Research on Privacy Preserving on *K*-anonymity

Yun Pan
School of Computer Science & Information Engineering, Zhejiang Gongshang University, Hangzhou, China
Email: panyun@mail.zjgsu.edu.cn

Xiao-ling Zhu
College of Computer Science & Information Engineering, Zhejiang Gongshang University, Hangzhou, China
Email: lynn_525@sina.com

Ting-gui Chen
Contemporary Business and Trade Research Center, Zhejiang Gongshang University, Hangzhou, China,
Email: ctgsimon@gmail.com (*corresponding author*)

***Abstract*—The disclosure of sensitive information has become prominent nowadays; privacy preservation has become a research hotspot in the field of data security. Among all the algorithms of privacy preservation in data mining, *K*-anonymity is a kind of common and valid algorithm in privacy preservation, which can effectively prevent the loss of sensitive information under linking attacks, and it is widely used in various fields recent years. This article based on the existing *K*-anonymity privacy preservation of the basic ideas and concepts, *K*-anonymity model, and enhanced the *K*-anonymity model, and gives a simple example to compare each algorithm; finally, it prospected the development direction of *K*-anonymity on privacy preservation.**

***Index Terms*—data mining; privacy preservation; *K*-anonymity; generalization & suppression; the enhanced *K*-anonymity**

I. INTRODUCTION

With the rapid development of information technology and the wide application of networks, large-scale of digital information is stored and published, and knowledge discovery and data mining applications in information retrieval have played an active role gradually, which greatly contributed to the various departments from the massive data mining of useful information needs. At the same time it also brings many problems regarding the privacy, the disclosure of sensitive information has become prominent nowadays, and privacy preservation has become a research hotspot in the field of data security. For example, the association among the illegal records in public security system, the customer's credit card transactions, telecommunications users' personal information, housing information, and so on. It is of great significance for government and business when make decisions, while it may destroy the citizens' personal privacy [1,2]. A reasonable and effective method of protection, which can protect the user's privacy and keep the data available at the same time, is the trend of developments in information security.

First of all, what need to definite is that privacy disclosure may not only due to the data mining technology itself, but the specific applications of data mining methods and the specific process. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information-information that can be used to increase revenue, cuts costs, or both. One of its important features is obtained models or rules from a large number of data mining, usually for the integrated data rather than the details of the data. So, is it possible for us to extract the patterns or rules based on the non-précised original data? To achieve both the reasonable privacy preservation approach of sensitive data and the extraction based on the statistical data patterns are the starting point of the study and the ultimate goal of privacy preservation.

Privacy preservation based on data mining can be applied widely in various industries and it has good prospect. The existing data mining methods are: heuristic-based privacy preservation technology, cryptography-based privacy preserving techniques, and privacy preservation technology based on the reconstruction, for different methods they applied well in corresponding fields, and can preserve user's privacy information to some extent [3]. This paper researched based on K-anonymity, a common approach in privacy preservation so far, and analyzed the main ideas and models of existing K-anonymity algorithm, and finally prospected the development direction of K-anonymity on privacy protection in the future.

II. THE MAIN ALGORITHM OF *K*-ANONYMITY MODEL

A. K-anonymity model

In real life, for the demands of research and statistical, some agencies often should publish some data, such as research on population statistics, medical and health. Although some personal identifiable information such as *name*, *ID number*, *telephone number* and other attributes, has been concealed when publishing that information, the attacker may get the sensitive information through other channels, for instance, linking attack is an effective way to deduce the sensitive information mainly works on quasi-identifier from two or more data table [4, 5]. Currently, K-anonymous is the most effective and common approach to avoid linking attack and protect t

doi:10.4304/jsw.7.7.1649-1656

private information from disclosure. Figure 1 is a demonstration of linking attack.

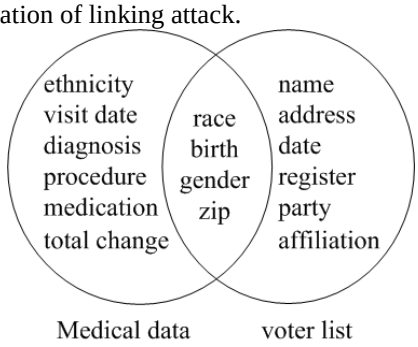


Figure 1 linking attack

K-anonymity was raised in 1998 by Samarati and L Sweeney, it requires the published data exists a certain number (at least for the *K*) whose records cannot be distinguished, so that an attacker cannot distinguish the respective privacy information of a specific individuals, thereby it prevents the leakage of personal privacy. User can specify a parameter *K* for the greatest risk of information leakage that they can receive in *K*-anonymous. It protects the privacy of individuals to some extent, while it also reduces the availability of data; the work of *K*-anonymity focuses mainly on the protection of private information and increase their availability. Since the proposed, *K*-anonymity has been the general concern of academia; many scholars at home and abroad research and develop the technology in different way. Since the proposed, *K*-anonymity has been the general concern of academia; many scholars at home and abroad research and develop the technology in different way.

