Privacy Preserving Mechanism for Multi-sensitive attributes Based on Time sequence

Qiuwei Yang
College of Information Science and Engineering
Hunan University
Changsha, China
yky_wenfeng@163.com

Yingting Li
College of Information Science and Engineering
Hunan University
Changsha, China
lyt365@hnu.edu.cn

***Abstract*—Current privacy preserving data publishing techniques mostly concentrate on attributes of a single releasing. Nevertheless , most of the practical application may contain multiple release attributes. Directly applying the existing single release attribute to privacy preserving techniques often leads to unexpected private information leakage.This paper is first published by the time sequence introduced to several published data set problems, combined with the published data set Markov chain thought towards preserve data privacy. Experimental results show that the algorithm can effectively prevent the information loss of privacy, and enhance the security of the data release.**

Keywords-Privacy preserving; Time sequence;Markov chain; Multi-sensitive attributes

I. Introduction

Most existing studies have focused on a single data set release behavior analysis, there is a logical correlation or data semantics associated segmentation, so as to achieve the purpose of privacy protection. It is worth noting that, in most practic application scenarios, the independent analysis of the release behavior of a single set of attributes has been unable to meet privacy requirement.This paper presents a multi-sensitive attributes with time sequence published privacy preserving, release time series exist in different physical attributes for the implementation of certain treatment strategies. The basic strategy is constructed with the temporal characteristics of the safety data released on the one hand to make anonymous data released by the non-disclosure of data of individual privacy and need to ensure the release of anonymous data available highly, that can still be based on published anonymous data is more accurate data analysis. Finally, through the experiments validates that it has small loss and high practicability.

Focus on a particular public attributes alone, will not disclose the intention reluctant public information, when there are two or more attributes published simultaneously.A malicious analysts are reasoning, mining, association analysis, background knowledge, do not want to get public entities. There has information on conservation value. When the data content changes, the data set must be released, if it is in accordance with the original static anonymous strategy, even though each release version anonymity requirements are met, an attacker can combine multiple data tables published formed between multiple versions inference channel, thus leading to privacy information leakage.

II. Preliminary Analysis

Sensitive attribute datasets $\{a_s,a_{s+1},\ldots,a_t\}$ often exist some closely relating with sensitive attributes. However, the attributes will not directly contain their own individual privacy information, and the association between these attributes and sensitive attributes are public, calling such properties are associated with sensitive properties. As shown in Table 1, the attribute of *Disease* is associated with attribute of *Physician*, which means the attribute values can be deduced *Physician* entity's illness information. Table 2 reflects a property set that exists several properties of the interaction.

Table 1 Multi-attributes sensitive private data relationships

ID	Name	Sex	Zip code	Physician	Disease
t_1	San	M	12000	John	Flu
t_2	Anne	F	21000	John	Pneumonia
t_3	Mike	M	11500	John	Cancer
t_4	Lily	M	66000	Bob	Flu
t_5	Harry	F	42000	Bob	Pneumonia
t_6	Lucy	M	16000	Sam	HIV

Table 2 Sensitive attributes associated information

Physician	Disease
John	Flu,Pneumonia,Cancer
Bob	Flu,Pneumonia
Sam	HIV

Assuming the set of attributes that is released for the i times is represented by $A(i)$, the next times with A (j). Supposing that the system is safe when released in the i times, then the j times when released, there are three relationship models as follows:

1) one to one relationship model

If $\exists\, a \in A(i), \exists\, b \in A(j)$, and a is a sensitive property of A, then $b \rightarrow a$.

2) many to one relationship model

If $\exists\, a \in A(i), \exists\, b \in A(j), \exists\, c \in A(j)...$,
and a is a sensitive property of A, then $(b \wedge c \wedge ...) \rightarrow a$.

3) many to many relationship model

If $\exists\, a \in A(i)$, $\exists\, b \in A(i)...$, $\exists\, p \in A(j)$, $\exists\, q \in A(j)...$, and a and b are sensitive attribute A, then $(p \wedge q...) \rightarrow (a \wedge b...)$.

