Kumari and Chakravarthy *Hum. Cent. Comput. Inf. Sci.* *(2016) 6:12*
DOI 10.1186/s13673-016-0069-y

Human-centric Computing
and Information Sciences

RESEARCH

Open Access

CrossMark

Cooperative privacy game: a novel strategy for preserving privacy in data publishing

Valli Kumari[1] and Srinivasa Chakravarthy[2*]

*Correspondence:
chakri.ls@gmail.com
[2] Department of Computer Science and Engineering, NS Raju Institute of Technology, Sontyam, Vishakapatnam 531 173, India
Full list of author information is available at the end of the article

Abstract

Achieving data privacy before publishing has been becoming an extreme concern of researchers, individuals and service providers. A novel methodology, Cooperative Privacy Game (*CoPG*), has been proposed to achieve data privacy in which Cooperative Game Theory is used to achieve the privacy and is named as Cooperative Privacy (*CoP*). The core idea of *CoP* is to play the best strategy for a player to preserve his privacy by himself which in turn contributes to preserving other players privacy. *CoP* considers each tuple as a player and tuples form coalitions as described in the procedure. The main objective of the *CoP* is to obtain individuals (player) privacy as a goal that is rationally interested in other individuals' (players) privacy. *CoP* is formally defined in terms of Nash equilibria, i.e., all the players are in their best coalition, to achieve *k*-anonymity. The cooperative values of the each tuple are measured using the characteristic function of the *CoPG* to identify the coalitions. As the underlying game is convex; the algorithm is efficient and yields high quality coalition formation with respect to intensity and disperse. The efficiency of anonymization process is calculated using information loss metric. The variations of the information loss with the parameters α (weight factor of nearness) and β (multiplicity) are analyzed and the obtained results are discussed.

Keywords: Privacy preserving data publishing, *k* - anonymity, Cooperative game theory, Shapley value

Background

Privacy concerns are rapidly increasing 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. There are different types of anonymization criterion like *k*-anonymity [1], ℓ-diversity [2] etc., (see [3] for some more mechanisms) proposed for temporally data base releases, however, still there are some issues in these methodologies to achieve privacy [3].

Motivation towards cooperative privacy

In the social network scenarios, the acceptance of an unknown friend request causes providing his personal information as well as his existing friend's information. In other words, the friend who doesn't aware about privacy and if he accepts the friend request,

Springer Open

then it may paves a way towards privacy theft of his private information as well as his friend's information. It is not just enough to preserve our personal privacy, the people circled around us should also take an action. Though many social network sites provide different levels of privacy control, in addition rational cooperation of the people is also necessary.

Domingo-Ferrer initiates epitome of cooperation in privacy and termed it as Co-Privacy [4, 5]. However, CoV (cooperative value) is modeled, that estimates the cooperation between the tuples using Cooperative Game Theory and it is titled as cooperative privacy. The following are the prime motivations towards the cooperative privacy (CoP) [5]:

- *To keep the information society growing on over a period of time, preservation of privacy is necessary* It is just like trying to solve the global issues (e.g. international terrorism, global warming etc.) to sustain the physical world. Now, information society gives importance to preservation of privacy as they understand its significance but are scared of using these services. The people are forced towards privacy preservation in information society, just like the importance given to *Go-Green* and *No Plastic* by the environmentalists in society.
- *As far as possible, privacy should be maintained by the rational cooperation of others, in absence of which the entire information system may be inconsistent* It is similar to the traffic rules. If a person doesn't follow the traffic rules, it causes a trouble to others and some times it may lead to deadlock. Even though the government has scaffold privacy of users as human rights, they still remain quite unrealistic. Just the setting of rules by the government is not enough to achieve privacy preservation, effort should be put by the technology people to enforce the users to maintain privacy world. At the same time there should be a rational cooperation among the users for societal usefulness.

This paper proposes a game named Cooperative Privacy Game (CoPG), using Coalitional Game Theory [6] to find the CoP of a data set which is to be published. In CoPG, each tuple is considered as a player and assigned a real value called cooperative value (CoV), which is formally defined as characteristic function. The CoV of each player in the data table is defined as stated by Shapley value [7] which assumes the compactness around it. CoPG is to cogitate the cooperation between the tuples (players) which is estimated based on the CoV. CoV is used to divide the given data table into groups, each called as coalition. Later, by applying anonymization techniques over these coalitions CoP is achieved in terms of Nash equilibria [6] for k-anonymity [1].

Since the underlying game CoPG is convex [8], the algorithm which is used in formation of coalitions, is efficient and yields high quality with respect to intensity and disperse. Here, intensity is the average distance between the point to the center and disperse is the average distance between point to point. The Shapley value of the characteristic function of the coalitional game is considered in this paper coincides with other solution concepts named Nucleolus, Gately point, τ-value. This was proved by Swapnil et al. [9]. It supports the adoption of the characteristic function, defined in the later section, for this game. Anonymization efficiency is calculated by using information loss metric and the advantages of proposed algorithms are discussed.

Related work

The notion of *k*-anonymity principle to protect privacy before publishing the data has been proposed by [1] Aggarwal [10], Bayardo et al. [11], LeFever et al. [12], Samarati et al. [13] employed and discussed suppression/generalization frameworks to achieve *k*-anonymity. To support the *k*-anonymity, new notions like *l*-diversity [2], *t*-closeness [14], (*α*, *k*)-anonymity [15] were proposed which improve the privacy protection mechanism. Giving these protected data sets to other parties for data mining does not raise the privacy issues but none of the existing methods are able to completely exhaust the risk of privacy protection.

Garg et al. [8] attained pattern clustering, an important methodology in data mining, by using game theory and proposed the use of Shapley value to give a good start to K-means. For clustering, Gupta and Ranganathan [16, 17] used a microeconomic game theoretic approach, which simultaneously optimizes two objectives, viz. compaction and equi-partitioning. Bulo and Pelillo [18] describes hypergraph clustering using evolutionary games. Chun and Hokari [19] proved the coincidence of Nucleolus and Shapley value for queueing problems.

Wang et al. [20] proposed efficient privacy preserving two-factor authentication schemes related to wireless sensor networks [21] presented a methodology using two-factor authentication to overcome the threat of de-synchronization attack of preserving anonymity [22, 23] initiated evaluation metric for anonymous—two factor authentication in distribution systems. Recent study in crime data publishing [24] achieved *k*-anonymity with constrained resources.

Generally, to estimate the trade-off, Game theory is one of the good methodologies. In Privacy Preserving Data Mining (PPDM) game theory is used to estimate the trade-off between utility measure and privacy level. Anderson [25] explains how the Game theory is applied and analyzed the privacy in legal issues. In Economical perspective, Bhome et al. [26], Kleinberg [27], Calzolari et al. [28], Preibusch [29] present many privacy issues. Calzolari [28] uses game theory techniques to explore the flow of customer's private information between two interested firms. Dwork [30] proposed differential privacy using mechanism design methodology of game theory. In the context of recommender systems Machanavajjhala [31] defines an accuracy metric for differential privacy which analyzes the trade-off between privacy and accuracy.

Kleinberg et al. [27] described three scenarios modelled as Coalitional Games (introduced in Osborne [32]) and the reward allocation exchange of private information is done according to the core and Shapley values. Chakravarthy et al. [33–35] described coalitional game theory mechanism to achieve k-anonymization for a data set.

