SBTAM – Balance between Privacy and Utility using k-anonymity

Prashant Gawande
HOD, Computer Dept
Sandip Polytechnic, Nashik
prashantvsl@gmail.com

ABSTRACT

p-sensitive k-anonymity and (p,α)S-sensitive k-anonymity model has been recently defined as a good privacy measure using k-anonymity. This new property requires that there be at least p distinct categories or values for each sensitive attribute within the equivalent class which are sharing a set of quasi-identifier attributes. In this paper, we identify the situations when the p-sensitive k-anonymity property is not enough for the sensitive attributes protection and (p,α)S-sensitive k-anonymity does not provide a balance between privacy and data utility. To overcome these shortcomings, we propose new enhanced privacy and utility requirements. SBTAM - Sensitivity Based Tuple Anonymity Method is one of the easy and efficient techniques to achieve privacy preserving for sensitive data. In this method first we consider the sensitivity of values in sensitive attribute and then only tuples having sensitive values are generalized, and the other tuples can be directly published. Experimental results on the Adult Database show the proposed methods not only can improve the privacy but also preserve the utility of the publishing data.

Keywords
Privacy preserving, k-Anonymity, Attribute disclosure sensitive tuples, Algorithms, Experiments.

1. INTRODUCTION

Many electronic databases are available today in the society, like medical patient data, census data, media related data and data collected by different government agencies. They would like to publish the data for the research purposes using data mining. When these organizations publish data it also contains a lot of sensitive information, so they would like to preserve the privacy of the individuals represented in the data[1]. But the problem with data mining output is that, it also reveals some information which is considered to be private and sensitive, so the privacy is becoming very important in many data mining applications.[2]. In order to avoid linking attacks using quasi-identifiers, Sweeney [2] proposed the k-anonymity model, where some of the quasi-identifier fields are suppressed or generalized. A table satisfies k-anonymity if every record in the table is indistinguishable from at least k-1 other records with respect to every set of quasi-identifier attributes. Such a table is called a k-anonymous table. Hence, for every combination of values of the quasi-identifiers in the k-anonymous table, there are at least k records that share those values. This ensures that individuals cannot be uniquely identified by linking attacks.

K-anonymity is the method that manipulates data in order to protect the privacy of the individual. In this method the attributes are classified into three types, key attribute, Quasi-identifier and sensitive attribute. Key attributes such as names, addresses, mobile number or pan card number which are unique are generally removed at time of release. Released information often contains other data called as Quasi-identifiers such as age, sex, and ZIP code[2], which can be linked to publicly available information to re-identify the individual, thus leaking information that was not intended for disclosure. Third class is sensitive attribute which is used by researcher and generally published directly.

The architecture of k-anonymity protection model against identity disclosure is as shown in figure 1.

Figure 1 Architecture of k-anonymity protection model

Suppose Table I is any raw data set which is to be released.

<table>
<thead>
<tr>
<th>NAME</th>
<th>ZIPCODE</th>
<th>AGE</th>
<th>SEX</th>
<th>DISEASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>423065</td>
<td>29</td>
<td>M</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>Smith</td>
<td>422036</td>
<td>32</td>
<td>F</td>
<td>Flu</td>
</tr>
<tr>
<td>Bob</td>
<td>423245</td>
<td>38</td>
<td>M</td>
<td>Cancer</td>
</tr>
<tr>
<td>Sachin</td>
<td>422035</td>
<td>37</td>
<td>F</td>
<td>HIV</td>
</tr>
<tr>
<td>Rama</td>
<td>423012</td>
<td>47</td>
<td>M</td>
<td>Headache</td>
</tr>
<tr>
<td>Sangita</td>
<td>423432</td>
<td>53</td>
<td>F</td>
<td>Viral</td>
</tr>
</tbody>
</table>

First step is to classify the attributes as shown in figure 1.

Table II Classification of Attributes for k-anonymity

<table>
<thead>
<tr>
<th>Key attribute</th>
<th>Quasi identifier</th>
<th>Sensitive attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Sex</td>
<td>Age</td>
</tr>
</tbody>
</table>
Then apply generalization on Quazi identifier using Domain and Value generalization hierarchies as shown in figure 2.

\[
S_1 = \{*, *\}
\]
\[
S_0 = \{\text{Male, Female}\}
\]

Figure 2 DGH and VGH of Sex Attribute

Table III shows a 2-anonymous view corresponding to Table I. The sensitive attributes (Disease) is retained without change in this example.

