(*a*, *d*)-Diversity: Privacy Protection Based on *l*-Diversity

Qian Wang Xiangling Shi
College of Computer Science, Chongqing University
Chongqing, China, 400030
{Wangq_cqu, sxl_430}@163.com

Abstract

In recent years, privacy protection has been emphasized while publishing data with sensitive information. Existing proposals for privacy protection can well avoid identity disclosure; however they do not provide sufficient protection for privacy under background knowledge attack. This paper analyzes the cause of attribute disclosure and proposes a novel idea for privacy protection based on l-Diversity. It takes the semantic meaning of the sensitive attributes into consideration and gives a stronger definition of privacy protection. 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. The experiment results shown in the paper demonstrate the feasibility of the proposal.

1. Introduction

With the widespread use of the Internet, more and more people rely on the Internet. However, they often inadvertently leave the basic personal information (such as Zip code, Birth date, Gender, etc.) on the Internet. With enhanced search engine, it makes information easy to obtain.

Meanwhile, the development of data mining has urged more and more companies or organizations to publish their data in order to discovery the potential rules among it. The published data usually contains private information of individuals, such as medical records released in order to study the spread of certain diseases.

Although companies and organizations are aware of the disclosure and remove the identifiers from the data before it's published, the problem still exists because a group of certain attributes, e.g., Zip-code, Birth date and Gender, may also leak the identity of individuals. A typical example is shown in [1].

Information disclosure can be divided into two categories: *Identity Disclosure* and *Attribute Disclosure*. Identity disclosure happens when an individual can be uniquely identified from the published data. Attribute disclosure happens when the information of an individual can be inferred from the published data.

Generally, there are three kinds of attribute in a data table:

Identifier: Attributes that uniquely identify an individual, e.g., Name, Social Security Number.

Quasi-identifier: Attributes whose values when taken together can potentially identify an individual, e.g., Zip-code, Birth date and Gender.

Sensitive Attribute: Attribute that is considered sensitive, e.g., Disease and Salary.

Latanya Sweeney proposes *k*-Anonymity, which requires each record of the published data table to be indistinguishable from at least *k*-1 other records with respect to the set of Quasi-identifiers. *k*-Anonymity well solves the problem of identity disclosure, but it doesn't take the sensitive attributes into consideration, and is unable to protect the privacy from *Homogeneity Attack* or *Background Knowledge Attack* [2].

l-Diversity [2] requires the sensitive attribute values to be well presented in each group with the same quasi-identifiers, and reduces the possibility of attribute disclosure, especially well solves problems caused by Homogeneity Attack. Similar methods include (α,*k*)-Anonymity [3]. *t*-Closeness [4] comes up with the consideration of the distribution of the sensitive attributes. It restricts the distance between the distribution of a sensitive attribute in each group with the same quasi-identifiers and the distribution of the same sensitive attribute in the whole table. (ε,*m*)-Anonymity [5] requires that, given a group *G* with the same quasi-identifiers, for every sensitive value *x* in *G*, at the most 1/*m* of the tuples in *G* can have sensitive values "similar" to *x*, where the similarity is controlled by ε. However, existing principles do not consider the sensitive attributes in the real world and ignore the relations among the sensitive attribute values.

978-0-7695-3570-8/09 $25.00 © 2009 IEEE
DOI 10.1109/WCSE.2009.362

Although (ε,m)-Anonymity defines the range of the sensitive values, it only works for continuous sensitive attributes.

In this paper, we have studied the previous works, and propose the restriction of privacy to lower the risk of information disclosure. The paper is organized as follows: Section 2 gives a description of the preliminaries. Section 3 presents the novel principle. Section 4 discusses the algorithm used to implement the idea. Section 5 lists the results using the algorithm discussed. We draw a conclusion in Section 6.