We first describe the assumptions on datasets and their releases, then we discuss possible inference rules that exist among multiple data releases and publishing requirements for preventing such inferences. Let the set of attributes $A=\{a_i,a_{i+1},...,a_s\}$, $B=\{a_j,a_{j+1},...,a_t\}$.Decision function associated with reasoning as follows:

$$\theta(A,B)=\begin{cases} true\ \ if\ \ A \rightarrow B \\ false\ \ else \end{cases} \qquad (1)$$

III. Privacy protection and safety assessment

A. *Based on Markov chain attributes published behavior description*

Assume the time sequence $T=\{t_1,t_2,...,t_i\}$,where $1 \le i \le n$, the sequence is defined as the release properties, random process $\{X(t), t \in T\}$, where the time sequence is defined as $T=\{0,1,2....\}$, and the state space is $I=\{0,1,2...\}$,if for any time n and any state i_0, i_1, $i_2,...,i_{n-1}$, i, j,there are $P\{X(n+1)=j|X(n)=i,\ X(n\text{-}1)=i_{n\text{-}1},...,X(1)=i_1,\ X(0)=i_0\ \}$ $=P\{X(n+1)=j|X(n)=i\}$, called $\{X(t), t \in T\}$ is called Markov chain, abbreviated as$\{X_n,\ n \ge 0\}$.

The Markov chain probability of $P_{ij}(n+1,n)=$ $P\{X_{n+1}=j|X_n=i\}$ is called the conditional or transitional probability which under the conditions of $X_{n+1}=j$, $X_n=i$. Indicating $X(t)$ in the state at n+1 times , $X\ (n+1) = j$ the probability distribution only maintains the state at n times X $(n) = i$, but not irrelevant with the previous state $X\ (n\text{-}1) = i_{n\text{-}1}$, $X\ (n\text{-}2) = i_{n\text{-}2}$ ..., $X\ (1) = i_1$, $X\ (0) = i_0$.

B. *The quantization of reasoning association*

1) Data release security description

Random variable X_n represents the n times data security release. X_n=0 indicates leakage, and X_n=1 indicates security, where n=0,1,2...., and $a_i(n)$ indicates the probability of the system is in state i when the n times release, where $i \in \{0,1\}$, i.e. $a_i(n)=P(X_n=i)$. P_{ij} represents that the current state is i, and the next release probability is j,where $i,j \in \{0,1\}$, i.e., $P_{ij}=P(X_{n+1}=j \mid X_n=i)$. $a_i(n)$ is the state of probability, and P_{ij} is called the state of transition probability.

Publishing information on the status of each description, the first release of the total sets of attributes is defined as set of *c_all*, at the time of each releasing, the total sets will be divided in three parts. There is a public attribute set collection *c_pub*, the collection *c_infer* that can be derived by public set using the associated rules, and the collection *c_rest* that represents unreleased attributes. These collections meet the condition of *c_pub* ∪ *c_infer* ∪ *c_rest* = *c_all*. When released, the system determines the current input to decide whether *c_pub* open set can be deduced to some sensitive attributes of the *c_infer set* . If there exists leakage during reasoning period, it will be processed for safety to ensure that the current information is security.

2) The calculation of attributes impact factor

The set of attributes, n available set of attributes collection $A=\{a_1,a_2,a_3,...a_i,...,a_n\}$, m privacy attributes collection $B=\{b_1,b_2,b_3,...\ b_i,...,b_m\}$. Assuming the set of A that discloses a single property do not contribute to privacy loss, meanwhile publicly disclosing more than two sets will appear reasoning relationship among attributes.

Assuming k total joint inference relations, each of reasoning rules has c_i attributes, then indicated as

$$f(a_i)=\sum_{i=1}^{k}\frac{1}{kc_i} \qquad (2)$$

The formula represents a_i impactive factor, where f satisfies the conditions as follow:

$$\text{i) } f(a_i) \ge 0 \qquad (3)$$

$$\text{ii)} \sum_{i=1}^{k} f(a_i)=1 \qquad (4)$$

C. *The measurement of privacy security*

Privacy security strategy includes system and release behavior of security. In some applications, it's required to preserve privacy from the whole perspective.

Viewed from the microscopic point of view, due to different impact factors of each property, the disclosure set may be the same threshold, and these properties may be from the same or different disclosure rules, which is difficult to distinguish. There is a setting that, when the attribute information must be released the same disclosure rules to determine which method to be used.

With disclosing of attributes $S_1,S_2,...,S_i,S_n$ represents a series of release states, we use disclosures of attributes set to describe the current behavior. When the i-1 times and the i times released, system security is $P(S_{i\text{-}1})$ and $P(S_i)$ respectively. According to the Markov chain probability , there is an association between the two releasing attribute information, which can be expressed as follows:

$$P(S_i \mid S_{i-1})=P(S_i * S_{i-1})/P(S_{i-1}) \qquad (5)$$

$$P(S_i * S_{i-1})=1-\sum_{k=1}^{i} f(k) \qquad (6)$$

$$P(S_{i-1})=1-\sum_{k=1}^{i-1} f(k) \qquad (7)$$

D. Phased release strategy with the time sequence and impactive rules

Defining the time sequence $t_1,t_2,...t_i$ ($1\leq i\leq N$) for recording each property released time respectively, i.e. a phased release property information, we decide whether there is a privacy leakage at some point released.Under the special circumstance,when the moment of t_i release strategy is safety, while the another moment t_j, gradually public with other attributes, combined with reasoning rules appearing between attributes, derived by determining the security of privacy disclosure of the current system, which takes the measurement of privacy protection policy.

