Systematic Clustering Method for l-diversity Model

Md Enamul Kabir[1], Hua Wang[1], Elisa Bertino[2] & Yunxiang Chi[3]

[1]Department of Mathematics and Computing
University of Southern Queensland,
Toowoomba, Queensland 4350, Australia,
Email: {kabir, wang}@usq.edu.au
[2]Department of Computer Science and CERIAS
Purdue University, West Lafayette, Indiana, USA
Email: bertino@cs.purdue.edu
[3]Toowoomba Pearl Company,
Toowoomba, Queensland 4350, Australia,
Email: toowoombapearls@yahoo.com.au

Abstract

Nowadays privacy becomes a major concern and many research efforts have been dedicated to the development of privacy protecting technology. Anonymization techniques provide an efficient approach to protect data privacy. We recently proposed a systematic clustering[1] method based on k-anonymization technique that minimizes the information loss and at the same time assures data quality. In this paper, we extended our previous work on the systematic clustering method to l-diversity model that assumes that every group of indistinguishable records contains at least l distinct sensitive attributes values. The proposed technique adopts to group similar data together with l-diverse sensitive values and then anonymizes each group individually. The structure of systematic clustering problem for l-diversity model is defined, investigated through paradigm and is implemented in two steps, namely clustering step for k-anonymization and l-diverse step. Finally, two algorithms of the proposed problem in two steps are developed and shown that the time complexity is in $O(\frac{n^2}{k})$ in the first step, where n is the total number of records containing individuals concerning their privacy and k is the anonymity parameter for k-anonymization.

Keywords: Privacy, k-anonymity, l-diversity, Systematic clustering.

1 Introduction

In recent years, the phenomenal advance technological developments in information technology have lead to an increase in the capability to store and record personal data about customers and individuals (Byun et al. 2006). Data mining is a common methodology to retrieve and discover useful hidden knowledge and information from the personal data. This has lead to concerns that the personal data may be breached and misused. Therefore it is necessary to protect personal data through some privacy preserving techniques before conducting data mining.

One of the most important concept for privacy is anonymity. Anonymity refers to a state where one's identity is completely hidden, and anonymity is oftentimes used as a synonym for privacy (Byun et al. 2007). Anonymous data can protect individuals in two ways, firstly to protect identity privacy for example it is not possible to learn about to whom a data record is related and secondly, attribute privacy for example not possible to know about particular property of individuals. In any databases specially where health records are collected by hospitals or government organizations, anonymity has a significant role to protect privacy as the information is linked to individuals could be highly sensitive. In commercial databases where organizations would like to disclose individual's data to third parties (e.g. external organizations), anonymity could be used to protect privacy of individuals as in such cases individual's privacy may not be respected. Thus within organizations individual's data should be restricted to access and anonymous by removing all information that can directly link data items to individuals via generalization or suppression before disclosing so that privacy is not beached. Such a process is referred to as data anonymization.

A contemporary approach dealing with the data privacy relies on the k-anonymity. The k-anonymity model proposed by Samarati (2001) and Sweeney (2002) is a simple and practical privacy-preserving approach to protect data from individual identification. The k-anonymity model works by ensuring that each record of a table is identical to at least $(k-1)$ other records with respect to a set of privacy-related features, called *quasi-identifiers*, that could be potentially used to identify individuals by linking these attributes to external data sets (Lin & Wei 2008). Therefore, privacy related information can't be revealed from the k-anonymity protected table during a data mining process. For example, consider the patient diagnosis records in a hospital in Table 1, where the attributes *ZipCode, Gender*, *Age* and *Education* are regarded as quasi-identifiers and *Disease* is a sensitive attribute. A diagnosis classifier can predict patients illness history based on attributes of *ZipCode, Gender*, *Age* and *Education* using these data. If the hospital simply publishes the table to other organizations for classifier development, the organizations might extract patients' disease history by joining this table with other tables (Chiu & Tsai 2007). On the contrary, Table 2 is a 3-anonymization version where data values of Table 1 in attributes *ZipCode, Gender*, *Age* and *Education* have been generalized as common values and the number of

[1]Clustering partitions record into clusters such that records within a cluster are similar to each other, while records in different clusters are most distinct from one another.

Table 1: Patients records in a hospital

	ZipCode	*Gender*	*Age*	*Education*	*Disease*	*Expense*
1	4350	Male	24	9th	Flue	2000
2	4351	Male	25	10th	Cancer	3500
3	4352	Male	26	9th	HIV+	6500
4	4350	Male	35	9th	Diabetes	2000
5	4350	Female	40	10th	Diabetes	3200
6	4350	Female	38	11th	Diabetes	2800
7	4352	male	41	9th	Flue	2700
8	4352	Female	42	10th	Heart disease	4800
9	4352	male	43	10th	Cancer	5200

Table 2: 3-Anonymization table

	ZipCode	*Gender*	*Age*	*Education*	*Disease*	*Expense*
1	435*	Person	[21-30]	Primary	Flue	2000
2	435*	Person	[21-30]	Primary	Cancer	3500
3	435*	Person	[21-30]	Primary	HIV+	6500
4	435*	Person	[31-40]	Secondary	Diabetes	2000
5	435*	Person	[31-40]	Secondary	Diabetes	3200
6	435*	Person	[31-40]	Secondary	Diabetes	2800
7	435*	Person	[41-50]	Primary	Flue	2700
8	435*	Person	[41-50]	Primary	Heart disease	4800
9	435*	Person	[41-50]	Primary	Cancer	5200

records in its two equivalence classes are both equal to three. It should be noted that the value of k in k-anonymity model is specified by users according to the purpose of their applications. By enforcing the k-anonymity requirement, it is guaranteed that even though an adversary knows that a k-anonymous table contains the record of a particular individual and also knows some of the quasi-identifier attribute values of the individual, he/she cannot determine which record in the table corresponds to the individual with a probability greater than $\frac{1}{k}$ (Byun et al. 2007). This indicates that the larger the values of k, the adversary has less chance of determining personal identifiable information and the data is more protected. On the other hand, if the k-values are too large it incurs more information loss. So, the k-value of the k-anonymization problem should not be too small or too large.

