Enhancing Informativeness in Data Publishing while Preserving Privacy using Coalitional Game Theory

Srinivasa L. Chakravarthy
Dept. of Computer Science and Systems Engineering
Andhra University
Andhra Pradesh, India 530-003
Email: chakri.ls@gmail.com

V. Valli Kumari
Dept. of Computer Science and Systems Engineering
Andhra University
Andhra Pradesh, India 530-003
Email: vallikutami@gmail.com

Abstract—k-Anonymity is one of the most popular conventional techniques for protecting the privacy of an individual. The shortcomings in the process of achieving k-Anonymity are presented and addressed by using Coalitional Game Theory (CGT) [1] and Concept Hierarchy Tree (CHT). The existing system considers information loss as a control parameter and provides anonymity level (k) as output. This paper proposes techniques to improve informativeness in preserving privacy using Coalitional Game Theory (CGT)

Here, each tuple is assigned a Weighted Pay Off (WPO) using CHT. This process is controlled by information loss parameter and provides better anonymization with improved informativeness. Empirical results illustrating the advantages of proposed scheme is discussed.

Keywords—Privacy Preserving Data Publishing, Anonymization, k-Anonymity, Co-Operative Game Theory, Weighted Concept Hierarchy Trees

I. INTRODUCTION

Organisations and agencies often release or publish their data for mining and research purposes. If individuals can be uniquely identified by this action then their personal information like age, disease, salary etc, would be disclosed. These issues lead to privacy concerns and there is a growing need for better privacy mechanisms to protect the privacy of individuals for different domains like Social Networks, Micro data releases etc. Just by removing uniquely identifying information of an individual like NAME, SSN etc., privacy may not be protected. The individuals can still be identified when published data is linked to external data. There are different types of techniques like k-anonymity [2], l-diversity [3] etc., (see [4] for some more mechanisms) proposed for data base releases temporally, but still there are issues in these methodologies in the process to achieve privacy [5].

Generally for these mechanisms different parameters k, l etc., are considered. For example, for k-anonymity, we fix the k value before anonymization. Information loss can be calculated only after the anonymization is done completely. [1] presented Coalitional Game Theoretic methodology to find k based on the affordable information loss value given. This paper also adopts the same methodology and includes Weighted Concept Hierarchy Tree (WCHT) to increase the informativeness in published data.

Usually attributes of a data are classified into four categories for Preserving Privacy of data table [6]. They are:

- Explicit Identifiers (EID) is the set of attributes which contain information that explicitly identifies a person and his sensitive information. Ex: NAME, SSN.
- Quasi Identifiers (QID) is a set of attributes that potentially identify individual’s information when linked with other external data. Ex: AGE, ZIP CODE.
- Sensitive Attributes (SA) can hold a person’s sensitive information. Ex: DISEASE, SALARY.
- Non-sensitive Attributes are other than above three.

k-anonymity principle is introduced by [2] [7] to overcome the issue of linking attack. It works on QID attribute set and the values under the QID are replaced with its more general values, such that the generalized values of each tuple are made indiscernible to at least k-1 other tuple in published data table. For example, if we consider QID = \{SEX, AGE, JOB\} then Table II is a 3-anonymized version of Table I

<table>
<thead>
<tr>
<th>Table I: Original Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job</td>
</tr>
<tr>
<td>Lawyer</td>
</tr>
<tr>
<td>Engineer</td>
</tr>
<tr>
<td>Doctor</td>
</tr>
<tr>
<td>Writer</td>
</tr>
<tr>
<td>Singer</td>
</tr>
<tr>
<td>Dancer</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table II: 3 - Anonymized Table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Job</td>
</tr>
<tr>
<td>Professional</td>
</tr>
<tr>
<td>Professional</td>
</tr>
<tr>
<td>Professional</td>
</tr>
<tr>
<td>Artist</td>
</tr>
<tr>
<td>Artist</td>
</tr>
<tr>
<td>Artist</td>
</tr>
</tbody>
</table>

This conventional approach for k-anonymity has some limitations and they are presented in [1]. It gives priority to
anonymization against information loss. This leads to over-
anonymization of the data in turn leading to high computa-
tional effort, as a result of which, we are bound to compromise
with more information loss. Secondly, data user is proposes
the value of \( k \), for \( k \)-anonymization, without proper knowledge
about \( k \). Thirdly, the information loss is computed only after
anonymization is done and finally, if the information loss
is more than what is expected by the data user the whole
anonymization process has to be repeated with new value of
\( k \).

