b3$ in the healthcare trajectory. It is easy to speculate that Henry has Hepatitis from Table 1.
- **Attribute linkage attack.** The attacker cannot lock to a specific record, but the SA distribution of the matched records is very concentrated. The attacker can infer that the data owner possess a certain attribute at a higher probability. For example, we assume that the attacker knows a sequence $c7 \rightarrow e9$. He can infer that the data owner may suffer Flu or $Hepatitis$ with the probability of $\frac{1}{2}$ or $\frac{1}{2}$ respectively because the 2nd and 8th records contain this sequence.

To resist these two attacks, we anonymize the original data set T into T^* to achieve l-diversity. Assuming that the attacker uses the trajectory sequence with an upper bound length of m as the background knowledge. *l-diversity* is defined as follows:

Definition 3 ***(l-diversity)***. *The anonymized dataset T^* satisfies l-diversity if for any sequence q that does not exceed m in length, all records that q matches contain at least l types of SA values: $\forall q \in T^*$, $|SV(q)| \geq l$, where $SV(q)$ represents all the SA values associated with q.*

4 Privacy Preservation in Publishing Electronic Health Records Based on Perturbation (PPHR)

Our goal is to protect sensitive information in publishing EHRs by implementing l-diversity and to provide the data utility. The notations commonly used in this section are listed in Table 2.

Table 2. Notations

Notations	Description
T	The original data before being published
m	Maximum length of the trajectory sequence as the adversary's background knowledge
CS	Set of sequences whose SA values do not satisfy l-diversity
$T(q)$	Records including q in T
$SV(q)$	All the SA values associated with q in T
SP	Set of sequences in CS that need subtraction operation
AP	Set of sequences in CS that need addition operation

4.1 Overview

Our PPHR aims to protect the privacy of sensitive attribute and resist the attacks based on the background knowledge of a part of healthcare trajectory. PPHR can be divided into two steps: identifying critical sequences in the trajectory data and anonymizing the dataset T. A critical sequence is one whose length does not exceed m and the number of SA values corresponding to this sequence does not satisfy l-diversity. To achieve l-diversity of SA values, we implement perturbation by adding or subtracting points in the trajectory sequences including the critical sequences as shown in Fig. 1:

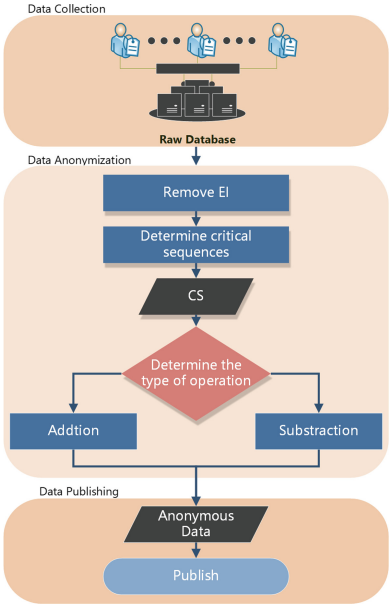


Fig. 1. Architecture of PPHR

4.2 Algorithm

In order to achieve l-diversity of SA values, we use perturbation to obscure the correlation between healthcare trajectory and SA. We first identify the critical sequences that are easy to reveal the privacy of patients, and then revise the matching records of these sequences to achieve l-diversity.

Determining Critical Sequences. In this process, we find those sequences whose length is equal to m and does not satisfy l-diversity. The steps to determine the critical sequence in the trajectory are listed as follows:

Step 1: First, the trajectory sequences of length m in each record are determined. In addition, if the whole length of a user's health trajectory is less than m, the trajectory will be checked whether it can be treated as a critical sequence in the next step.

Step 2: For each sequence q got in **Step** 1, if the number of types of the corresponding SA values matched by q in T is less than l, i.e. $|SV(q)| < l$, q will be regarded as a critical sequence and be added into CS, where $|SV(q)|$ represents all the SA values associated with q in T and CS is the set of sequences whose SA values do not satisfy l-diversity.