Usually, there are two methods to accomplish in k-anonymizing a dataset. The first one is suppression which involves not releasing entire tuple or a value at all to the third party, just like deleting them. The other one is generalization which involves replacing the value or tuple with less specific but semantically consistent value. For example, suppose there exists following five ages of individuals 51, 52, 53, 53, 55. We can generalize attribute *Age* to a age groups 50-55. On the other hand, we can also generalize them to other set 5*. However, we can suppress the age values by $\star$. Intuitively, generalization is better than suppression because of extracting some information. Undoubtedly, anonymization is accompanied with information loss. In order to be useful in practice, the dataset should keep as much informative as possible. Hence, it is necessary to consider deeply the tradeoff between privacy and information loss. To minimize the information loss due to k- anonymization, all records are partitioned into several groups such that each group contains at least k similar records with respect to the quasi-identifiers and then the records in each group are generalized or suppressed such that the values at each quasi-identifier are the same. Such similar groups are known as clusters. In the context of data mining, clustering is a useful technique that partitions records into clusters such that records within a cluster are similar to each other, while records in different clusters are most distinct from one another (Lin & Wei 2008). So k-anonymity model can be addressed from the viewpoint of clustering and recently Kabir et al. (2009) proposed systematic clustering method for k-anonymization. The experimental results showed that the proposed method outperforms over the recent clustering based k-anonymization techniques. However k-anonymity model may reveal sensitive information under the two attacks, namely homogeneity attack and background knowledge attack (Machanavajjhala et al. 2006). For example, Jak and Ron are two antagonistic neighbors. Jak knows that Ron goes to hospital recently and tries to find out the disease Ron suffers. Jak finds the 3-anonymous table as in Table 2. He knows that Ron is 39 years's old and lives in the suburb with postcode 4350. Ron must be record 4, 5 or 6. All three patients are suffering from diabetes. Jak knows for sure that Ron suffers from diabetes. Thus homogeneous values in the sensitive attribute of a k-anonymous group escape private information. Similarly k-anonymity does not protect individuals from a background knowledge attack. To overcome this problem, Machanavajjhala et al. (2006) presented an l-diversity model to enhance the k-anonymity model. The l-diversity model assumes that a private dataset contains some sensitive attribute(s) which cannot be modified. Such a sensitive attribute is then considered disclosed when the association between a sensitive attribute value and a particular individual can be inferred with a significant probability. In order to prevent such inferences, the l-diversity model requires that every group of indistinguishable records contains at least l distinct sensitive attributes values; thereby the risk of attribute disclosure is kept under $\frac{1}{l}$. For example, records 4, 5 and 6 in Table 2 form a 3-diverse group. The records contain three values with equal frequencies of 33.33%, and no value is dominant. If we assume that $l = 2$, then although Table 2 is 3-anonymized but it is not a 2-diverse table as in the second equivalence class the num-

ber of sensitive attribute value is only one (Diabetes).

As discussed, a key difficulty of data anonymization comes from the fact that data quality and privacy are conflicting goals. Although it is possible to enhance data privacy by hiding more data values, it decreases data quality. On the contrary, disclosing more data values increases data quality but decreases data privacy. Thus it is necessary to devise new enhanced k-anonymization approaches (for example l-diversity) that best address both the quality and the privacy of the data. In the previous paper (Kabir et al. 2009), we developed systematic clustering method for k-anonymization. However, as l-diversity is more primitive and protected model than k-anonymization, it is necessary to extend the systematic clustering algorithm in l-diversity model. This extension of systematic clustering method to l-diversity model is presented in this paper. It has done in two steps. In the first steps it develops some clusters that satisfy the k-anonymity requirements, called clustering step for k-anonymization. According to this step, first exclude the number of records containing individuals who don't bother about the disclosure of personal identification information. Sort all records by their quasi-identifiers and partitions all records into $[\frac{n}{k}]$ groups. Randomly select a record r from first group to form the first cluster and the first records of the subsequent clusters will form in a systematic way. Then adjusts the records in each group in a systematic way such that each group contains at least k records. Finally distribute the records of individuals who don't bother about the disclosure to their closest clusters or these records constitute another cluster/clusters depending on the number of such records and the k- value. Note that the process of including of such records cause no information loss. In the second step, it develops clusters that satisfy the l-diverse requirement on the sensitive attributes, called l-diverse step. According to this step, first remove clusters in the first step that does not satisfy l-diversity requirement. Then add the records containing in these clusters to other clusters that already satisfy l-diversity requirement where cause least information loss. Note that inclusion of new records to other clusters does not violate l-diversity requirement. There are many clustering based k-anonymization techniques (Byun et al. 2007, Loukides & Shao 2007, Chiu & Tsai 2007, Lin & Wei 2008, Gonzalez 1985) are available but to the best of our knowledge there is no such approaches exist for l-diversity model in the literature. Based on the leakages, this work is devoted a systematic clustering method for l-diversity model.

The reminder of this paper is organized as follows. We present some concepts relating to information loss and a brief overview of the clustering based approaches for k-anonymization in Section 2. In Section 3 we present proposed systematic clustering method for l-diversity model that consists in two steps. Important properties of the proposed algorithm are discussed in Section 4. We compare our proposed algorithms with the most recent clustering based algorithm in Section 5. Finally, concluding remarks are included in Section 6.