B. Description of the model

Definition 1 *QI* (Quasi-identifier): Given a table *U*, a table $T(A_1,...,A_n)$, $f_c : U \to T$ and $f_g : T \to U'$, where $U \subseteq U'$. A quasi-identifier of T, written Q_T, is a set of attributes $\{A_i,...,A_j\} \subseteq \{A_1,...,A_n\}$ where: $\exists p_i \in U$ such that $f_g(f_c(p_i)[Q_T]) = p_i$ [6].

Definition 2 (*K*-anonymity)**:** A table T satisfies *K*-anonymity if for every tuple $t \in T$ there exist *k-1* other tuples $t_{i_1}, t_{i_2}, \ldots, t_{i_{k-1}} \in T$ such that $t[C] = t_{i_1}[C] = t_{i_2}[C] = \cdots = t_{i_{k-1}}[C]$ for all $C \in QT$ [7].

The aim of the data mining is focused on how to set effectively the original data by the method of anonymization, and to achieve the best anonymity, the maximum data availability, the minimum spending of time and space at the same time. Different with data perturbation techniques interference methods such as distortion, disturbance and randomization, *K*-anonymity can maintain the authenticity of the data. The common use to achieve *K*-anonymity is through generalization and suppression.

The idea of generalizing an attribute is a simple concept. A value is replaced by a less specific value that is faithful to the original. Generalization can be divided into value generalization and domain generalization; usually they can be achieved in the form of a generalized tree. For example, one of the generalization of integer 5 can be [3, 6], because 5 is in the interval of [3, 6, 8].

The number of different generalizations of a table *T*, when generalization is enforced at the attribute level, is equal to the number of different combinations of domains that the attributes in the table can assume. Given domain generalization hierarchies DGH_{D_i} for attribute $\{A_1,...,A_n\}$ the number of generalizations, enforced at the attribute level, for table $T(A_1,...,A_n)$ is:

$$\prod_{i=1}^{n}\left(\left|DGH_{D_i}\right|+1\right) \quad (1)$$

Table I shows a medical information table, after generalized; it does not include sensitive information such as their names, Medicare numbers, home address, and *ID* numbers and so on [9]. However, as we can see, there exists quasi-identifier information such as *gender*, *age*, *zip*, etc, through the collection of these attributes, attackers can distinguish the personal information indirectly, and so it is possible to disclose patient medical information.

TABLE I
Original medical information

ID	quasi-identifier			other	sensitive information
	gender	age	zip		
1	male	25	644***	……	gastric ulcer
2	male	27	650***	……	AIDS
3	male	22	671***	……	flu
4	male	29	675***	……	neurasthenia
5	female	34	671***	……	flu
6	female	31	671***	……	hepatitis
7	female	37	650***	……	neurasthenia
8	female	36	650***	……	flu

Suppression is to delete or hide some of the attributes in the data table directly to protect patient privacy. Suppression involves not releasing a value at all, which is one special case of generalization. While there are numerous techniques available, combining these two offers several advantages. Sometimes known values are replaced by unknown values "*", such as zip in table II.

In order to prevent the disclosure of private information and protect patient privacy, after generalized the data, the patients' age information in the table is specified in a large range, and zip the code in the "*" indicates any number, the rationale is as follows in Table II :

TABLE II
Generalization of the data

ID	quasi-identifier			other	sensitive information
	gender	age	zip		
1	male	25	644***	……	gastric ulcer
2	male	27	650***	……	AIDS
3	male	22	671***	……	flu
4	male	29	675***	……	neurasthenia

5	female	34	671***	……	flu
6	female	31	671***	……	hepatitis
7	female	37	650***	……	neurasthenia
8	female	36	650***	……	flu

Obviously, it becomes difficult for others to infer the user's private information after generalization and suppression, and it prevents the identity disclosure.

C. The enhanced K-anonymity model

(1) *L*-diversity model

The original *K*-anonymity is to prevent identity disclosure, but through the property will still bring the disclosure of information. In order to solve the shortcomings in *K*-anonymity, Machanavajjhala proposed *K*-anonymity model for the two attack methods, homogeneity attack and background knowledge attack. Homogeneity attack is the attacker derived K-anonymous table information of a sensitive individual; background knowledge inference attacks is that attackers use some additional information in advance to carry out attacks. The two attacks will result in disclosure of sensitive property in *K*-anonymity, and *K*-anonymity cannot protect personal information against background knowledge attacks. For these two attacks, Machanavajjhala gives the anonymous group by increasing the sensitivity of a diversity of methods (*L*-diversity) to reduce the loss of privacy.