Preliminaries

This section outlines the information available in literature for *k*-anonymity and concise information about coalitional game theory concepts viz. Convex game, Shapley value, Core [32] and the related are given.

k-anonymity

Burnett et al. [36], presented the classification of attributes of a data table $\mathcal{D}$. Explicit Identifiers (EID), Quasi Identifiers (QID), Sensitive Attributes (SA) and Non-Sensitive Attributes (NSA) are different classifiers of the attributes. EID is set of attributes which explicitly identify a person and his possible sensitive information, whereas the set of attributes which can potentially identify the sensitive information of a person by associating other external sources is QID. The set containing attributes like *Disease*, *Salary* etc., which holds sensitive information of a person is given by SA and remaining that do not fall into the above three are categorized as NSA.

If every data tuple in a data table $\mathcal{D}$ is indiscernible, under QID set of attributes, with at least k-1 other tuples then the table is said to k-anonymized. For example, Table 1 is 3-anonymized version of Table 2.

Cooperative game

A Cooperative game $\mathcal{G}$ with transferable utility (TU) [37] consists of two parameters **N** and ν. **N** is a set of n players i.e., **N**= {1,2,..., n} and ν is a real valued function defined over power set of **N**, $\mathcal{P}(\mathbf{N})$ i.e., $\nu : \mathcal{P}(\mathbf{N}) \rightarrow \mathbb{R}$, $\nu(\phi) = 0$ is called characteristic function or value function. For any subset S of **N**, $\nu(S)$ is called as value or worth of the coalition S and this is explained with a simple example [38].

Example There are there players i.e. $\mathbf{N} = \{1,2,3\}$. Player 1 is a seller, players 2 and 3 are buyers. Player 1 has a single unit to sell and its cost is $4. Each buyer is interested to buy the unit. Players 2 and 3 'willingness-to-pay' are $9 and $11 respectively. Now the game is characterized as follows.

Table 1 Data records after anonymization and it is 3-anonymized data table

Job	Sex	Age	Disease
Professional	Person	[25–30]	Cancer
Professional	Person	[25–30]	HIV
Professional	Person	[25–30]	Asthma
Artist	Female	[30–35]	HIV
Artist	Female	[30–35]	Hepatitis
Artist	Female	[30–35]	Flu

Table 2 Sample data records before anonymization

Job	Sex	Age	Disease
Lawyer	Male	28	Cancer
Engineer	Male	25	HIV
Doctor	Female	30	Asthma
Writer	Female	34	HIV
Singer	Female	32	Hepatitis
Dancer	Female	35	Flu

The characteristic function ν is defined as

$$\begin{aligned}
\nu(\{1\}) &= \$0\\
\nu(\{2\}) &= \$0\\
\nu(\{3\}) &= \$0\\
\nu(\{1,2\}) &= \$9 - \$4 = \$5\\
\nu(\{1,3\}) &= \$11 - \$4 = \$7\\
\nu(\{2,3\}) &= \$0\\
\nu(\{1,2,3\}) &= \$7
\end{aligned}$$

The intuition of ν is pretty simple. If there is no coalition for transact then the pay-off is zero and this shows first three definitions. Now if Player 1 and 2 come together and transact then the total gain of this coalition is the difference between buyer's willingness-to-pay and sell's cost price and hence it is \$5. Similarly worth of the coalition of Player 1 and 3 is \$7. These two are represented by 4th and 5th relations. Players 2 and 3 cannot come together as each is trying for seller but not the buyer and therefore the worth is \$0. Finally, $\nu(\{1,2,3\}) = \$7$ not $\$5 + \$7 = \$12$, because Player 1 has only one unit to sell and so he can transacts with only one buyer either Player 2 or Player 3. Obviously, Player 1 transact with the higher willingness-to-pay to maximize his worth, henceforth, $\nu(\{1,2,3\}) = \$7$ rather than \$5.

Convex cooperative game

A cooperative game $\mathcal{G}$ is Convex [35] if for any $S, T \subseteq \mathbf{N}$, $\nu(S \cup T) = \nu(S) + \nu(T) - \nu(S \cap T)$. It means that the marginal contribution of a player t_i is more for $S \supseteq T$ i.e. larger coalitions and formally:

$$\forall T, T \subseteq S \subseteq \mathbf{N}\backslash\{t_i\}, \quad \textit{for}\ t_i \in \mathbf{N}(\nu(S \cup \{t_i\}) - \nu(S) \geq (\nu(T \cup \{t_i\}) - \nu(T)) \tag{1}$$

Any coalitional game can be analyzed by using solution concepts, which describes the distribution patterns of the total value of the game among individual players. The following are some of the solution concepts.

The core

Let $x = (x_1, x_2, \ldots x_n)$ be a payoff allocation vector, where x_i is the payoff of ith player. The core is the set of all payoff allocation vectors which satisfy the following properties.

- Individual rationality: $\forall i \in \mathbf{N}$, $x_i \geq \nu(\{i\})$
- Collective rationality: $\sum\limits_{i \in \mathbf{N}} x_i = \nu(\mathbf{N})$
- Coalitional rationality: $\forall S \subseteq \mathbf{N}$, $\sum\limits_{i \in \mathbf{N}} x_i \geq \nu(S)$

Every payoff allocation in the core of the game is 'stable', intuitively, no player will get benefit by unilaterally deviating from a given payoff allocation of the core. A payoff allocation which holds Individual's rationality and collective rationality is called *Imputation*.

Shapley value

The Shapley value of coalitional game is a solution concept. It explains the expected payoff allocation for the Cooperative Privacy Game $\mathcal{G}$. It formalizes a fair distribution of the total payoff among the players of the coalition formation. The payoff allocation, based on this solution concept, is fair as it is including the information of each player's contribution to the total value i.e., it assumes the relative importance of the each player in coalition formation [39].

Let Π be set of all permutations over $\mathbf{N}$ and x_i^π be contribution of player t_i to permutation π of CoPG $\mathcal{G}$. Any imputation $cov = (cov_1, cov_2, \ldots cov_n)$ is a Shapley value fairly distribution if it follows the axioms of Lloyd Shapley [7]. The Shapley value of each player i in the game $\mathcal{G}$, is formally given by

$$cov_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \{\nu(S \cup \{i\}) - \nu(S)\} \tag{2}$$

To overcome the rigidness of computation of the Eq. 2, [8] provided an equivalent equation stated as follows:

$$cov_i = \frac{1}{n!} \sum_{i \in S} (|S| - 1)!(n - |S|)!\{\nu(S) - \nu(S - i)\} = \frac{1}{n!} \sum_{\pi \in \Pi} x_i^\pi \tag{3}$$

In the evaluation of CoV of each tuple underlying the solution concept, Shapley value is the only mapping in the distribution of payoff's of the players in a coalitional game which follows the properties like linearity, symmetry and carrier property [8]. This is one of the reasons, why we take on Shapley value in the process of computing cooperative value (CoV) which is used in the proposed method.

Cooperative Privacy Game Model

This Game Model provides a mechanism to find out the privacy level, k-anonymity [1], of the given data set by using the cooperation between the tuples. The underlying cooperation between every pair of tuples is estimated and termed as CoV. CoV takes advantage of Shapley value of each tuple. The data is segregated into groups based on the CoV.

