An Enhanced K-Anonymity Model against Homogeneity Attack†

Qian Wang
College of Computer Science of Chongqing University, Chongqing, China
Email: wangqian@cqu.edu.cn

Zhiwei Xu and Shengzhi Qu
College of Computer Science of Chongqing University, Chongqing, China
Email: {xuzhiwei1986, qushengzhi}@126.com

***Abstract*—k-anonymity is an important model in the field of privacy protection and it is an effective method to prevent privacy disclosure in micro-data release. However, it is ineffective for the attribute disclosure by the homogeneity attack. The existing models based on k-anonymity have solved this problem to a certain extent, but they did not distinguish the different values of the sensitive attribute, processed a series of unnecessary generalization and expanded the information loss when they protect the sensitive attribute. Based on k-anonymity, this paper proposed a model based on average leakage probability and probability difference of sensitive attribute value. It is not only an effective method to deal with the problem of attributes disclosure that k-anonymity cannot deal, but also to realize different levels of protection to the various sensitive attribute values. It has reduced the generalization to the data in the most possibility during the procedure and ensures the most effectiveness of quasi-identifier attributes. Greedy generalization algorithm based on the generalization information loss is also proposed in this paper. To choose the generalization attributes, the information loss is considered and the importance of generalization attribute to sensitive attribute is accounted as well. Comparison experiment and performance experiment are made to the proposed model. The experiment results show that the model is feasible.**

***Index Terms*—privacy preservation, identity disclosure, attributes disclosure, k-anonymity, homogeneity Attack**

I. Introduction

With the coming of information age, information sharing has reached an unprecedented level. It brings convenience for people to access information; however it causes the disclosure of personal information as well which becomes more and more prominent. Patients' medical condition published online in the hospitals, the basic cost of living allowances information of people bulletined by government, the wages of employees announced by Company, and so on are the actions may lead to privacy disclosure. Once the attacker gets two or more groups of information, they may infer the privacy of personal information by linking the information. For example: through the voters' registration information and the patients' medical condition, attacker can infer the medical condition of a voter. This can result in the leakage of private information and eventually hurts their life and affect their employment.

There are two types of privacy information disclosure: Identity Disclosure and Attribute Disclosure. Identity disclosure happens when an individual can be uniquely identified from the release data. Data reconstruction approach [1] can deal with identity disclosure well. Attribute disclosure happens when the information of an individual can be inferred from the release data.

For privacy disclosure under linking attack, the main solution is k-anonymity [2] [3], proposed by L.Sweeney. It is simply to understand, and also protects privacy from identity disclosure well. However it cannot prevent the sensitive attribute disclosure due to the fact that it does not put restriction on sensitive attributes. For example, Homogeneity Attack causes the disclosure of private information when all the values of sensitive attribute are the same in one equivalence class. The main solutions include: p-sensitive k-anonymity [4] requires each equivalence class contains p sensitive attribute values at least. (p+, α)-sensitive k-anonymity [5] requires each equivalence class contains at least p distinct sensitive attribute values with its total weight at least α. L-Diversity [6] requires every equivalence class contains at least L well-presented values of sensitive attribute. Attacker has at most 1/L probability to confirm the individual's sensitive attribute. (L, α)-diversity[7] requires data table satisfying L-diversity and the weight of each equivalence class at leasht α. (a, d)-Diversity [8] takes the semantic meaning of the sensitive attributes into consideration. First, the sensitive attribute values are divided into groups, and then the records are grouped according to the sensitive attribute, finally the table is anonymized. (α, k)-anonymity [9] requires the frequency of each sensitive attribute value in each equivalence class not exceed α. (α, β, K)-anonymity[10] requires that data table satisfy k-anonymity, each sensitive attribute have β

†Supported by Natural science Foundation Project of CQ CSTC. Project Number: CSTCC2009BB2046.

doi:10.4304/jsw.6.10.1945-1952

different values and the number of no similar sensitive attribute values must be α bigger than the number of similar sensitive attribute values in equivalence class. (ε, m)-anonymity [11] is an anonymous method for the numeric sensitive attributes. It requires each sensitive attribute value in equivalence class, at most 1/m probability sensitive attribute values similar to it. Skyline (B, t)-privacy [12] is proposed to prevent the background knowledge attack. In all tuples, the privacy of the biggest disclosure risk is not greater than t for a series of background knowledge. On the base of monotonic generalization principle, security k-anonymity algorithm[13] has been proposed for the re-publishing of the incremental datasets.