Definition 3(*L*-Diversity): If an equivalent group contains at least l acceptable values of sensitivity attribute, then the equivalence group satisfies the *L*-Diversity nature; and if all of the equivalence groups in data table satisfy *L*-Diversity, then the data sheet satisfies *L*-Diversity.

In a published table, a *K*-anonymous group contains at least *L* sensitive properties that on behalf of a good sense of representative. For example, in Table II, tuples whose *ID* are 1, 2, 3, 4, 6 form a group with 5 species diversity, their frequency was 12.5%, 12.5%, 37.5%, 25% and 12.5% in value, and no one has predominant function, so it can be set by *L*-diversity model.

An effective and practical utility of the model will usually achieve more effective data protection. *L*-diversity model is effective, but not necessarily useful. If we try to protect each sensitive attribute value, then it may not release sensitive property or quasi-identifier. However, in this model, it is difficult to speculate how much background knowledge the attacker knows about, any posted data will become unsafe if the other knows a lot about the patient's background knowledge, there is not a good way on how set the parameters in *L*-diversity model.

(2) (*α, K*)- anonymity model

A more practical approach is not to consider the value of each sensitive attribute, for example, people may regard the failing grades as privacy, while for those excellent or good test results not as privacy. In (*α, K*)-anonymity model, property with higher degree of sensitivity has been better protected, by constraining the frequency of anonymity property values in the sensitive group less than a given parameter *α*, so it avoids the situation that the frequency of certain sensitive information too high , increases the diversity of sensitive values, and prevents the consistency attack [10].

Definition 4((*α, K*)-anonymity). A view of a table is said to be an (*α, K*)-anonymity of the table if the view modifies the table such that the view satisfies both *K*-anonymity and *α*, where $0<\alpha< 0$ [11].

For example, when a patient with different medical records, some patients' illness are more sensitive and need protection, such as *AIDS*; there are many diseases that are very common and not need to protect, such as flu. In (*α, K*)-anonymity model, only the relevant and sensitive property values are necessary to protect, so only consider the sensitive attribute value. This model allows inference between credibility to the sensitive lower than *α*, it is simple and effective way to prevent sensitive to the value for the homogeneity attack. Table III is shown as follow:

TABLE III
(0.25, 3)-anonymity

ID	quasi-identifier			other	sensitive information
	gender	age	zip		
1	male	25	644***	……	flu
2	male	27	650***	……	neurasthenia
3	male	22	671***	……	flu
4	male	29	675***	……	neurasthenia
5	female	34	671***	……	flu
6	female	31	671***	……	flu
7	female	37	650***	……	neurasthenia
8	female	36	650***	……	flu

Table III provides a (*α, K*)-anonymity form. In the table, flu and neurasthenia is not considered as sensitive information, from the (female 30-39, 671***) to the reasoning of neurosis confidence level is 25%.

The attackers couldn't see the value of property with higher degree of sensitivity after processing by (*α, K*)-anonymity model, so it can protect the security of this kind of information effectively.

There are still shortcomings although (*α, K*)-anonymity model can resist the homogeneity attack and background attack to a certain extent. While (*α, K*)-anonymity model only consider the sensitive attribute of the highest level-sensitive property value, there is no other level of sensitive property to process property values, and does not take into account the sensitivity of the same property value, so the same level of sensitive attributes property values or the existence of other levels of privacy disclosure.

(3) (*α, L*)-diversification *K*-anonymity model

(*α, K*)-anonymity model considers only the highest level of the sensitive property value, but neither other level of sensitive property values, nor take into account the sensitivity property values of the same issues, *K*-anonymity model in (*α, L*)-diversification [12].

Definition 5 ((*α*, *L*)-diversification *K*-anonymity model): Given a data table *T*, it meets both *K*-anonymity and *α*-distribution, and the number of *Sid* in the sensitive group is no less than *L* at the same time. *α*-distribution constraint is that all of the sensitive attributes of privacy frequency of *Sid* from equivalent class less than *α* which is a given data, that is sensitive to all the anonymous group of private property when the degree of the frequency *Sid* ≤*α*, where a is user-defined number, and 0 <*α*<1.