With the time sequences $t_1,t_2,...,t_i$ ($1\leq i\leq N$), there are two privacy disclosing cases, the first one is when the *i* time release, due to the relationship between the internal elements of a collection of properties caused privacy disclosure; second before the *i*-1 time release leakage caused the total set at the *i* time release caused.

IV. The algorithms

Data privacy preserving can be considered a special type of optimization problem where the cost of data modification must be minimized while respecting anonymity constraints. Thus, the key components of privacy preserving technique includes the determination strategy of leakage privacy and is based on privacy security measure of privacy preserving.

A. The determination strategy of leaked privacy

Algorithm 1:

```
Input:the original dataset is A, safety threshold Th_val,
        is_leakage=true
Output: the value of is_leakage
 Begin
  Set c_pub=Ø,c_infer=Ø,c_rest=A;
    For each $a_i \in c\_rest$
        $f(a_i)=0$;
    End For
    Define the number of disclosure rule is k;
    For each inference rules total of k
        Process the properties appearing exiting in reasoning rules;
        $c\_pub=c\_pub+a_i$;$c\_rest=c\_rest-a_i$;
      If $a_i$ can be generated by the inference rules
        then $c\_infer=c\_infer+a_i$;
          Calculate and update the properties of factor $f(a_i)$;
      End If
    End for
    For calculate security for each disclosure rules
            $f(a_i)=f(a_i)+1/kc_i$;
     Denoted by $\beta_1,\beta_2,...\beta_i$ ,respectively;
       $\beta=1-min\{\beta_1,\beta_2,...\beta_i\}$ //to calculate the minimum safety of current disclosure rules;
    End For
    If $Th\_val>\beta$
        then is_leakage=true;
    Else
        is_leakage=false;
    End If
End Begin
```

B. Based security measure of privacy preserving

Algorithm 2:

```
Input:the original dataset is A, safety threshold
        Th_val,is_leakage
Output: release sequence after anonymizing
Begin
  For each time disclosure rules
      If contains is_leakage=false in the current disclosure rules
      Donated as existing privacy leakage at i-th time, $\beta_{max}=max\{\beta_1,\beta_2,...\beta_i\}$;
      The properties of this disclosure rules involved denoted as collection B;
        If $a_i \in B$ && maximum impactive factor of $f(a_i)$
            Anonymize the value and split the inference rule of attribute $a_i$;
        End If
      is_leakage=true;
    End If
    Disclosure of the existence of the use of inference rules paradigm regroup;
  End For
  Print the release of the sequence output anonymous;
End Begin
```

V. Experiment Results

In this section, we describe our experimental settings and report the results in details.

A. Experimental setup

The experiments were performed on a 1.8GHz Intel IV processor machine with 2GB of RAM. The operating system on the machine was Microsoft Windows 7 Professional Edition, and the implementation was built and run in Java2 Platform, Standard Edition 5.0. For this experiments, we used the Adult dataset from the UC Irvine Machine Learning Repository , which is considered a defacto benchmark for evaluating the performance of anonymization algorithms. We remove records with missing values and retain only seven of the original attributes. In our experiments, we consider {*Gender, Sex, Age, race, zip code*} as the quasi-identifier, and {*Disease*, *Physician*} as the sensitive attributes.

B. Data Quality

The first question we discuss is how vulnerable datasets are to inferences when they are statically anonymized.In the experiment, we first anonymize 10K records and generate the first "published" dataset. We then generate twenty more subsequent datasets by anonymizing 1,000 more records each time. From figure 1, under different safety threshold β, with time series, the average information loss is presented according to the situation.