Algorithm 1. Determining critical sequences

Require:
Original dataset of EHRs: T
Ensure:
Critical sequences: CS

$CS \leftarrow Null$ ▷ set of sequences whose SA values do not satisfy l-diversity.
for each trajectory $t \in T$ **do**
  if $length(t) \leq m$ **then**
    if $\mid SV(t) \mid < l$ **then**
      $add\ t \rightarrow CS$
    end if
  else
    $Q \leftarrow Null$ ▷ set of sequences of length m
    Add all the sequences q of length m in t to Q
    for each sequence$q \in Q$ **do**
      if $\mid SV(t) \mid < l$ **then**
        $add\ q \rightarrow CS$
      end if
    end for
  end if
end for
return CS

Performing the Anonymization. In this process, we execute addition or substraction operation to make SA satisfy l-diversity. For each sequence q in CS, we first determine the addition or substraction operation by evaluating the data utility. Then, l-diversity of SA values corresponding to q will be satisfied by adding or substracting points in the healthcare trajectory of records corresponding to q.

Algorithm 2. Performing the anonymization

Require:
Original dataset of EHRs:T
Ensure:
Anonymous dataset of EHRs: T^*

$T^* \leftarrow T$
$CS \leftarrow Null$ ▷ set of sequences whose SA values do not satisfy l-diversity.
$SP \leftarrow Null$ ▷ Set of sequences in CS that need substraction operation.
$AP \leftarrow Null$ ▷ Set of sequences in CS that need addition operation.
for each sequence $q \in CS$ **do**
  if $|T(q)| \leq (l - |SV(q)|) * |q|$ **then**
    $add\ q \rightarrow SP$
  else
    $add\ q \rightarrow AP$
  end if
end for
for each sequence $q \in SP$ **do**
  for each point $p \in q$ **do**
    if no new critical sequence caused by subtracting p **then**
      $subtracting\ p\ from\ T^*(q)$
    end if
  end for
end for
for each $sequence \in AD$ **do**
  $AlterRec \leftarrow Null$ ▷ The records can be constructed q
  Add the records whose SA values are not in $SV(q)$ and there is no location conflict at the corresponding timestamp into $AlterRec$
  sort $AlterRec$ by LCS
  constructed q in first $l - |SV(q)|$ records of $AlterRec$ in T^*
end for
$return\ T^*$

Step 1: We define the following criteria to determine addition or substraction operation,

$$CR(q) = |T(q)| - (l - |SV(q)|) * |q|, \tag{3}$$

where $|T(q)|$ represents the number of records that include q, $SV(q)$ all the SA values associated with q, and $|q|$ the length of q. When l-diversity is not satisfied, $l - |SV(q)|$ indicates the number of different SA values that need to be added in order to satisfy l-diversity. $(l - |SV(q)|) * |q|$ represents the upper limit of number of points to be add to achieve l-diversity. $CR(q) \leq 0$ means the number of points

modified in the anonymized data will be smaller if the subtraction operation is executed. In this case, a better data utility can be provided. q will be add to the set SP. Otherwise, if $CR(q) > 0$ holds, the addition operation is necessary and q will be add to the set AP.

Step 2: For each critical sequence q in SP, we use the subtraction method to eliminate q from T, but a new critical sequence cannot be generated. When a special point is moved from all the records in $T(q)$, q will not appear any more in the published data. Consequently, there is no any privacy leakage caused by q. If a new critical sequence is caused by executing the subtraction operation, q will be added into AD.

For example, 2-diversity is not satisfied for the sequence $q = f6 \rightarrow e9$ in Table 1, because there is only one SA value such as Ben's disease. To achieve 2-diversity, we execute subtraction to process q. If $f6$ is moved from $q = f6 \rightarrow e9$, $e9$ can achieve 2-diversity such as the 2nd and 8th records in Table 1. But, $c5 \rightarrow f6$ will be a new critical sequence, because only one value for this sequence such as the 5th record exists. If $e9$ is moved from $q = f6 \rightarrow e9$, there is no new critical sequence generated. Finally, we will subtract $e9$ to eliminate the privacy threat of the original sequence $q = f6 \rightarrow e9$.

Step 3: For each sequence q in AP, we use addition operation to construct q on the selected records to satisfy l-diversity.

First of all, we select the records whose SA values are not in $SV(q)$ because we must increase the variety of SA values in order to achieve l-diversity. In addition, we need to add points at some timestamps to construct q. When adding a point, we must ensure that there is no same point at the corresponding timestamp in T.