Assume a data set $\mathcal{D}$ having an attribute set $\mathcal{A}$, and among them $\mathcal{A}_{QID}$ is collection of QID's of $\mathcal{D}$, i.e. $\mathcal{A}_{QID} = \{A_{QID_1}, A_{QID_2}, \ldots A_{QID_q}\}$. Let $\mathcal{D}_{QID} \subseteq \mathcal{D}$ be the set of possible tuples of $\mathcal{D}$ under $\mathcal{A}_{QID}$. Consider $\mathcal{D}_{QID} = \{t_1, t_2, ...t_n\}$ is the projection of $\mathcal{D}$ under $\mathcal{A}_{QID}$ of n input instances. A real valued function d, called distance function (for instance Euclidean Distance), defined as $d : \mathcal{D}_{QID} X \mathcal{D}_{QID} \rightarrow \{0\} \cup \mathfrak{R}^+$, where $d(t_i, t_j) \forall t_i, t_j \in \mathcal{D}_{QID}$ gives the distance between t_i and t_j, and also it is clear that $d(t_i, t_i) = 0$.

To set up a CoPG among the players(tuples) CoV is a function defined as $f : \{0\} \cup \mathfrak{R}^+ \rightarrow [0, 1]$. Insightfully, if two data tuples, namely, t_i and t_j are very similar then the $f(d(t_i, t_j))$ reaches 1.

$$f(d(t_i, t_j)) = 1 - \frac{d(t_i, t_j)}{d_{max}} \tag{4}$$

where d_{max} is maximum of the distances between all pairs of points in the data set. It is used to normalize the distances.

The following assumptions are made to establish the CoPG $\mathcal{G} = (\mathbf{N}, \nu)$:

- Each tuple is a player and $\mathbf{N} = \mathcal{D}_{QID}$, so $|\mathbf{N}| = n$.
- Every player interacts with other players and tries to maximize their CoV as it depends on the 'average increase in their worth' across all valid subsets.
- The characteristic function ν is defined as follows for all coalitions $S \subseteq \mathbf{N}$

$$\nu(S) = \frac{1}{2} \sum_{t_i,t_j \in S, t_i \neq t_j} f(d(t_i, t_j)) \tag{5}$$

Equation 5, computes the total worth of the coalition S and it has quadratic computation complexity which is proved in later section. The worth of the coalition is calculated as the sum of pairwise coordinations between the players; consequently this formulation smartly forms groups, each being called coalition which fulfil the property that the points having more CoV will be in the same coalition. These are formed based on the similarities between the players, leading to seclusion of data set into groups. These groups further under go anonymization process, which is discussed in the following sections.

Convexity of CoPG

In the process of proving that CoPG is convex, here are some propositions stated and proved.

Proposition 1 *The Cooperative game $\mathcal{G} = (\mathbf{N}, \nu)$ is convex where ν is defined as*

$$\begin{aligned} \nu(S) = {} & 0 \quad \textit{if } S = \{t_i\} \quad \textit{where } i = \{1, 2, \ldots, n\} \\ = {} & \frac{1}{2} \sum_{t_i,t_j \in S, t_i \neq t_j} f(d(t_i, t_j)) \qquad \textit{otherwise} \end{aligned}$$

Proof According to the definition of Convex game 1, for any player $t_k \in \mathbf{N}$, if we consider two coalitions S and T such that $T \subseteq S \subseteq N \setminus \{t_k\}$ then

$$\begin{aligned} &\{\nu(S \cup \{t_k\}) - \nu(S)\} - \{\nu(T \cup \{t_k\}) - \nu(T)\} \\ &\quad = \{\nu(S \cup \{t_k\}) - \nu(T \cup \{t_k\})\} - \{\nu(S) - \nu(T)\} \\ &\quad = \{\nu(S) + \sum_{t_i \in S} f(d(t_i, t_k))\} - \{\nu(T) + \sum_{t_i \in T} f(d(t_i, t_k))\} - \{\nu(S) - \nu(T)\} \\ &\quad = \sum_{t_i \in S \setminus T} f(d(t_i, t_k)) \geq 0 \quad (\because \textit{Rangeoffis}[0, 1]) \end{aligned}$$

Every convex Cooperative game has non-empty core [6] and also Shapley value belonging to core. From the Proposition 1, our CoPG with characteristic function stated in the Eq. 5 is a convex game and hence it has a solution.

Complexity of calculating cooperative value

This section presents the calculation process of CoV. The CoV for each tuple in the data table is computed using Eq. 3, but the computation is hard because it includes $n!$ as a factor. The following proposition overcomes the computational infeasibility and provides a relation for CoV to compute in polynomial time.

Proposition 2 *Computational complexity of cooperative value (CoV), i.e. Eq. 3, is quadratic.*

Proof According to Eq. 3 we have

$$cov_i = \frac{1}{n!}\sum_{\pi\in\Pi} x_i^{\pi}$$

That implies CoV is the summation of contribution of the player t_i for each coalition over all possible permutations. But for specific t_i, x_i^{π} is equal to summation of all similarities with other players whose position is less than the position of t_i with respect to a specific permutation.

$$cov_i = \frac{1}{n!}\sum_{\pi\in\Pi}\sum_{\pi(t_j)<\pi(t_i)} f(d(t_i,t_j))$$

Now for specific t_i and t_j the total number of possible permutations is $(n-2)!$. So, the second summation in above equation contains $(n-2)!$ terms. Also if t_i takes first position in a permutation then there is no possibility of t_j, if t_i takes second position then one possibility is there for t_j. If we metric then we have the following and hence the result.

$$cov_i = \frac{(n-2)!}{n!}\{1+2+\cdots+(n-1)\}\sum_{t_j\in\mathbf{N},t_i\neq t_j} f(d(t_i,t_j)))$$

If we adopt the above argument, the CoV of a tuple t_i can be found with O(n) complexity. So, in quadratic time we can estimate the CoV of all tuples of given data table. In experimentations it is observed that evolution of CoVs takes about linear time as the actual participation of t_i is very less than the possible permutations $n!$.

Proposition 3 *In convex CoPG setting for given* $\epsilon > 0$ *and* $d(t_i,t_j) \leq \epsilon \rightarrow 0, \forall t_i, t_j \in \mathbf{N}$ *then Cooperative values of* t_i *and* t_j *are almost same.*

Proof From Proposition 3:

$$\begin{aligned}
cov_i - cov_j &= \frac{1}{2}\sum_{t_k \in \mathbf{N}, t_i \neq t_k} f(d(t_i, t_k)) - \frac{1}{2}\sum_{t_k \in \mathbf{N}, t_j \neq t_k} f(d(t_k, t_j) \\
&= \frac{1}{2}\sum_{t_k \in \mathbf{N}, t_i, t_j \neq t_k} f(d(t_i, t_k)) - f(d(t_k, t_j)) \\
&= \frac{1}{2}\sum_{t_k \in \mathbf{N}, t_i, t_j \neq t_k} \left(\left[1 - \frac{d(t_i, t_k)}{d_{max}}\right] - \left[1 - \frac{d(t_k, t_j)}{d_{max}}\right]\right) && (\because \textit{From } 4) \\
&= \frac{1}{2 * d_{max}} \sum_{t_k \in \mathbf{N}, t_i, t_j \neq t_k} (d(t_k, t_j) - d(t_i, t_k)) \\
&\leq \frac{1}{2 * d_{max}} \sum_{t_k \in \mathbf{N}, t_i, t_j \neq t_k} d(t_i, t_j) && (\because d(.,.)\ \textit{is a metric on its domain}) \\
&= \frac{(\mathbf{N} - 2)}{2 * d_{max}} d(t_i, t_j) \\
&\leq \frac{(\mathbf{N} - 2) * \epsilon}{2 * d_{max}}
\end{aligned}$$