[1] adopted Concept Hierarchy Trees (CHTs) of QID set,
for their methodology. When this scheme is used for gen-
eralization, the generalization near the root of the CHT leads
to more distortion as compared with the generalization far
from the root. Hence we see more information loss after
anonymization. This is the motivation of the present work. This
paper proposes the use of Concept Hierarchy Tree (CHT) with
weighted levels, which in turn reduces the information loss.

II. RELATED WORK

Privacy protection of an individual has been emphasized
seriously in the past. According to [8] Statistical commu-

nity proposed methods like randomization to resolve the re-
identification of the individuals on statistical databases which
were used in fraud detection. This added noise to the data
before the data release. However, these methods failed in
providing an effective solution for anonymity that lead to the
failure of data integrity [9].

[2] has proposed the notion of \( k \)-anonymity to protect the
privacy in publishing the data. Even though the idea of \( k
\)-anonymity is straightforward conceptually, but the compu-
tational complexity of finding optimal solution is NP-hard
[10]. A number of considerable algorithms have been proposed
[11] [12] [13] [14] [7]. These algorithms employ suppression
or generalization frameworks to replace the QID-values with
less specific ones to their generalized values such a way that
after anonymization every QID-group must contain at least \( k
\)-tuples having the same QID-values. These solutions, however,
suffer from high cost of information loss due to pre-defined
generalization hierarchies. To aid the \( k \)-anonymity new notions
like \( l \)-diversity [3], \( \ell \)-closeness [15], \( (\alpha, k) \)-anonymity [16]were
proposed to enhance the privacy protection mechanism.

A more general view of \( k \)-anonymization is clustering with
a constraint of minimum number of objects in every cluster
[17] [18]. A number of methods use identity protection by
custering, specially, numerical attributes [10] [19]. For ordinal
attributes, clustering based method is presented [20], but it does
not deal with attributes in hierarchical structure.

Information loss of an anonymized table can be considered
as similarity between the original data and privacy protected
data. In the anonymization process all the existing methods
provide data privacy at the cost of losing some information in
the data. A number of methods are proposed for calculating the
tradeoff between privacy and utility. In some of these methods
the tradeoff is calculated during anonymization process while
on the other side this can be done after completion of
the process.

[21] proposed average size of equivalence classes and
discernibility metric respectively, to calculate the utility of
an anonymized data set which takes equivalence class size
into account in \( k \)-anonymity. Different measures of utility
such as information-gain-privacy-loss ratio [5], clustering and
partitioning based measure, and risk return trade-off [15] have
been proposed to find the next generalization step in the
anonymization algorithms. [22] developed a framework for data
mining, which considers trade-off between utility measure and
privacy. But, it measures the trade-off after anonymization
process has been completed.

In the course of calculating the trade-off between utility
measure and privacy level, game theory is one of the methodol-
dy. [14] explains how the Game theory is applied and analyzes
the privacy in legal issues. In economical perspective [23]
theory techniques to explore the flow of customer’s private
information between two interested firms.

[27] proposed differential privacy using mechanism design
methodology of Game theory. In the context of recommender
systems [28] defines an accuracy metric for differential privacy
which analyzes the trade-off between privacy and accuracy.
[24] described three scenarios modeled as coalitional games
[29] and the reward allocation exchange of private information
is done based on core and shapely values. [30] proposed coal-
tional game theory mechanism to achieve \( k \)-anonymization for
a data set. The results have been proved based on the CHT of
attributes which includes high distortion. This paper is an effort
to overcome the shortcomings by modifying and including
Concept Hierarchy Trees with weighted levels.

III. PRELIMINARIES

\( k \)-anonymization’s main objective is to make each data
tuple of sensitive attributes in a published table indiscernible to
at least \( k-1 \) other tuple and as a result no sensitive information
can be easily inferred. This section presents formal definitions
to achieve \( k \)-anonymity. Let \( T \) be a data table, \( A_i, 1 \leq i \leq m \)
be the attributes of the table and \( D_{A_i} \) domain values of each
attribute. QID is the set of attributes which can potentially
identify an individual’s information. Typical examples include
Age and Zip code. In the process of achieving anonymization
of the given data table, generalizations of the domain values
of QID set of attributes, are adopted.