2. Preliminaries

Basic Concept

$T = \{t_1, t_2, \ldots, t_n\}$ is a table with attributes $A_1, \ldots, A_m$. t_i denotes the i^{th} record in table T. A represents the set of all attributes $\{A_1, A_2, \ldots, A_m\}$ and $t[A_i]$ is the value of attribute A_i for record t. S represents the set of all sensitive attributes, and QI denotes the set of all quasi-identifiers.

Equivalent Class

A block of records with the same quasi-identifier values, also called q^*-block.

Generalization

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

***l*-Diversity**

A q^*-block is l-diverse if it contains at least l "well-represented" values for the sensitive attribute S. A table is l-diverse if every q^*-block is l-diverse.

Background Knowledge Attack

Background knowledge attack assumes that the attacker knows that an individual's information is contained in an equivalent class, and based on certain knowledge that the attacker gains somewhere else, e.g., the relationship between an attribute in QI and S, the attacker is convinced to identify the sensitive attribute value of the individual.

3. (*a*,*d*)-Diversity

3.1. Problem

Table 1 shows an original medical record table with identities removed. Age, Gender and Work Class are Quasi-identifier. Condition is considered as sensitive attribute. Table 2 is a 4-Diversity table derived from Table 1. The records are divided into three equivalent classes.

Table 1. Original medical record table

	Age	Gender	Work Class	Condition
1	38	Female	Local-gov	Gastric Ulcer
2	29	Male	Private	Gastritis
3	18	Male	Self-emp-not-inc	Paludism
4	46	Female	Federal-gov	Gastritis
5	27	Female	Local-gov	Influenza
6	25	Male	Self-emp-inc	Fracture
7	59	Female	State-gov	Fracture
8	22	Male	Self-emp-not-inc	Pneumonia
9	19	Male	Self-emp-not-inc	Influenza
10	25	Male	Private	Dyspepsia
11	20	Male	Self-emp-not-inc	Dislocation
12	54	Female	Federal-Gov	Dislocation

Table 2. 4-Diversity table

	Age	Gender	Work Class	Condition
1	18-22	Male	Self-emp-not-inc	Influenza
2	18-22	Male	Self-emp-not-inc	Paludism
3	18-22	Male	Self-emp-not-inc	Pneumonia
4	18-22	Male	Self-emp-not-inc	Dislocation
5	25-38	*	*	Gastritis
6	25-38	*	*	Dyspepsia
7	25-38	*	*	Fracture
8	25-38	*	*	Gastric Ulcer
9	27-59	Female	Gov	Influenza
10	27-59	Female	Gov	Gastritis
11	27-59	Female	Gov	Dislocation
12	27-59	Female	Gov	Fracture

Suppose Tom is a neighbor of Alice. One day, Alice gets sick, and is sent to the hospital. Tom wants to find out the condition of Alice. He looks into the 4-diversity table of the current medical records published by the hospital as it contains information about Alice. As a neighbor he knows that Alice is a 39-year-old woman with a local government job. From Table 2, it's obvious to see Alice's data is in the last equivalent class. There are four values of the sensitive attribute, so Tom needs more information to determine which value belongs to Alice.

However, the day before Alice was sent to the hospital, Tom found that Alice had a fever. Since only one of the diseases, influenza, has the symptom of fever, Tom is convinced that Alice catches influenza.

This example suggests that l-Diversity has to be improved to reduce the risk of attribute disclosure.

3.2 Premise

Before analysis, see the two extreme scenarios below based on the example described in 3.1:

(1) Tom is an experienced doctor, and gets along well with Alice. They often talk to each other. The day before Alice was sent to the hospital, she has told Tom

about her feelings. According to years' experiences, Tom has already had an idea of what disease she catches, there's no need for him to look into the data published by the hospital.

(2) Tom has just moved here, and is not familiar with his neighbor Alice. He knows nothing except her age, gender and work class. After the data is published, he can only guess what disease she catches.

In scenario (1), Tom has strong background knowledge, while in scenario (2), Tom has no background knowledge. For those with strong background knowledge, whatever the publish table is like, they still may come to the right answer. Therefore, it's practical to protect privacy from attackers with non-strong background knowledge. This paper based on the premise of non-strong background knowledge.