2 Preliminaries Relating to Anonymization

The k-anonymity model has drawn a considerable interest in the research community for the last few years and a number of algorithms have been proposed (Ciriani et al. 2008, Bayardo 2005, Fung et al. 2005, LeFevre et al. 2005, 2006, Sweeney 2002, Sun et al. 2008). However, these way out suffer from high information loss mainly due to reliance on pre-defined generalization hierarchies (Bayardo 2005, Fung et al. 2005, LeFevre et al. 2005, Sweeney 2002) or total order (Ciriani et al. 2008, LeFevre et al. 2006) imposed on each attribute domain. Some existing work on k-anonymization has attempted to capture usefulness by measuring the number of total suppressions (Meyerson 2004), the size of the anonymized group (Bayardo 2005, LeFevre et al. 2006), the height of generalisation hierarchies (Samarati 2001, Byun et al. 2007), or information loss through anonymization (Xu al. 2006). However, such metrics fail to detain security. In other works Machanavajjhala et al. (2006), Truta & Vinary (2006) attempts have been made to enhance protection by enforcing anonymized groups. The intuition behind this is that if the values of a sensitive attribute of an anonymized group are quite diverse, then it is difficult for an attacker to breach privacy. However, these frequency-based criteria treat numerical attributes as categorical and thus protection is not captured adequately. For instance, l-diversity proposed by Machanavajjhala et al. (2006) requires a sensitive attribute to have at least l distinct values in an anonymized group. Please refer to Ciriani et al. (2008) and Machanavajjhala et al. (2006) for a survey of various k anonymization and l-diverse approaches.

2.1 Information Loss

Anonymization via generalization or suppression usually causes information loss. Now a natural question arise, how much information is lost due to anonymization. Thus the idea of information loss is used to measure the amount of information loss due to k-anonymization. There are various methods of conniving information loss (Bayardo 2005, Byun et al. 2007, Lin & Wei 2008, Solanas et al. 2008, Iyengar2002 2002). The measurement of information loss in this article is based on the description given by Byun et al. (2007). Please also refer to Byun et al. (2007) for more details.

Let η denote a set of records with r numeric quasi-identifiers $N_1, N_2, ..., N_r$ and s categorical quasi-identifiers $C_1, C_2, ..., C_s$. Let $\Im = \{\Omega_1, \Omega_2, ..., \Omega_p\}$ be a partitioning of η, such that $\cup_{i=1}^{p} \Omega_i = \eta$, Ω_i and Ω_j $(i \neq j)$ are pair wise mutually exclusive. To generalize the values of each categorical attribute $C_i (i = 1, 2, ..., s)$, let τ_{C_i} be the taxonomy tree defined for the domain of C_i.

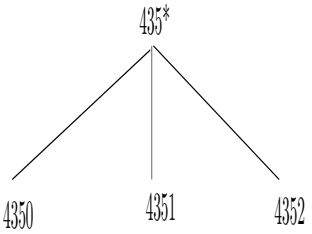

Figure 1: Taxonomy tree of *ZipCode*.

Consider a cluster Ω in η which consists of some numerical and categorical attributes. Let $N_{i_{max}}$, $N_{i_{min}}$ be the maximum and minimum values of the records in Ω and $\eta_{N_{i_{max}}}, \eta_{N_{i_{min}}}$ be the maximum and minimum values of the records in η with respect to numeric attribute $N_i (i = 1, 2, ..., r)$ and $\cup_{C_i}$ be the union set of values in Ω with respect to the categorical attribute $C_i (i = 1, 2, ..., s)$. Then the amount of information loss due to generalizing Ω, denoted by

$IL(\Omega)$ is defined as

$$IL(\Omega) =\mid \Omega \mid .(\sum_{i=1}^{r} \frac{N_{i_{max}} - N_{i_{min}}}{\eta_{N_{i_{max}}} - \eta_{N_{i_{min}}}} + \sum_{j=1}^{s} \frac{H(\Lambda(\cup_{C_j}))}{H(\tau_{C_j})}). \quad (1)$$

where $\mid \Omega \mid$ is the number of records in Ω, $\tau(\cup_{C_j})$ is the subtree rooted at the lowest common ancestor of every value in $\cup_{C_j}$ and $H(\tau)$ is the height of taxonomy tree τ.

Suppose that the total number of records in η is partitioned into p clusters, namely $\Im = \{\Omega_1, \Omega_2, ..., \Omega_p\}$. The total information loss of η is the sum of the information loss of each $\Omega_i (i = 1, 2, ..., p)$. So the total information loss will be:

$$\begin{aligned} IL(\eta) &= \sum_{i=1}^{p} IL(\Omega_i) \\ &= \sum_{i=1}^{p} \mid \Omega_i \mid .(\sum_{k=1}^{r} \frac{N_{ik_{max}} - N_{ik_{min}}}{\eta_{N_{ik_{max}}} - \eta_{N_{ik_{min}}}} \\ &+ \sum_{j=1}^{s} \frac{H(\Lambda(\cup_{C_{ij}}))}{H(\tau_{C_{ij}})}) \end{aligned} \quad (2)$$

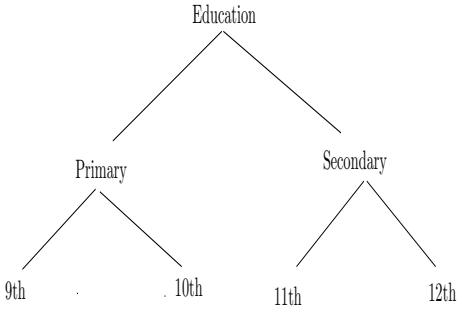

Figure 2: Taxonomy tree of *Education*.

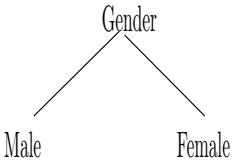

Figure 3: Taxonomy tree of *Gender*.

The main objective of clustering techniques is to construct the clusters in such a way that the total information loss of η will be minimum.

