Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/070546

L-Diversity Algorithm for Incremental Data Release

Pingshui Wang[1,*] *and Jiandong Wang*[2]

[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[2] Department of Computer Science and Technology, Anhui University of Finance and Economics, Bengbu 233030, China

Received: 8 Feb. 2013, Revised: 9 Jun. 2013, Accepted: 12 Jun. 2013
Published online: 1 Sep. 2013

Abstract: At present most privacy preserving algorithms based on l-diversity model are limited only to static data release. It is low efficiency and vulnerable to inference attack if these anonymous algorithms are directly applied to dynamic data publishing. To address this issue, this paper analyzes various inference channels that possibly exist between multiple anonymized datasets and discusses how to avoid such inferences and provides an effective approach to securely anonymize a dynamic dataset based on incremental clustering: incremental l-diversity algorithm. Theory analysis and experiment results show that the proposed method is effective and efficient.

Keywords: Data release, privacy preservation, anonymization, k-anonymity, l-diversity

1 Introduction

Rapid development in Internet, data storage and data processing technologies enables organizations to collect individual privacy information for research purposes. For example, a hospital may release patients' diagnosis records so that researchers can study the characteristics of various diseases. However, if individuals can be uniquely identified in the released data then their private information would be disclosed. Releasing data about individuals without revealing their sensitive information is an important problem. To avoid the identification of records in released data, uniquely identifying information like names and social security numbers are removed from the table. However, this first sanitization still does not ensure the privacy of individuals in the data. Anonymization provides a relative guarantee that the identity of individuals cannot be discovered. In recent years, a new definition of privacy called k-anonymity has gained popularity. The k-anonymity model, proposed by Sweeney [1, 2], is a simple and practical privacy-preserving approach and has drawn considerable interest from research community, and a number of effective algorithms have been proposed [3–9]. The k-anonymity model ensures that each record in the table is identical to at least k-1 other records with respect to the quasi-identifier attributes. Therefore, no privacy related information can be inferred from the k-anonymity protected table with high confidence, but k-anonymity algorithm is reluctant to background knowledge attack and homogeneity attack. Machanavajjhala et al. [10] proposed l-diverse anonymity model, which requires each equivalence class contain at least l “well-represented” values for the sensitive attribute and can effectively prevent background knowledge attack and homogeneity attack. However, most current l-diverse anonymity methods based on generalization and suppression techniques suffer from high information loss mainly due to reliance on pre-defined generalization hierarchies or total order imposed on each attribute domain so that the released dataset is low utility [11]. Moreover, most l-diverse anonymity methods are limited only to static data release, which assume that the entire dataset is available at the time of release. In many applications data collection is rather a continual process, which means that new data are collected and added, and old data are updated or purged. Processing a large dataset to achieve l-diverse anonymity is time-consuming and vulnerable to inference attack if we simply re-anonymize the entire dataset without considering previous releases of the dataset. For this, Byun et al. [12] proposed an approach to securely anonymize a continuously growing dataset by means of postponing data release. Xiao et al. [13] developed a new generalization principle named m-invariance that effectively limits the risk of privacy disclosure in re-publication, but it needs to insert some fake records to some extent, which has bad effect on data analysis. Wu et al. [14] proposed an important monotonic

[*] Corresponding author e-mail: pshwang@163.com

generalization principle that effectively prevents privacy breach in re-publication. However, these methods don't consider the background knowledge that the attackers gained, so the privacy disclosure is unavoidable. To address these issues, we analyze the inference channels existing between the anonymized tables in the application of dynamic data releasing while incorporating the attackers' background knowledge, and discuss how to avoid these inference attacks. Furthermore, we develop a novel l-diverse anonymity algorithm based on incremental clustering techniques. Extensive experimental results show that our method is practical and effective.

The rest of this paper is organized as follows. In section 2, we introduce the related concepts. In section 3, we present our method of l-diverse anonymity based on incremental clustering techniques. In section 4, we analyze the performance of our method through extensive experiments. Section 5 contains the conclusions and future work.