A. Concept Hierarchy Trees

1) Generalization Relationship: If QID is a quasi-
identifier and \( D_{QID_i} \) the set of all distinct possible domain
values of QID table of T, then a relationship \(<_{QID_i} \) between
\( D_{QID_i} \) and \( D'_{QID_i} \) is said to be generalization relationship
if it is a partial order i.e. \( D_{QID_i}, <_{QID_i}, D'_{QID_i} \). Here \( D_{QID_i} \)
is the set of generalized domain values of the attribute \( QID_i \).
For example, from Figure 3, Professional is a generalized value
and Any is the more generalized value of \{ Lawyer, Engineer,
Doctor \}.

2) Definition (Generalization Hierarchy): Let \( V_{QID_i} \) be
the collection of all attribute values satisfying the relationship
\(<_{QID_i} \) corresponding to \( QID_i \). The generalization hierarchy
is used to organize the values of \( V_{QID_i} \) as follows [1].

- If there are one or more values of \( V_{QID_i} \) which
makes it impossible that \( V_{QID_i} <_{QID_i} V_{QID_i} \)

IJCTA | May-June 2014
Available online@www.ijcta.com
ISSN:2229-6093

844


844
may hold on for any $V_{QID_i}(q <> p)$, where $V_{QID_i}, V_{QID_q} \subseteq V_{QID}$, then this kind of value is represented as leaf node in generalization hierarchy.

- If there is no value $V_{QID_i}$, which makes it impossible that $V_{QID_i} <_{QID_i} V_{QID_q}$ may hold on for any $V_{QID_i}(q <> r)$, where $V_{QID_i}, V_{QID_q} \subseteq V_{QID}$, then this kind of value is represented as root node in the hierarchy.

- If the values satisfy the relation $V_{QID_i}(q) <_{QID_i} V_{QID_q}$, then $V_{QID_i}$ always exists below $V_{QID_q}$ level.

In generalization process the specific values are replaced by its more general values in taxonomy tree of an attribute. For example, Figure 1. Person is a more general value than the child nodes {Male, Female}. Figure 1 2 3 represents generalization hierarchy trees of the attributes ‘SEX’, ‘AGE’, ‘JOB’ respectively.

B. k-anonymity

Let $T(A_i), 1 \leq i \leq m$ be a table associated with quasi identifier $QID$ set. The $QID$ set of attributes which when put together would identify an individual and the number of quasi identifiers of the table $T$ can be taken as cardinality of $QID$ i.e. $|QID|$. $T$ is said to satisfy $k$-anonymity with respect to a particular $QID$ attribute set if and only if each tuple under $QID$ attribute set i.e., t.$[QID]$ is indiscernible with at least $k$ tuples in the table $T$.

For anonymization of the table concept hierarchy trees (CHT) (Generalization Hierarchies) are adopted for three $QID$ attributes SEX, AGE, JOB as shown in Figure 1 2 3. As the given concept is generalized, the number of distinct concepts gets reduced and many tuples become redundant according to their attribute instances. These redundant tuples will be grouped and each such group is called as an equivalence class.

C. Coalitional Game Theory

Privacy is achieved by moving a tuple up along the CHT of each QID attributes. Such a situation can be modeled by using game theory techniques. [31] explains how the game theory strategies help in decision making process in complex situations where two or more agents have conflicting interests. Agents usually maximize or minimize their objective function. A cooperative anonymization game is one where a group of tuples (players) are forced to behave in a cooperative manner. In this process the players collude to maximize privacy and minimize information loss. Coalitional games are applied to achieve these objectives. It is a known fact that there is always a tradeoff between privacy and utility. If privacy increases then utility decreases and vice versa. Game theory is one of the best approaches to find out the tradeoff between them. This section discusses preliminary essentials about Coalition Games. Based on the cooperative game strategy we define a cooperative anonymization game and its building components.

1) Co-operative Anonymization Game: It is assumed that the game $G$ consists of set $P$ of $N$ players, means that $P = \{1, 2, ..., N\}$. A non-empty subset $C$ of $P$ (i.e. $C \subseteq P$) is called coalition of players. A standard condition can be imposed that coalition should contain at least one player and that there is no payoff for null coalition. It is assumed that payoff of the coalition is transferable between the members, i.e. the payoffs of the coalition are freely redistributed among themselves. According to [29] there is a mapping $payoff : 2^N \rightarrow \mathbb{R}$ where, $payoff(C) \in 2^N$ is maximum coalition payoff.