Hence the hypothesis follows with the argument: as $\epsilon \rightarrow 0$ implies $cov_i \rightarrow cov_j$ □

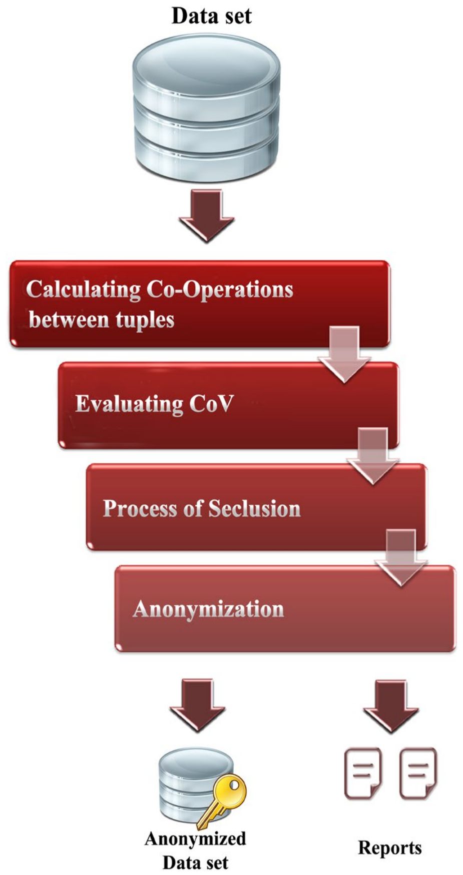

Fig. 1 Anonymization process architecture: cooperation, between pair of records, is estimated as first step and as a second step CoV of each record is calculated. In third and fourth steps, respectively, seclusion and anonymization are performed

Insightfully, the above proposition states that the cooperative values of tuples which are more similar i.e., the distance between them is almost zero, are nearly equal. It results that the tuples having almost equal CoV will be in same coalition.

Achieving cooperative privacy

This section describes the mechanism adopted by the data protector who is taking action about privacy of sensitive information in his data releases. Figure 1, shows the possible steps involved in the process of anonymization for a given data set $\mathcal{D}$ to achieve cooperative privacy.

The methodology of the process is explained in the following steps:

1. **Calculate Cooperation value between each pair** The similarity between every pair of tuples (players) is estimated as Cooperation value of the pair in the given data set $\mathcal{D}_{\mathcal{QID}}$ using Eq. 4.
2. **Evaluating CoV** For each tuple CoV is assigned a value using Eq. 6 and Proposition 2.
3. **Process of seclusion** The tuples are secluded into groups based on CoV, which undergo anonymization process.
4. **Anonymization** Each secluded group of given data table is anonymized and the k-anonymized data along with information loss and k value of the data table $\mathcal{D}$ is published.

Calculating values of cooperation

In step 1, a data set $\mathcal{D}$ is considered with set of attributes $\mathcal{A}$. By choosing the *QID* attributes, we have set of quasi identifiers $\mathcal{A}_{\mathcal{QID}}$. The projection of $\mathcal{D}$ under $\mathcal{A}_{\mathcal{QID}}$ is $\mathcal{D}_{\mathcal{QID}}$. By using Eq. 4, the CoV between every pair of tuples of $\mathcal{D}_{\mathcal{QID}}$ is found. A symmetric matrix of order n (as $\mathcal{D}_{\mathcal{QID}}$ is having n tuples) called CoMatrix can be constructed using the cooperative values. According to proposition 2, this CoMatrix can be constructed in quadratic polynomial time. For simplicity Manhattan distance is chosen as distance function in Eq. 4 and $\mathcal{A}_{\mathcal{QID}}$ with only numerical attributes. d_{max}, which is maximum of all possible distances, is used for normalization in the formalization itself. Algorithm, presented in the Table 3, explains the calculation of CoMatrix in $O(n^2)$ time and it is given as input to step 2.

Evaluation of CoV

In this section, the evaluation process of CoV of each tuple of $\mathcal{D}_{\mathcal{QID}}$ is discussed. It is a hard problem to compute the CoV, using Eq. 5, of each tuple as it includes $n!$ permutation orderings. Nevertheless, the game setting $\mathcal{G}$ is convex, the underlying CoV is Shapley value gives the center of gravity of the extreme points of the non-empty core [8]. The selection of characteristic function of this game model is shown in Eq. 3. As laid down

Table 3 Algorithm

```
Algorithm: CoMatrix Calculation
Input : $\mathcal{D}_{\mathcal{QID}}$ of size n
Output: CoMatrix of order n
Assumptions:
(i) $t_i$ is $i^{th}$ tuple of $\mathcal{D}_{\mathcal{QID}}$.
(ii) d(. , .) represents Manhattan distance between $i^{th}$ and $j^{th}$ tuples of $\mathcal{D}_{\mathcal{QID}}$
(iii) $d_{max}$ is the normalization factor of set of all possible domain values of each attribute of $\mathcal{D}_{\mathcal{QID}}$
Method :
Begin
    for i ← 1 to n do
        for j ← i+1 to n do
            $com_{i,j} = 1 - \frac{d(t_i,t_j)}{d_{max}}$        // Using Equation 4
        End for
    End for
    return $CoMatrix = [com_{i,j}]_{nXn}$
End
```

by Proposition 2 the CoV of each tuple can also be estimated using the following relation, quadratic time:

$$cov_i = \frac{1}{2} \sum_{t_j \in \mathcal{D}_{\mathcal{QID}}, t_i \neq t_j} f(d(t_i, t_j)) \tag{6}$$

Algorithm, presented in the Table 4, describes how each tuple will be assigned CoV. It assumes the CoMatrix evaluated in previous step as input and returns an array of CoVs of size n, corresponding to $\mathcal{D}_{\mathcal{QID}}$. This Algorithm takes $O(n^2)$ complexity.

Process of seclusion

This process describes how to seclude the tuples of the data set $\mathcal{D}_{\mathcal{QID}}$ into groups based on their CoVs, the inner sense is that, the density of tuples around a tuple will form a group. The basic idea is to start with a tuple whose CoV is maximum at the initial core point and collect all the tuples having 'very near' CoVs as core point and put them into one group is named as coalition group. The parameter α is called cooperative parameter which governs this 'very near' in the process.

The CoVs of tuples gradually decreases when they are far away from the center of the coalition and hence α decreases accordingly. So, in order to degrade α in terms of CoVs,

Table 4 Algorithm

```
Algorithm: Calculation of CoV's
Input : CoMatrix of order n, whose entries $com_{i,j}$
Output: An array SV of size n
Method :
Begin
SV[] ← $\phi$
    for i ← 1 to n do
        for j ← 1 to n do
            if $(i \neq j)$
                $sv_i \leftarrow sv_i + com_{i,j}$
            End if
        End for
        $SV[i] \leftarrow sv_i$
    End for
    return SV[]
End
```

a non-linear decreasing function has been considered. For this, $\alpha = \beta * h(l_{max})$ is taken into account where h is defined over the set of all CoVs and $\beta \in [0, 1]$ is a weight factor. In practice, $\alpha = \beta * \sqrt{\frac{l_{max}}{g_{max}+1}}$ is considered. Here, g_{max} is global maximum of CoV used for normalization of the CoVs and l_{max} is local maximum of coalition group. However, any degradation function α can be chosen over these CoVs based on the domain values of the given data set and by the same token β also.