These methods are based on k-anonymity, and prevent attribute disclosure to a certain extent. However, they all require sensitive attribute values in each equivalence class satisfy the same condition, which results in excessive generalization. Meanwhile, none of those methods considers that different sensitive attribute values can get different protection according to the practical situation. To solve this problem, the paper extends k-anonymity into a new model based on the average leakage probability and probability difference of sensitive attribute values. It does not require all the sensitive attribute values satisfy the same condition, but asks the average leakage probability of different sensitive attribute values can meet different requirements. The different standards are set for the difference between the average leakage probability and the leakage probability of different sensitive attribute values in an equivalence class. The average leakage probability reflects a general sensitive attribute value's security level. Different probability differences may provide personalized protection, and can avoid excessive generalization to some extent. With the greed algorithm, the method saved plenty of time to generate a better result compared with seeking for the best solution.

The rest of the paper is organized as follows: section 2 presents an introduction to k-anonymity, including relevant concepts and classical algorithms to achieve k-anonymity. Section 3 introduces the anonymity model based on the average leakage probability and probability difference of sensitive attributes, and it is extended from k-anonymity. It not only avoids identity disclosure, but also protects the privacy from attribute disclosure. Furthermore, the description of algorithm for ahceiving (alp, dif)-Anonymity is shown in this section. The experiment results in Section 4 demonstrate the feasibility of the model. Section 5 draws the conclusion.

II. K-ANONYMITY MODEL

A. *Relevant Concepts and Definitions*

Generally, there are three kinds of attribute in a data table: identifier, quasi-identifier and sensitive attribute.

Identifier (ID): Attributes that can uniquely identify an individual directly, i.e., Name, ID number, Social Security Number.

Quasi-identifier (QI): Attributes that can be linked with external data to re-identify individual records. i.e., Zip-code, Birth date and Gender.

Let U denote entity set, T(A_1, A_2, …, A_n) denote entity table, mapping: f_c : U→T and f_g: T→U', U⊆U', a quasi-identifier QI_T of T is a set of attributes {A_i, …, A_j} , {A_i, …, A_j} ⊆ {A_1, …, A_n} where: $\exists p_i \in U$ such that $f_g(f_c(p_i)[QI_T]) = p_i$

Sensitive Attribute (SA): Attributes that individual want to conceal. i.e., Disease and Salary

k-anonymity: Let T(A1, ..., An) be a table and QIT be the quasi-identifier associated with it. T is said to satisfy k-anonymity if and only if for each quasi-identifier QI∈QI_T each sequence of values in T[QI] appears at least with k occurrences in T[QI].

Equivalent Class: A block of records with the same quasi-identifier values.

Generalization: A value is replaced by a less specific, more general value that is faithful to the original.

A generalization for an attribute A is a function on A. That is, f: $A \to B$ is a generalization. $A_0 \xrightarrow{f_0} A_1 \xrightarrow{f_1} \ldots\ldots \xrightarrow{f_{n-1}} A_n$ is a generalization sequence.

Generalization tree: is a hieratical tree formed according to the semantic meaning of the attributes. The most bottom leaves denote attribute value in the data table. To be the closer to the top, the more general the value's meaning is.

For numerical attribute, to construct the generalization tree is simple. The node in the upper level includes all the values of the nodes below it. The root node contains all values of the attribute. As shown in figure 1.

For categorical attributes, first consider the classification of each attribute value. Then, divide attributes into serveral classes and form the generalization tree. As shown in figure 2.

Generalization lattice: For two or more attributes generalization, different attributes have different levels of generalization. In this situation, there will be different attribute generalization sequences. These sequences form

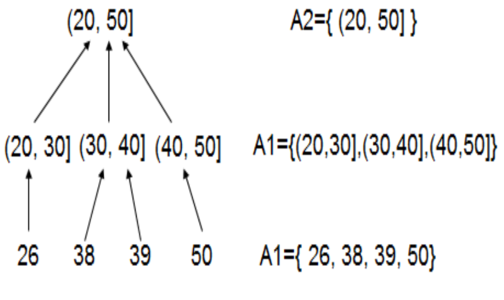

Figure 1. Generalization tree of numerical attribute

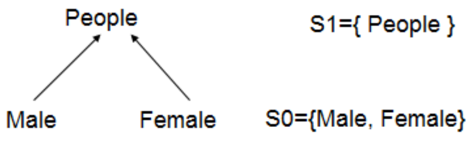

Figure 2 Generalization tree of categorical attribute

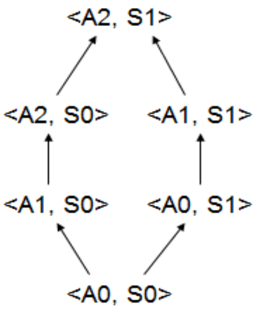

Figure 3. Generalization tree of categorical attribute

the generalization lattice based on quasi-identifier generalization level. For example: generalization lattice constituted with age and gender is shown in figure 3.

B. Classical algorithm for k-anonymity model

DataFly algorithm is a classical algorithm has realized k-anonymity, which is proposed by L. Sweeney. This algorithm counts each attribute frequency according to the requirement of k-anonymity, and choose the attribute which has the most distinct values to generalize until the dataset satisfy k-anonymity. DataFly algorithm is simple and easy to implement. However, it does not take into account different attribute has different weight and information loss. It causes that the data usually distorts much after being generalized.