K-anonymity model can determine flexibility to protect the privacy or not according to the extent of protection. At the same time, have special treatment to the high-level sensitive property values of privacy preservation, and with better privacy protect effect. People of high-income need higher privacy preservation degree for salary than people of low-income. The model can not only solve the problem of imbalance in the distribution of sensitive attributes, but also proves the feasibility of the method by the experiments, and has a lower information loss Table IV is a Health categories table (where *Sid* is sensitive to the privacy level of property value) is *Sid* the data in Table II for the classification. $D[S]=\{S_1,...S_k\}$ is the attribute values of the sensitive poverty *S*, $weight(S_i)$ is the weight of attribute S_i.

TABLE IV
Health categories

ID	Value	Sid
1	AIDS, hepatitis	1
2	gastric ulcer	2
3	Neurasthenia, flu	3

$$\begin{cases} weight(S_i)=0 \\ weight(S_i)=\dfrac{i-1}{k-1};1<i<k \\ weight(S_k)=1 \end{cases} \quad (2)$$

(*α*, *L*)-diversification *K*-anonymity model is a data table to meet *K*-anonymity, *α*-distribution, and the number of *Sid* in the sensitive group is no less than *L* at the same time. *α*-distribution constraint is that all of the sensitive attributes of privacy frequency of *Sid* from equivalent class less than α which is a given data, that is sensitive to all the anonymous group of private property when the degree of the frequency $Sid \le \alpha$, where a is user-defined number, and 0 <*α*<1.

Assuming the *Sid* need to protect equals to 1, privacy preservation level is as classification in Table IV, while Table V is a constraint to meet the 0.5 distribution of a data set. In Table V, there are two anonymous group: {1,2,3,4} and {5,6,7,8}, in the first anonymous group, the frequency is 0.25 when *Sid*=1, and in the second one the frequency is 0.25, so for all anonymous groups when *Sid* equals to 1 the frequency of *Sid* ≤ 0.25. Then it meets to (0.25, 3)-diversification 4-anonymous as shown in Table V:

TABLE V
(0.25, 3)-diversification 4-anonymous

ID	quasi-identifier			other	sensitive information
	gender	age	zip		
1	male	25	644***	……	2
2	male	27	650***	……	1
3	male	22	671***	……	3
4	male	29	675***	……	3
5	female	34	671***	……	3
6	female	31	671***	……	1
7	female	37	650***	……	3
8	female	36	650***	……	3

Table V provides a data table satisfying (0.25, 3)-diversification 4-anonymous model, according to the foregoing, the distribution of this data sheet meet to the constraints of 0.25, and the number of each anonymous tuple is no less than 4, the different number of the *Sid* values in the table equals to 3, so Table IV satisfy (0.25, 3)-diversification 4-anonymous model.

Construct (*α*, *L*)-diversification *K*-anonymity model algorithm as follows:

Input: data set *T*;

Output: data table *T* * that meet (*α*, *L*)-diversification *K*-anonymity model.

(1)According to the health status in table V, the value of the sensitive property in table *T* was replaced by Sid, who represent the sensitive level, then table *T* turns into table *T1*.

(2) Construct a data table *T2* that consistent with (*α*, *K*) data tables anonymous model, in which *Sid* is regarded as the sensitive property, and the generalization in accordance with top-down algorithm.

(3) For each anonymous group, check the privacy level Sid for the number of different values.

(4) IF (3*L*).

(5) Return the final table *T**.

(6) Else

(7) Of all the anonymous groups that does not meet the requirement, have them further generalization or exchange tuples to make sure that the value of *Sid* is greater than L.

(8) Returns the final data table *T* *.

First, the value of the sensitive property in table *T* was replaced by *Sid* that represent the sensitive level, according to the health status in table V, then table *T* turns into table *T1*, and then turn *T1* into anonymity, so as to meet *K*-anonymous and *α*-distribution, in this step, top-down local generalization algorithm has been used. And then check whether the generalization of the data sheet meets (*α*, *L*)-diversification *K*-anonymity model conditions, that is the privacy degree of different values greater than or equal to *L*. If all anonymous groups met for The condition, the entire data table is the final meet (*α*, *L*)-diversification *K*-anonymity model data tables, and if not, have further generalization or

suppression to make sure that the different values of privacy degree number is greater than *L*.

In (*α*, *L*)-diversification *K*-anonymity model, the parameter α can be set by users themselves according to their privacy preservation needs. It provides an effective solution to the problem of imbalance distribution of sensitive attributes, divides the attribute values on the sensitive level of privacy preservation, and protects the privacy effectively.