Last, we use the Longest Common Subsequence (LCS) to sort the selected records. A sequence will be the longest common subsequence if it is a subsequence of two or more sequences and is the longest of all subsequences. For example, the LCS of $a1 \rightarrow d2 \rightarrow c5 \rightarrow f6 \rightarrow c7$ and $c5 \rightarrow f6 \rightarrow e9$ is $c5 \rightarrow f6$. We choose $l - |SV(q)|$ records which have longer LCS to construct q to satisfy l-diversity.

4.3 Privacy Analysis

In our algorithm, we only need to consider sequences of length m. We use perturbation to process the sequences of length m to achieve l-diversity of SA. In this section, we aim to prove that those sequences of length less than m make l-diversity be satisfied if l-diversity is met for all sequences of length m.

For each sequence q of length less than m, we assume q is the subsequence of n parent sequences $q_1, q_2 ..., q_n$ which q_i represents a sequence of length m. The records in $SV(q)$ are composed of all records in $SV(q_i)$ for $i = 1 \cdots n$. We can get the following equations:

$$\begin{aligned} SV(q) &= SV(q_1) \cup SV(q_2) \cdots \cup SV(q_n) \\ |SV(q)| &= |SV(q_1) \cup \cdots \cup SV(q_n)| \\ &\geq |SV(q_i)| \\ &\geq l \end{aligned} \tag{4}$$

Consequently, we can prove that all the sequences of length no more than m can make *l-diversity* be satisfied in the anonymized data T^*. l-diversity of the SA values is achieved in T^*.

5 Performance Evaluations

To evaluate the performance of our PPHR, we use a real-world dataset **MIMIC-III** dataset [11]. MIMIC is a publicly available data set which includes identified health data associated with approximately 40,000 patients. It includes personal information, diagnostic information, medication information, measurement results, etc. We selected the health data of 11,047 patients. *Disease* as SA contains 32 possible values and 8 of them are considered as sensitive values. The health measurement history of these 11,047 patients contains 90 types of disease and 24 different timestamps. We implement our PPHR algorithm in Python. We evaluate the performance on a PC with an Intel Core i7 2.5 GHz CPU and 8 GB RAM.

We compare our PPHR with **PPTD** [14], and **KCL-Local** [4]. **KCL-Local** combines local suppression and global suppression to implement $(k, C)_m$-privacy model. $(k, C)_m$-privacy model can implement k-anonymity to resist record linkage attack and implement C confidence to resist attribute linkage attack. m is the upper limit of the attacker's background knowledge as defined in this paper. **PPTD** achieve personalized privacy with sensitive attribute generalization and trajectory local suppression which also resist record linkage attack and attribute linkage attack.

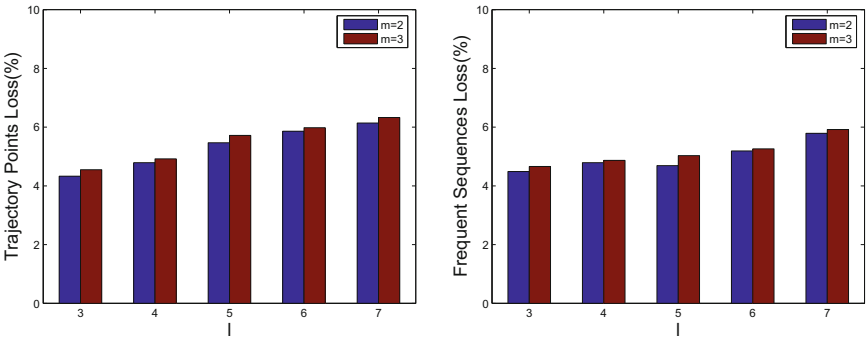
5.1 Utility Loss

In this section, we the following metrics to evaluate the performance of data utility [4,14].

- **Trajectory Points Loss (TPL)**, the loss rate of trajectory points data after anonymization which contains ratios for increasing and decreasing trajectory points, is defined as $\frac{|P(T^*)-P(T)|+|P(T)-P(T^*)|}{|P(T)|}$, where $P(T^*)$ and $P(T)$ are the sets of trajectory points in T^* and T.
- **Frequent Sequences Loss (FSL)**, the loss rate of frequent sequences which contains ratios for increasing and decreasing frequent sequences, is defined as $\frac{|F(T^*)-F(T)|+|F(T)-F(T^*)|}{|F(T)|}$, where $F(T^*)$ and $F(T)$ are the sets of frequent sequences in T^* and T.