Throughout this paper a coalitional game, $G = (P, payoff)$ with transferable payoff is only considered. ‘payoff’ is called characteristic function of the game. For example, in a three-player majority game the three players can obtain one unit of payoff, any two of them can obtain payoff $\in [0, 1]$ independent of the actions of the third, and each player can alone obtain nothing independent of the actions of the remaining two players. This scenario can be modelled with CGT [32].

2) Convex Games: A co-ordinate game $G = (P, payoff)$ is known as Convex Co-ordinate game when for any two coalitions $C_i, C_j \in 2^N$,

$$payoff(C_i \cup C_j) \geq payoff(C_i) + payoff(C_j) - payoff(C_i \cap C_j)$$ (1)

Example: For the above mentioned three player game we can model a situation as the coalitional game $(N, payoff)$, for which $N = \{1, 2, 3\}$, $payoff(N) = 1$, $v(S) = \alpha \text{ whenever } |S| = 2$, and $v(i) = 0 \forall i \in N$. Here $S$ is subset of $N$ and $payoff(S)$ is payoff of $S$. The core of this game is the set of all nonnegative payoff profiles $(x_1, x_2, x_3)$ for which $x(N) =$
1 and \( x(S) \geq \alpha \) for every two-player coalition \( S \). Hence the core is nonempty if and only if \( \alpha \leq 2/3 \) [32].

A table \( T \) is named as optimally anonymous if it contains minimum information loss as well as it satisfies the given privacy requirement. [33] states that achieving optimal \( k \)-anonymity is very expensive and is NP-hard. But all researchers concentrated on suboptimal \( k \)-anonymity. In this connection Information loss metrics for anonymized data sets are used to calculate information loss measures.

The following sections present in detail the procedure involved in finding the best value for \( k \) and hence minimizing the chance of revealing the identity of the tuple owner to at most \( 1/k \), thus minimizing the information loss and hence maximizing the utility of the published data.

### IV. PROPOSED SYSTEM

Srinivasa et al. [1] has proposed a model to achieve \( k \)-anonymity using coalitional game theory. In this process, to attain \( k \)-anonymization, concept hierarchy tree of each \( QID \) attribute was used. The values of tuples under \( QID \) set 'climb' over the respective concept hierarchy trees in such a way that the information loss of the data set incurred by these tuples will be not more than the given threshold. Then the data will be segregated according to the generalization of the tuples.

The shortcomings of the existing system are listed:

- If the length of the \( CHT \) is two i.e., all possible domain values of \( QID \) are generalized to root in single 'climb' then it incurs more information loss of data set while undergoing anonymization. For example, if \( Gender \) is considered as one of \( QID \) attribute whose possible domain values are \{ Male, Female \}, then in the process there is more information loss. Since the \( CHT \) of \( Gender \) has only two levels for generalization, the possible contributions to information loss are 0.5 and 1. It may cause more information loss.

- The generalization near the top of \( CHT \) gives more information loss when compared to generalization far from the top. For example, from Figure 3, climbing from Lawyer to Any gives more information loss than from Lawyer to Professional.

The existed system in [1], considered user defined/ required information loss as control parameter. Anonymization process has been performed based on this control parameter. Finally, anonymization level \( k \) of the input data set will be given as output. Due to the above issues more distortion is incurred in the published data set and as a result less anonymization level is possible. Hence, with less anonymization level there is a possibility of privacy breach.

To minimize the information loss and in turn to enhance the informativeness this paper presents new methodology. In this methodology Concept Hierarchy Trees(\( CHT \)) with weighted levels is adopted to minimize the information loss.

**Definition 1:** Let \( A_i, (1 \leq i \leq m) \) be the attributes of the table and \( D_{A_i} \) be set of all possible domain values of each attribute \( A_i \). Let \( L_i \) be the maximum possible level of \( CHT \) of an attribute \( A_i \) and \( 1, 2, ..., L_i - 1, L_i \) be the domain levels from most specific to most general respectively. In the anonymization process if a tuple value, under a \( QID_i \), is generalized to level \( q \), \( 1 \leq q \leq L_i \) over its associated \( CHT \), then the weighted payoff (\( WPO \)) is given by:

\[
WPO(t/QID) = \prod_{i=1}^{(|QID|)} \frac{\sum_{l=1}^{q} w_l}{\sum_{l=1}^{L_i} w_l}
\]

Here \( w_l \) represents the user defined weighted distances between level \( l \) to \( l+1 \), where \( 1 \leq l \leq L_i \).