Example: Consider patients records in Table 1 and the 3-anonymization table in Table 2. The anonymized table consists of three clusters. The first cluster consists of first three records, the second clusters consists of middle three records and the last cluster consists of last three records. Consider attributes ZipCode, Gender, Age, Education, where Age is a quantitative variable and the others are categorical variable. Also consider the taxonomy tree of ZipCode, Education and Gender in Figure 1, Figure 2 and Figure 3 respectively. In the table the number of clusters are 3 and the size of each cluster is also 3. In the first cluster the maximum and minimum values respectively as 26 and 24, in the second cluster these values are respectively as 40 and 35 and finally in the last cluster these values are respectively as 43 and 41. Also the maximum and minimum values of all records respectively as 43 and 24. Then the total information Loss of the anonymized table in Table 2 will be

$$\begin{aligned} IL(\eta) &= \mid 3 \mid (\frac{26-24}{43-24} + 1 + 1 + \frac{1}{2}) + \mid 3 \mid (\frac{40-35}{43-24} \\ &+ 1 + 1 + \frac{2}{2}) + \mid 3 \mid (\frac{43-41}{43-24} + 1 + 1 + \frac{1}{2}) \\ &\approx 25.44. \end{aligned} \quad (3)$$

2.2 Clustering based techniques

Clustering based techniques are now using in anonymization to protect the privacy of sensitive attributes and there are various k-anonymization clustering techniques in the literature (LeFevre et al. 2006, Byun et al. 2007, Loukides & Shao 2007, Chiu & Tsai 2007, Lin & Wei 2008). Byun et al. (2007) introduced clustering techniques instead of equivalence class on k anonymization and proposed the greedy k-member clustering algorithm. This algorithm works by first randomly selecting a record r as the seed to start building a cluster, and subsequently selecting and adding more records to the cluster such that the added records incur the least information loss within the cluster. Once the number of records in this cluster reaches k, this algorithm selects a new record that is the furthest from r, and repeats the same process to build the next cluster. When there are fewer than k records not assigned to any cluster yet, this algorithm then individually assigns these records to their closest clusters. This algorithm has two drawbacks. First, it is slow. Second, it is sensitive to outliers. To build a new cluster, this algorithm chooses a new record that is the furthest from the first record selected for previous cluster. If the data contains outliers, it is likely that outliers have a great chance of being selected. If a cluster contains outliers, the information loss of this cluster increases. The time complexity of the algorithm is $O(n^2)$, where n is the number of records in the data set to be anonymized. Their experimental results showed that the k-member algorithm causes significantly less information loss than another k-anonymization technique called "Mondrian" proposed by LeFevre et al. (2006).

Loukides & Shao (2007) proposed another clustering technique for k- anonymization. Similar to k-member this algorithm forms one cluster at a time. But, unlike the k- member algorithm, this algorithm chooses the seed of each cluster randomly. Also, when building a cluster, this algorithm keeps selecting and adding records to the cluster until the information loss exceeds a user defined threshold. If the number of records of a particular class is less than k, the entire cluster is deleted. With the help of the user-defined threshold, this algorithm is less sensitive to outliers. The time complexity of the algorithm is $O(\frac{n^2 log(n)}{c})$, where c is the average number of records in each cluster. However, this algorithm also has two drawbacks. First, it is difficult to decide a proper value for the user-defied threshold. Second, this algorithm might delete many records, which in turn cause a significant information loss. Chiu and Tsai (Chiu & Tsai 2007) proposed another algorithm for k-anonymization that adapts the weighted feature c-means clustering. Unlike the previous two algorithms, this algorithm attempts to build all clusters simultaneously by first randomly selecting $\lfloor \frac{n}{k} \rfloor$ records as seeds. Then this algorithm allocates all records in the data set to their respective

closest cluster and consequently updates feature weights to minimize information loss. This process is continued until the assignment of records to cluster stops changing. If some clusters contain fewer than k records, merge those clusters with other large clusters to satisfy the k-anonymity requirement. One of the main drawback of this algorithm is that it can only be used for quantitative quasi-identifier. The time complexity of this algorithm is $O(\frac{t^2}{k})$, where t is the number of iterations needed for the assignment of records to clusters to converge.

To reduce the information loss and execution time recently Lin & Wei (2008), proposed an efficient one-pass k-mean clustering problem that runs in $O(\frac{n^2}{k})$. They showed that their algorithm performs better than the proposed algorithm of Byun et al. (2007) with respect to both execution time and information loss. Like Chiu & Tsai (2007)'s algorithm, this algorithm forms all clusters at a time. According to their methods first sort all records by their quasi-identifiers, determine approximate number of clusters, by $p = \frac{n}{k}$, where k is the cluster size. Then randomly select p records as seeds to build p clusters. For each record r the algorithm finds the cluster that is closet to r, assign r to that cluster and subsequently updates the center point. Finally, if some clusters contain more than k records remove excess records from those clusters that are dissimilar to most of the records and then add these records to other similar clusters (whose size less than k). Although this method has less execution time there is still chance of being affected by extreme values. Again if this algorithm first selects p records that come from same equivalent class then the total information loss will be higher. All of these clustering techniques are based on k-anonymization techniques. However there is no such approach is available in the literature for l-diversity.

Very recently Kabir et al. (2009) proposed systematic clustering method for k-anonymization that run in $O(\frac{n^2}{k})$, where n is the total number of records containing individuals concerning their privacy. It has shown by experiment that the method attains a reasonable dominance with respect to both information loss and execution time over the recent clustering algorithm. The proposed systematic clustering method differs from previously proposed clustering based k-anonymization methods in four different ways. First, it endeavour to make all clusters simultaneously. On the contrary, the methods proposed by Byun et al. (2007) and Loukides & Shao (2007) build one cluster at a time. Second, it takes less time than the previous two methods as only the first record randomly selects and the subsequently records from in a systematic way. Third, since first record of each clusters contains non identical value, this method easily captures if there are any extreme values and lastly the total information loss will be reduced as in the final step the process of incurring no information loss. Based on the performance of this algorithm, in this paper it is implemented in l-diversity model.

