Anonymity: An Assessment and Perspective in Privacy Preserving Data Mining

Sumana M
M.S.Ramaiah Institute of Technology
ISE Department
Bangalore

Dr Hareesh K S
Associate Professor
Department of Computer Science and Engg,
Manipal Institute of Technology
Manipal

ABSTRACT
Privacy Preserving Data mining techniques depends on privacy, which captures what information is sensitive in the original data and should therefore be protected from either direct or indirect disclosure. Secrecy and anonymity are useful ways of thinking about privacy. This privacy should be measureable and entity to be considered private should be valuable. In this paper, we discuss the various anonymization techniques that can be used for privatizing data. The goal of anonymization is to secure access to confidential information while at the same time releasing aggregate information to the public. The challenge in each of the techniques is to protect data so that they can be published without revealing confidential information that can be linked to specific individuals. Also protection is to be achieved with minimum loss of the accuracy sought by database users. Different approaches of anonymization have been discussed and a comparison of the same has been provided.

General Terms
Data mining, privacy, anonymity.

Keywords
Data preprocessing,k-anonymity, quasi-identifier,

1. INTRODUCTION
Privacy Preserving Data Mining performs data mining on the private data. Different methods such as anonymization, perturbation[4] or cryptographical approaches have been used for privatizing the data. All the variations of the anonymization approach it is required that, in the released table the tuples/respondents are indistinguishable (within a set of individuals with respect a set of attributes, called quasi-identifier. The k-anonymity privacy requirement for publishing microdata requires that each equivalence class (i.e., a set of records that are indistinguishable from each other with respect to certain "identifying" attributes) contains at least k records. Recently, several authors have recognized that k-anonymity cannot prevent attribute disclosure. The notion of ℓ-diversity has been proposed to address this; ℓ-diversity requires that each equivalence class has at least ℓ well-represented values for each sensitive attribute. ℓ-diversity is complex and not sufficient to prevent attribute disclosure. An approach called t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). General (α,k)-anonymity model is an effective approach to protecting individual privacy before microdata are released. But it has some defects on privacy preservation and data distortion when the distribution of sensitive values is not well-proportioned. To solve the problem, a complete (α,k)-anonymity model is proposed which can implement sensitive values individuation preservation by setting the frequency constraints for each sensitive value in all the equivalence classes.

Table 1: 3-anonymity

TID	Color	Birth	Gender	ZIP	Income
t1	Black	1965	M	560054	15000
t2	Black	1965	M	560054	17000
t3	Black	1964	M	560054	25000
t4	White	1964	F	540064	24000
t5	White	1965	F	540064	15000
t6	White	1964	F	540064	16000
t7	Black	1965	F	540064	15600

2. K-ANONYMITY
K-anonymity[2] is a privacy model developed for the linking attack. The concept of *k*-anonymity tries to capture, on the private table PT to be released, one of the main requirements that has been followed by the statistical community and by agencies releasing the data, and according to which the released data should be indistinguishably related to no less than a certain number of respondents.

Definition 1. Quasi-Identifier: *Given a table T with attributes* $(A_1,\ldots\ldots,A_n)$*, a quasi –identifier is a minimal set of attributes* $(A_{i1},\ldots\ldots,A_{il})$ $(1<= i1<\ldots\ldots < il <= n)$ *in T that can be joined with external information to re-identify individual records.*

For example in Table 1 the Quasi-identifier is (color,Gender,Zip).

Definition 2. K-Anonymity *: A table T is said to be k-anonymous given a parameter k and the quasi-identifier QI =* $(A_{i1},\ldots\ldots,A_{il})$ *if for each tuple* $t_i \in T$*, there exist at least another (k-1) tuples* $t_1,\ldots\ldots,t_{k-1}$ *such that those k tuples have the same projection on the quasi-identifier. Tuple t and all other tuples*

indistinguishable from t on the quasi-identifier form an equivalence class.

K-anonymity focuses on 2 techniques in particular: Generalization and Suppression. Suppression is masking the attribute value with a special value in the domain. Generalization is replacing a specific value with a more generalized one.