In addition, the origin data table is derived from the real world, and the attributes always have real meaning. Therefore, there're relationships among the attribute values. For example, diseases with the stomach problem always cause stomachache. This paper is also based on the premise of relations among attribute values.

3.3 Analysis

l-Diversity essentially concerns about the number of sensitive attribute values in each equivalent class. Suppose in the situation that an attacker A knows a patient B's information is contained in an equivalent class of a *l*-Diversity table, n_i is the number of records with the sensitive value S_i, and n represents the number of records in the equivalent class and k is the background knowledge. We define the risk of the disclosure of S_i as:

$$P(S_i \mid k) = \frac{P(S_i) \cdot P(k \mid S_i)}{\sum_{j=1}^{c} P(S_j) \cdot P(k \mid S_j)} \quad (1)$$

Where $P(S_i)$ is the risk of the disclosure of S_i without background knowledge:

$$P(S_i) = \frac{n_i}{n} \quad (2)$$

Suppose A is Tom, and B is Alice, now A knows B is in the last equivalent class of Table 2 and B had a fever the day before she was sent to the hospital. As we know, influenza usually causes fever while the other three seldom do. Now k denotes the symptom of diseases. Given $P(k|S_1) = 0.9$, $P(k|S_2) = 0.08$, $P(k|S_3) = 0.03$, $P(k|S_4) = 0.02$, the risk of disclosure is:

$$P(S_1 \mid k) = \frac{P(S_1) \cdot P(k \mid S_1)}{\sum_{j=1}^{c} P(S_j) \cdot P(k \mid S_j)} = 0.874$$

Similarly $P(S_2|k) = 0.078$, $P(S_3|k) = 0.029$, $P(S_4|k) = 0.019$. $P(S_1|k)$ is greater than the others, so it will be easy for Tom to conclude that Alice has caught influenza with a high probability.

(α,k)-Anonymity is similar to *l*-Diversity. *t*-Closeness considers the distribution of each sensitive attribute value to be close to the distribution of that sensitive attribute values in the whole table, however it doesn't make much improvement in the above situation. (ε,m)-Anonymity notices the similarity of the sensitive attribute values, but only considers the continuous sensitive attribute. The high probability of such disclosure mainly depends on the real meaning of the sensitive attributes. None of existing principles takes this into consideration.

3.4 (*a*,*d*)-Diversity

Based on the premise of relations among attribute values, there exist same or similar characteristics among sensitive attribute values, e.g., some of the diseases share the same symptoms that are different from other diseases. If an equivalent class mentioned above contains another disease that has the symptom of fever, Tom still needs more knowledge to get the right answer, therefore the risk of information disclosure will be lowered. On the other hand, since diseases with the same symptoms are usually caused by the same or similar problem, e.g., diseases with the symptoms of cough are usually caused by lung problems, if all the diseases in an equivalent class have the same symptom, the attacker will be aware of what's wrong with the individual with little or no background knowledge at all.

Taking the relations among the real meaning of the sensitive attributes into consideration, we derive (a, d)-diversity principle as follow:

(*a*, *d*)-Diversity: An equivalent class is said to satisfy (a, d)-diversity if it contains at least a analogous values for sensitive attributes S, and at least d dissimilar values for S. A table is said to satisfy (a, d)-diversity if every equivalent class satisfies (a, d)-diversity.

Analogous Values: Given certain characteristic, two values are analogous if they have the most in common.

Dissimilar Values: Given certain characteristic, two values are dissimilar, if they have nothing or little in common.

Analogous and dissimilar values are measured by $\delta = |P(k|S_1) - P(k|S_2)|$, where k represents the specific characteristic.

Still in the example above, if we use the symptoms of the diseases to tell whether they are analogous or dissimilar, we will get a (2, 2)-diversity table shown in Table 3. Now Alice is in the second equivalent class. Since both influenza and pneumonia have the symptom of fever, it's not easy to tell which disease Alice has caught based on the same background knowledge.