Table III 2-Anonymous view of Table II

<table>
<thead>
<tr>
<th>ID</th>
<th>ZIP CODE</th>
<th>AGE</th>
<th>SEX</th>
<th>DISEASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>423***</td>
<td>&gt;25</td>
<td>M</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>2</td>
<td>423***</td>
<td>&gt;25</td>
<td>M</td>
<td>Cancer</td>
</tr>
<tr>
<td>3</td>
<td>422***</td>
<td>3*</td>
<td>F</td>
<td>Flu</td>
</tr>
<tr>
<td>4</td>
<td>422***</td>
<td>3*</td>
<td>F</td>
<td>HIV</td>
</tr>
<tr>
<td>5</td>
<td>423***</td>
<td>&gt;40 *</td>
<td></td>
<td>Headache</td>
</tr>
<tr>
<td>6</td>
<td>423***</td>
<td>&gt;40 *</td>
<td></td>
<td>Viral</td>
</tr>
</tbody>
</table>

1.1 Motivation

A k-anonymity model provides protection against identity disclosure but cannot provide a safeguard against attribute disclosure in all cases. A simple case of attribute disclosure will be when all the sensitive attributes within an equivalence class have same value.

The two tables show the original and anonymous version of the dataset. In the second table we have 3 equivalent classes. We have achieved 3-anonymity by generalization. The Disease attribute is sensitive. Let us assume that the attacker can get from public information Sam’s age, say 39 and Zip code, say 27688. Suppose XYZ person also knows that Sam’s record is among one of the records in the releasing table. From second table, attacker can figure out that Sam’s record is from first equivalence class and can thus conclude that he is suffering from Heart Cancer. This attack called as homogeneity attack, which predicts the sensitive attribute because the sensitive values in an equivalence class lack diversity.

Attribute disclosure occurs when the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data release. Several models such as p-sensitive k-anonymity [12], l-diversity [5] and t-closeness [3] are introduced to overcome from this attack.

**Limitation of p-sensitive k-anonymity:** The purpose of p-sensitive k-anonymity [12],[13] is to protect against attribute disclosure by requiring that there should be at least p different values for each sensitive attribute within the records sharing a combination of quasi-identifier. It considers only different sensitive values per equivalent class but does not consider the sensitive category per equivalent class.

![Figure 4 Similar Sensitive Category attack](image)

As shown in figure 4 when attacker link the sangita’s record with anonymized table, it is linked with the first equivalent class. It is having two sensitive values so no issue of homogeneity attack but as the both sensitive values belongs to top secret group, attacker can figure out that sangita is suffering from a top secret disease.

\((p,\alpha)\) sensitive k-anonymity [13] were proposed to overcome from this shortcoming. In this model the sensitive values are categories into m groups and weight is given to each group. Then the modified table \(T^*\) will satisfies \((p,\alpha)\)-sensitive k-anonymity property if it satisfies k-anonymity, and for each QI-group in \(T^*\), the number of distinct categories for each sensitive attribute is at least p with its total weight greater than or equal to \(\alpha\).

**Limitations of \((p,\alpha)\) sensitive k-anonymity:** Suppose disease is sensitive attribute and it is categorize i.4 category as shown in table IV and weight to S(1)=0,S(2)=0.5 and S(3)=1. Then \((2,1)\) sensitive 3 anonymous table is as shown in table V.

![Figure 3 Homogeneity Attack](image)
This model overcomes the problem of p-sensitive k anonymity model. But the main problem with this model is that

- Optimal solution is an NP hard problem. [13]
- All tuples are involved in the generalization process even if it is non sensitive and so the data loss is more and utility is the major issue with this model.

2. RELATED WORK
In recent years, numerous algorithms have been proposed for implementing k-anonymity via generalization and suppression. Samarati [4] presents an algorithm which uses a binary search on the domain generalization hierarchy to find minimal k-anonymous table. Model such as l-diversity proposed in 2006 by A. Machanavajjhala [5] solve k-anonymity problem. It puts constraints on minimum number of distinct values seen within a equivalence class for any sensitive attribute means the sensitive attribute should be l diversified in every equivalent class, S. Venkatasubramanian in 2007 [3] present a t-closeness model was introduced to overcome attacks possible on l-diversity like similarity attack. This model considers the semantic meaning of sensitive values and also the ratio of distribution of sensitive values in equivalent class and in table. But this model having some shortcomings like the data utility is hampered. R. Wong, J. Li, A. Fu, K. Wang [7] propose an (a, k)-anonymity model to protect both identifications and relationships to sensitive information in data were proposed in the literature in order to deal with the problem of k-anonymity. Bayardo and Agrawal [9] present an optimal algorithm that starts from a fully generalized table and specializes the dataset in a minimal k-anonymous table. Fung et al. [8] present a top-down approach to make a table satisfied k-anonymous LeFevre et al. [6] describes an algorithm that uses a bottom-up technique also finds the generalization lattice of QI and how to find a node in lattice which give minimum loss while satisfying anonymity. Pei[10] discuss the approaches for multiple constraints and incremental updates in k-anonymity.