The idea of generalizing an attribute is a simple concept. A value is replaced by a less specific, more general value that is faithful to the original. In Table 1 the original ZIP codes {560054,540059} can be generalized to 56005* thereby stripping the rightmost digit and semantically indicating a larger geographical area. So there is mapping from Z_o={560054,540059} to Z_1 = 56005*.

Given an attribute *A* of a private table PT, we can define a domain generalization hierarchy DGH_A for *A* as a set of functions f_h : *h*=0,…,*n*-1 such that: $f_0(A_0)$ -> A_1 ,$f_1(A_1)$ -> A_2,……, $f_{n-1}(A_{n-1})$ -> A_n where A= A_0 and I A_n I =1. Hence $DGH_A = \bigcup_{h=0}^{n} Ah$. Such a relationship implies the existence of a value generalization hierarchy VGH_A for attribute *A*.

For example domain R0 = {Black, White} can be generalized to domain R1 = {Person} which can be further generalized to domain R2 = {******}. Similarly zip code values fall in the domain Z0 = {560054,560059,560064} can be generalized to domain Z1 = { 56005*, 56006*} which is further generalized to Z3 = { 5600**} and Z4 = { ******}. Another method adopted to be applied in conjunction with generalization to obtain *k*-anonymity is *tuple suppression*. The intuition behind the introduction of suppression is that this additional method can reduce the amount of generalization necessary to satisfy the *k*-anonymity constraint. The application of generalization and suppression to a private table PT produces more general (less precise) and less complete (if some tuples are suppressed) tables that provide better protection of the respondents' identities. Generalized tables are then defined as follows.

Definition 3 (Generalized table - with suppression). Let Ti and Tj be two tables defined on the same set of attributes. Table Tj is said to be a generalization (with tuple suppression) of table Ti, denoted Ti ≤ Tj , iff:

1. Size of Ti <= size of Tj

2. the domain dom(A, Tj) of each attribute A in Tj is equal to, or a generalization of, the domain dom(A, Ti) of attribute A in Ti;

3. it is possible to define an one-to-one function associating each tuple tj in Tj with a tuple ti in Ti, such that the value of each attribute in tj is equal to, or a generalization of, the value of the corresponding attribute in ti.

Note that like for generalization[3], it is possible to adopt different suppression solutions for guaranteeing k-anonymity without removing more tuples than necessary (i.e., ensuring minimality of the suppression), at a given level of generalization. The joint use of generalization and suppression helps in maintaining as much information as possible in the process of k-anonymization. The question is whether it is better to generalize, loosing data precision, or to suppress, loosing completeness. Samarati assumes that the data holder establishes a threshold, denoted MaxSup, specifying the maximum number of tuples that can be suppressed. The concept of k-minimal generalization with suppression is then formally defined as follows.

Definition 4 (k-minimal generalization - with suppression). Let Ti and Tj be two tables such that Ti ≤ Tj , and let MaxSup be the specified threshold of acceptable suppression. Tj is said to be a k-minimal generalization of table Ti iff:

1. Tj satisfies k-anonymity enforcing minimal required suppression, that is, Tj satisfies k-anonymity and $\forall T_z$: Ti ≤ T_z; $DV_{i,z} = DV_{i,j}$; T_z satisfies k-anonymity => lTjl ≥ lTzl

2. lTil - lTjl ≤ MaxSup

3. $\forall T_z$: Ti ≤ Tz and Tz satisfies conditions 1 and 2 => $\sim(DV_{i,z} < DV_{i,j})$.

Intuitively as shown in [5], this definition states that a generalization Tj is k-minimal iff it satisfies k-anonymity, it does not enforce more suppression than it is allowed lTil - lTjl ≤ MaxSup, , and there does not exist another generalization satisfying these conditions with a distance vector smaller than that of Tj . Where $DV_{i,j}$ is the distance vector from table i to table j. Some of the most important advantages of k-anonymity are that No additional noise or artificial perturbation is added into the original data and also protects identity disclosure.