To study the effectiveness of PPHR, we evaluate the utility loss by varying l and m. For frequent sequences loss, we choose $K' = 70$ which is the frequency threshold of frequent trajectory sequences.

Effect of *l*. Figure 2 shows that the impact of l on TPL and FSL. As l varies from 3 to 7, both types of utility loss increase slowly because as privacy requirements increase, more points need to be added or subtracted in our PPHR. For different sequences, we take n appropriate addition or substraction operation to achieve *l-diversity*, which can effectively reduce the utility loss. Addition is more conducive to protect the frequent sequences. In addition, as m increases, utility loss also increases.



(a) TPL vs. l (b) FSL vs. l

Fig. 2. Utility Loss vs. l-diversity.

5.2 Leakage Probability

We use the leakage probability as a measure of the probability that each sequence could cause a privacy breach. The leakage probability of a sequence q is defined as:

$$Pr_{leak}(q) = max(\frac{1}{|SV(q)|}, \frac{max_{SA}}{|T(q)|}),$$

where $\frac{1}{|SV(q)|}$ and $\frac{max_{SA}}{|T(q)|}$ represent the probability of identity disclosure and that of attribute disclosure respectively.

We randomly sample 20k sequences whose length is not more than m to calculate the leakage probability of each sequence. The average leakage probability is shown in Fig. 3. As l increases, leakage probability gradually decreases because both the number of records that q matches and the types of SA increase.

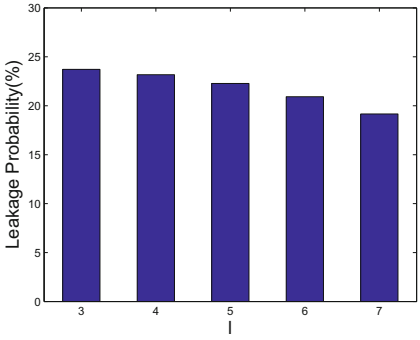
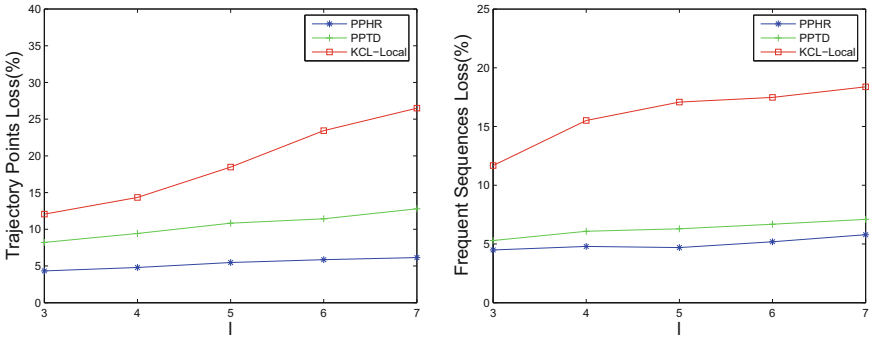


Fig. 3. Leakage probability vs. l

5.3 Comparison

We compare our PPHR with **KCL-Local** and **PPTD** on both types of utility loss and runtime. Our PPHR achieves l-diversity to defend against both attacks while KCL-Local and PPTD achieve $(k, C)_m$-privacy model. Though these kinds of schemes implement different privacy models, they can resist record linkage attack and attribute linkage attack. Consequently, we compare them by evaluating the leakage probability as the privacy protection degree. Then, we compare the utility loss under the same level of privacy protection. For example, leakage probability of 3-diversity and $(5, 0.5)$-privacy model means the same level of privacy protection.



(a) TPL (b) FSL

Fig. 4. Utility loss

We vary l from 3 to 7 with $m = 3$ to compare the effect of l on PPHR. We set the values of k and C to ensure these 3 schemes can provide the same privacy level as l varies. Figure 4 shows KCL-Local has the worst performance because

a large number of points are subtracted by the global suppression and therefore utility loss is caused severely. The performance of PPTD is better than KCL-Local, because PPTD only handles records that may cause privacy breaches. Our PPHR chooses addition or subtraction options to achieve the best data utility by trying to change fewer points.