Table 3. (2, 2)-Diversity table

	Age	Gender	Work Class	Condition
1	18-29	Male	Non-gov	Influenza
2	18-29	Male	Non-gov	Paludism
3	18-29	Male	Non-gov	Gastritis
4	18-29	Male	Non-gov	Dyspepsia
5	20-27	*	*	Dislocation
6	20-27	*	*	Fracture
7	20-27	*	*	Pneumonia
8	20-27	*	*	Influenza
9	38-59	Female	Gov	Gastric Ulcer
10	38-59	Female	Gov	Gastritis
11	38-59	Female	Gov	Dislocation
12	38-59	Female	Gov	Fracture

If E is an equivalent class, and an individual is in E. The goal of (a,d)-Diversity is prevent $P(S_i|k)$ from being too much bigger or smaller than the others in E while keeping balance of each $P(S_i)$ in E.

If $P(S_x)$ is much bigger than the others in E, the risk of disclosure of the sensitive attribute S_x will be higher even with little background knowledge. In order to prevent this phenomenon, we keep every $P(S)$ around $\overline{P(S)}$.

$$\overline{P(S)} = \frac{1}{c}\sum_{i=1}^{c} P(S_i) \tag{3}$$

In order to make every $P(S_i)$ close to $\overline{P(S)}$, the number of records with different sensitive attribute values will be adjusted.

According to Equation (1), the bigger difference between $P(k|S_i)$ will result in the bigger difference between $P(S_i|k)$. However, continue to adjust the number of records will break the balance of $P(S)$.

(a, d)-Diversity considers both distribution of $P(S_i)$ and $P(S_i|k)$. First, records are divided into groups. In each group, for every sensitive attribute value, there should be at least a-1 analogous value, and there should be at least d blocks of records with analogous value.

To get the sensitive attribute value of an individual in the specific equivalent class E with background knowledge k, the attacker should determine which block the individual belongs to first, and then picks out the sensitive attribute value. The risk of attribute disclosure will be affected by a and d. If d becomes bigger, the risk for the attacker to find out which block an individual belongs to will be lower. If a becomes bigger, the risk for the attacker to find out the exact sensitive attribute value will be lower.

4. Implementation

In this section we build an algorithm to achieve (a, d)-diversity table from an original one using generalization. It contains the following steps:

(1) Analyze the origin table

Given a table T, we have to break down the attributes into quasi-identifiers (QI), sensitive attributes (S) and other attributes. For each QI, if it is a discrete or categorical attribute, then we use a generalization hierarchy specified by the publisher. And for each pair of sensitive attributes, we compute the similarity (Sim) between them as follow:

$$Sim(S_1, S_2) = \begin{cases} 1 & S1, S2 \text{ are analogous} \\ 0 & S1, S2 \text{ are dissimilar} \end{cases} \tag{4}$$

With the definition of similarity above, we divide the origin table into several groups, and in each group, the similarity between every two sensitive attribute values is 1.

(2) Get a-Similarity blocks

For each record in each group derived from step (1), find at least a-1 records with different sensitive attribute value and distance ($Dist$) between each two record's quasi-identifier as smaller as possible. These a records form an a-Similarity block. In each block, change the values into a range from the minimum to maximum in the block, if the attribute is continuous. Generalize the values to the same node in the generalization hierarchy with the least operations if the attribute is categorical.

Using the idea of clustering [8], when computing the distance between two quasi-identifiers, we use the quasi-identifier attributes to define a metric space and then each record is viewed as a point in the space. Since the ultimate table is used for data mining [9], different attributes in quasi-identifiers may have different usages. We assign each attribute an argument w indicating its usages for analysis. Therefore, $Dist$ can be defined as follow:

$$Dist(QI_1, QI_2) = \sum_{i=1}^{c} w_i \cdot dist(QI_1[i], QI_2[i]) \tag{5}$$

where $\sum_{i=1}^{c} w_i = 1$, and

$$dist(qi_1, qi_2) = \begin{cases} \dfrac{|qi_2 - qi_1|}{qi_{\max} - qi_{\min}} & qi_1, qi_2 \text{ are continuous} \\ \dfrac{St(qi_1, qi_2)}{St(qi_1, r) + St(qi_2, r)} & qi_1, qi_2 \text{ are categorical} \end{cases}$$

$St(qi_1, qi_2)$ denotes the least generalization operation needed to convert qi_1 and qi_2 to the same node in the generation hierarchy, and r represents the top most node in the same generation hierarchy.