This paper fully considers the significance of generalization attributes to sensitive attributes and information loss when choosing generalization attributes, and chooses the attribute that will cause smaller information loss and is less significant to the sensitive attributes to generalize.

III. (ALP, DIF)-ANONYMITY MODEL

A. Model Description

k-anonymity requires quasi-identifiers of every record in the table are the same with the quasi-identifiers of other k-1 records, which can solve the problem of identity disclosure. But the distribution of sensitive attribute values is not taken into account by k-anonymity. It can not solve the problem of leakage of sensitive attribute. Hence some improved methods have been proposed, and the most typical method is the P-sensitive, it not only requires the release data to meets the k-anonymity, but also needs each equivalence class has at least p different sensitive attribute values. i.e., table I as follows meets the 2 – Anonymity, but if one person is known as a male at the age of 40-50, and owns a doctor's degree, it can be easily referred that he is suffering from HIV. This is a typical situation of leaking the sensitive attribute. If the table is further generalized to meet the 2-sensitive, any equivalent class at least has 2 different sensitive attribute values, Table II can be formed.

TABLE I.
2 – ANONYMITY

Age	*Education*	*Sex*	*Illness*
(40, 50]	Dr.	M	HIV
(40, 50]	Dr.	M	HIV
[30, 40]	Master	M	cancer
[30, 40]	Master	M	cold
[20, 30)	Secondary education	F	fever
[20, 30)	Secondary education	F	fever

TABLE II.
2-SENSITIVE

Age	*Education*	*Sex*	*Illness*
(20, 50]	*	person	HIV
(20, 50]	*	person	HIV
(20, 50]	*	person	cancer
(20, 50]	*	person	cold
(20, 50]	*	person	fever
(20, 50]	*	person	fever

TABLE III.
A Release Table

Age	*Education*	*Sex*	*Illness*
(40, 50]	tertiary education	M	HIV
(40, 50]	tertiary education	M	HIV
(40, 50]	tertiary education	M	CANCER
(40, 50]	tertiary education	M	COLD
[20, 30)	secondary education	F	FEVER
[20, 30)	secondary education	F	FEVER

P-sensitive solves the problem of k-anonymity attribute leakage to a certain extent, however, by observing Table II, it can be found that the dataset eventually becomes an equivalence class and the data is generalized to a higher level. It has caused more loss of information, and made the usability of data greatly reduced. The reason of inducing this situation is that it is not considered fully about different sensitive attribute values can be given different degree of protection. i.e., fever is a very common disease and is not needed too much protection; while the HIV, which is serious to most of the people, must be well protected. Therefore, according to the actual situation, the data is released as in Table III, reducing loss of information, and making more usability of release data.

Release data in Table 3.3 reduces information loss and gives different protections to different sensitive attribute values. It has successfully achieved a better balance between privacy protection and data usability.

It has been taken account of the leakage of identity and sensitive information attributes in privacy protection with (alp, dif)-anonymity model. According to the actual situations, different protection degree for different sensitive attribute values can be set, and to some extent over-generalizing of attribute values can be prevented. (alp, dif)-anonymity model primarily requires the release data to meet the k-anonymity, and then ALP, the average leakage probability of sensitive attribute values, is no greater than the given threshold alp, finally the difference of leakage probability of sensitive attribute value in each equivalence class with average leakage probability, DIF which is no greater than the given threshold dif. alp generally reflects the level of data protection; the dif reflects the similarity between the alp and the leakage probability of sensitive attribute value in equivalence class. The smaller the dif is, the lower the degree of leakage is. If the DIF of an equivalence class is less than zero, it means that the leakage probability of the equivalence class is less than the average leakage probability and satisfies the model requirements. If an equivalence class with a DIF is greater than zero, the DIF can be made less than a given threshold dif through the generalization, so that the sensitive attribute values in the same equivalence class can be various. More numbers of sensitive attribute values are, the smaller the corresponding dif is. By setting different values of alp and dif to different sensitive attribute values, different protections for different sensitive attribute values can be realized.

B. Model Analysis

Suppose, in data table a sensitive attribute value has a value S which needs to be protected. The data table has m equivalence classes and y tuples with sensitive attribute value s. The numbers of tuples in equivalence classes are respectively x_1, x_2…x_m, while the numbers of tuples with sensitive attribute value s are respectively y_1, y_2...y_m, and

$$y = \sum_{i=1}^{m} yi\,. \tag{1}$$

Thus the average leakage probability ALP_s of sensitive attribute value s is:

$$ALP_S = \left(y_1 \times \frac{y_1}{x_1} + y_2 \times \frac{y_2}{x_2} + \cdots y_m \times \frac{y_m}{x_m} \right) \div y = \left(\sum_{i=1}^{m} \frac{y_i}{x_i} \times y_i \right) \div y. \tag{2}$$

To the sensitive attribute value s, if the average leakage probability $ALP_s \leq alp_s$, and the leakage probability LP_s when s in any equivalence class satisfies $LP_s - ALP_s \leq dif_s$, the S meet the requirements of (alp, dif)-anonymity mode.