3 Clustering for l-diversity

As discussed before, clustering escorts to better data quality of the disclosed dataset as it partitions a set of records into groups such that records in the same group are more similar to each other than records to other groups. If the records in a particular group are more similar, the group leads to a minimal generalization and thus incurs less information loss. In this respect, the problem of k-anonymization can also be considered as a clustering problem, where each equivalent class is a cluster and the size of each cluster is at least k. But the requirement for l-diversity model to satisfy at least l distinct sensitive attribute values in each equivalent class. So the optimal solution of clustering problem is to construct a set of clusters such that it satisfies both k-anonymity and l-diversity requirement and the total information loss will be as minimum as possible. In this section, we formally define and present our systematic clustering algorithm that minimizes the information loss and respects the k-anonymity and l-diversity requirement.

3.1 Systematic clustering problem

There are various clustering problems in the literature. Among them, the k-center clustering problem proposed by Gonzalez (1985) aims to find k clusters from a given dataset such that the maximum inter-cluster distance (or radius) is minimized. Thus the optimum solution is to constitute p clusters $\{\Omega_1, \Omega_2, ..., \Omega_p\}$ in such a way that minimizes the cost metric

$$MAX_{i=1,.....}MAX_{j,k=1,...,|\Omega_i|}D(r_{i,j}, r_{i,k}), \quad (4)$$

where $r_{i,j}$ represents a data point in cluster Ω_i and $D(x, y)$ is a distance between two data points, x and y.

In the k-anonymity problem the restriction is that the number of records in each equivalence class should be at least k and in the l-diversity model the restriction is that the number of sensitive attribute values in each equivalence class must be at least l distinct values but there is no such restriction about the number of clusters in both cases. So a clustering problem is to form in such a way that each cluster contains at least k similar records, l distinct sensitive records and the sum of information losses of all clusters is as small as possible. For applying systematic method to l-diversity model of selecting records we need to follow two steps. The first one is the clustering step for k-anonymization and the second one is the l-diverse step. Suppose that we would like to apply systematic clustering method to l-diversity model for Table 1. Then in the clustering step for k-annualization first sort all records in the whole data set with respect to quasi-identifiers. There are 9 records in Table 1 and suppose that dataset already sorted according to the quasi identifier attributes *ZipCode*, *Gender*, *Age* and *Education*. If the anonymized table follows 3-anonymity requirements, then the number of clusters should be $\frac{9}{3} = 3$. Select a record (say, 2th record) from the first 3 records to form the first cluster. Then select $(2 + 3)th = 5$th and $(2 + 2 \times 3)th = 8$th records in a systematic way to form the second and third cluster respectively. Now again select another record from the first 3 records (say, 3rd not 2th as it already selected) and calculate the information loss with all of the three clusters using the equation (1). The information losses are respectively as 5.10, 6.47 and 6.68, if this record includes in the first, second and third cluster. So, 3rd record will be included in the first cluster as it causes least information loss. Similarly select $(3 + 3)th = 6$th and $(3 + 2 \times 3)th = 9$th record in a systematic way and include them in the second and third cluster respectively. Finally select 1st, $(1 + 3)th = 4$th and $(1 + 2 \times 3)th = 7$th record and include these records to the first, second and third cluster respectively as they will then cause least information loss. If the total number of records is not exactly divisible by

the k-anonymity parameter, then rest records will be included to the similar clusters where information loss is minimum and this process continues until the number of records in a particular cluster is k to satisfy k-anonymity requirement. Thus in the clustering step for k-anonymization as set of clusters are built that satisfy the k-anonymity requirement. In the l-diverse step, the clusters will be formed in such a way that the number of distinct sensitive attribute values in each cluster is at least l. Note that if in the clustering step the table already satisfies l-diversity requirement, next step is not required. Suppose that $l = 3$, in this particular example, so the clusters that are obtained in the first step does not satisfy l diversity requirement as the second cluster consists only one distinct sensitive attribute value. So in the l-diverse step remove this cluster and the records containing in this cluster to other similar clusters that causes least information loss. All of the three records in this cluster will be included in the third cluster as these records will then incur less information loss. Thus we get a table in Table 3 that satisfy both the 3-anonymity and 3-diversity requirements. The process of building the table by using systematic method protects individuals private information as well as sensitive attributes.

Definition 1 (Systematic clustering problem for l-diversity) *The systematic clustering problem is to find a set of clusters from a given set of n records such that each cluster contains at least k ($k \leq n$) records (where the records select in a systematic way and include in a cluster that cause least information loss), the number of distinct sensitive attribute values is at least $l(l \geq 2)$ and that the sum of all intra-cluster distances is as minimum as possible. More specifically, if η be a set of n records and k & l are the specified anonymization and diversity parameter, the optimal solution of the systematic clustering problem is a set of clusters $\Im = \{\Omega_1, \Omega_2, ...\}$ such that:*

1. $\Omega_i \cap \Omega_j = \Phi$, *for all* $i \neq j = 1, 2, ...$,
2. $\cup_{i=1,.....} = \eta$,
3. *for all* $\Omega_i \in \Im, \mid \Omega_i \mid \geq k$ *&* $l \geq 2$, *and*
4. *the total information loss obtained by using equation (2) is minimized.*

In Definition 1, a set of clusters are constructed in such a way that the clusters are mutually exclusive, the sum of records of all clusters is equal to the total number of records, the size of each cluster is at least k and the number of distinct sensitive attribute values is at least l that satisfies both the criteria of k-anonymization and l-diversity. The problem tries to minimize the sum of all intra-cluster distances, where an intra-cluster distance of a cluster is defined as the maximum distance between any two records in the cluster. In the following subsection we formally design a systematic clustering algorithm for l-diversity.

3.2 Systematic clustering algorithm

Based on the information loss in Subsection (2.1) and the definition of systematic clustering problem, we are now ready to discuss a systematic clustering algorithm for l-diversity model. As discussed, the whole procedure consists of the two steps, namely clustering step for k-anonymization and l-diverse step.