Growth Control Queue (GCQ) is an array introduced in the Algorithm (see Table 5). The advantage of using this queue is to add tuple indexes to the queue, if their Shapley value is at least γ-multiple of center of the coalition. Here, γ is multiplicity of CoV. It senses that, GCQ contains all unallocated points which has very low CoV value as compared to the density around the coalition group. These points do not take part in the further growth of the coalition group and it provides the uniform distribution of density throughout the coalition and the density does not vary beyond the threshold [9]. The GCQ grabs all this information and the empty queue manages the growth of the coalition.

Anonymization

This phase assumes the set of cooperative groups (CoG) as an input which is obtained from the third phase and it returns the anonymized data for the purpose of publishing by using anonymization algorithms [3]. Hierarchy free generalization of numerical attributes [12] are used to attain k-anonymization and information loss of the anonymized data is also measured.

Table 5 Algorithm

```
Algorithm: Process of Seclusion
Input: $\mathcal{D}_{\mathcal{QID}}$, $SV$ of size $n$, Weight factor $\beta \in [0,1]$, Multiplicity $\gamma \in [0,1]$.
Output: Set of Seclude groups CoG      // Set of Cooperative Groups
Method :
Begin
    Sort($SV$)     // Sort the array $SV$ as non-increasing order
    $g_{max} \leftarrow SV[1]$     // Choose $1^{st}$ value of $SV$ as Global Maximum
    $GCQ[] \leftarrow \phi$     // Create a dynamic array for Growth Control Queue
    $CoG \leftarrow \phi$     // Used to store Cooperative groups
    $X \leftarrow \mathcal{D}_{\mathcal{QID}}$     // Keeping data set in X.
    Repeat
        $m \leftarrow argmax_{i, t_i \in X}\{sv_i\}$     // choose index whose CoV is maximum
        $l_{max} \leftarrow SV[m]$     // Assign local maximum
        $\alpha \leftarrow \beta\sqrt{\frac{l_{max}}{g_{max}+1}}$     // used to capture the rate of degradation
        $CoG_m \leftarrow t_m$     // Initialization of Cooperative group with its center $t_m$
        $GCQ \leftarrow t_m$     // Initialization of Growth Control Queue
        Repeat
            if $(f(d(t_m, t_j)) \geqslant \alpha)$     // Allocate $t_j$'s to $CoG_m$
                $X \leftarrow X \setminus \{t_j\}$
                if $(SV[t_j] \geq \gamma * l_{max})$
                    $GCQ \leftarrow GCQ \cup \{t_j\}$
                Else
                    $CoG_m \leftarrow CoG_m \cup \{t_j\}$
                End if
            End if
        Untill $GCQ \neq \phi$
        $CoG \leftarrow CoG \cup CoG_m$
        $CoG_m \leftarrow \phi$
    Untill   $X \neq \phi$
    Return CoG     // Collection of all partitions of the $\mathcal{D}$ under $QID$
End
```

In the process, for every coalition and for every QID attribute, *max* and *min* of all possible domain values are found and all these values are replaced under the QID in that particular group with [*min*, *max*]. Finally, *k* is calculated as *Min* of sizes of all possible partitions after the process.

The data user who is collecting the data from data collector, typically wants to get more information from it. 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*-anonymization of a given data set, typically, calculate how much quality has been lost 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 IL [3], however, the following relation is adopted to calculate IL of the numerical attributes after anonymization:

$$IL(D) = \frac{1}{|D| * |QID|} \sum_{i=1}^{|CoG|} |CoG_i| \sum_{j=1}^{|QID|} \frac{Max[Dom(CoG_{i,j})] - Min[Dom(CoG_{i,j})]}{Max[Dom(QID_j)] - Min[Dom(QID_j)]} \quad (7)$$

Here, $Max[Dom(QID_j)] - Min[Dom(QID_j)]$ is the spread in domain of $\mathcal{D}_{\mathcal{QID}}$ under QID_j, and $Max[Dom(CoG_{i,j})] - Min[Dom(CoG_{i,j})]$ is the spread of the domain of QID_j in the specific coalition group CoG_i. So, we can consider the IL as sum of all ratios of the spreads weighted by the ratio of group size and data set size.

Algorithm, described in the Table 6 explores the process of anonymization of the coordination groups. It also explains the computation of the IL as well as finding *k* value for *k*-anonymization. It assumes the output of Algorithm 5 as input and returns IL of anonymized data, *k* value of *k*-anonymization and published data $\mathcal{D}'$.

Experimentation and empirical analysis

Experiments have been performed on Intel Core @ 2.93 GHz with 4GB RAM out of it 2GB of RAM has been exclusively allocated for the Net Beans platform. Experiments are conducted on Adult Data set available at UCI Machine Learning Repository [40]. 1000 records are selected randomly from the preprocessed Adult Data set which has 36,282 data records. Age, Fnlwgt, Hours-per-week, the numerical attributes, are chosen as Quasi Identifiers for our experimentation and number of coalitions, anonymity level, number of outliers, IL (using Eq. 7) are calculated over different values of similarity weight factor (β) and multiplicity factor (γ). As a state-of-art study, CoV algorithm is compared with Mondrain Multidimensional methodology [12] and K-member clustering for *k*-anonymity [41]. See Fig. 2.

Number of coalitions vs β and γ

The variations of number of coalitions over different γ values are given in Fig. 3. As multiplicity factor (γ) is increased, the number of coalitions increases, because, when multiplicity factor is relaxed then more number of tuples are included in the coalition which leads less number of coalitions i.e., if γ value is increased then there is a possibility for tight segregation which causes more number of coalitions. The number of coalitions is constant until some fixed value γ which relatively depends upon the weight factor β. Another observation is that there is a sudden climb after certain value (sum of β and γ is around 1.75 for our sample data set) and the growth rate of number of coalitions decreases according to decrease in the weight factor β (See Fig. 3).

Table 6 Algorithm

```
Algorithm: Anonymization
Input : CoG // Collection of partitions of coordination groups of $\mathcal{D}$
Output: IL(Information Loss), k(Anonymization level), $\mathcal{D}'$ (Anonymized data set under QID)
Assumptions:
i) $Max_{i,j}$ and $Min_{i,j}$ represents maximum and minimum of domain values
   of quasi identifier $QID_j$ in the coaltion group $CoG_i$.
ii)$Max_j$ and $Min_j$ are maximum and minimum of domain values of $QID_j$ in the entire data set.
iii) spread is used to find difference between Max and Min.
iv) $t_k[QID_j]$ represents the possible value of the tuple $t_k$ under $QID_j$
Method :
Begin
    $IL \leftarrow 0$
    Repeat for each $QID_j$
        Repeat for each $CoG_i$
            $Max_{i,j} \leftarrow argmax_{t_k \in CoG_i} \; t_k[QID_j]$
            $Min_{i,j} \leftarrow argmin_{t_k \in CoG_i} \; t_k[QID_j]$
            $spread_{i,j} \leftarrow Max_{i,j} - Min_{i,j}$
        Until $CoG \neq \phi$
        $Max_j \leftarrow argmax_{\forall i} \; \{Max_{i,j}\}$
        $Min_j \leftarrow argmin_{\forall i} \; \{Min_{i,j}\}$
        $spread_j \leftarrow Max_j - Min_j$
        $ratio_i \leftarrow \frac{spread_{i,j}}{spread_j}$
        Repeat for each $CoG_i$
            $IL_i \leftarrow 0$
            $count_l \leftarrow 0$
            Repeat for each $t_k \in CoG_i \subseteq \mathcal{D}$
                replace $t_k[QID_j] \leftarrow [Max_{i,j}, Min_{i,j}]$
                $count_l \leftarrow count_l + 1$
            Until $CoG_i \neq \phi$
            $IL_i \leftarrow count_l * ratio_i$
        Until $CoG \neq \phi$
        $IL \leftarrow IL + \frac{1}{n} * IL_i$
        $k \leftarrow argmin_{\forall l}\{count_l\}$
    Until $QID \neq \phi$
    return $IL, k, \mathcal{D}'$
End
```