But consider a *k*-anonymized table, where there is a sensitive attribute and suppose that all tuples with a specific value for the quasi-identifier have the same sensitive attribute value. Machanavajjhala, Gehrke, and Kifer describe two possible attacks. As shown in [4] One, is Homogeneity Attack where an attacker knows both the quasi-identifier value of an entity and knows that this entity is represented in the table, then the attacker can infer the sensitive value associated with certainty. Two, the background knowledge attack is instead based on a prior knowledge of the attacker of some additional external information. For instance, suppose that Alice knows that Hellen is a white female. Alice can then infer that Hellen suffers of chest pain or short breath. Suppose now that Alice knows that Hellen runs for two hours every day. Since a person that suffers of short breath cannot run for a long period, Alice can infer with probability equal to 1 that Hellen suffers of chest pain.

3. ℓ-DIVERSITY

To avoid the above problems of k-anonymity attacks, Machanavajjhala, Gehrke, and Kifer introduce the notion of ℓ-diversity as shown in [6].

Definition 5 (The ℓ -diversity Principle): An equivalence class is said to have ℓ-diversity if there are at least ℓ "well-represented" values for the sensitive attribute. A table is said to have ℓ -diversity if every equivalence class of the table has ℓ -diversity.

In Distinct ℓ -diversity the simplest understanding of "well represented" would be to ensure there are atleast **ℓ** *distinct* values for the sensitive attribute in each equivalence class.

Distinct **ℓ** -diversity does not prevent probabilistic inference attacks. The entropy of an equivalence class E is defined to be

$Entropy(E) = -\sum_{s \in S} p(E,s) \log p(E,s)$

in which S is the domain of the sensitive attribute, and $p(E, s)$ is the fraction of records in E that have sensitive value s. A table is said to have entropy **ℓ** -diversity if for every equivalence class E, $Entropy\ (E) \geq \log \ell$. Sometimes this may too restrictive, as the entropy of the entire table may be low if a few values are very common. This leads to the following less conservative notion of **ℓ** -diversity. Recursive (c, **ℓ**)-diversity makes sure that the most frequent value does not appear too frequently, and the less frequent values do not appear too rarely. Recursive (c, **ℓ**)-diversity makes sure that the most frequent value does not appear too frequently, and the less frequent values do not appear too rarely. Let m be the number of values in an equivalence class, and r_i, $1 \leq i \leq m$ be the number of times that the i^{th} most frequent sensitive value appears in an equivalence class E. Then E is said to have recursive(c, **ℓ**)-diversity if $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$.

While the ℓ-diversity principle represents an important step beyond k-anonymity in protecting against attribute disclosure, it has several shortcomings. One, ℓ-diversity may be difficult and unnecessary to achieve. Also, ℓ-diversity is insufficient to prevent attribute disclosure. Attacks on ℓ-diversity can be described as follows. One, *Skewness Attack:* When the overall distribution is skewed, satisfying ℓ-diversity does not prevent attribute disclosure. Another, *Similarity Attack:* When the sensitive attribute values in an equivalence class are distinct but semantically similar, an adversary can learn important information.

Table 2. Original Salary table

TID	ZIP Code	Age	Salary	Disease
1	560077	29	3K	gastric ulcer
2	560002	22	4K	gastritis
3	560078	27	5K	stomach cancer
4	560005	43	6K	gastritis
5	560009	52	11K	flu
6	560006	47	8K	bronchitis
7	560005	30	7K	bronchitis
8	560073	36	9K	pneumonia
9	560007	32	10K	stomach cancer

Table 3: A 3-diverse version of Table 2

TID	ZIP Code	Age	Salary	Disease
1	5600**	2*	3K	gastric ulcer
2	5600**	2*	4K	gastritis
3	5600**	2*	5K	stomach cancer
4	56000*	≥40	6K	gastritis
5	56000*	≥40	11K	flu
6	56000*	≥40	8K	bronchitis
7	5600**	3*	7K	bronchitis
8	5600**	3*	9K	pneumonia
9	5600**	3*	10K	stomach cancer

Table 2 is the original table, and Table 3 shows an anonymized version satisfying distinct and entropy 3-diversity. There are two sensitive attributes: *Salary* and *Disease.* Suppose one knows that Bob's record corresponds to one of the first three records, then one knows that Bob's salary is in the range [3K–5K] and can infer that Bob's salary is relatively low. This attack applies not only to numeric attributes like "Salary", but also to categorical attributes like "Disease". Knowing that Bob's record belongs to the first equivalence class enables one to conclude that Bob has some stomach-related problems, because all three diseases in the class are stomach-related. This leakage of sensitive information occurs because while ℓ-diversity requirement ensures "diversity" of sensitive values in each group, it does not take into account the semantical closeness of these values.