Theorem 1: The average leakage probability of sensitive attribute value is non-increasing as attributes are generalized.

If K keeps unchanged, there will be no equivalence classes combine, and it can be easily proved that ALP keeps the same; if K increases, there will be equivalence classes combine. The following proves the ALP is non-increasing during the combination of equivalence classes.

Proof: Suppose that there are two equivalence classes in a data table, according to the formula (2), the average leakage probability of sensitive attribute value S is:

$$ALP_1 = \frac{y_1 \times \frac{y_1}{x_1} + y_2 \times \frac{y_2}{x_2}}{\sum_{i=1}^{2} y_i} \tag{3}$$

After generalizing, two equivalence classes combine into one , and the average leakage probability of sensitive attribute value S is:

$$ALP_2 = \frac{(y_1 + y_2) \times \frac{y_1 + y_2}{x_1 + x_2}}{\sum_{i=1}^{2} y_i}. \tag{4}$$

So there is:

$$ALP_1 - ALP_2 = \frac{(y_1 x_2 - y_2 x_1)^2}{x_1 x_2 y (x_1 + x_2)}. \tag{5}$$

x_1, x_2, y ≥ 0, so $ALP_1 \geq ALP_2$.

Now suppose that the conclusion is form when there are m-1 equivalence classes, namely:

$$y_1 \times \frac{y_1}{x_1} + y_2 \times \frac{y_2}{x_2} + \cdots + y_{m-1} \times \frac{y_{m-1}}{x_{m-1}} \geq \frac{(y_1 + y_2 + \cdots + y_{m-1})^2}{x_1 + x_2 + \cdots + x_{m-1}}.(6)$$

Next proves that conclusion is also form when there are m equivalence classes:

$$y_1 \times \frac{y_1}{x_1} + y_2 \times \frac{y_2}{x_2} + \cdots + y_m \times \frac{y_m}{x_m} \geq \frac{(y_1 + y_2 + \cdots + y_m)^2}{x_1 + x_2 + \cdots + x_m}.(7)$$

It can be simplified as follows:

$$\left[(y_1 + y_2 + \cdots + y_{m-1}) \times x_m - (x_1 + x_2 + \cdots + x_{m-1}) \times y_m \right]^2 \geq 0.(8)$$

All the values above are greater than 0, thus the formula is permanently satisfied and the conclusion is proved.

After the combination of equivalence classes, the average leakage probability of sensitive attribute value is no greater than the previous value, and when the numbers of sensitive attribute values in each combined equivalence class stay the same, the new average leakage probability is also equal to the previous value.

Theorem 2: The range of the average leakage probability of sensitive attribute value is($\frac{y}{\sum_{i=1}^{m} x_i}$, 1)

Proof: When generalized to the highest level, the number of equivalence class is 1, the number of sensitive attribute value is y, the leakage probability is $\frac{y}{\sum_{i=1}^{m} x_i}$. Each equivalence class only has one record, the leak probability is 1. So the range is established.

Theorem 3: The DIF of the sensitive attribute value whose leakage probability is higher than others is non-increasing when some attributes are generalized.

If K keeps unchanged, there is no equivalence class combine, it can be easily proved that DIF stays the same; if K increased, and there will be equivalence classes combination. The following proves the DIF is non-increasing during the combination of equivalence classes.

Proof: Suppose that two equivalence classes for combination are g_1, g_2, in which:

g_1: the number of the tuples whose sensitive attribute values s is y_1, the number of all tuples is x_1.

g_2: the number of the tuples whose sensitive attribute values s is y_2, the number of all tuples is x_2.

Using the combination as a watershed, before it the ALP is e_1 and after it the ALP becomes e_2. Evidently, theorem 1 tells that $e_1 \geq e_2$.

Suppose that the leakage probability of g_1 is higher, which is to say:

$y_1/x_1>=y_2/x_2$ be equivalent to: $y_1x_2-y_2x_1>=0$

Thus the formula above is permanently satisfied and the conclusion is proved.

When the two equivalence classes combined have the same sensitive attribute values, the formula becomes equal, but in reality, , the situation is rare, so the DIF of the sensitive attribute value whose leakage probability is higher than others is non-increasing by generlization. During the process of the classes combination, the number of equivalence classes that does not meet the LP-SP ≤ DIF is decreased. For different sensitive attribute values, we can set different ALP and DIF threshold values, the release data can meet different requirments of protection according to the actural situation. It realizes the personalized anonymity, but also decrease the unnecessary information loss.