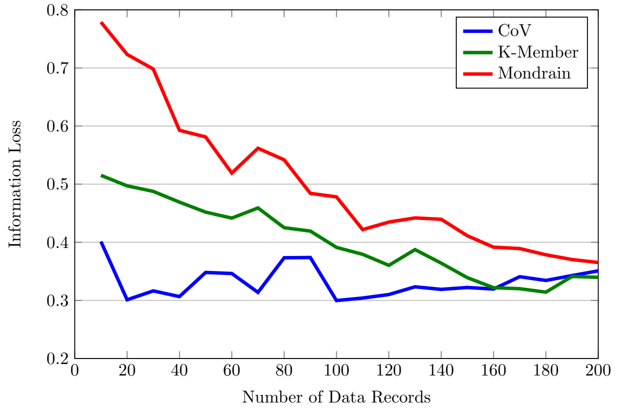

Fig. 2 Comparison of CoV with K-Member clustering, Mondrain Multi-dimensional methods. Information loss is estimated using the three methods and variations are presented. Our method shows the consistency with the size of the data records where as the other two gradually decrease. Dramatically, when data size is largeref:adam.1996 marg the methods give approximate equality

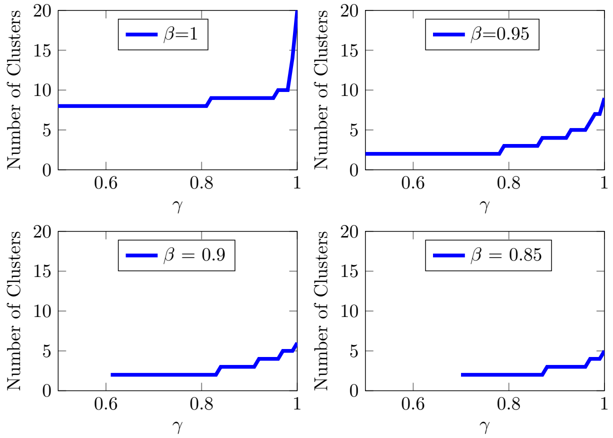

Fig. 3 Number of coalitions vs γ

Figure 4 depicts the relation between number of coalitions and β. It is observed that the kind of variations is almost same as above, but growth rate in number of coalitions is more as compared with the former one. So, it can be said that the influence of β is more than that of γ in the process of Seclusion.

Number of outliers vs β and γ

The coalitions having single record are marked as outliers in Algorithm (see Table 5), and the number of outliers for different values of β and γ are established. Figure 5 depicts the variations of number of outliers with γ. It shows that there is no possibility of outliers for lesser values of γ. The relaxation of γ, includes the tuples which are defined as outliers in the case of more values of γ.

Intuitively, the records which are far away, in distance point of view, from the coalitions are also included when γ is reduced. It can be observed that the number of outliers decrease as β decreases. A similar observation can be seen in Fig. 6, then graphs are drawn for number of outliers and β. The presence of outliers are more than the previous case as β is influenced more than γ.

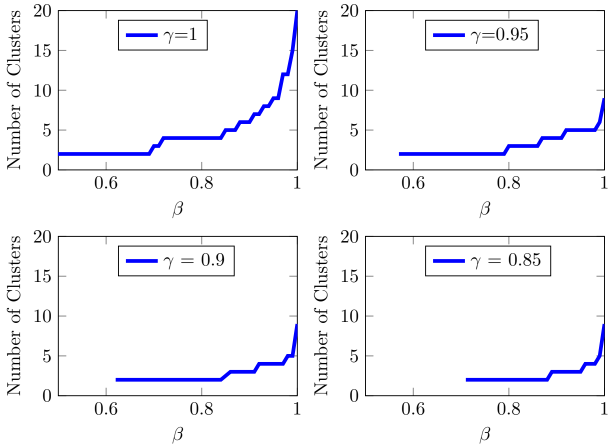

Fig. 4 Number of coalitions vs β

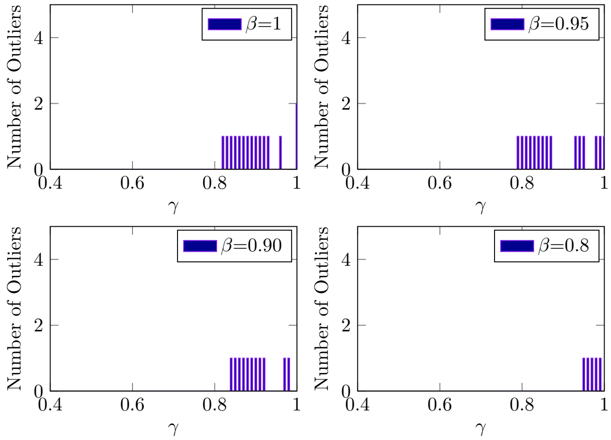

Fig. 5 Outliers vs γ

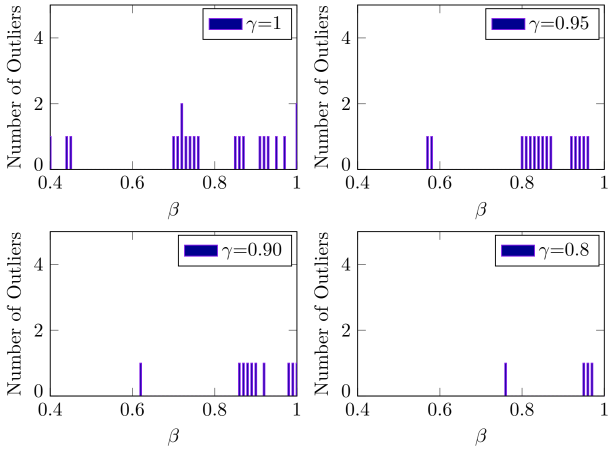

Fig. 6 Outliers vs β

Information loss vs β and γ

This section presents how the IL varies over the parameters weight factor β and multiplicity factor γ. Figure 7 describes the changes in the IL with different γ values. The IL is calculated using Eq. 7. As γ increases it doesn't allow to include more number of records into the groups. So, IL calculated by using Eq. 7 implies that the coalitions having more similar data records, have less information loss. Insightfully, when we relax the γ then the far away tuple are also included into the groups.