#### A. Two Schemes for Weights

Even though user is allowed to give weights to the system, in this section we presented two different schemes. Scheme 1 is based on uniform weights and Scheme 2 is based on level weights.

1) Uniform Weights: \( (w_l = 1, 1 \leq l \leq L_i) \) : In this scheme all weights are adjusted to 1, so insightfully the \( WPO \) of a tuple over the \( QID \) is given by the number of levels being generalized by the tuple value over all possible steps. It is equivalent to the definition for the payoff given in [1]. For example, from table I and table II the first tuple \{Lawyer, Male, 28\}, under \( QID = \{JOB, SEX, AGE\} \), is generalized to \{Profession, Person, [25 - 30]\}. In this process the tuple value Lawyer takes two levels of generalization, so it assumes the payoff 2/3 corresponding to JOB. Similarly under \( SEX \) and \( AGE \) it assumes 2/3, 2/3 respectively. So, according to definition 1 the tuple will be assigned a \( WPO \) value 4/9.

| TABLE III: The WPO values with all possible combinations of \( QID \) using Scheme 1 |
|------------------|---|---|---|---|---|---|
| \( t_1 \) | \( t_2 \) | \( t_3 \) | \( t_4 \) | \( t_5 \) | \( t_6 \) |
| J | S | A | J | S | A | J | A | J | A |
| 2/3 | 2/3 | 2/3 | 2/3 | 2/3 | 2/3 | 5/11 | 4/9 | 4/9 | 4/9 |

2) Level Weights: \( (w_l = \frac{1}{(L_i - l + 1)\alpha}, 1 \leq l \leq L_i, \alpha \geq 1) \) : This scheme provides the weights as above said relation where \( L_i \) is the maximum possible length of \( CHT \) associated with \( QID_i \) and \( l \) is level generalization tree. \( \alpha \) is a positive real value and it is used to regulate the variation between the weights. The intuition of this scheme is that the generalization near the top causes more information loss against the generalization far from the top. Therefore we formulate this scheme such that generalization near the top will be assigned more weight and far from the top will be assigned less weight. For example, if \( \alpha = 1 \) and the generalization for the tuple value Lawyer takes two levels then \( WPO \) is 5/11 corresponding to JOB. Similarly, under \( SEX \) and \( AGE \), it is 1, 5/11 respectively.

So, according to definition 1 the tuple will be assigned a \( WPO \) value 25/121. Table III provides all possible \( WPO \) of Table I.
TABLE IV: The WPO values with all possible combinations of QID using Scheme 2

<table>
<thead>
<tr>
<th>J</th>
<th>S</th>
<th>A</th>
<th>J</th>
<th>S</th>
<th>A</th>
<th>J</th>
<th>A</th>
<th>S'</th>
</tr>
</thead>
<tbody>
<tr>
<td>t₁</td>
<td>5/11</td>
<td>1</td>
<td>5/11</td>
<td>5/11</td>
<td>1</td>
<td>5/11</td>
<td>25/121</td>
<td>25/121</td>
</tr>
<tr>
<td>t₂</td>
<td>5/11</td>
<td>1</td>
<td>5/11</td>
<td>5/11</td>
<td>1</td>
<td>5/11</td>
<td>25/121</td>
<td>25/121</td>
</tr>
<tr>
<td>t₃</td>
<td>5/11</td>
<td>1</td>
<td>5/11</td>
<td>5/11</td>
<td>1</td>
<td>5/11</td>
<td>25/121</td>
<td>25/121</td>
</tr>
<tr>
<td>t₄</td>
<td>5/11</td>
<td>1/3</td>
<td>5/11</td>
<td>5/33</td>
<td>5/33</td>
<td>25/121</td>
<td>25/263</td>
<td></td>
</tr>
<tr>
<td>t₅</td>
<td>5/11</td>
<td>1/3</td>
<td>5/11</td>
<td>5/33</td>
<td>5/33</td>
<td>25/121</td>
<td>25/263</td>
<td></td>
</tr>
<tr>
<td>t₆</td>
<td>5/11</td>
<td>1/3</td>
<td>5/11</td>
<td>5/33</td>
<td>5/33</td>
<td>25/121</td>
<td>25/263</td>
<td></td>
</tr>
</tbody>
</table>