3. SBTAM – A NEW MODEL
The objective of proposed model is to provide with Minimum information loss so that the data is utilized for data mining purpose. The traditional k-anonymity models take all tuples in publishing table T as sensitive tuples. So they are to be generalized and the publishing data lost a lot of useful information. In this proposed method, firstly an algorithm called Sensitivity Based Tuple Anonymity Method (SBTAM) is to be presented. The kernel idea is to protect individuals’ privacy as well as only the high sensitive tuples should be generalized with a satisfied parameter k. The other tuples should not be generalized and can be published directly.[14][15]

Basic Notation. Let PT\{A1,A2,..,An\} be a any table. Let Ai can be K,QI or S. Where K is key attribute which is to be removed at time of release. Let QI denote the quasi-identifier specified by the application (administrator). Let S denote the sensitive attribute. A sensitive attribute is an attribute whose value must be kept secret from people who have no direct access to the original data. Let t[S] denote the value of attribute x for tuple t. \{PT\} denote the number records or size of table PT.

Definition 1. (Quasi-identifier): A set of non-sensitive attributes \{Q1, ...,Qp\} of table T, if these attributes can be linked with external data will uniquely identify at least one individual in the general population.

Definition 2. (k-Anonymity): Let RT(A1,...,An) be a table and QI\{r\} be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in RT|QI\{r\} appears with at least k occurrences in RT|QI\{r\}.

Definition 3. (Sensitive-values Set): A Set M consists of values which the user selected as Top secret sensitive values from set S. Denote by M. [15]

Definition 4. (Sensitive tuple): Let t \in T, if t[S] \notin M, we called t is a sensitive tuple.

Definition 5 (Global recoding): The act of generalizing an attribute in PT, through all the tuples, to the same level in the respective generalization hierarchy of that attributes.

Definition 6 (Equivalent Class/QI-group): A Equivalent class consists of several subsets of T, such that each QI group in T belongs to exactly one subset. We refer to these subsets as Equivalent classes, and denote them as EC1, EC2, ..., ECM.

3.1 Model and Algorithm
Algorithm for Sensitivity Based Tuple Anonymity Method (SBTAM):

SBTAM Algorithm

| Input – Table T, set of Quazi-identifier QI, Sensitive Attribute S, Anonymity parameter k, Categories of Sensitive Values M and L. |
| Output – Anonymized table T* |

Step1: Select Input table and QI set of quasi-identifier attributes

Step2: Select sensitive attribute S.

Step3: Classify sensitive values in two classes, Sensitive as M and Non-sensitive as L.

Step4: For each tuple whose sensitive value belongs to set M i.e. if t[S] \in M then move all these tuples to Table T1, and apply generalization on Quazi attribute so that tuples get anonymized

Step 5: If t[S] \in L then move all these tuples to Table T2

Step 6: Apply Generalization on tuples of Table T1.

Step 7: Ensure k-anonymity of T1

Step 8: Check for Utility

Step 9: Append rows of table T1,T2.

T*=T1+T2 which is table ready to release.

Step 10: Check Table T* for Linking Attack.

Step 11: Release Table T*

The proposed output of this method is like below shown in Table VI. Only tuples containing sensitive values from set A is generalized rest all are published directly so information loss is reduced and so data is utilized more precisely

<p>| Table VI Sensitivity based tuple anonymization |</p>
<table>
<thead>
<tr>
<th>Zip Code</th>
<th>Age</th>
<th>Sex</th>
<th>Diagnostic Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>423061</td>
<td>27</td>
<td>M</td>
<td>Flu</td>
</tr>
<tr>
<td>453063</td>
<td>26</td>
<td>F</td>
<td>Headache</td>
</tr>
<tr>
<td>453052</td>
<td>32</td>
<td>F</td>
<td>Skin Infection</td>
</tr>
<tr>
<td>493051</td>
<td>36</td>
<td>F</td>
<td>Heart Disease</td>
</tr>
<tr>
<td>41305*</td>
<td>4*</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>41305*</td>
<td>4*</td>
<td>*</td>
<td>HIV</td>
</tr>
</tbody>
</table>
3.2 Mathematical Model of SBTAM

\( S = \{ I, O, \varphi, C \} \)

Where,

- \( I \) is set of inputs
- \( O \) is set of Outputs
- \( \varphi \) is set of constraint
- \( C \) is constraint = Size of \( |QI| = 2 \) and \( |k| = 2 \)
- \( I = \{ T, QI, DGH_{hi}, S, H, L \} \)

Where,

- \( T \) - is input table which is to be anonymized.
- \( QI \) - is set of attributes which are involved in anonymization process and are to be generalized
- \( K \) - is anonymity parameter which is generally an
- \( S \) - is sensitive attribute.
- \( H \) - Highly sensitive values from S
- \( L \) - Non or less sensitive values from S

Such that

\( H \cup L = S \), where \( H \in S \) and \( L \in S \) and \( H \cap L = \emptyset \)

For all tuple \( t \in T \) \( \forall i=I,...,n \), if \( t(S) \in H \) then apply generalization using \( DGH \) until \( k \) is not satisfied.