In the present work for the anonymization process over these groups hierarchy free construction is used. In this methodology the values of an attribute are generalised in a group by *min*, *max*. While implementing this process if a far way tuple is included in the group then unnecessarily more generalization is required which in turn increases the IL. This implies that the IL increases with the increase in γ.

The behaviour of the graphs plotted for IL and different cases of β are almost same. The IL is constant as γ increases until some point, then there is a sudden decline at which the sum of β and γ assumes some fixed value (It is around 1.75 for our sample data set).

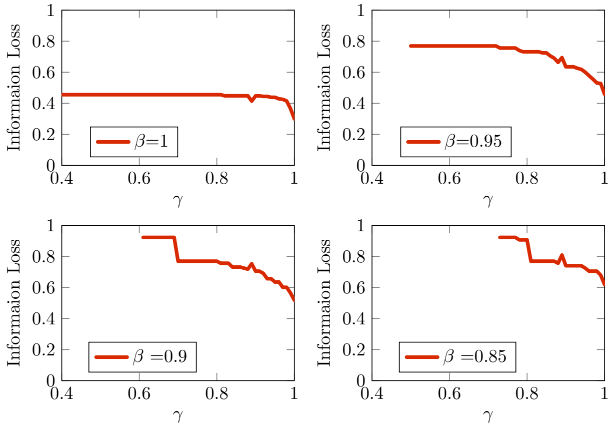

Fig. 7 Information loss vs γ

Figure 8, shows the relation between IL and β. Similar patterns shown above are seen but the rate of decrease is less than the previous cases and thence it can be concluded that the algorithm is more influenced by β.

Information loss vs size of data set

In this section the variation of IL value with size of data set is explained. Figure 9 depicts the IL values corresponding to different sizes of data for different β and γ. For all cases, it shows that there are fluctuations up to certain size depending upon the data set. After that IL increases but the rate of growth is less compared to the rate of growth of size of the data set. When β and γ are equal to 1, the interpolated curve for IL is shown in Fig. 10.

Representation of information loss, β and γ

Figure 11 shows that the change in IL over the variation in β and γ. The graph shows that the IL is minimum when β and γ are equal to 1. IL value increases as β or γ increases, but simultaneously the number of outliers decrease as shown in Figs. 5 and 6.

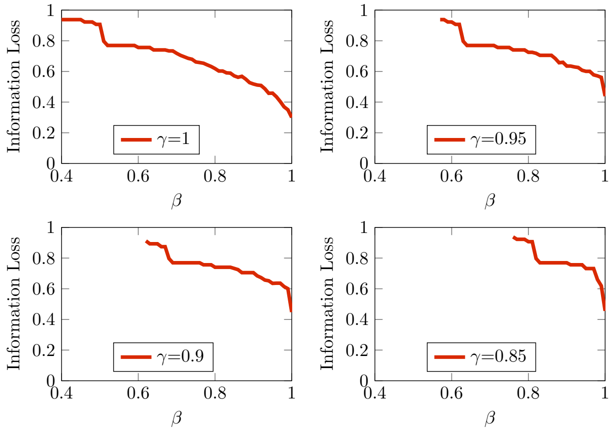

Fig. 8 Information loss vs β

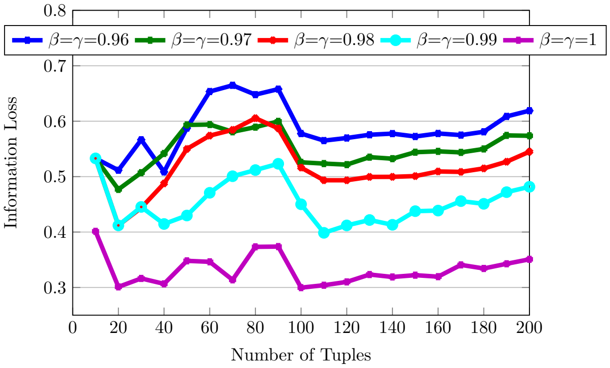

Fig. 9 Relation between information loss with size of the data set similar pattern can be observed for Information loss with respect to data size, by varying β and γ

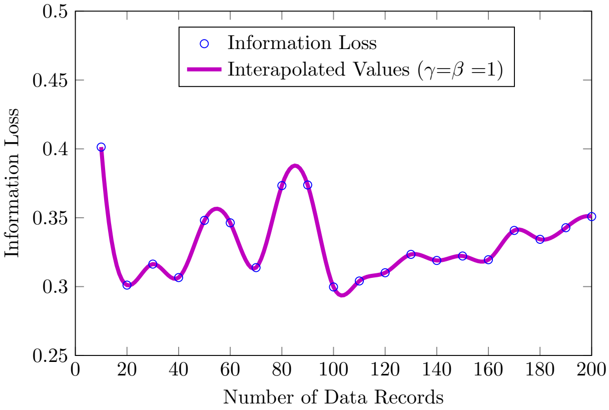

Fig. 10 Interpolate curve between information loss and size of the data set: estimated Information loss according to number of data records for given β and γ equal to one. More fluctuations can be seen for less number of records because of occurrence of more outliers. In contrast, if data size increases information loss is not much varying.

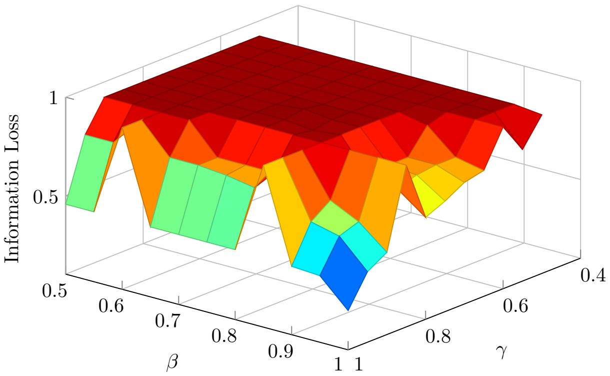

Fig. 11 Relation between information loss, β and γ: variation of information loss is shown for given β and γ. The best choice for least Information Loss is for both β and γ equal to 1

Conclusions and future work

Different mechanisms are required to protect the privacy in information society, where people are forced to give private information. Rational cooperation between the people who are involved in the information society is required inspite of several rules imposed by the governments. This motivation led towards this novel strategy of Cooperative Privacy for Privacy Preserving Data Publishing using Cooperative Game theory.

To achieve privacy in Data releases a Cooperative Privacy Game (CoPG) is set up, in which each tuple in the data table behaves as a player, trying to preserve their privacy and in turn helps in preserving other's privacy. This is formalized with characteristic function ν. All the tuples(players) segregate themselves to form groups called coalitions based on Cooperative value (CoV). The CoV of each tuple is calculated based on the solution concept Shapley value. The CoV fairly distributes the worth through out the group of the tuples and the separation is also unbiased. The separation process and hierarchy free anonymization process are described. Algorithms which required for the processes are presented. Experimentation results and insightful observations are reported.

As a future work, the following directions are provided and these are non-exhaustive:

- Expanding this approach to incorporate the security functionalities which are obtained from the players involved in the game.
- Expanding the theory to design a game model with mixed strategies rather than pure strategies for Cooperative privacy because the user in the game may act differently with other players.
- In experimentation process outliers were obtained. To decrease the possible outliers appropriate mechanisms are to be incorporated to the model.
- Extensive study of theory is necessary for the choice of β and γ.

Authors' contributions

SC investigated the state-of-art, designed and developed the model and took suggestions as and when necessary from VK. Both authors read and approved the final manuscript.