(3) Retrieve (a, d)-Diversity table

For each block in each group, find d-1 blocks with distance (*Dist*) of quasi-identifiers as close as possible and each block belongs to different group. Finally, we derive the (a, d)-Diversity result of the origin table.

The algorithm is described as follow:

Input: Data table T, Sensitive attribute groups $SG=\{S_1, S_2, \dots, S_b\}$, Weights of quasi-identifier attributes $W=\{w_1, w_2, \dots, w_c\}$, a, d.
Output: Data table satisfying (a, d)-Diversity T'.
Procedure:
1. $T_1=T$, $G=\{\}$.
2. For each record t in T_1.
3. $G_{tmp}=\{t\}$, $T_1=T_1/\{t\}$.
4. While $|G_{tmp}| \neq a$.
5. Find t' with minimum value of $dist(t[QI],t'[QI])$ in T_1, and $t[S]$, $t'[S]$ belongs to the same S_t.
6. $G_{tmp}= G_{tmp} \cup \{t'\}$, $T_1=T_1/\{t'\}$.
7. EndWhile.
8. $G= G \cup G_{tmp}$.
9. EndFor.
10. Generalize each group in G..
11. $T'=\{\}$.
12. For each group G' in G.
13. $G_{tmp}=\{\}$, fetch a record t from G'.
14. While $|G_{tmp}| \neq d$.
15. find G_{min} in G, such that for the first record t' in G_{min}, $dist(t[QI],t'[QI])$ is minimum, and t[S], t'[S] belong to different S_t.
16. $G_{tmp}=G_{tmp} \cup G_{min}$, $G=G/G_{min}$.
17. EndWhile.
18. $T'= T' \cup G_{tmp}$.
19. EndFor.
20. Generalize each group in T'.
21. Output T'.

In step 1, it takes $O(n)$ time to go through the records in order to divide them into groups. Step 2 generates a-Similarity blocks, and under the worst condition, for each record, it searches the rest n-1 records for a-1 times. The time used will not exceed $O(n^2)$. The procedure in step 3 is similar to step 2, and the scale is smaller than step 2. Step 2 is the most time consuming and therefore the time complexity of the algorithm is $O(n^2)$.

5. Experiments

In our experiments, we use the algorithm described in section 4, and it's implemented in Java with data stored in MS SQL Server. All experiments are run under Windows XP on a machine with a 1.6GHz Core Duel processor T2300E and 1GB RAM.

We use the Adult Database from the UCI Machine Learning Repository [6]. The database contains 32,561 records from US Census data. According to the hierarchy described in [7], we test 8 attributes from the data. Table 4 is a description of the 8 attributes. Occupation will easily reveal the income which is considered private; therefore occupation is viewed as sensitive attribute. For the other attributes, we pick out 3 to 7 attributes from them to compose quasi-identifiers in each experiment.

Table 4. Attributes description

	Attribute	Distinct values	Generalization type
1	Age	73	Continuous
2	Gender	2	Suppression(1)
3	Race	5	Suppression(1)
4	Marital Status	4	Hierarchy(2)
5	Education	16	Hierarchy(3)
6	Native Country	41	Hierarchy(2)
7	Work Class	7	Hierarchy(2)
8	Occupation	14	Sensitive Attribute

According to the performance and behavior of people with different occupation, we divided sensitive attribute values into 3 groups shown in Table 5.