4. T-CLOSENESS

Definition 6 (The *t*-closeness Principle :) An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.

As described in [6] , By knowing the quasi-identifier values of the individual, the observer is able to identify the equivalence class that the individual's record is in, and learns the distribution P of sensitive attribute values in this class. We assume that Q is the distribution of the sensitive attribute in the overall population in the table. We require that P and Q are close. Now the problem is to measure the distance between these two probabilistic distributions. There are a number of ways to define the distance between them. The variational distance, Kullback-Leibler (KL) distance and Earth Mover's Distance measures have been used. Selecting and using a distance measure in t-closeness is a major drawback in this approach. While EMD measure combines distance –estimation properties but does not include the scaling nature of the KL distance measure.

5. COMPLETE (α,K)-ANONYMITY MODEL

Usually, different sensitive values have different sensitivities and should have different protection requirements. Complete (α,k)-anonymity model[1] sets a specific frequency constraint α for each sensitive value. Different sensitive values have different frequency constraints α, which can implement that sensitive values with high sensitivity have low frequency in each

equivalence class. For example the attribute disease is a sensitive attribute with value "HIV" more sensitive than value "Fever" or "Flu".

Definition 7 (Complete (*α*,*k*)-anonymity). Given an anonymity table T', a quasi-identifier attributes set Q and a sensitive attribute domain S. For each sensitive value s (s ∈ S), let α_s be a user-specified threshold of s. T' is said to be a complete (α,k)-anonymization if T' satisfies k-anonymity and also satisfies simple α_s-deassociation property for each s with respect to Q and S.

Complete (*α*,*k*)-anonymity model, which requires that each sensitive value *s* ($s \in S$) satisfies corresponding simple (α_s,*k*)-anonymity model, is more flexible compared with general (*α*,*k*)-anonymity model and simple (*α*,*k*)-anonymity model.

Definition 8 (*α*-Deassociation). Given a dataset D, an attribute set Q and a sensitive value s in the domain of attribute S ∈ Q . Let (E, s) be the set of tuples in equivalence class E containing s for S and α be a user-specified threshold, where $0 < \alpha < 1$. Dataset D is α-deassociated with respect to attribute set Q and the sensitive value s if the relative frequency of s in every equivalence class is less than or equal to α. That is, $|(E, s)|/|E| \leq \alpha$ for all equivalence classes E.

Definition 9 (Simple (*α*,*k*)-anonymity). Given an anonymity table T', a quasi-identifier Q and a sensitive value s in the domain of sensitive attribute. T' is said to be a simple (α,k)-anonymity if T' satisfies both k-anonymity and α-deassociation properties with respect to Q and s.

The constraint *α* in the simple (*α*,*k*)-anonymity model is only oriented to one specific sensitive value, so simple (*α*,*k*)-anonymity tables cannot protect other sensitive values.

Definition 10 (*α*-Rare). Given an equivalence class E, a sensitive attribute domain X and an attribute value x € X . Let (E, x) be the set of tuples containing x in E and α be a user-specified threshold, where $0 \leq \alpha \leq 1$. Equivalence class E is α-rare with respect to sensitive attribute set X if the proportion of every attribute value of X in the dataset is not greater than α, i.e. $|(E, x)|/|E| \leq \alpha$ for x ∈ X .

Definition 11 (General *α*-deassociation Property). Given an anonymity table T', a quasi-identifier Q and a sensitive attribute X. Let α be a user-specified threshold, where $0 \leq \alpha \leq 1$. Dataset T' is generally α-deassociated with respect to Q and X if, for any equivalent classes,

E∈T ' , E is α-rare with respect to X.