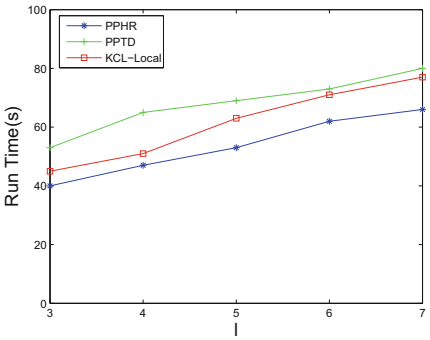


Fig. 5. Runtime

Runtime. Figure 5 shows the runtime increases with l because more sequences are processed, causing more time. PPTD has the longest running time because it takes some time to determine new critical sequences. Our PPHR has the shortest running time because we only process the sequences of length m but KCL-Local and PPTD should deal with the sequences of length no more than m. Besides, it is no necessary to consider the influence of new critical sequences during the addition operation in our PPHR.

6 Conclusion

We design and implement an anonymous technique to protect the sensitive attribute during publishing the EHRs. In our scheme, we first determine the critical sequences based on which some specific patients are easy to be identified. To resist the record linkage attack and attribute linkage attack, we adopt perturbation to process these critical sequences by adding or deleting some points to make the SA values in the published data satisfy l-diversity. Our performance studies based on a comprehensive set of real-world data demonstrate that our scheme can provide higher data utility compared to peer schemes. In the future work, we plan to optimization our algorithm to resist other linkage attacks.

Acknowledgment. This research is sponsored in part by National Key Research and Development Program of China (2017YFC0704200), the National Natural Science Foundation of China (contract/grant numbers: 61772113 and 61872053).

References

1. Ahmed, F., Liu, A.X., Jin, R.: Social graph publishing with privacy guarantees. In: IEEE 36th International Conference on Distributed Computing Systems (ICDCS), pp. 447–456. IEEE (2016)
2. Brito, F.T., Neto, A.C.A., Costa, C.F., Mendonça, A.L., Machado, J.C.: A distributed approach for privacy preservation in the publication of trajectory data. In: Proceedings of the 2nd Workshop on Privacy in Geographic Information Collection and Analysis, p. 5. ACM (2015)
3. Cano, I., Torra, V.: Edit constraints on microaggregation and additive noise. In: Dimitrakakis, C., Gkoulalas-Divanis, A., Mitrokotsa, A., Verykios, V.S., Saygin, Y. (eds.) PSDML 2010. LNCS (LNAI), vol. 6549, pp. 1–14. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19896-0_1
4. Chen, R., Fung, B.C., Mohammed, N., Desai, B.C., Wang, K.: Privacy-preserving trajectory data publishing by local suppression. Inform. Sci. **231**, 83–97 (2013)
5. Domingo-Ferrer, J., Torra, V.: A quantitative comparison of disclosure control methods for microdata. In: Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, pp. 111–134 (2001)
6. Fung, B., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: a survey of recent developments. ACM Comput. Surv. (CSUR) **42**(4), 14 (2010)
7. Gao, S., Ma, J., Sun, C., Li, X.: Balancing trajectory privacy and data utility using a personalized anonymization model. J. Netw. Comput. Appl. **38**(1), 125–134 (2014)
8. Gkoulalas-Divanis, A., Loukides, G., Sun, J.: Publishing data from electronic health records while preserving privacy: a survey of algorithms. J. Biomed. Inform. **50**(8), 4–19 (2014)
9. Harnsamut, N., Natwichai, J.: Privacy preservation for trajectory data publishing and heuristic approach. In: Barolli, L., Enokido, T., Takizawa, M. (eds.) NBiS 2017. LNDECT, vol. 7, pp. 787–797. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-65521-5_71
10. Harnsamut, N., Natwichai, J., Riyana, S.: Privacy preservation for trajectory data publishing by look-up table generalization. In: Wang, J., Cong, G., Chen, J., Qi, J. (eds.) ADC 2018. LNCS, vol. 10837, pp. 15–27. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-92013-9_2
11. Johnson, A.E.W., et al.: MIMIC-III, a freely accessible critical care database. Sci. Data **3**, 160035 (2016)
12. Kiyomoto, S., Tanaka, T.: A user-oriented anonymization mechanism for public data. In: Garcia-Alfaro, J., Navarro-Arribas, G., Cavalli, A., Leneutre, J. (eds.) DPM/SETOP -2010. LNCS, vol. 6514, pp. 22–35. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19348-4_3
13. Komishani, E.G., Abadi, M.: A generalization-based approach for personalized privacy preservation in trajectory data publishing. In: Sixth International Symposium on Telecommunications, pp. 1129–1135 (2012)
14. Komishani, E.G., Abadi, M., Deldar, F.: PPTD: preserving personalized privacy in trajectory data publishing by sensitive attribute generalization and trajectory local suppression. Knowl. Based Syst. **94**, 43–59 (2016)
15. Lang, L., Lang, L.: Hipaa privacy rule and negative influence on health research. Gastroenterology **134**(1), 6–6 (2008)
16. LeFevre, K., DeWitt, D.J., Ramakrishnan, R.: Incognito: efficient full-domain k-anonymity. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp. 49–60. ACM (2005)