The enhanced *K*-anonymity models are mainly based on *K*-anonymity and to make the information security. *L*-diversity model in which properties are divided into groups, by increasing the variety in groups, it can prevent attackers from locating the information; (*α*, *K*)-anonymous model, by processing to the higher level of sensitive attribute and make its feasible degree smaller than *α*, can effectively protect sensitive information of higher degree; (*α*, *L*)-diversification *K*-anonymity model divides the properties according to the level of sensitive information which is determined by the users themselves flexibility, and for the sensitive attributes with higher degree value require special treatment [13-15].

III. APPLICATIONS OF *K*-ANONYMITY

K-anonymity algorithm is a kind of popular model about privacy preservation. The following we will compare and analyze each algorithm through an example. Table VI is part of the workers' basic information table in an enterprise, which consists of seven properties:{ *Name*, *Sex*, *Education*, *Birth*, *Occupation*, *Phone-number and Salary*}, take *Phone-number* and *Salary* as sensitive attributes. If publish the data in table VI to the Internet directly, this will result in serious personal information leakage, which may brings a series of troublesome that affects people's daily life, once used maliciously by other, which may bring out a series of troublesome that may even affect people's daily life. The following is a process that dealing with data in table VI through the algorithms describe above.

TABLE VI
Workers' basic information

NO	Name	Sex	Education	Birth	Occupation	Phone-number	Salary
1	Wang-Lin	Male	College	1976-12-27	staff	12345678	4000
2	Luo-Jia	Male	Graduate	1981-04-06	director	12345678	6000
3	Shui-Qiang	Male	Graduate	1981-01-22	manager	12345678	9000
4	Liu-Ming	Male	College	1979-03-09	director	12475678	4000
5	Zhang-Hang	Female	College	1979-06-17	staff	12435678	6000
6	Chen-Xi	Female	College	1976-11-01	manager	12455678	9000
7	Ma-Zhuo	Female	Graduate	1984-08-31	staff	12425678	4000
8	Yao-Ting	Female	Graduate	1982-07-16	staff	12505678	4000

A. Generalization & suppression

Table VII is a data table processed by the method of generalization; it does not include the sensitive attributes such as *Name*, *Phone-number*, and *Salary*. While as we can see, there is still some Quasi-identifier information such as *Sex* and *Birth* in the table, attacker may distinguish the personal information through the attributes left in the table, so it may disclose information.

TABLE VII
Data after anonymization

NO	Sex	Education	Birth	Salary
1	Male	College	19******	4000
2	Male	Graduate	19******	6000
3	Male	Graduate	19******	9000
4	Male	College	19******	4000
5	Female	College	19******	6000
6	Female	College	19******	9000
7	Female	Graduate	19******	4000
8	Female	Graduate	19******	4000

In table VI, there are sensitive attributes that require higher degree of protection, such as *Name* and *Phone-number*, which may lead to a greater disclosure of personal information, while they have little use in studies and analysis or other purposes, for example, when analyzing the relationships between income and working age and education, so we can anonymize it directly. After generalization, specified the values in *Birth* in a range of interval, "*" indicates any number of processed data in Table VIII is shown as follows:

TABLE VIII
Data after suppression

NO	Sex	Education	Birth	Salary
1	Male	College	19******	4000
2	Male	Graduate	19******	6000
3	Male	Graduate	19******	9000
4	Male	College	19******	4000
5	Female	College	19******	6000
6	Female	College	19******	9000
7	Female	Graduate	19******	4000
8	Female	Graduate	19******	4000

Obviously, after generalization and suppression, it is difficult for an attacker to infer personal information from the published quasi-identifier according to information in the table VIII. However, attackers can still obtain the sensitive information by linking attacks or combing with other tables.

B. *(α, K)-anonymity model*

(*α*, *K*)-anonymity model is an algorithm that mainly for the situation when there are few sensitive values in a data table, that is, only consider the property value with high-degree of sensitive in the sensitive attribute. It allows the confidence level, from *K*-anonymous group to the reasoning of certain sensitive property value, less than a given value *α*. Table IX is a (*α*, *K*)-anonymity table as follows, in the table, it is not viewed as sensitive information when *Salary* = {4000, 6000}, from the (female, College, 197*****) to the reasoning of *Salary*= {4000} confidence level is 25%.

TABLE IX
(0.25, 3)-anonymity model

NO	Sex	Education	Birth	Salary
1	Male	College	19******	4000
2	Male	Graduate	19******	6000
3	Male	Graduate	19******	4000
4	Male	College	19******	4000
5	Female	College	19******	6000
6	Female	College	19******	6000
7	Female	Graduate	19******	4000
8	Female	Graduate	19******	4000

In this kind of algorithm, attackers could not see the value of the corresponding high-sensitive properties, so it can protect the high-sensitive information effectively. However, there are still limitations in (*α*, *K*)-anonymity model, for sometimes it is difficult for one to judge a property value is sensitive or not and no exact standard make judgment, so it is of poor adaptability.