Definition 12 (General (*α*,*k*)-anonymity). Given an anonymity table T', a quasi-identifier Q and a sensitive attribute domain X. T' is said to be a general (α,k)- anonymity if T' satisfies both k-anonymity and general α- deassociation properties with respect to Q and X.

General (*α*,*k*)-anonymity model, which sets one *α* for all sensitive values, is an extension of simple (*α*,*k*)- anonymity model. It lacks flexibility, for it makes all sensitive values use one uniform *α*. Actually, different sensitive values generally have different sensitivities and should use different *α*.

Complete (*α*,*k*)-anonymity model as an extension of general (*α*,*k*)-anonymity model or simple (*α*,*k*)-anonymity model. When an *α*-threshold is set for one sensitive value, i.e. let $\alpha_s = \alpha$ ($0<\alpha<1$) and *s'*(*s* '€{*S* − *s*}), α_s'=1, it becomes a simple (*α*,*k*)-anonymity model. When *α*-threshold is set for all sensitive values, i.e. let ∈*s* ($s \in S$), $\alpha_s = \alpha$ ($0<\alpha<1$), it becomes a general (*α*,*k*)-anonymity model. When α_s=1 for every *s* in *S*, it becomes a *k*-anonymity model. To implement complete (*α*,*k*)-anonymity model, we must set parameter *α* for each sensitive value according to its sensitivity. Parameter *α* should satisfy 2 constraints: one is that *α* should no less that the sensitive value frequency in dataset, or else it is impossible to generate the dataset satisfying (*α*,*k*)-anonymity constraint; the other is that *α* should no smaller than the smallest frequency in the equivalence class. For example in a 2-anonymity dataset, the size of the optimal equivalence class is between *2* and 2*2-1, so *α* cannot be less than 1/3, or else it is impossible to generate an optimal equivalence class satisfying (*α*,*k*)-anonymity constraint.

Aggarwal has proved that the size of the optimal equivalence class should be between *k* and 2*k*-1. So the two classes C_1, C_2 can be clustered into one class should satisfy two conditions: one is that the size of the merged class should no more than 2*k*, namely rule 1. The other is that it should satisfy complete (*α*,*k*)-anonymity model frequency constraint, namely rule 2.

rule 1: $|C_1|+|C_2| \leq 2k-1$

rule 2: Let $n=\max\{k, |C_1+C_2|\}$, *x* to be sensitive value in C_1+C_2, $|C_1+C_2, x|$ to be the number of record which sensitive value is *x* in C_1+C_2, α_x to be frequency constraint of value *x*, then $\in x, |C_1+C_2,x|/n \leq \alpha_x$

The equivalence classes which satisfy rule 1 and rule 2 are named can-be-merged equivalence class set, and one of them is called compatible equivalence class to others in the same can-be-merged equivalence class set.

The complete (*α*,*k*)-anonymity clustering algorithm's idea is: repeat selecting a class C_1 whose size is less than *k*, finding the compatible equivalence class C_2 with the nearest distance to C_1, merge C_1 and C_2 to C_{12}, until the size of all the classes is between *k* and 2*k*-1. The algorithm is showed below.

Table 4 also satisfies complete (*α*,*k*)-anonymity, when we let both *α* for "HIV" and *α* for"Cancer" be 0.4 and let both *α* for "flu" and *α* for "fever" be 0.9.

Table 4: Complete(0.4,3) anonymity table

Job	Birth	Postcode	Illness
*	1975.*.*	154*	HIV
*	1975.*.*	154*	flu
*	1975.*.*	154*	fever
*	1975.*.*	154*	Cancer
*	1975.1.*	1542	Cancer
*	1975.1.*	1542	flu
*	1975.1.*	1542	HIV

6. INFORMATION LOSS

As in [1] Figure1 plots the performance curves of the information loss over various *k* values with the four models. From figure1 we can see that the information loss of the four models increases as the *k* increasing. This is because that with *k* increasing, we require more tuples to be identical, more distortions will be generated. We can also see that at the same *k*, the information loss of *k*-anonymity model is the least, the second is simple (α,*k*)-anonymity model, then is complete (α,*k*)-anonymity model and general (α,*k*)-anonymity model's information loss is the most.