Clustering step for k-anonymization

Table 4: Clustering step for k-anonymization algorithm

Input: a set η of n records containing individuals concerning their privacy, where $\eta_1, \eta_2, ..., \eta_n \in \eta$; the value k for k-anonymity
Output: a partitioning $\Im = \{\Omega_1, \Omega_2, ..., \Omega_p\}$ of τ
1. Sort all records in η by their quasi-identifiers;
2. Let p:= int$\lfloor \frac{n}{k} \rfloor$;
3. Get randomly k distinct records $r_1, r_2, ..., r_k$ from first 1 to k;
4. Let p_{ij} is the jth element in the ith cluster;
5. For $i = 1$ to p;
6. Let $p_{i1} := \eta_{[r_1+k(i-1)]}$;
7. Next i;
8. For $j := 2$ to k;
9. For $i := 1$ to p;
10. Let IL_i: = InfoLoss($\eta_{[r_j+k(i-1)]}$);
11. Let X:= Find cluster number with lowest IL_i;
12. where cluster size $\leq k$;
13. Add $\eta_{[r_j+p(i-1)]}$ to p_x;
14. Next i;
15. Next j;
16. Let $e := (n - pk)$;
17. Find extra element $E_1, E_2, ..., E_e \in E$;
18. For $k := 1$ to e;
19. For $m := 1$ to p;
20. Let IL_m := InfoLoss(E_k) in cluster m;
21. Next m;
22. Let X := Find cluster number with lowest IL;
23. Add E_k to p_x;
24. Next k;

The aim of this step is to develop a set of clusters from a given set of n records that satisfy the k-anonymity requirement. The general ida if this step is as follows:

Note that for collecting medical data from patients it may be expected that some patients are not concerned about the privacy of their medical records and the other attributes. We would like to explore this opportunities because unnecessary anonimization may produce more information loss. Let q be the probability that a particular patient is not concerned about the privacy of medical records. Then out of n patients we can expect that on an average nq patients are not concerned about their privacy. According to this step first exclude the records of individuals who are not concerned about the privacy. Then sort all records by their quasi-identifiers and identify the equivalence class and the number of clusters by, $p = \frac{(n-nq)}{k}$, where k is anonymity parameter for k-anonymization and round this as integer. Randomly select a record r_i from first k records as seeds to form the first cluster. If there are p clusters to be formed then select the $(r_i + k)$th, $(r_i + 2k)$th,..., $\{r_i + (p-1)k\}$th records in a systematic way to form 2nd, 3rd, ..., pth cluster respectively. Select another record $r_j (j \neq i)$ from the first k records and add this record to the cluster which causes least information loss. Similar in a systematic way select $(r_j + k)$th, $(r_j + 2k)$th,..., $\{r_j + (p-1)k\}$th records and add these records to their respective clusters that cause least information loss. If any cluster size is exactly k, stop adding records to that cluster and continue the same process until all records of first k records finish. If $(n - nq)$ is not exactly divisible by k and still there are some records left, add this records to their closest clusters that incur least information loss. Finally distribute the nq records to their closest clusters or these nq records constitute separate cluster/clusters depending on their size. Note that these nq records do not incur any information loss. Since only the first record randomly selects and the subsequent records from in a systematic way, it has less execution time. Again usually the first record of each cluster

Table 3: 3-diversity table

	ZipCode	*Gender*	*Age*	*Education*	*Disease*	*Expense*
1	435⋆	Person	[21-30]	Primary	Flue	2000
2	435⋆	Person	[21-30]	Primary	Cancer	3500
3	435⋆	Person	[21-30]	Primary	HIV+	6500
4	435⋆	Person	[31-50]	Educated	Diabetes	2000
5	435⋆	Person	[31-50]	Educated	Diabetes	3200
6	435⋆	Person	[31-50]	Educated	Diabetes	2800
7	435⋆	Person	[31-50]	Educated	Flue	2700
8	435⋆	Person	[31-50]	Educated	Heart disease	4800
9	435⋆	Person	[31-50]	Educated	Cancer	5200

contains non identical value, so this algorithm easily captures if there are any extreme values. Moreover, this algorithm is adding some records that contain no information loss, so it is a natural expectation that the total information loss will be reduced. The clustering step for k-anonymization algorithm is shown in Table 4. In the algorithm it is assumed that all n individuals are concerned about their privacy. Thus in this step, we get some clusters that satisfy the k-anonymity requirement and the total information loss of all of these clusters will be minimum.

l-diverse step

As discussed in the previous section, we have some clusters that satisfy k-anonymity requirement but may or may not satisfy l-diversity requirement. Note that l-diverse step is invoked only if in the first step, some of the clusters in the k-anonymization table does not satisfy l-diversity requirement. If for a certain l-value, all clusters in the anonymized table satisfy the l-diversity requirement, the l-diverse step of the table is not required. According to this step, remove the clusters that do not satisfy l-diversity requirements and add the records containing in these clusters to other clusters that cause least information loss. As the existing clusters already satisfy k-anonymity and l-diversity requirement, inclusion of new records don't violate these requirement. The algorithm of the l-diverse step is illustrated Table 5.