C. *(α, L)-diversification K-anonymity model*

In (*α*, *L*)-diversification *K*-anonymity model, there are some improvement compares to (*α*, *K*)-anonymity model. For it considers not only the high-sensitive property, but also the low-sensitive property, for example, as to *Salary*, take different approach to protect the privacy of high-income earners and low-income earners. Besides, users can set the degree of privacy preservation for sensitive property and the value of parameter *α*.

TABLE X
Degree of salary privacy

ID	Value	Sid
1	9000	1
2	6000	2
3	4000	3

Table X is a classification based on the degree of privacy preservation in salary, according to the value Sid (the degree of protection it requires), *Sid* = {1, 2, 3}, and *Sid* = 1 is the degree that requires protection. Table XI is a data set meets to the constraints of a 0.5 distribution, and there are two anonymous groups: {1, 2, 3, 4}, {5, 6, 7, 8}; when *Sid*=1, in the first group the frequency equals to 0.25, and in the second one the frequency equals to 0.25, so for all anonymous groups *Sid*=1, frequency≤0.25; The number of each anonymous tuples is at least 4, the number of different value of Sid equal to 3, so table XI is (0.25, 3)-diversification 4-anonymity. Table XI is shown as follows:

TABLE XI
(0.25, 3)-diversification 4-anonymity model

NO	Sex	Education	Birth	Salary
1	Male	College	19******	3
2	Male	Graduate	19******	3
3	Male	Graduate	19******	1
4	Male	College	19******	3
5	Female	College	19******	2
6	Female	College	19******	1
7	Female	Graduate	19******	2
8	Female	Graduate	19******	3

Obviously, in (*α*, *L*)-diversification *K*-anonymity model, information of different degrees sensitive are processed corresponding, so it is more flexible in application than that of (*α*, *L*)-anonymity model.

IV. ANONYMITY METRICS

Evaluation of performance is to estimate the time complexity of the algorithm or algorithms in the average number of basic operations. Algorithm used to evaluate the scalability of data capacity increases in the efficiency of the trend, which requires *K*-anonymity algorithm for large or very large databases are scalable. Data availability refers to the *K*-anonymity of the data after the loss of the amount of information, which is mainly embodied by the precision.

A. *The precision metric*

Definition 5(Precision metric *Prec*): Let $PT(A_1,...,A_{Na})$ be a table, $t_{Pj} \in PT$, $RT(A_1,...A_{Na})$ be a generalization of PT, $t_{Pj} \in PT$, each DGH_A be the domain generalization hierarchy for attribute A, and $f_i's$ be generalizations on *A*. The precision of RT, written $\Pr ec(RT)$, based on generalization and suppression [16], he describes as follows:

$$\Pr ec(RT) = 1 - \frac{\sum_{i=1}^{N_a}\sum_{j=1}^{N}\frac{h}{DGH_{Ai}}}{|T| * |\mathrm{N_a}|} \qquad (3)$$

B. Measure of information loss

All anonymous algorithms are based on specific information loss index metrics to guide anonymous strategy and to measure the pros and cons of the algorithm. If use normalized certainty penalty as the measure of information loss index, then, as algorithm that related in reference 25, given property *A*, if A is a numeric attribute, the information loss is:

$$NCP_A(x) = \frac{Range(x)}{Range(R_A)} \tag{4}$$

And if *A* is categorical attribute, the information loss is:

$$NCP_A(x) = \frac{|Sub(x)|}{|Sub(R_A)|} \tag{5}$$

The loss of property information will be extended to tuples *t* and the data table *T*, the calculation methods are:

$$NCP_A(t) = \sum_{i=1}^{n} NCP_{Ai}(t[A_i]) \tag{6}$$

$$NCP_A(t) = \sum_{i \in T} NCP(t) \tag{7}$$

Information loss of $NCT_A(T')$ is as small as possible here [17].

However, Bayardo and Agrawal consider that through the cost in generalization and suppression operation to measure the availability of the anonymity table. Suppose the generalized cost in each record of Equivalence class *E* is $|E|(|E| \setminus K)$, which is the number of records, and $|D|$ is the cost of the suppression of a record, which is the size of a database [18]. So the total cost of getting the anonymous table is:

$$C = \sum_{|E| \setminus K} |E|^2 + \sum_{|E| < K} |D||E| \tag{8}$$

Obviously, the greater the suppression of equivalence classes and more records, the higher the cost of anonymization, accordingly, the smaller availability of the anonymity table is.