Table 5. Sensitive attribute groups

Category	Values
1	Adm-clerical, Exec-managerial, Prof-specialty, Sales, Tech-support
2	Armed-Forces, Protective-serv, Other-service
3	Craft-repair, Farming-fishing, Handlers-cleaners, Machine-op-inspct, Priv-house-serv, Transport-moving

We assign a weight to each attribute in quasi-identifier, and the weight varies when the size of quasi-identifier changes. Table 6 shows the weight assigned in different phases in our experiment.

Table 6 Weights of attributes

	3	4	5	6	7
Age	0.2	0.4	0.05	0.1	0.05
Gender	0.7	0.15	0.35	0.3	0.1
Race	0.1	0.15	0.05	0.05	0.02
Marital Status	×	0.3	0.1	0.05	0.12
Education	×	×	0.45	0.1	0.08
Native Country	×	×	×	0.4	0.03
Work Class	×	×	×	×	0.6

The size of quasi-identifier varies from 3 to 7, and the chart below shows the time used to achieve the (*a*, *d*)-Diversity table.

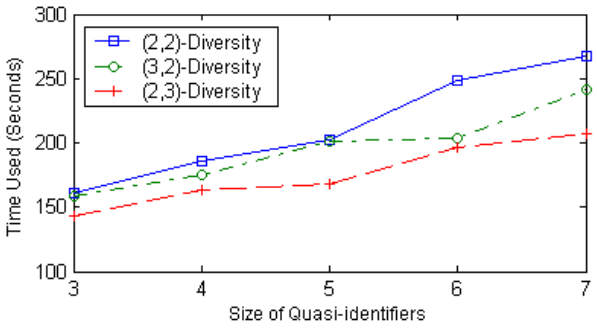

Figure 1. Time used

6. Conclusion

While privacy protection becomes more and more important, various solutions have been proposed. *k*-Anonymity well protects privacy against identity disclosure; however it doesn't protect it against attribute disclosure. *l*-Diversity and (*α*,*k*)-Anonymity are proposed to protect privacy from attribute disclosure. They require an equivalent class to contain several well-represented values for each sensitive attribute. *t*-Closeness goes deeper for the distribution of the sensitive attribute values to be close to the distribution in the whole table. (*ε*,*m*)-Anonymity requires that for every sensitive value *x* in an equivalent class *E*, at the most 1/*m* of the tuples in *E* can have sensitive values "similar" to *x*, where the similarity is controlled by *ε*. However, they can not provide sufficient protection against the disclosure. We propose a principle called (*a*, *d*)-Diversity based on *l*-Diversity and take the real meaning of sensitive attribute and relations among the sensitive attribute values into consideration, since background knowledge is always based on the relationship among the attributes. The advantage of the principle is that it concerns more about the real meaning of the sensitive attributes which the attackers concern the most.

7. References

[1] Latanya Sweeney. k-anonymity : a model for protecting privacy. International Journal of Uncertainty, 2002, pp. 557–570.

[2] PAshwin Machanavajjhala, Daniel Kifer, PJohannes Gehrke, Muthuramakrishnan Venkitasubramaniam. *l*-diversity : Privacy beyond k-anonymity. Proc of the International Conference on Data Engineer, Atlanta, 2007.

[3] Raymond Chi-Wing Wong, Jiuyong Li, Ada Wai-Chee Fu, Ke Wang. (α,k)-Anonymity: An Enhanced k-Anonymity Model for Privacy Preserving Data Publishing, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, 2006.

[4] Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian. *t*-Closeness: Privacy Beyond k-Anonymity and l-Diversity. IEEE 23rd International Conference on Data Engineering, 2007.

[5] Jiexing Li, Yufei Tao, Xiaokui Xiao. Preservation of Proximity Privacy in Publishing Numerical Sensitive Data, Proceedings of the 2008 ACM SIGMOD international conference on Management of data, Vancouver, 2008.

[6] U.C.I Machine Learning Repository. http://archive.ics.uci.edu/ml/.

[7] Vijay S. Iyengar. Transforming data to satisfy privacy constraints. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, 2002, pp. 279-288.