Table 5: l-diverse algorithm

Input: a partitioning $\Im_1 = \{\Omega_1, \Omega_2, ..., \Omega_{p_1}\}$ of τ that satisfy k-anonymity requirement, a partitioning $\Im_1^\star = \{\Omega_1^\star, \Omega_2^\star, ..., \Omega_{p_2}^\star\}$ of τ that satisfy both the k-anonymity and the l-diversity requirement, a set of sensitive attributes $S_i (i = 1, 2, 3, ...)$, and the value of l for l-diversity.
Output: a partitioning $\Im = \{\Omega_1, \Omega_2, ...\}$ of τ that satisfy the l-diversity requirement.
1. Let $\Upsilon = \{r_1, r_2, ...\}$={all records of $\Im_1$}
2. Let r_j is the jth record of Υ;
3. For $j = 1, 2, ...$;
4. For $i = 1, 2, ..., p_2$;
5. Let $\mathrm{IL}_i^\star$:=InfoLoss($\Omega_i^\star$), $i = 1, 2, ...p_2^\star$;
6. Find the cluster $\Omega_i^\star$ with lowest $\mathrm{IL}_i^\star$;
7. Add r_j to $\Omega_i^\star$;
8. Next j;

Definition 2 (Systematic clustering decision problem for l-diversity) *In a given data set of n records, there is a clustering scheme $\Im = \{\Omega_1, \Omega_2, ...\}$ such that*

1. *$\mid \Omega_i \mid \geq k, 1 < k \leq n$: the size of each cluster is greater than or equal to a positive integer k,*
2. *$l \geq 2$, the number of distinct sensitive attribute values in each cluster is at least 2, and*
3. *$\sum_{i=1} IL(\Omega_i) < c, c > 0$: the total information loss of the clustering scheme is less than a positive integer c.*

where each cluster $\Omega_i (i = 1, 2, ...)$ contains the records that are more similar to each other with respect to k and l such that require minimum generalization and thus causes least information loss. In the following subsection we are going to discuss some properties of the proposed systematic clustering algorithm.

4 Analysis of the new algorithm

As discussed before, the proposed algorithm is designed in such a way that finds a solution of l-diversity model. In the first step, this algorithm stops adding records in a particular cluster if the cluster size exactly k. Again it always keep in mind in adding records that incur less information loss. Moreover, the records are selected in a systematic way that makes the algorithm faster. With respect to this, this algorithm has the following desirable properties.

Theorem 1. *Let n be the total number of input records and k be the specified k anonymity parameter. The time complexity of the clustering step for k-anonymization is in $O(\frac{n^2}{k})$.*

Proof. In the clustering step for k-anonymization, after sorting the records with respect to the quasi-identifiers the algorithm determine the numbers of clusters by $p = \frac{n}{k}$. Then it selects the records as seeds in a systematic way to form all p clusters simultaneously. Thus for each tuple in the dataset, the algorithm needs to assign it to one of the p clusters, which has a complexity of $O(p)$. As a result, the assignment of all tuples to the clusters has a time complexity of

$$\begin{aligned} T &= O(Number\,of\,tuples * Number\,of\,clusters) \\ &= O(n * p) = O(n * \frac{n}{k}) = O(\frac{n^2}{k}). \end{aligned} \quad (5)$$

Therefore, the total execution time is in $O(\frac{n^2}{k})$.

As discussed, in the first step the algorithm develops clusters that satisfy the k-anonymity requirement. The second step is required only if in the first step some clusters does not satisfy l-diversity requirement. The time complexity in the second step thus depends on number of such clusters and the reassignment of records to other clusters. So the time complexity in the second step is not straight forward.

Theorem 2. *Let n be the total number of input records and q be the probability that a particular individual doesn't bother about the disclosure. Then the algorithms in fact work out the information loss of $(n-nq)$ individuals instead of all n individuals.*

Proof. If q be the probability that a particular individual doesn't bother about the disclosure. Then out of n individuals, nq individuals are not bothered about the disclosure. Assume that these nq records are in one separate cluster that causes no information loss. Also let $IL(\eta)$ and $IL(\eta_{all})$ are the total information loss due to l-diversity model and any other clustering algorithm respectively. According to the systematic clustering algorithm, the total information loss will be:

$$\begin{aligned} IL(\eta) &= IL(n) \\ &= IL(nq) + IL(n-nq) \\ &= 0 + IL(n-nq) = IL(n-nq). \quad (6) \end{aligned}$$

So in the systematic clustering algorithm for l-diversity model, it actually calculate the information loss of $(n-nq)$ records instead of calculating the information loss of all n records.

Theorem 3. *Let n be the total number of input records and k be the anonymity parameter in k-anonymization. Then according to the algorithm in first step, the cluster size of any cluster is at least k but no more than $(2k-1)$.*

Proof. Let n be the total number of input records. According to clustering step for k-anonymization, first select the initial seeds of all clusters in a systematic way and subsequently selecting adding more records to the clusters such that the added records incur least information loss. Again this algorithm stops adding record to a particular cluster if the number of records is exactly k. So in the worst case, if there are $(k-1)$ records left and if these all records include in a cluster that already contains k records, the total number of records in that cluster will be $(k+k-1) = (2k-1)$. Therefore the maximum size of a cluster will be $(2k-1)$.

5 Comparison

As discussed before, to the best of our knowledge no clustering methods exit in the literature for l-diversity model. Most of the clustering based approaches (LeFevre et al. 2006, Byun et al. 2007, Loukides & Shao 2007, Chiu & Tsai 2007, Lin & Wei 2008, Kabir et al. 2009) are based on k-anonymization techniques. However, the closest works related to this paper is the systematic clustering method for k-anonymization proposed by Kabir et al. (2009) and the one pass $k-$means algorithm (OKA) proposed by Lin & Wei (2008). In this section we compare our proposed algorithms with these two algorithms.

Kabir et al. (2009) developed an algorithm that selects records in a systematic way to form the clusters and endeavors to make all clusters simultaneously. Experimental results showed that systematic clustering method is the best fit for k-anonymization with respect to both information loss and execution time over the recent clustering algorithms. The first step of the clustering technique proposed in this paper is exactly the same as Kabir et al. (2009), however in the second step we developed an algorithm for l-diversity model that has significantly improve the work of Kabir et al. (2009). According to the algorithm of Kabir et al. (2009) clusters are formed that satisfy the k-anonymity requirement. By contrast, by using the algorithms proposed in this paper clusters are formed that satisfy both the k-anonymity and l-diversity requirement that their work did not provide.