Author details

[1] Department Computer Science and Systems Engineering, Andhra University, Visakhapatnam 530 003, India. [2] Department of Computer Science and Engineering, NS Raju Institute of Technology, Sontyam, Vishakapatnam 531 173, India.

Acknowledgements

Authors would like to acknowledge Future Technology Research Association International (FTRA) who sponsored this article for publication.

Competing interests

The authors declare that they have no competing interests.

Received: 22 May 2015 Accepted: 16 June 2016
Published online: 14 July 2016

References

1. Sweeney J (2002) k-anonymity: a model for protecting privacy. Int J Uncertain Fuzz Knowl Syst 10(5):571–588
2. Gehrke AM, Kifer D, Venkatasubramaniam, M (2006) *l*-diversity: privacy beyond *k*-anonymity. In: Proceedings of the 22nd IEEE International Conference on Data Engineering: 3-8 April 2006, IEEE, Atlanta, p 24
3. Wang BF, Chen K, Philip R (2010) Privacy-preserving data publishing: a survey on recent developments. ACM Computing Surveys, New York
4. Domingo-Ferrer J (2010) Coprivacy: towards a theory of sustainable privacy. In: Proceedings in International Conference PSD: 22-24 Spetember 2010, Corfu, pp 258–268
5. Domingo-Ferrer, J (2011) Coprivacy: an introduction to the theory and applications of co-operative privacy. SORT. 2011, Special issue: Privacy in statistical databases
6. Roughgarden NN, Tardos T, Vazirani E (2007) Algorithmic game theory. Cambridge University Press, Cambridge

7. Shapley LS (1971) Cores of convex games. Int J Game Theory 1(1):11–26
8. Garg VK, Narahari Y, Murty MN (2013) Novel biobjective clustering (bigc) based on cooperative game theory. IEEE Trans Knowl Data Eng 25(5):1070–1082
9. Swapnil D, Satyanath B, Anoop KR, Varun R (2011) Pattern custering cooperative game theroy. In: Proceedings in Centenary Conference, Indian Institute of Science, Bangalore, pp 1–6
10. Aggarwal CC (2005) On *k*-anonymity and the curse of dimensionality. In: Proceedings of the 31st International Conference on Very Large Data Bases: 30th Aug to 2nd Sept' 2005, VLDB Endowment, Norway, pp 901–909
11. Bayardo RJ, Agrawal R (2005) Data privacy through optimal k-anonymization. In: Proceedings of the 21st International Conference on Data Engineering: 5-8 April 2005, IEEE, California, pp 217–228
12. Lefevre K, Dewitt DJ, Ramakrishnan R (2006) Mondrian multidimensional k-anonymity. In: In 22nd International Conference on Data Engineering, ICDE, p 25
13. Samarati P, Sweeney L (1998) Generalizing data to provide anonymity when disclosing information (abstract). ACM PODS 98: 188
14. Li N, Li T (2007) t-closeness: privacy beyond *k*-anonymity and *l*-diversity. In: In Proceeding of IEEE 23rd Intl Confeference on Data Engineering (ICDE07)
15. Wong RCW, Li J, Fu AWC, Wang K (2006) (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp 754–759
16. Ranganathan N (2008) A microeconomic approach to multi obeciective spatial clustering. In: Proceedings of the 19th International Conference on Pattern Recognition: 8–11 Dec'2008, IEEE
17. Gupta U, Ranganathan N (2010) A game theoretic approach for simultaneous compaction and equipartitioning of spatial data sets. IEEE Trans Knowl Data Eng 22(4):465–478
18. Bulò SR, Pelillo M (2009) A game-theoretic approach to hypergraph clustering. Adv Neural Inform Process Syst 35:1571–1579
19. Chun Y, Hokari T (2007) On the coincidence of the shapley value and the nucleolus in queueing problems. Seoul J Econ 20:223–238
20. Wang D, Wang P, Wang N, Qing S (2015) Preserving privacy for free: efficient and provably secure two-factor authentication scheme with user anonymity. Inform Sci 321:162–178
21. Wang D, Wang P (2014) On the anonymity of two-factor authentication schemes for wireless sensor networks: attacks, principle and solutions. Comput Netw 20:1–15
22. Wang D, Debiao H et al (2014) An enhanced privacy preserving remote user authentication scheme with provable security. IEEE Trans Depend Secur Comput 12(4):428–442
23. Chaudhry SA, Farash MS, Naqvi H, Kumari S, Khan MK (2015) An enhanced privacy preserving remote user authentication scheme with provable security. Secur Commun Netw 8(18):3782–3795
24. Burke M-J, Kayem AVDM (2014) K-anonymity for privacy preserving crime data publishing in resource constrained environments. In: Proceedings of the 2014 28th International Conference on Advanced Information Networking and Applications Workshops, WAINA '14IEEE Computer Society, Washington, DC, pp 833–840
25. Anderson HE (2006) The privacy gambit: towards a game theoretic approach to international data protection. Vanderbilt J Entertain Technol Law 9(1):44
26. Böhme R, Koble S (2007) On the viability of privacy-enhancing technologies in a self-regulated business-to-consumer market: will privacy remain a luxury good? In: Workshop on economics of information security, pp 21–27
27. Kleinberg J, Papadimitriou CH, Raghavan P (2001) On the value of private information. In: Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge, TARK '01Morgan Kaufmann Publishers Inc., San Francisco, pp 249–257
28. calzolari G, Pavan A (2001) Optimal design of privacy policies: technical report. Gremaq, University of Toulouse, France
29. Preibusch S (2006) Implementing privacy negotiations in e-commerce. In: Frontiers of WWW Research and Development -APWeb. LNCS'06, pp 604–615
30. Dwork C (2006) Differential privacy. In: In ICALP, pp 1–12
31. Machanavajjhala A, Korolova A, Sarma AD (2011) Personalized social recommendations: accurate or private. Proc VLDB Endow 4(7):440–450
32. Osborne MJ (2003) An introduction to game theory. Oxford University Press, Oxford
33. Srinivasa L, Chakravarthy V, Kumari V (2011) Preserving data privacy using coalitional game theory. In: ECML PKDD, Workshop on KD-HCM, pp 53–65
34. Srinivasa L, Chakravarthy V, Kumari V, Sarojini C (2012) A coalitional game theoretic mechanism for privacy preserving publishing based on k-anonymity. Procedia Technol 6:889–896
35. Srinivasa L, Chakravarthy V, Kumari V (2014) Privacy preserving data publishing: a coalitional game theory perspective. Int J Comput Intell Stud 3(2):196–220
36. Burnett L, Barlow-Stewart K, Pros A, Aizenberg H (2008) The gene trustee: a universal identification system that ensures privacy and confidentiality for human genetic databases. J Law Med 10:506–513
37. Myerson BR (1997) Game theory: analysis of conflict. Harvard University Press, Cambridge
38. Brandenburger AM, Stuart HW (1996) Value-based business strategy. J Econ Manag Strategy 5:5–24
39. Straffin PD (1993) Game theory and strategy. The Mathemactical Association of America, USA
40. UCI Repository of Machine Learning Databases. www.ics.uci.edu/ mlearn/MLRepository.html
41. Byun JW, Kamra EB A, Li N (2007) Efficient k-anonymization using clustering techniques. In: DASFAA, Springer, Berlin, p 188