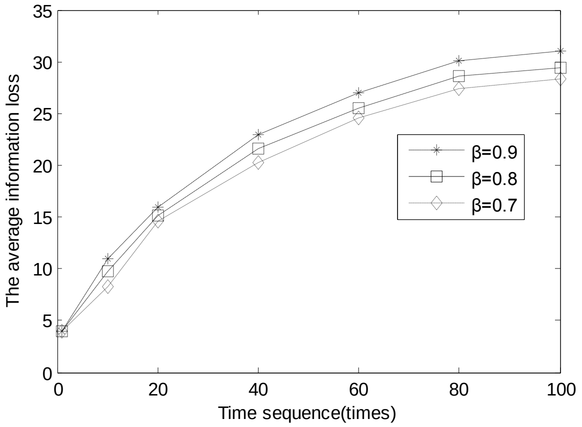

Figure 1. Data Quality of different threshold

C. Execution Time

Figure 2 shows the execution times of anonymizing various sizes of datasets. As shown, we compare the execution time of different size of datasets,respectively 3k,5k,10k.

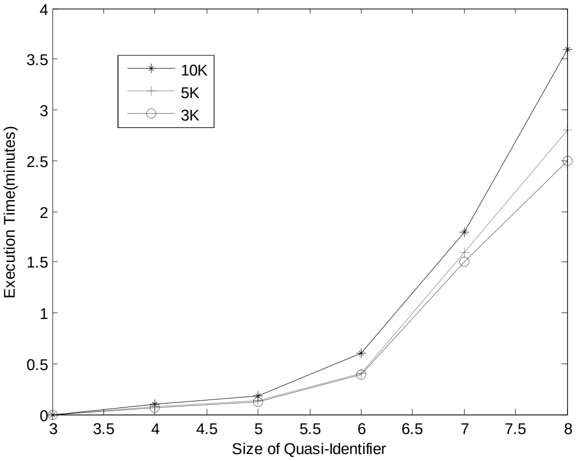

Figure 2. Execution Time

VI. Conclusion

In this paper, we present a multi-sensitive attributes with time sequence published privacy preserving, release time series existing in different physical attributes for the implementation of certain treatment strategies. As a second step, we provide an efficient algorithm that improves the limitations of previous studies, which protects privacy adequately and has high data quality and low information loss metric.

References

Traditional the "one-time" release, there are lots of models proposed. The traditional model is k-anonymity[1, 2, 3]which requires that each quasi-identifier value appears at least k times in the release table. However, it has been found that k-anonymity is vulnerable to homogeneity attack and background knowledge attack. So a much stronger model is proposed that is l-diversity[4], which requires each QI-group to contain at least l 'well-represented' sensitive values.

For the dynamic data publication, Byun [5] first proposes a solution to effectively prevent adversaries from inferring individuals' sensitive information by serial releasing datasets[6] . However this method has two drawbacks: the one is that it only considered insertions; the second is that the anonymized data will not be published until the inserted tuples themselves could satisfy l-diversity, but it will need infinite time. The References[7,8], also consider insertions only. The literature[9] develops m-Invariance, initiating a formal study of anonymous re-publication of dynamic datasets with records updates, which proposes a new idea to maintain the indistinguishability of sensitive values in each separate publication. But it also has itself drawback, it doesn't consider the effect of permanent sensitive values (e g. Lung cancer). Bu [10] presents a first consider that privacy preserving serial data publishing with permanent sensitive values and dynamic registration lists.

[1] Litai, Y., Tang, C., Wu, J., Zhou, M.: Two clustering based on k-anonymity privacy protection. Jilin University (Information Science) (February 2009).

[2] L.Sweeney, "K-anonymity: a model for protecting privacy," International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems. 2002, pp. 571-588.

[3] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression," International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems. 2002, pp. 571-588.

[4] A. Machanavajjhala, J.Gehrke, and D. Kifer, "l-diversity: Privacy beyond k-anonymity," in ICDE. 2006.

[5] J. Byun, Y. Sohn, E.Bertino, and N.li, "Secure anonymization for incremental datasets," in Secure Data Management. 2006, pp.48-63.

[6] Xin Jin ,Nan Zhang and Gautam Das ASAP: Eliminating algorithm-based disclosure in privacy-preserving data publishing.In Information Systems Volume 36, Issue 5, July 2011, Pages 859-880.

[7] J. Pei, J. Xu, Z. Wang, W. wang, and K. Wang, "Maintaining k-anonymity against incremental updates," in SSDBM. 2007.

[8] B. C. M. Fung, K. wang, A. Fu, and J. Pei, "Anonymity for continuous data publishing," in EDBT. 2008.

[9] F. Li, and S. Zhou, "Challenging More Updates: Towards Anonymous Re-publication of Fully Dynamic Datasets," in CoRR. 2008.

[10] Yingyi. Bu, Ada. Wai-Chee. Fu, Raymond. Chi-Wing.Wong, Lei. Chen, and Jiuyong. Li, "Privacy Preserving Serial Data Publishing By Role Composition," in VLDB. 2008.