2 The related concepts

2.1 k-Anonymity

In order to preserve privacy, Sweeney [1] proposed the k-anonymity model which achieves k-anonymity using generalization and suppression techniques, so that, any individual is indistinguishable from at least k-1 other ones with respect to the quasi-identifier attributes in the released dataset. For example, table 2 is a 2-anonymous table of table 1. Generalization involves replacing a value with a less specific but semantically consistent value. For example, the date of birth could be generalized to a range such as year of birth, so as to reduce the risk of identification. Suppression involves not releasing a value at all. In recent years, numerous algorithms have been proposed for implementing k-anonymity via generalization and suppression. Usually, the record group that contains the same quasi-identifier attributes value is called an equivalence class. For example, record 1 and record 2 in table 2 constitute an equivalence class, record 3 and record 4 too.

Table 1: Original table.

Name	Race	Birth	Sex	Zip	Disease
Alice	black	1965-3-18	F	02141	gastric ulcer
Helen	black	1965-5-1	F	02142	dyspepsia
David	black	1966-6-10	M	02135	pneumonia
Bob	black	1966-7-15	M	02137	bronchitis
Jane	white	1968-3-20	F	02139	flu
Paul	white	1968-4-1	F	02138	cancer

Table 2: Anonymized table of table 1.

Race	Birth	Sex	Zip	Disease
black	1965	F	0214*	gastric ulcer
black	1965	F	0214*	dyspepsia
black	1966	M	0213*	pneumonia
black	1966	M	0213*	bronchitis
white	1968	F	0213*	flu
white	1968	F	0213*	cancer

2.2 l-Diversity

Since k-anonymity algorithm is reluctant to background knowledge attack and homogeneity attack, Machanavajjhala et al. [10] proposed l-diverse anonymity model, which requires each equivalence class contain at least l "well-represented" values for the sensitive attribute. As a simple and direct interpretation, l-diverse anonymity means that each equivalence class contains at least l different sensitive attribute values. For example, table 2 is also a 2-diverse table of table 1.

2.3 m-Invariance

Xiao et al. [13] proposed a new generalization principle named *m*-invariance that effectively limits the risk of privacy disclosure in re-publication. An anonymized table $T^*(j)$ $(1 \leq j \leq n)$ is *m*-unique, if each equivalence class in $T^*(j)$ contains at least *m* records, and all the records in the equivalence class have different sensitive values. The rationale of *m*-invariance is that, if a record *t* is published several times, and all its generalized hosting equivalence classes must contain the same sensitive values.

2.4 Incremental clustering

Clustering is the problem of partitioning a set of objects into groups such that objects in the same group are more similar to each other than objects in other groups with respect to some defined similarity criteria. Large amounts of data are created everyday for various purposes. They dynamically change because modifications such as insertion or deletion might occur over time. However, traditional cluster analysis focuses on static datasets in which objects are kept unchanged after being processed. When the dataset is modified, the previously learned patterns have to be updated accordingly. In the case that modifications occur frequently, re-clustering the whole dataset from beginning is not a good choice, especially when the number of the data objects is large and the out-of-service time is limited. Incremental clustering algorithms, which only update the clusters that are affected by the changed data, are therefore highly desirable.

3 L-diversity algorithm based on incremental clustering

To describe our algorithm, some concepts are defined in the following. For simplicity, only the case of records increment is considered. Firstly, we present the formal definition of inference channel in incremental data release. Secondly, we analyze the potential inference attacks in the process of data anonymization. Finally, we discuss the scheme on preventing the inference attacks and present an incremental l-diversity algorithm based on incremental clustering.

3.1 *Inference channel*

Inference channel existing between different equivalence classes is the primary reason for privacy disclosure. For this, we give its formal definition in the following.