C. Relationship between information loss and α

When publishing a data table to the internet, if the attribute of tuple is not generalized, there will be no information loss, but if the property value is generalized to a higher level of generalization level tree, then it will be information loss. The more information generalized (such as a generalization to the level of the tree root), the greater information loss will be. Attribute values are defined as the value of the loss of a high degree of generalization. Here is a figure describes the relationships between information loss and parameter α, when *K* and *L* are given values, *K*=2 and *α*=0.2, 0.4, 0.6, 0. 8, 1, and get the relationship between information loss and α of a *K*-anonymous table. Calculate separately and get figure 2 as follows:

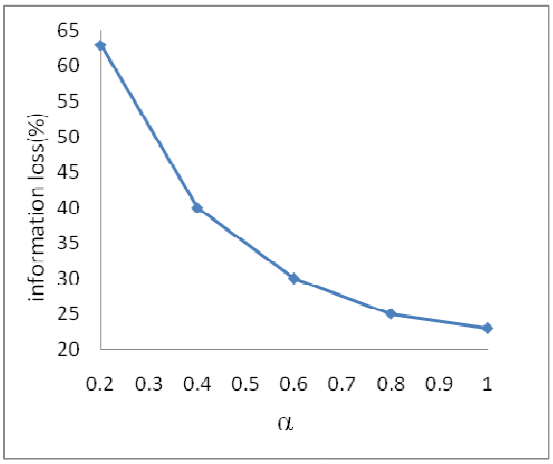


Figure 2 relationships between information loss and α

As we can see, the loss of information reduces with the incensement of *α*, in fact, the greater *α* is, the less to meet the requirements of *α*-distribution constraint. Data sets require less generalization, so information loss reduced [19].

V. CONCLUSIONS

Because *K*-anonymity can prevent users' private information from being leaked in the released environment, ensure the authenticity of the published data, *K*-anonymity is an effective way to protect data privacy under data distribution environment, it applies widely in the industry and attracted widespread attention. This paper analyzes and compares some of the existing *K*-anonymity model and its applications, and overcome these problems with the strengthening of the model, summarizes some of the *K*-anonymous used to achieve the main technology.

The research of *K*-anonymous privacy preservation is of very widely application, and the development in the future is facing more and more challenges, there are still some problems to be discussed. However, nowadays the majority of *K*-anonymity algorithms are based on static data sets, and in the real world, data is constantly changing, including changes in forms of data, attribute changes, adding new data, and deleting the old data. Besides, the data between data sets are likely to be interrelated, how to achieve privacy preservation in a much more complex environment with dynamic data, still need further study [20-24].

ACKNOWLEDGEMENT

This research is supported by National Natural Science Foundation of China (Grant No. 71071141 and 71071140), Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20103326110001 and 20103326120001), Humanity and Sociology Foundation of Ministry of Education of China (Grant No. 11YJC630019), Zhejiang Provincial Natural Science Foundation of China (No.Z1091224 and Y7100673), Zhejiang Provincial Social Science Foundation of China (Grant No. 10JDSM03YB), the Scientific Research Fund of Zhejiang Province, China (Grand No. 2011C23008),

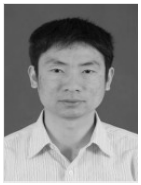
Research Project of Department of Education of Zhejiang Province (No.Y200907458 and Y201016434), the Contemporary Business and Trade Research Center of Zhejiang Gongshang University (No. 1130KUSM09013 and 11JDSM02Z). We also gratefully acknowledge the support of Science and Technology Innovative project (No. 1130XJ1710215).