C. Implementation

a. Selecting generalization attribute

During the process of selecting generalization attribute, the usefulness weight of the attribute for the data release should be considered. In the condition of the same information loss (adopt the calculation method of information loss in the reference [14]), previously generalize the smaller weight attribute, to make the release data with higher usability. QI_i indicates the ith attribute of Quasi-identifier, VQI_i indicates a value of QI_i, domain(VQI_i) indicates the domain that VQI_i belongs to, NUM(domain(VQI_i)) indicates the number of values of the domain that VQI_i belongs to, ContainNum(VQI_i) indicates the number of attribute values of VQI_i contained in the generalization tree, hence the information loss of VQI_i generalized to VQI_i' is：

$$\mathrm{Loss}(\mathrm{VQI}_i')=\frac{\mathrm{ContainNum}(\mathrm{VQI}_i')\text{-}1}{\mathrm{NUM}(\mathrm{domain}(\mathrm{VQI}_i))}. \quad (9)$$

Lattice(Node, Edge), as a generalization lattice, its information loss from node_i to node_j is：

$$\mathrm{Loss}(\mathrm{node}_j)=\left(1\text{-}W_{\mathrm{QI}_i}\right)\sum_{j=1}^{n}\mathrm{Loss}\left(\mathrm{VQI}_i'\right)_j. \quad (10)$$

VQI_i' is a value of $node_j$, n is a gross number, W_{QIi} is the importance weight of QI_i, and $W_{QIi}\in[0, 1]$. In each step, the attribute with the least information loss is chosen for generalization.

b. Generalization algorithm

The essence of achieving (alp, dif)-anonymity model is: building a generalization tree and a generalization lattice (GL) for each Quasi-Identifiers attribute. From the bottom node of generalizaion lattice, searching the node with least information loss to generalize employing greedy method, until the micro-data satisfying the given K and all the sensitive attribute values satisfy the given threshold value of alpi and the relative difi.

Algorithm description:

Input: data table T(node0), parameter K, the threshold of average leakage probability of sensitive attribute value alp1...alpm, and its relevant leakage probability difference dif1...difm, Quasi-Identifiers QI={QI_1, QI_2, ..., QI_n}, hierarchical structure of domain generalization DGH(QI_i) and the generalization lattice of each Quasi-Identifier, and suppression threshold v. node0 indicates the most bottom node in the generalization lattice, namely original data table.

Output: data table T* that satisfies the given threshold value K, alp_i, dif_i.

```
Begin{
Table=table(node0); H= the height of generalization lattice
Step1 //Satisfying k-anonymity
While (! All of the equivalence classes satisfy k-anonymity){
    Node=getMinLossAttribute(lattice);    //acquire the attribute with least information loss in generalization lattice
    Table(node);    //generalize the attribute
    If(the number of turples unsatisfies k-anonymity≤v)
    Suppression these turples
}
Step2: Satisfy alp and dif
While(!($ALP_i$ ≤ alpi&&$DIF_i$≤$dif_i$ for each sensitive attribute value in all of the equivalence classes)){
    Node=getMinLossAttribute(lattice);    //acquire the attribute with least information loss in generalization lattice
    Table(node);    //generalize the attribute
}
Return Table (node)
}
```

IV. Experiments

A. Experiment Environment

The dataset used in the experiments is the adult dataset from the UC Irvine machine learning repository [15], which is the typical database for k-anonymity research. The size of the database is 5.5MB. After eliminating some records missing some attribute values, we select 21411 tuples with 7 attributes and choose the attribute Martial-Status as the sensitive attribute. Table IV describes the datasets.

B. Data Precision Criterion

Using the equation from literature [3], the cost measure method of anonymization based on generalization hierarchy of data, to evaluate the precision of release data (equation (11)).

$$\Pr ec(RT)=1-\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{H(A_{ij},A_j)}{H_{A_j}}}{n\times m}. \quad (11)$$

TABLE IV.
DESCRIPTION OF ADULT DATA SET

No.	Attribute	Type	Generalizations	Distinct Value
1	Age	Int	5, 10, 20years ranges	73
2	Education	Class	Taxonomy Tree	16
3	Sex	Class	Taxonomy Tree	2
4	Occupation	Class	Taxonomy Tree	14
5	Country	Class	Taxonomy Tree	32
6	Salary	Class	Taxonomy Tree	2
7	Martial-Status	Class	Taxonomy Tree	4

In this formula, the H denotes the height of Taxonomy tree, Prec(RT) is the data precision of the release data, which shows the generalization cost for m tuples generalize on n Quasi-identifier attributes. The smaller the Prec(RT) is, the more the data generalization cost. For calculation of anonymous costs by Prec, it is necessary to construct the generalization hierarchy tree of each Quasi-identifier attribute. As to full-domain generalization, because all the values of each quasi-identifier attribute are generalized to the same height, the above formula can be simplified as follows:

$$\Pr ec(T) = 1 - \frac{\sum_{i=1}^{N} \frac{h_i}{|DGH_i|}}{N}. \tag{12}$$

In the simplified formula, N represents the number of Quasi-identifier attributes; DGHi is the domain generalization height of the ith attribute; hi represents the generalization height of the ith attribute; Prec(T) represents the release dataset precision.