Definition 1 (**Inference channel**). Let $T(QI_1,QI_2,...,QI_m,S)$ be an original table, where $QI_1,QI_2,...,QI_m$is the quasi-identifiers and $S$is the sensitive attribute. LetΔT_1,ΔT_2,... be the incremental data table. Assuming that T_i is a table released at time$i$, we denote$T_1 = T$, $T_2 = T_1+\Delta T_1$, ...,$T_n = T_{n-1}+\Delta T_{n-1}$, and the corresponding anonymized table is $T_1*, T_2*, ..., T_n*$. We say that there exists an inference channel between T_i* and T_j* if there are two integers $i, j (i \neq j)$ so that a sensitive attribute value can be inferred with a high confidence by comparing T_i* and T_j* together.

3.2 *Inference attacks*

We assume that the attacker keeps track of all the released tables. That is to say, the attacker possesses all the released tables. We also assume that the attacker grasps the particular individual information (sensitive attribute value not be included) and he has the knowledge of who is and who is not contained in each table. The following inference attacks possibly exist.

Definition 2 (**Difference attack**). Let $T_i* = \{e_{i1}, e_{i2}, ..., e_{im}\}$and $T_j* = \{e_{j1}, e_{j2}, ..., e_{jn}\}$ be the set of equivalence class belonging to the released tables at time i and time j respectively, and $I(e)$ be the set of individuals in equivalence class$e$, $S(e)$ be the set of sensitive attribute values in equivalence class$e$. We say that there exists a difference attack between e_{ik} and e_{jl} if there exists$e_{ik} \in T_i*$,$e_{jl} \in T_j*$and$e_{ik} \subset e_{jl}$ so that the number of sensitive attribute values in $S_D = S(e_{jl}) - S(e_{ik})$ is greater than zero and less than l, that is to say, the attack can infer the individual's sensitive attribute value with a probability greater than $1/l$.

Definition 3 (**Intersection attack**). Let $T_i* = \{e_{i1}, e_{i2}, ..., e_{im}\}$and $T_j* = \{e_{j1}, e_{j2}, ..., e_{jn}\}$ be the set of equivalence class belonging to the released tables at time i and time j respectively, and $I(e)$ be the set of individuals in equivalence class$e$, $S(e)$ be the set of sensitive attribute values in equivalence class$e$. We say that there exists an intersection attack between e_{ik} and e_{jl} if there exists$e_{ik} \in T_i*$,$e_{jl} \in T_j*$ so that the number of sensitive attribute values in $I_D = I(e_{jl}) \cap I(e_{ik})$ is greater than zero and less than$l$.

3.3 *Inference check*

For the inference attacks existing between different equivalence classes, the methods we used ensure that all the equivalence classes not only satisfy the l-diversity requirement, but also hold the property of m-invariance, which is named incremental l-diversity anonymity, that is to say, each equivalence class has the same sensitive attribute values set before and after update, therefore, the inference attacks maybe avoidable. More details are described in the next section.

3.4 *Candidate equivalence class*

To ensure our algorithm satisfy the property of m-invariance, we define the concept of candidate equivalence class as follows.

Definition 4 (**Candidate equivalence class**). Let $T* = \{e_1, e_2, ..., e_n\}$ be an anonymized table, r be an added record, $r.s$ denote the sensitive attribute value for record$r$, $S(e)$ denote the set of different sensitive attribute values. Then, the candidate equivalence class for record r with respect to the anonymized table $T*$ is defined as:

$$C_r = \{e | e \in T*, r.s \in S(e)\}. \tag{1}$$

3.5 *Information loss metric*

The information loss is a very important problem for data anonymization algorithm. In this section, we describe the information loss function for our algorithm to achieve l-diversity based on incremental clustering techniques. In a microdata set, there are two types of data: numeric and categorical data. Therefore, we need different distance functions to measure numeric data and categorical data respectively [11].