Lin & Wei (2008) proposed a two-stage algorithm, called one pass k-means algorithm (OKA). During the first stage, the algorithm clusters data using the k-means algorithm, but only for the first iteration. During the second stage, records are moved from clusters with more than k records (called the shrinking clusters) to clusters with fewer than k records (called the growing clusters) to ensure that each cluster eventually contains no fewer than k records. The algorithm is designed for k-anonymization but the authors did not implement the algorithm for l-diversity model. By contrast, the proposed algorithms presented in this paper are intended for l-diversity model.

6 Conclusion and future works

In this paper, we propose algorithms for l- diversity model as an enhanced of k-anonymity model. The proposed technique uses the idea of clustering and is implemented in two steps, namely clustering step for k-anonymization and l-diverse step. The basic concepts of the proposed algorithms are discussed and investigated through example and properties. The time complexity of the developed algorithm is in $O(\frac{n^2}{k})$, where n is the total number of records containing individuals concerning their privacy in the first step. The effect of the proposed approach can be useful for protecting private information of individuals as l-diversity model is one of the most popular approaches for privacy preserving techniques.

The proposed approach in this paper is implemented through paradigm. Our future work is to conduct an experimental study to show the efficiency and the effectiveness of the proposed algorithms. Again recently there are many disparities of the k-anonymity model have been proposed in the literature to further protect the private information, e.g., t-closeness (Li 2007), (α, k)-anonymity (Wong et al. 2006). Our further work is also to extend the algorithms to these models.

References

Samrati, P. (2001), Protecting respondent's privacy in microdata release, *in* 'TKDE'.

Sweeney, L. (2002), 'K-anonymity: a model for protecting privacy', *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* **10**(5), 557–570.

Ciriani, V., di Vimercati, S.D.C., Foresti, S. & Samarti, P. (2008), k-anonymous data mining: A aurvey, *in* 'Privacy-preserving data mining: Models and algorithms', C.C. Aggarwal and P.S. Yu, Eds. Boston: Kluwer Academic Publishers, 103–134.

Bayardo, R.J. & Agrawal, R. (2005), Data privacy through optimal k-anonymization, *in* 'ICDE'.

Fung,B.C.M., Wang, K.& Yu, P.S. (2005), Top-down specialization for information and privacy preservation, *in* 'ICDE'.

LeFevre, K., DeWitt, D. & Ramakrishnan, R. (2005), Incogniti: Efficient full-domain k-anonymity, *in*

'ACM International Conference on Management of Data'.

LeFevre, K., DeWitt, D. & Ramakrishnan, R. (2006), Mondrian multidimensional k-anonymity, *in* 'ICDE'.

Sweeney, L. (2002), 'Achieving k-anonymity privacy protection using generalization and supression', *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* **10**(5), 571–588, 2002.

Byun, J.W. & Bertino, E. (2006), 'Micro-views, or on how to protect privacy while enhancing data usability: concepts and challenges', *SIGMOD* **35**(1), 9–13, 2006.

Byun, J.W., Sohn, Y., Bertino, E. & Li, N. (2006), Secure anonymization for incremental datasets, *in* '3rd VLDB Workshop on Secure Data Management'.

Byun, J.W., Kamra, A., Bertino, E. & Li, N. (2006), Efficient k-anonymization using clustering techniques, *in* 'International Conference on Database Systems for Advanced Applications'.

Loukides, G. & Shao, J. (2007), Capturing data usefulness and privacy protection in k-anonymisation, *in* 'Proceedings of the 2007 ACM symposium on Applied Computing'.

Chiu, C.C. & Tsai, C.Y. (2007), A k- anonymity clustering method for effective data privacy preservation, *in* 'Third International Conference on Advanced Data Mining and Applications'.

Lin, J.L. & Wei, M.C. (2008), An efficient clustering method for k-anonymization, *in* 'Proceedings of the 2008 international workshop on Privacy and anonymity in information society'.

Meyerson, A. & Williams, R. (2004), On the complexity of optimal k-anonymity, *in* 'PODS'.

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B. & Fu, A.W.C. (2006), Utility-based anonymization using local recording, *in* 'KDD'.

Machanavajjhala, A., Gehrke, J., Kifer, D. & VenkitasubramanianXu, M. (2006), l-diversity: Privacy beyond k-anonymity, *in* 'ICEDE'.

Truta, T. & Vinary, B. (2006), Privacy protection: p-sensitive k-anonymity property, *in* 'International Workshop on Privacy Data Management'.

Sun, X., Li, M., Wang, H. & Plank, A. (2008), An efficient hash-based algorithm for minimal k-anonymity, *in* 'ACSC'.

Sun, X., Wang, H. & Li, J. (2008), Priority Driven K-Anonymisation for Privacy Protection, *in* 'AusDM'.

Hettich, C.B.S. & Merz, C. (1998), UCI repository of machine learning databases.

Gonzalez, T.Z. (2002), 'Clustering to minimize the maximum intercluster distance', *Theoretical Computer Science* **38**, 293–306, 1985.

Li, N. & Li, T. (2007), t-closeness: Privacy beyond k-anonymity and l-diversity, *in* 'ICDE'.

Wong, R.C.W., Li, J., Fu, A.W.C. & Wang, K. (2006), (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing, *in* 'Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining'.

Solanas, A., Sebe, F. & Domingo-FerrerWong, J. (2006), Micro-aggregation-based heuristics for p-sensitive k-anonymity: One step beyond, *in* 'International Work-shop on Privacy and Anonymity in the Information Society'.

Iyengar, V.S. (2002), Transforming data to satisfy privacy constraints, *in* 'SIGKDD'.

Kabir, M.E., Wang, H. & Bertino, E. (2009), Efficient Systematic Clustering Method for k-Anonymization, *in* 'Working Paper Series', No SC-MC-0905, Faculty of Sciences, University of Southern Queensland, Australia.

Li, J., Wang, H., Jin, H. & Yong, J. (2008), 'Current Developments of k-Anonymous Data Releasing', *Electronic Journal of Health Informatics* **3**(1), e6, 2008.