C. *Comparative Experiment*

(alp, dif)-anonymity model enhances the protection of sensitive attributes based on the k-anonymity, and it not only prevents identity disclosure , but also provides different level of protections for different sensitive attributes values. We compare the (alp, dif)-anonymity model with k-anonymity model respectively in the execution time, the resistance of homogeneity attacks and the release data precision. The results show that, comparing to the k-anonymity model, the (alp, dif)-anonymity model can evidently reduce the number of records that suffer from Homogeneity attacks in the situation of moderately execution time increase and data precision decrease.

The sensitive attribute Martial-Status has four distinct values: Divorced, Widowed, Married-civ-spouse and Separated. Because Divorced and Widowed are relatively sensitive, we set their alp values as 0.43 and 0.42, set their dif as 0.27 and 0.31 relevantly. For the Separated, we set the alp=0.5 and dif=0.6; while there can be no protection for the Married-civ-spouse, thus alp=1, dif=1. We randomly select 6, 000 records each time, repeat the experiment three times, and use the average of the three results as a final result.

a. *Homogeneity Attack*

Figure 4 gives that the number of records suffering from homogeneity attack varies with the k (k-anonymity vs. (alp, dif)-anonymity):

Figure 4 shows that when k increases, the number of records sufferring from homogeneity attack decreases, moreover the greater the k, the more the decreasing number. From the figure 4, compared with the k-anonymity model, the attacked records number of (alp, dif)-anonymity model is evidently smaller. That is because the (alp, dif)-anonymity model further restrict the value of the sensitive attributes in the equivalence class, reducing the probability of being attacked and improving the level of protection.

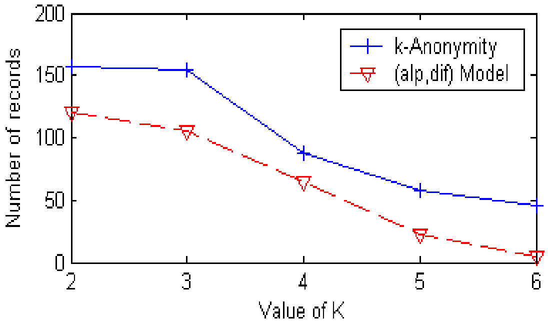

Figure 4. homogeneity attack versus K

b. *Execution Time*

Figure 5 gives the graphs that execution time changes with K. (k-anonymity vs. (alp, dif)-anonymity):

In figure 5, when k increases, the execution time of both models increase. This is because when k increases, the tuples number with the same Quasi-identifier in each equivalence class increases. To meet the requirement above, the Quasi-identifier attribute needs to be generalized to a higher level and thus the execution time increases. The (alp, dif)-anonymity model run time is longer, because comparing to k-anonymity it has an additional requirement for the distribution of sensitive attributes value in equivalence class. The figure 4.2 shows that the superior execution time is in an acceptable range. Furthermore, the different sensitive attribute values can be set to different alp and dif threshold, thus we can achieve the balance between privacy protection and execution time.

c. *Data precision*

Figure 6 gives the data precision changes with k. (k-anonymity vs. (alp, dif)-anonymity):

Figure 6 shows that when k increases, the data precision decreases. This is because with k increasing, the generalization level gets higher, and the number of suppression tuples increase, so.the information loss of data increase. (alp, dif)-anonymity model weaken the data precision compared with the k-anonymity model. This is because the (alp, dif)-anonymity model enhances the privacy protection based on the k-anonymity model while the cost is the decrease of the data precision. Traditionally, during processing, we need to set the dif according to the average leakage probability, the range of the value is [0, alp-dif]. The dif of different sensitive attribute values can be set differently based on the actual situation so that to ensure the data usability.

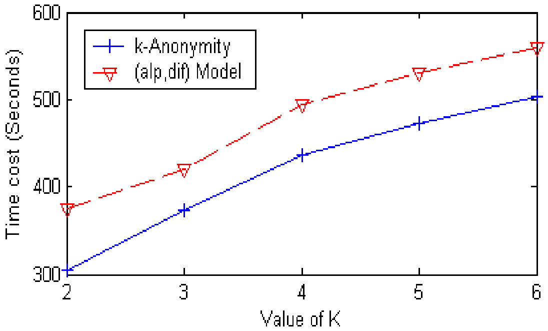

Figure 5. Execution time versus K

D. Performance Experiment

This set of experiments tests the performances of (alp, dif)-anonymity model, in the aspects of execution time, precision of release data, and resistance to homogeneity attack. Specifically, we choose divorce in marital status, the sensitive attribute. Based on the selection above, we do experiments under two different situations. Both of them have a fixed dif value 0.3 but the alp has different values of 0.43 and 0.41. The experiments are to test the influence of alp to the performances of (alp, dif)-anonymity model. Randomly select 6, 000 records each time, repeat the experiment three times, and use the average of the three results as a final result.

a. Changes of execution time with k

Figure 7 gives that execution time changes with k based on two different dif of (alp, dif)-anonymity model.