Definition 5 (**Information loss**). Let $e = \{r_1, ..., r_c\}$ be a cluster where the quasi-identifiers consist of numeric attributes $N_1, ..., N_m$ and categorical attributes$C_1, ..., C_n$. Let T_{C_i} be the taxonomy tree defined for the domain of categorical attributeC_i. Let $MIN_{N_i}$and MAX_{N_i} be the min and max values in e with respect to attributeN_i, and let $\cup_{C_i}$ be the union set of values in e with respect to attributeC_i. Then the amount of information loss occurred by generalizing$e$, denoted by $IL(e)$, is defined as:

$$IL(e) = |e| \cdot (\sum_{i=1,...,m} w_i \frac{(MAX_{N_i} - MIN_{N_i})}{|N_i|} + \sum_{j=1,...,n} w_j \frac{H(\Lambda(U_{C_j}))}{H(T_{C_j})}) \tag{2}$$

where $|e|$ is the number of records in e, $|N_i|$ represents the size of numeric domain N_i, w_i represents the weigh of attribute N_i, $\Lambda(U_{C_j})$ is the subtree rooted at the lowest common ancestor of every value in U_{C_j}, and $H(T)$ is the height of taxonomy tree T.

Definition 6 (**Total information loss**). Let E be the set of all equivalence classes in the anonymized tableT^*. Then the amount of total information loss of T^* is defined as:

$$Total - IL(T^*) = \sum_{e \in E} IL(e). \qquad (3)$$

3.6 L-Diversity algorithm based on incremental clustering

Based on the above concepts, we propose a new l-diversity algorithm based on incremental clustering technique for incremental data release with low information loss and high execution efficiency. Our algorithm includes three steps, as is shown in the following.

Firstly, insert the independent l-diverse equivalence classes into the previous anonymized table.

Secondly, process the rest records by their candidate equivalence classes subject to lower information loss.

Finally, divide the larger equivalence classes if no inference channels are generated.

Algorithm: *l-Diversity algorithm based on incremental clustering*

Input: a releasable dataset $T_{n-1}*$, an incremental dataset ΔT_{n-1}, and a diversity threshold value l.

Output: a releasable dataset T_n*, which ensures that each equivalence class has the same sensitive attribute values set before and after update and has minimal information loss.

```
1.Go to step 5 if the number of sensitive attribute values in $\Delta T_{n-1}$is less than $l$.
2.$T_n* = T_{n-1}*$.
3.Merge the independent $l$-diverse equivalence classes generated from $\Delta T_{n-1}$with$T_n*$.
4.Remove the corresponding records from$\Delta T_{n-1}$.
5.For each record $r$ in $\Delta T_{n-1}$
6.Generate the candidate equivalence classes $C_r$ in $T_n*$ according to its sensitive attribute value;
7.Insert the record $r$ into a selected candidate equivalence class, which results the minimal information loss;
8.$\Delta T_{n-1} = \Delta T_{n-1} - r$.
9.For each equivalence class whose size is more than$2l-1$and each sensitive attribute value exists at least two times
10.Divide the equivalence class if no inference channels are generated.
11.Return$T_n*$.
```

4 Experimental results

The main goal of the experiments was to investigate the performance of our algorithm in terms of privacy disclosure, information loss and execution efficiency. To accurately evaluate our algorithm (denoted by incremental l-diversity algorithm), we compared our implementation with two other methods. The one is using the l-diversity algorithm based full-domain generalization [3] re-anonymize the entire dataset (denoted by l-diversity algorithm 1), the other is using the same algorithm anonymize the incremental section (denoted by l-diversity algorithm 2).

4.1 Experimental setup

In our experiments, we adopted the publicly available dataset, Adult Database, from the UC Irvine Machine Learning Repository, which is considered a de facto benchmark for evaluating the performance of anonymization algorithms. We also used a configuration similar to [3], using nine of the attributes, as shown in Table 3, and eliminating records with unknown values. The resulting dataset contains 45,222 records. We considered {age, gender, race, education, marital status, native country, work class, salary class} as quasi-identifiers, and occupation attribute as sensitive attribute, which has 14 different sensitive attribute values. About 30K records were randomly chosen as the experimental dataset, half of which as the initial dataset, the rest as the incremental data.

The experiments were performed on a machine with Intel(R) Core(TM)2 Duo CPU T5450 1.67GHz(Double Kernel), 2.0GB RAM, Windows XP, MATLAB7.0, and Visual C + + 6.0.