17. Loukides, G., Liagouris, J., Gkoulalas-Divanis, A., Terrovitis, M.: Disassociation for electronic health record privacy. J. Biomed. Inform. **50**(8), 46–61 (2014)
18. Machanavajjhala, A., Gehrke, J., Kifer, D.: l-diversity: privacy beyond k-anonymity. In: International Conference on Data Engineering, p. 24 (2006)
19. Okkalioglu, B.D., Koc, M., Koc, M., Polat, H.: A survey: deriving private information from perturbed data. Artif. Intell. Rev. **44**(4), 547–569 (2015)
20. Ozkaynak, M., Reeder, B., Hoffecker, L., Makic, M.B., Sousa, K.: Use of electronic health records by nurses for symptom management in inpatient settings: a systematic review. Comput. Inform. Nurs. (CIN) 1 (2017)
21. Rajaei, M., Haghjoo, M.S., Miyaneh, E.K.: Ambiguity in social network data for presence, sensitive-attribute, degree and relationship privacy protection. PLoS ONE **10**(6), 1–23 (2015)
22. Tai-Seale, M., Wilson, C.J., Stone, A., Durbin, M., Luft, H.S.: Patients body mass index and blood pressure over time: diagnoses, treatments, and the effects of comorbidities. Med. Care **52**, S110–S117 (2014)
23. Tavares, J., Oliveira, T.: Electronic health record patient portal adoption by health care consumers: an acceptance model and survey. J. Med. Internet Res. **18**(3), e49 (2016)
24. Terrovitis, M., Poulis, G., Mamoulis, N., Skiadopoulos, S.: Local suppression and splitting techniques for privacy preserving publication of trajectories. IEEE Trans. Knowl. Data Eng. **29**(7), 1466–1479 (2017)
25. Victor, N., Lopez, D., Abawajy, J.H.: Privacy models for big data: a survey. Int. J. Big Data Intell. **3**(1), 61–75 (2016)
26. Xu, Y., Ma, T., Tang, M., Tian, W.: A survey of privacy preserving data publishing using generalization and suppression. Appl. Math. Inf. Sci. **8**(3), 1103 (2014)
27. Xue, M., Karras, P., Chedy, R., Kalnis, P., Pung, H.K.: Delineating social network data anonymization via random edge perturbation. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 475–484. ACM (2012)
28. Yao, L., Liu, D., Wang, X., Wu, G.: Preserving the relationship privacy of the published social-network data based on compressive sensing. In: IEEE/ACM International Symposium on Quality of Service, pp. 1–10 (2017)
29. Yarovoy, R., Bonchi, F., Lakshmanan, L.V.S., Wang, W.H.: Anonymizing moving objects: how to hide a mob in a crowd? In: International Conference on Extending Database Technology, EDBT 2009, Saint Petersburg, Russia, 24–26 March 2009, Proceedings, pp. 72–83 (2009)
30. Zaman, A.N.K., Obimbo, C., Dara, R.A.: An improved data sanitization algorithm for privacy preserving medical data publishing. In: Canadian Conference on Artificial Intelligence, pp. 64–70 (2017)
31. Zhang, X., Liu, C., Nepal, S., Chen, J.: An efficient quasi-identifier index based approach for privacy preservation over incremental data sets on cloud. J. Comput. Syst. Sci. **79**(5), 542–555 (2013)