\( DGH_{hi} \) is generalization hierarchy of each \( QI \) attribute \( A_i \) \( \forall i=2,...,n \). Given an attribute \( A \), then a generalization for an attribute is a function on \( A \). That is, each \( f : A \rightarrow B \) is a generalization. Can also represent as

\[ A_0 \rightarrow A_1 \rightarrow \ldots \rightarrow A_n \]  

\( DGH_4 \) for \( A \) as a set of functions \( f_{i} : h=0,...,n-1 \). \( DGH_4 \) is over

\[ \frac{\sum_{i=0}^{n} A_{h}}{n} \]

A generalization function on tuple \( t \) with respect to \( A \) starting with level zero where all values are as it is to level \( n \) where all values are suppressed

\( O = \{ T^*, Prec, Cavg \} \) is output

Where,

- \( T^* \) is \( k \) minimal anonymized table ready to release to data miner
- \( Prec \) [2] is precision which is utility metric for calculating information loss after anonymity

\[ Prec = 1 - \frac{\sum_{i=1}^{n} |DGH_{ai}|}{N^p} \]

Where,

- \( N \) - Total no of records involved in the generalization
- \( p \) - Total no of attributes involved in generalization
- \( h \) - Height of generalization used
- \( |DGH_{ai}| \) - Size os generalization for attribute \( Ai \)

\( Cavg \) is - Normalized average \( QI \)-group.

\[ Cavg = \frac{\text{Total No of records}}{\text{No of Equivalents classes}} \]

4. EXPERIMENTAL RESULTS

This method is computed on the Adult Database from the UCI Machine Learning Repository [11]. The Adult Database contains 32561 tuples from US Census data. After preprocessing data and removing tuples containing missing values 30162 tuples are selected. This database contains 11 attributes from that only 9 attributes are used. From that 8 attributes are consider as quazi-identifier and one attribute ‘occupation’ as a sensitive attribute. Table VII provides a brief description of the data including the attributes used in method, the number of distinct values for each attribute, the type of generalization that was used for Quazi identifier attributes and the height of the generalization hierarchy for each attribute.
Table VII Description of Adults Data Set

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Type</th>
<th>Distinct values</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Numeric</td>
<td>72</td>
<td>3</td>
</tr>
<tr>
<td>Sex</td>
<td>Categorical</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Workclass</td>
<td>Categorical</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Education</td>
<td>Categorical</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>Marital_status</td>
<td>Categorical</td>
<td>7</td>
<td>3</td>
</tr>
<tr>
<td>Income</td>
<td>Categorical</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Race</td>
<td>Categorical</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Native_Country</td>
<td>Categorical</td>
<td>41</td>
<td>3</td>
</tr>
<tr>
<td>Occupation</td>
<td>Sensitive</td>
<td>14</td>
<td>-</td>
</tr>
</tbody>
</table>

Sensitive values from occupation are considered as Adm-clerical, Priv-house-serv, Sales and Transport-moving the total rows are 7448 for experimental purpose. Fig. 6 shows the precision of publishing data varies along with QI when anonymity parameter \( k \) is 2 using equation (1) and fig 7 shows the normalized average QI group size varying along with \( k \) when no of QI attributes are 2 using equation(2).

From fig 6 and 7 it is clear that the traditional k-anonymity algorithm and \((p,\alpha)\)sensitive k anonymity yields more information loss because all tuples are involved in generalized. Only sensitive tuples are generalized in SBTAM. Also running time is less as shown in fig 8.

5. CONCLUSION AND FUTURE WORK

p-sensitive k-anonymity is a novel property that can help to increase the privacy of the respondents whose data is being used but this property is not enough for protecting sensitive attributes. \((p,\alpha)\)sensitive k-anonymity model enhance the previous p-sensitive k-anonymity model but compromising with data utility and precision. Our experimental results show that our proposed model SBTAM have advantages in terms of effectiveness, efficiency and data utility. As we stated before, both privacy principles have their own limitations, and we could adopt this new introduced method to overcome from these limitations and to achieve a better balance between data utility and privacy level.

6. REFERENCES