Table 3: Experimental data information.

	Attribute	Distinct Values	Generalisations	Tree High
1	Age	74	5-,10-,20-year	4
2	Gender	2	Suppression	1
3	Race	5	Suppression	1
4	Education	16	Taxonomy Tree	3
5	Martial Status	7	Taxonomy Tree	2
6	Native Country	41	Taxonomy Tree	3
7	Work Class	7	Taxonomy Tree	2
8	Occupation	14	**Sensitive Attribute**	/
9	Salary Class	2	Suppression	1

4.2 Privacy disclosure risk

In this section, we report experimental results on our algorithm, the l-diversity algorithm 1 and the l-diversity algorithm 2 for privacy disclosure risk. Figure 1 shows the number of possibly disclosed records of the three algorithms for incremental dataset, which is increased by 10 percent every time. As the figure illustrating, zero record is disclosed for our algorithm and the l-diversity algorithm 2, while the l-diversity algorithm 1 has high privacy disclosure risk. The reason is that the l-diversity algorithm 1 results in amount of inference attacks, e.g. background knowledge attack, difference attack and intersection attack etc. It is clear that the inference channels existing between anonymized tables in incremental data release have seriously affected the security of the released data.

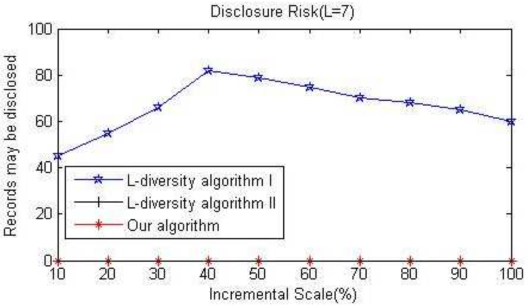

Fig. 1: Disclosure risk

4.3 *Information loss*

Figure 2 shows the information loss costs of the three algorithms for incremental dataset, which was increased by 10 percent every time. As the figure illustrating, our algorithm and the l-diversity algorithm?result in lower cost of the Total-IL than that of the l-diversity algorithm 2. The reason is that the l-diversity algorithm 2 only independently anonymizes the incremental data, which produces high information loss. Our algorithm also anonymizes the incremental data, but it executes based on the previous anonymized results, therefore resulting in low information loss.

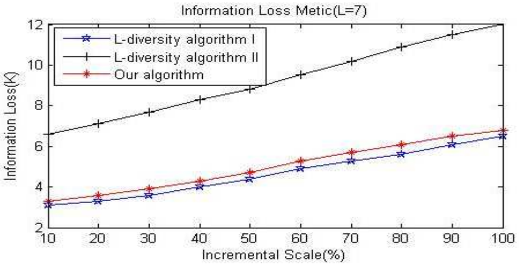

Fig. 2: Information loss.

4.4 *Execution time*

The execution time of the three algorithms for incremental dataset is shown in Figure 3. The execution time of our algorithm and the l-diversity algorithm 2 is always dramatically less than that of the l-diversity algorithm 1, especially for the lower incremental scale. The reason is that, l-diversity algorithm 1 needs to anonymize the entire dataset, while our algorithm and the l-diversity algorithm 2 only process the incremental data based on the previous anonymous results.

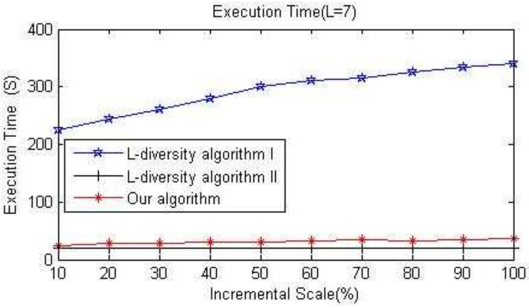

Fig. 3: Execution time

5 Conclusions and future work

The most existing anonymity methods based on l-diversity model are limited to static data release. However, in many applications data collection is rather a continual process. Processing a large dataset to achieve l-diversity is time-consuming and vulnerable to inference attack if we simply re-anonymize the entire dataset without considering previous publication of the dataset. To address these issues, we analyze the inference channels existing between the anonymized tables in the application of dynamic data release while incorporating the attackers' background knowledge, and discuss how to avoid these inference attacks, and propose a new l-diversity algorithm based on incremental clustering techniques. Extensive experimental results show that our method is effective and efficient. However, our algorithm can only be applied in incremental data publication, in the future, we will further study and develop more efficient methods for the datasets that are added, updated or purged.