1 (J = JOB, S = SEX, A = AGE)

C. Information Loss

The data user who colllects the data from data collector, typically wants to get more information from the data. When anonymized data set is published, some information is lost due to the algorithm applied over the data. The user needs more qualitative data for his purposes like data mining etc. The quality of k-anonymity of a given data set is typically calculated by the quality loss in process of anonymization. The utilization of the data set after completion of the anonymization is measured using information loss (IL). There are different measures to estimate Information Loss [4]. This paper proposes the following relation using WPO of the tuple to calculate information loss of the table T after anonymization:

\[ IL(T) = \frac{1}{|T|} \sum_{i=1}^{n} \sum_{t \in T} WPO(t/QID_i) \]  

Here, WPO(t/QID_i) is weighted payoff of the tuple t_i under QID_i attribute. CoG_i is coalition group in which t is a member. |CoG_i| is the size of the group and |CoG_j|, the number of such groups. For example, the information loss of the anonymized data table II can be evaluated, using Scheme 1 as (1/6 * 3) [3 * [(2/3) + (2/3) + (2/3)] + 3 * [(2/3) + (1/2) + (2/3)]] = 0.5244, whereas by using Scheme 2 it will be equal to 0.5252. This shows that the Scheme 2 will give less distortion than the Scheme 1.

V. ALGORITHM AND ITS COMPLEXITY

A. Algorithm Complexity

The anonymization procedure is explained in the Anonymizer algorithm. Initially the algorithm considers original data table T, user given information loss as threshold (Th), and concept hierarchy trees (CHT) of specified QID set of the data table as input. For each tuple t_j in the table T for each QID_i attribute, if the information loss does not meet the threshold level then one level is shifted up over the CHT of QID_i. After that we assign weighted payoffs for each tuple with respect to each attribute as shown in step 10. The final payoff of each tuple can be calculated as shown in step 12. Now for each pair of tuples in T, if payoffs are equal and also if instances are equal under each QID_i then the tuples will be considered as an equivalence class (step 16-23). The minimum number of tuples for all equivalence classes k and information loss of anonymized table is calculated according to the definition 1 (step 25-26).

B. Complexity

Let N be the number of records and q be number of quasi identifiers and n_1, n_2, ..., n_q be the distinct number of domain values of each QID. For constructing the concept hierarchy tree for each QID attribute the complexity is O(q.n.logn) and if we consider, without any loss of generality, n is the maximum of n_1, n_2, ..., n_q. The complexity for constructing the concept hierarchy trees for all QID is O(q.n.logn). Now for the payoff calculation of each record under QID set the complexity is O(Nq). The complexity for dividing the data
Algorithm : Anonymizer

Input:
(i) Original Data Table T,
(ii) threshold (Th),
(iii) Concept hierarchy trees (CHT) of QID’s.
Output: (i) Anonymized table (T’) (ii) k value of the table.

Assumptions:
(i) QID = \{A_1, A_2, ..., A_{|QID|}\} // set of all QID
(ii) EQ = Set of equivalence classes such that
\[ T' = \bigcup_{j} EQ_j, \cap_j EQ_j = \emptyset \]
(iii) \[ w_{i,j+1} = 1 \text{ (Scheme 1) (OR)} \frac{1}{(L_i-1)} \text{ (Scheme 2)} \]

Method:
01 Begin
02 For each tuple \( t_j \) in T do
03 Begin
04 For each \( t_j, [A_i] \) in T
05 \( q_j = 1 \)
06 While (infloss \((t_j, [A_i])\) < (Th) 
07 Climb (CHT([A_i]));
// Shift one level over CHT of \( A_i \)
08 \( q_j = q_j + 1 \);
09 End While
10 \[ WPO(t_j, [A_i]) = \prod_{j=1}^{q_j} w_{i,j+1} \sum_{i=1}^{l_{j+1}} w_{i,j+1} \]
11 End For
12 \[ WPO(t_j) = \prod_{i=1}^{|QID|} WPO(t_j, [A_i]) \]
13 End Begin
14 End For
15 For each tuple \( t_j \) in T do
16 Begin
17 \( EQ_j = t_j; \)
18 For each tuple \( t_k \) in T
19 If (WPO\((t_j) = WPO(t_k)\)) \& \&
(\( t_j, [A_1, A_2, ..., A_n] = t_k, [A_1, A_2, ..., A_n] \))
20 \( EQ_j = EQ_j \cup \{ t_k \} \)
21 End If
22 End for
23 End Begin
24 End for
25 \[ k = kfinder(T') \] // Minimum number of tuples among all equivalence classes
26 IL = infloss(T') // Using Equation 3
27 End Begin