Figure 7 shows that with k increasing, execution time increases. When the k value is fixed, the smaller the alp is, the longer the execution time. It is because smaller alp requires higher level of privacy protection, which leads the corresponding higher level of generalization, and the more time consuming.

b. Changes of data precision with K

Figure 8 gives the release data precision changes with k's increase.

Figure 8 shows that when k increases, the data precision decreases steadily. When the k is fixed, the larger the alp is, the higher the Prec(RT). Because the larger average leakage probability makes the release data to meet the model requirement more easily, and avoids the higher level of data generalization and suppression.

c. Changes of number of records suffering from homogeneity attack with k

Figure 9 shows that when k increases, the records number decreases steadily, and alp also has obvious influence to the records number. The alp is smaller, the level of protection to the data is higher and the records number is smaller. As showed in figure 4.6, when alp=0.41 and k=5 or 6, the number of records suffering from homogeneity attack is already small enough, however, it can be seen from the previous experiments, information loss at this situation is also large, so we need to find the right balance when actually release the data, which is to say, to avoid information loss too large, to ensure the number of records suffering from homogeneity attack as small as possible, and both of them are in the acceptable ranges.

d. Change of execution time with records number

This experiment is to test the changes of the execution time with the records number when satisfy the requirements of (alp, dif)-anonymity model. In this experiment, the records numbers are respectively selected 1000, 2000, 3000, 4000, 5000, 6000, and the parameters of (alp, dif)-anonymity models are respectively selected k=2、alp=0.4、dif=0.3 and k=3, alp=0.4, dif=0.3. When records number is 1000, we randomly select 1000 records from dataset for experiment, repeat the experiment three times, and use the average of the three results as a final result. Dealing the numbers of records as 2000, 3000, 4000, 5000, 6000 with the same method, the execution time changes with the records numbers and they are given in Figure 10.

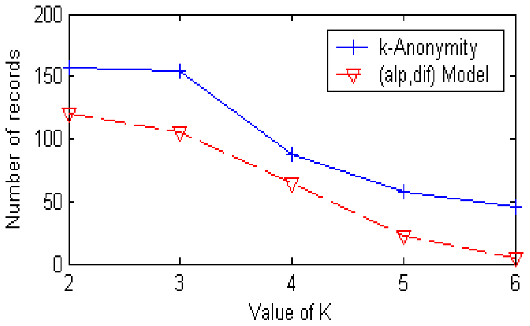

Figure 6. the data precision versus k

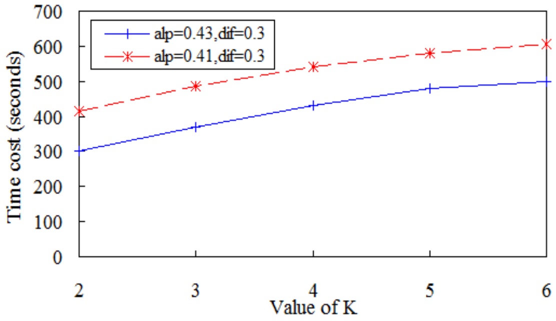

Figure 7. Execution time with different K

Figure 10 shows that execution time gradually increases with the increase of the records number, and the increasing rate has a accelerating trend. because when the records is more, the time for selecting generalization attributes is longer, and it requires to generalize to higher level from bottom to top in generalization lattice. When