Acknowledgments

We thank the anonymous reviewers and editors for their very constructive comments. This work was supported by Anhui Provincial Natural Science Foundation of China (No. 11040606M140).

References

[1] L. Sweeney, k-Anonymity: A model for protecting privacy, Journal of Uncertainty, Fuzziness and Knowledge-based Systems, **10**, 557-570 (2002).

[2] L. Sweeney, Achieving k-anonymity privacy protection using generalization and suppression, Journal on Uncertainty, Fuzziness and Knowledge-based Systems, **10**, 571-588 (2002).

[3] K. Lefevre, D. J. Dewitt and R. Ramakrishnan, Incognito: Efficient full-domain k-anonymity, Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data.New York: ACM Press, 49-60 (2005).

[4] K. LeFevre, D. J. DeWitt and R. Ramakrishnan, Mondrian multidimensional k-anonymity, Proceedings of the 22nd International Conference on Data Engineering. Los Alamitos: IEEE Computer Society, 25-36 (2006).

[5] X. Xiao and Y. Tao, Anatomy: Simple and effective privacy preservation, Proceedings of the Very Large Data Bases (VLDB) Conference, 139-150 (2006).

[6] A. Herath, Y. Al-Bastaki, S. Herath, Task based Interdisciplinary E-Commerce Course with UML Sequence Diagrams, Algorithm Transformations and Spatial Circuits to Boost Learning Information Security Concepts, International Journal of Computing and Digital Systems, **2**, 79-87 (2013).

[7] A. A. A. Radwan, M. H. Mohamed, M. A. Mofaddel, H. El-Sayed, A Study of Critical Transmission Range for Connectivity in Ad Hoc Network, Information Sciences Letters, **2**, 77-87 (2013).

[8] Y. Hong, Using the smartphone as a personal information security center, Applied Mathematics & Information Sciences, **6**, 573S-578S (2012).

[9] Y. Khmelevsky, V. A. Ustimenko, Practical Aspects of the Information System Reengineering, The South Pacific Journal of Natural Science, **21**, 75-81 (2003).

[10] A. Machanavajjhala, J. Gehrke and D. Kifer, l-Diversity: Privacy beyond k-anonymity, Proceedings of the 22nd International Conference on Data Engineering. Los Alamitos: IEEE Computer Society, 24-35 (2006).

[11] J. Byun, A. Kamra, E. Bertino and N. H. Li, Efficient k-anonymization using clustering techniques, Proceedings of the 12th International Conference on Database Systems for Advanced Applications. Springer-Verlag Berlin Heidelberg, LNCS, **4443**, 188-200 (2007).

[12] J. Byun, Y. Sohn and E. Bertino, Secure anonymization for incremental datasets, Proceedings of the 3rd VLDB Workshop on Secure Data Management, 48-63 (2006).

[13] X. Xiao and Y. Tao, M-invariance: Towards privacy preserving re-publication of dynamic datasets, Proceedings of the ACM Conference on Management of Data (SIGMOD), 689-700 (2007).

[14] Y. Wu, Z. Sun and X. Wang, Privacy preserving k-anonymity for re-publication of incremental datasets, Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering, 53-60 (2009).

Pingshui Wang is currently a Ph.D. candidate at Nanjing University of Aeronautics and Astronautics. His major research interests include Privacy Preserving, Information Security and Data Mining.

Jiandong Wang is currently a professor and Ph.D. supervisor at Nanjing University of Aeronautics and Astronautics. His major research interests are Machine Learning, Data Mining and Information Security.