The complexity of the algorithm is \( O(q.N^2) \). Then the overall complexity is \( O(q.n.logn) + O(N.q) + O(q.N^2) \approx O(q.N^2) \).

The complexity is evaluated keeping in mind the number of QIDs and size of data set. The constructions of CHT also play a role in the evaluation of cost or complexity. The cost of re-evaluations of CHT’s may be affected by the inclusion of a chunk of data records with values out of range of CHT. We aspire to address this issue in further studies.

VI. EXPERIMENTATION AND EMPIRICAL ANALYSIS

Experiments were performed on Intel Core2 @ 2.93 GHz with 2GB RAM being allocated separately for the Net beans platform for two different data sets. The synthesized data set, for experimentation, is collected from [1]. The data set contains a set of 239 records with three quasi identifiers namely age, sex and job. Disease was taken as the sensitive attribute. This data set with 239 records considered for the experimentation can be grouped into not less than two; so, we have chosen \( k = 5,10,15,...,120 \) (at intervals of 5) and information loss was calculated using these two schemes and then compared. Figure 5 shows that there is a fast growth in information loss in the first half and becomes almost stable in the second half for two schemes, except for a few cases. Experimentation also shows that Scheme 2 gives better results than Scheme 1.

Second step of the experimentation is confined to determining the value of \( k \) for the given information loss. We calculated \( k \) values for Information loss ranging from 0.1 to 1 at intervals of 0.05 and plotted the same on the graph, shown in Figure 6. The progress rate of \( k \) value is slow up to 0.5 of information loss, beyond that the rate increased, which means that the choice of information loss should not be more than 0.5 for obtaining better anonymization. Figure 6 also confirms that Scheme 2 results in more anonymization level \( k \) than Scheme 1. Anonymization level is more for Scheme 2 as it gives less information loss which is the control parameter of the algorithm. So, for given information loss as a threshold, Scheme 2 provides more anonymization level, hence less possibility of privacy breach. It shows the efficiency of Scheme 2 over Scheme 1.

A. Choice of \( \alpha \) Value

As a third step of this experimentation we thrive to find the optimal \( \alpha \). This experimentation has been done over the same data set as described above. Information loss values are calculated for \( \alpha \) ranging from 0 to 10 at intervals of 1. Figure 7, shows relation between \( \alpha \) and information loss. It is observed that as \( \alpha \) increases the information loss value decreases, and for \( \alpha \) more than 5 the information loss values almost remains stable. So, for this data set this value would be optimal value. As a result we can improve the anonymization level of data set which in turn decreases the privacy breach probability.

VII. CONCLUSION AND FUTURE DIRECTION

To overcome shortcomings in \( k \)-anonymization for achieving privacy of a published table two different schemes using Coalitional Game theory are presented. Concept Hierarchy Tree (CHT) methodology is adopted to attain \( k \)-anonymization. This system allows user specified information loss of the published data, and it controls the proposed system. Hence, data user will receive an anonymized data table for better utilization.

The distinct tuples under QID set collate into equivalence classes in such way that the size of equivalence class is at least \( k \). The further scope of the study is to find the optimal way of collating the tuple in the equivalence classes. To improve informativeness in published data two Schemes based on the parameter \( \alpha \) are proposed. The underpinned limitations of Scheme 1 are discovered and resolved in Scheme 2 using Weighted Pay off (WPO). Thus, the information loss obtaining in Scheme 2 is less and hence more anonymization is attained. Experimentation also proves the same. This paper also suggests optimal \( \alpha \) for sample data set.
As future work, instead of static CHTs, we intend to use Dynamic Concept Hierarchy Tree (DCHT). The construction of DCHT for Numerical attributes is somewhat computable compared to categorical attributes. One may use Ontological or NLP methodologies in the construction of DCHT for categorical attributes.

REFERENCES