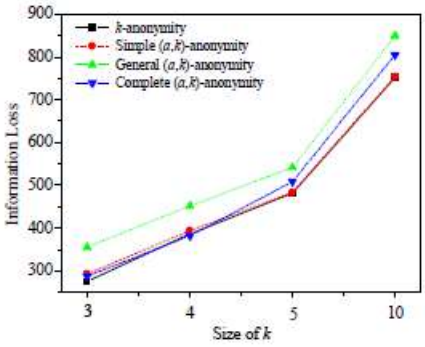

Figure 1: Information loss comparisons for varying k

Figure2 plots the performance curves of information loss over various quasi-identifier sizes with the four models. We can see that the distortion ratio increases with the quasi- identifier size increasing. This is because that as the quasi-identifier size increases, more distortion is needed. We can also see that at the same quasi-identifier size, the information loss of k-anonymity model is the least, the second is simple (α,k)-anonymity model, then is complete (α,k)-anonymity model, general (α,k)-anonymity model's information loss is the most.

So from fig.1 and fig. 2, we can conclude that the complete (α,k)-anonymity model can provide effective protection for all the sensitive values with low information loss.

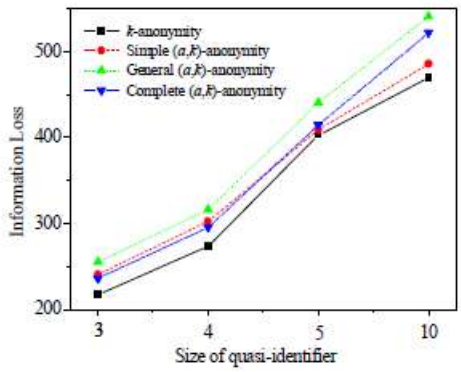

Figure 2: Information Loss comparisons based on quasi-identifiers

7. CONCLUSION

Data mining techniques are used to find patterns in large databases of information. But sometimes these patterns can reveal sensitive information about the data holder or individuals whose information are the subject of the patterns. The notion of privacy-preserving data mining is to identify and disallow such revelations as evident in the kinds of patterns learned using traditional data mining techniques. Due to the varying privacy needs of different individuals only one single approach is not realistic in many privacy preserving data mining tasks.

K-anonymity has recently been investigated as an interesting approach to protect microdata undergoing public or semi-public release from linking attacks. In this paper the original *k*-anonymity proposal and its enforcement via generalization and suppression as means to protect respondents' identities while releasing truthful information. While *k*-anonymity protects against identity disclosure, it does not provide sufficient protection against attribute disclosure. The notion of ℓ-diversity attempts to solve this problem by requiring that each equivalence class has at least ℓ well-represented values for each sensitive attribute. We have shown that ℓ-diversity has a number of limitations and have proposed a privacy notion called *t*-closeness which requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold *t*). A complete (α,*k*) anonymity model has been proposed which can satisfy sensitive values individuation secure requirement by setting frequency constraint for each sensitive value. The design of a complete (α,*k*) clustering algorithm has also been discussed. Experimental results show that the complete (α,*k*)-anonymity model can preserve privacy effectively with less data distortion.

8. REFERENCES

[1]. Han Jian-min, Yu Hui-qun, Yu Juan, Cen Ting-ting "A Complete (α,*k*)-Anonymity Model for Sensitive Values Individuation Preservation", 2008 IEEE DOI 10.1109/ISECS.2008.92

[2]. Latanya Sweeney "Achieving *k*-anonymity Privacy Protection Using Generalization and Suppression",May 2002, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 571-588.

[3]. V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, *k*-Anonymity, Springer US, Advances in Information Security (2007).

[4]. Li Liu, Murat Kantarcioglu and Bhavani Thuraisingham, Privacy Preserving Decision Tree Mining from Perturbed Data, Proceedings of the 42nd Hawaii International Conference on System Sciences – 2009

[5]. Charu C Aggarwal, Philip S Yu, Privacy Preserving Data Mining: Models and Algorithms, Springer Publication, 2007.

[6]. Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian, *t*-Closeness: Privacy Beyond *k*-Anonymity and ℓ-Diversity, Citiseer 2007