REFERENCES

[1] Q. Long, "Privacy preservation based on *K*-anonymity," Science & Technology association forum, vol. 3, no. 5, pp. 41-43, 2010.
[2] Q. Long, "A *K*-anonymity Study of the Student-Score Publishing", Journal of Yunan University of Nationalities (Natural Sciences Edition), vol.3, NO. 2, pp. 144-148, March, 2011.
[3] P. Lü, N. Chen, and W. Dong, "Study of Data Mining Technique in Presence of Privacy Preserving", Computer Technology and Develppment, 2006, 16(7).
[4] T. Cen, J. Han, J. Wang and X. Li, "Survey of *K*-anonymity research on privacy preservation," Computer Engineering and Applications, vol. 44, no. 4, pp. 130-134, 2008.
[5] K. Yin, Z. Xiong and J. Wu. "Survey of Privacy Preserving in Personalization Service," Application Research of Computers, vol. 25, NO. 7, pp. 123-140, 2008.
[6] L. Sweeney, "Achieving *K*-anonymity privacy preservation using generalization and suppression," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 54, no. 5, pp. 571-588, 2002.
[7] A. Machanavajjhala, J. Gehrke and D. Kifer, C, "*l*-Diversity: Privacy beyond *K*-Anonymity," ACM Transactions on Knowledge Discovery from Data, vol. 1. No. 1, pp: 24-35, 2007.
[8] J. Li, G. Liu , J. Xi, and Y. Lu, "An anonymity approach satisfying demand of maximum privacy disclosure rate"; Journal of YanShan University, 2010, 34(3).
[9] X. Qin, A. Men and Y. Zou, "Privacy preservation based on *K*-anonymity algorithms," Journal of ChiFeng university, vol. 26, no. 5, pp. 14-16, 2010.
[10] H. Jin, Z. Zhang, S. Liu, S. Ju, (α_i, *K*)-anonymity Privacy Preservation Based on Sensitivity Grading, Computer Engineering, Vol.37 No.14, pp.12-17.
[11] R. Wong, J. Li, A. Fu, et al "(α, *K*)-anonymity: An enhanced *K*-anonymity model for privacy preserving data publishing," Inter-national Conference on knowledge Discovery and Data Mining, 2006: 754-759.
[12] Y. Kan and T. Cao, "Enhanced privacy preserving K-anonymity model: (α, *L*)-diversity *K*-anonymity," Computer Engineering and Applications, vol. 46, no. 21, pp. 148-151, 2010.
[13] Y. Tong, Y. Tao, S. Tang and B. Yang, "Identity-Reserved anonymity in privacy preserving data publishing," Journal of Software, 2010, 21(4):771-781.
[14] T. Ma, M. Tang, "Data Mining Based on Privacy Preserving," Computer Engineering, 2008, 34(9).
[15] S. Zhou, F. Li, Y. Tao, X. Xiao, "Privacy Preservation in Database Applications: A Survey", Chinese Journal of Computers, 2009, 32(5).
[16] C. Huang, Y. Fei, M. Li, Y. Dai and J. Liu, "*K*-anonymity Algorithms Based on Multi-Dimensional Generalization Path". Computer Engineering, 2009, 35(2): 154-156.
[17] Z. Zhu, Z. Wang, and W. Wang, "Personal Privacy Constraints Based *K*-anonymity Model", Journal of Computer Research and Development; 2010, 47(Suppl.): 271-278.
[18] P. Wang and J. Wang, "Progress of research on *K*-anonymity privacy2preserving techniques," Journal of ChiFeng University (Natural Science Edition), vol. 27, no. 6, pp. 2016-2019, 2010.
[19] R. Wong, J. Li, A. Fu, K. Wang. "(alpha, k)-Anonymous Data Publishing, Journal of Intelligent Information Systems, Vol. 33, No. 2, Oct., pages 209-234, 2009.
[20] H. Luo and G. Liu, "(*L*, *K*)-anonymity for privacy preserving," Application Research of Computers, vol. 25, no. 2, pp. 564-574, 2008.
[21] R. Wang, J. Liu, "Research of privacy preserving association rules mining algorithm," Computer Engineering and Applications, 2009, 45(26):126-130.
[22] G. Li, Y. Wang and X. Su, "Privacy Preserving Data Ming on Decision Tree," Dianzi Xuebao (Acta Electronic Sinica). Vol. 38, no. 1, pp. 204-212. Jan. 2010.
[23] X. Yang, X. Liu, B. Wang, G. Yu, "*K*-anonymization approaches for supporting multiple constraints," Journal of Software, 2006, 17(5): 1222—1231.
[24] R.C.W. Wong, A. Fu, K. Wang, Y. Xu, J. Pei, and P. Yu, "Probabilistic Inference Protection on Anonymized Data", The 2010 IEEE International Conference on Data Mining (ICDM), Sydney, Australia on 14-17 Dec, 2010.

Yun Pan is a lecturer a lecturer of Information System, College of Computer and Information Engineering at Zhejiang Gongshang University. His research interests include decision support systems, XML data processing and cloud data management. He has published six papers in conferences and journals, together with five published books.

Xiao-ling Zhu is a junior student of College of Computer Science & Information Engineering, Zhejiang Gongshang University. Her current research interest focuses on management decision theory and decision support systems. She has published two papers in journals or proceedings.



Ting-gui Chen is a lecturer of Information System, College of Computer and Information Engineering at Zhejiang Gongshang University. His current research interest focuses on management decision theory and decision support systems, swarm intelligence and complexity science. He has published over 20 publications in academic journals and conference proceedings.