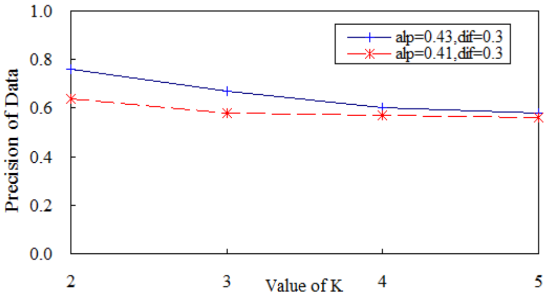

Figure 8. Data precision with different K

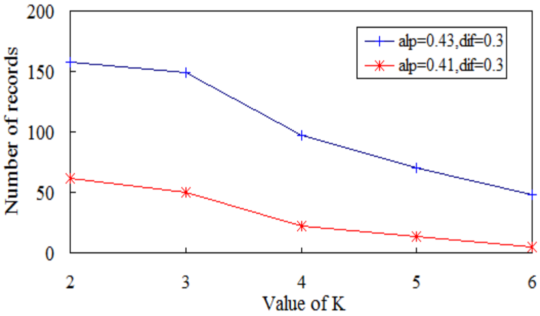

Figure 9. Records Suffered By Homogeneity Attack

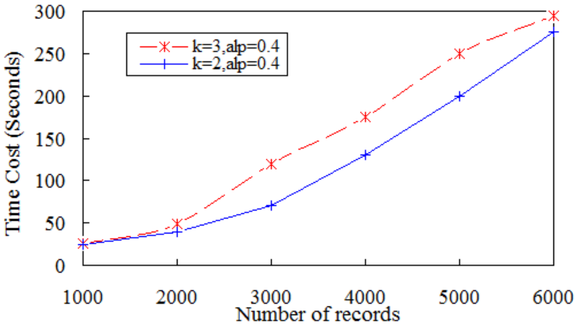

Figure 10. Change of execution time with Records number increasing

the alp is fixed and the records number is constant, the larger the k is, the longer the execution time needs. Because under the situation of satisfying alp, the data needs to be generalized to a higher level and to meet the requirements of better privacy protection for a larger k and it will take more time.

V. Conclusion

This paper proposed the (alp, dif)-anonymity model to make up the shortage of k-anonymity in protection of attribute disclosure. It can prevent attribute disclosure by controlling average leakage probability and probability difference of sensitive attribute value. The characteristic of the model has been analyzed in theory by using mathematical method. Comparison experiment and performance experiment has been made in the aspects of homogeneity attack, execution time and dada precision. The results of experiments show that comparing with the k-anonymity, (alp, dif)-Anonymity model has improved the security and precision of release data under moderately decreasing the algorithm time efficiency. This model is method with high utility can provide a better privacy protection. The model proposed in this paper cannot support the application of multi-sensitive attributes and increment datasets released, and in the follow-up research, the model will be reformed to satisfy this application requirement so that its utility can be improved.

References

[1] Dan Zhu, Xiao-Bai Li, Shuning Wu. “Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining”. Decision Support Systems. December 2009, Volume 48, Pages 133-140

[2] Sweeney L. “k-anonymity: a model for protecting privacy”. International Journal on Uncertainty, Fuzziness and Kno-wledge-based Systems, 2002, 10(5): 557-570.

[3] Sweeney L. “Achieving k-anonymity privacy protection using generalization and suppression”. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 571-588

[4] T.M.Truta and B.Vinay. “Privacy protection: p-sensitive k-anonymity property”. In Proceedings of the 22nd Internati-onal Conference on Data Engineering Workshops, the Second Internation Work on Privccy Data Management (PDM’06), 2006, 94

[5] X.xun, S.Hua and W.Jiuyong Li. “(p+, α)-sensitive k-anonymity: A new enhanced privacy protection model”. In Pr-oceedings 2008 IEEE 8th International Conference on Computer and Information Technology, 2008, 59-64

[6] A.Machanavajjhala, J.Gehrke, and D.Kifer. “l-diversity: Privacy beyond k-anonymity”. Proc.22nd Intnl. Conf.DataEngg. (ICDE), 2006, 24

[7] Xiao sun, Min Li, Hua Wang. “A family of enhanced models for privacy preserving data publishing”. Future Generation Computer Systems. 2011, Pages 348-356

[8] Qian Wang, Xiangling Shi. “(a, d)-Diversity: Privacy Protection Based on l-Diversity”. 2009 WRI World Congress on Software Engineering, WCSE (2009)

[9] W.Chiwing, L.Jiuyong and W.Ke. “(a, k)-Anonymity: an enhanced k-anonymity model for privacy preserving data publishing”. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Da-ta mining, 2006, 754-759

[10] Yan Zhao, Jian Wang, Yongcheng Luo. “(α, β, k)-anonymity: An effective privacy preserving model for databases”. Test and Measurement, 2009. ICTM '09. International Conference on. Page(s): 412 – 415

[11] J. Li, Y. Tao, and X. Xiao. “Preservation of proximity privacy in publishing numeric sensitive data”. In SIGMOD, 2008, 473–486

[12] Tiancheng Li, Ninghui Li, jian Zhang. “Modeling and integrating background knowledge in data anonymization”. Proceedings - International Conference on Data Engineering, 2009, 6-17, Proceedings - 25th IEEE International Conference on Data Engineering, ICDE (2009)

[13] Yingjie Wu, Zhihui Sun, Xiaodong Wang. “Privacy Preserving k-anonymity for Re-publication of Incremental Datasets”. Computer Science and Information Engineering, 2009 WRI World Congress on. March 31 2009-April 2 2009. page(s): 53 – 60

[14] X.Xiao and Y. Tao. “Personalized privacy preservation”.[C].New York:ACM Press, 2006, 229-240.

[15] S.Hettich, C.L.Blake, and C.J.Merz, UCI repository of machine learning databases, Available at http://archive.ics.uci.edu/ml/datasets/Adult, University of California, Irvine , 2008